Pratique de l'Analyse des Données Numériques et Textuelles avec Dtm–Vic

Ludovic Lebart
Marie Piron

Pratique de l'analyse des données numériques et textuelles avec Dtm-Vic

(Seconde édition, Octobre 2012)

Ludovic Lebart

Marie Piron

ISBN 978-2-9537772-0-8

Sommaire

Dtm-Vic

***D**ata and **t**ext **m**ining*

***V**isualization, **i**nference, **C**lassification*

Logiciel d'analyse exploratoire multidimensionnelle de données numériques et textuelles

Librement téléchargeable sur : www.dtm-vic.com

Introduction

Dtm-Vic est un logiciel consacré à l'**analyse exploratoire multidimensionnelle des données numériques et textuelles.**

L'**analyse exploratoire**, comme son nom le suggère, est une démarche préliminaire de contact avec un recueil de données, contact suivi d'investigations, de description, sans se limiter à un protocole fixé à l'avance. L'exploration suppose que les données sont complexes, que les connaissances *a priori* sur ces données sont limitées.

L'analyse **multidimensionnelle**, elle, s'attache au cas où les dimensions (le plus souvent: les variables) sont nombreuses, ce qui est un facteur de complexité, et par conséquent une incitation à commencer par une démarche exploratoire. Une autre incitation plus technique à utiliser cette démarche concerne le caractère peu réaliste des hypothèses statistiques distributionnelles dans le cas multidimensionnel, qui rend malaisée l'utilisation codifiée des tests d'hypothèses.

L'**analyse exploratoire multidimensionnelle des données numériques** sera un volet important du logiciel Dtm-Vic. Les outils de base en sont d'une part les méthodes factorielles (ou analyses en axes principaux) telles que l'analyse en composantes principales, les analyses des correspondances simples et multiples, d'autre part les méthodes de classification (classification hiérarchique, méthodes de partitionnement, cartes auto-organisées). Ces techniques ne s'excluent pas mutuellement, elles sont au contraire systématiquement utilisées comme des techniques complémentaires apportant chacune des points de vue indispensables sur la réalité statistique. L'ouvrage de base qui accompagne les méthodes mises en oeuvre dans ce volet du logiciel Dtm-Vic a pour titre: "*Statistique Exploratoire Multidimensionnelle*"[1].

Les **données textuelles** sont, en particulier, des données à la fois

[1] *Statistique Exploratoire Multidimensionnelle. Visualisation et Inférence en Fouille de Données.* Ludovic Lebart, Marie Piron, Alain Morineau (2006). 4ème ed. Dunod, Paris.

multidimensionnelles et complexes. Elles sont donc des candidats possibles aux traitements proposés par les analyses exploratoires. Elles sont souvent associées à des données numériques. C'est le cas emblématique des enquêtes par sondage comportant à la fois des questions fermées (données numériques continues et variables nominales) et des questions ouvertes (données textuelles). Ces données d'enquêtes constituent l'exemple-type autour duquel s'est développé Dtm-Vic. Une partie importante des méthodes mises en oeuvre dans le volet textuel du logiciel Dtm-Vic sont présentées et commentées dans l'ouvrage "*Statistique textuelle*"[2].

L'**analyse exploratoire multidimensionnelle des données numériques et textuelles** apparaît comme une phase incontournable du traitement de ces recueils complexes.

On sait, et les exemples sont célèbres, que les explorateurs découvrent souvent autre chose que ce qu'ils cherchent. Les utilisateurs de Dtm-Vic ont souvent l'occasion de le vérifier, de façon pas forcément plaisante pour tout le monde : les analyses réalisées constituent de redoutables tests de cohérence et de qualité de l'information de base, que n'apprécient pas toujours ceux qui ont recueilli cette information, ni ceux qui l'ont utilisée trop vite.

Mais, pour les utilisateurs chevronnés, notamment en sciences sociales, ces épreuves de cohérence globales ne sont pas des retombées accidentelles des explorations mais bien un de leurs objectifs fondamentaux, explicitement inséré dans une démarche critique qui voit le recueil comme une construction et même dans une certaine mesure, une fabrication de l'information.

* *
*

Dans la version 5 de Dtm-Vic à laquelle ce manuel d'utilisation se réfère principalement, l'interface du logiciel est en Anglais (mots-clés, rubriques d'aide, noms des analyses), option qui tient compte du fait que les deux tiers des utilisateurs du logiciel sont non francophones. Le public francophone de chercheurs et de chargés d'étude n'aura cependant pas

[2] *Statistique textuelle.* Ludovic Lebart, André Salem (1994), Dunod, Paris. La version anglaise: *Exploring Textual Data* (L. Lebart, A. Salem, E. Berry, 1998, Kluwer, Dordrecht) inclut des exemples utilisés dans ce manuel.

de mal à piloter le logiciel dans ces conditions. Il est difficile pour une petite équipe, et pour un logiciel dont l'accès est libre, non subventionné, de maintenir plusieurs versions dans des langues différentes. Une version française est toutefois projetée à moyen terme.

Les limites actuelles du logiciel (révisables) en ce qui concerne la taille des données d'entrée sont les suivantes : 30 000 lignes (ces lignes sont des individus ou observations), 1200 colonnes (variables numériques continues, variables numériques codant des variables nominales – une variable nominale = une colonne), 100 000 caractères pour les "réponses textuelles" d'un individu/observation, mais pas de limite pour un texte non associé à un fichier numérique. Ce format correspond à la grande majorité des applications aux enquêtes socio-économiques, aux fichiers issus des enquêtes de gestion ou de satisfaction, aux relevés écologiques, aux analyses sensorielles, etc.

* *
*

On a choisi, dans ce manuel, après une brève présentation du logiciel (chapitre I), de présenter six exemples de traitement sur des données déjà préparées, c'est-à-dire présentée dans un format convenable, et fournies avec le logiciel (chapitre II et III). Ces exemples correspondent à des utilisations fréquentes de Dtm-Vic. L'utilisateur apprendra à créer lui-même un fichier de commande à partir de l'interface proposée. On trouvera successivement une analyse en composantes principales (enchaînée avec une classification et, pour les classes, un positionnement factoriel et une description automatique), une analyse des correspondances, une analyse des correspondances multiples (également complétée par une classification), une analyse factorielle lexicale d'une série de texte, puis, dans le cadre d'une enquête, une analyse des correspondances d'une table lexicale construite à partir d'une question ouverte et d'une question fermée, enfin une analyse et une classification directe des réponses à une question ouverte. Les cinq premières applications donnent lieu à des visualisations validées par la technique du *bootstrap*.

En espérant avoir motivé le lecteur par cette première présentation des fonctionnalités du logiciel, on aborde au chapitre IV les procédures d'importation des données. On conçoit facilement que traiter des unités

statistiques aussi disparates qu'un nombre, une catégorie, une réponse laconique à une question ouverte, ou un roman de Zola peut parfois être compliqué. La transparence totale des fichiers d'entrée ou produits par Dtm-Vic (tous les fichiers sont en format texte non propriétaire) devrait cependant rassurer l'utilisateur et limiter la complexité du processus.

Arrivé au seuil du quatrième chapitre, la lectrice ou le lecteur dispose déjà d'une certaine autonomie. Quelques procédures élémentaires d'archivage ou de recodage sont proposées au chapitre V pour permettre d'affiner ou d'approfondir les analyses précédentes.

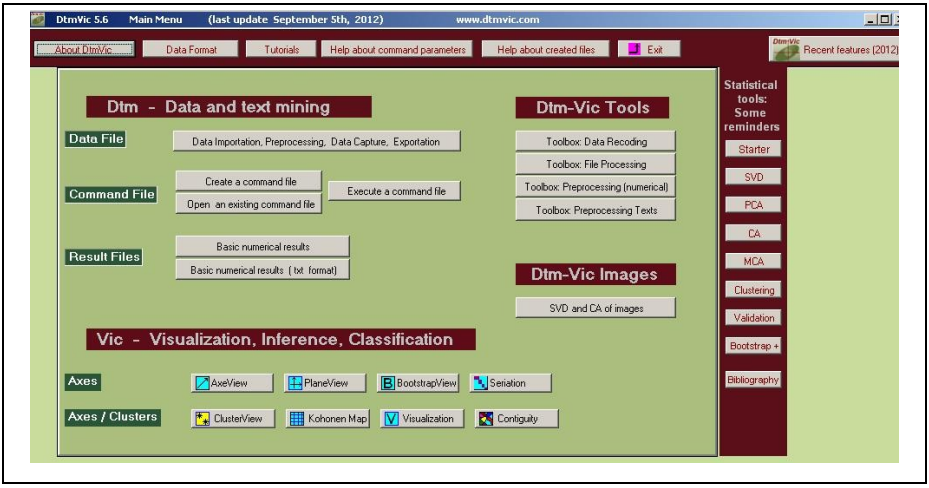
Enfin, le sixième et dernier chapitre présente des applications plus approfondies, mettant notamment en œuvre de nouvelles options des procédures de visualisation. Ce chapitre VI aborde aussi les analyses de contiguïté, les descriptions de graphes, et illustre les capacités de compression des techniques factorielles.

Toutes ces phases de l'apprentissage supposent que le logiciel et le recueil d'exemples aient été copiés ou téléchargés, ce qui est possible à partir du site[3]: http://www. Dtm-Vic.com.

[3] On pourra également télécharger sur ce site l'ouvrage précité "Statistique textuelle" (L. Lebart et A. Salem) et l'ouvrage "La sémiométrie, Essai de Statistique structurale" (L. Lebart, M. Piron, J.-F. Steiner. 2003, Dunod, Paris), d'où sont extraits certains jeux de données utilisés ici. Les autres ouvrages cités ne sont pas libres de droit à cette date et doivent être consultés en bibliothèque ou acquis dans le réseau des librairies.

I. Présentation générale de Dtm-Vic

Pour lancer l'exécution de *Dtm-Vic*, il suffit de cliquer sur l'icône du raccourci placé sur le bureau de *Windows* par le programme d'installation ou par l'utilisateur. On obtient l'écran d'accueil suivant:



Dtm-Vic est structuré en deux étapes :

I – La première étape *Dtm – Data and Text mining* comprend les procédures de mise en place des données (importation, saisie, exportation) et les procédures d'analyses des données (création, puis exécution du fichier de commande).

II – La seconde étape *Vic – Visualization, Inference, Classification* fournit les outils de visualisation, de validation et d'interprétation des résultats.

On peut également voir sur l'écran d'accueil deux rubriques optionnelles : la "boîte à outils", *DtmVic Tools* qui propose différents types de recodage, de stockage des données, et la rubrique *DtmVic Images* consacrée à certaines analyses d'images.

Ce manuel doit permettre de procéder à une mise en oeuvre de ces

étapes de calcul et de visualisation. Certaines d'entre elles, les plus spécifiques du logiciel (mentionnées dans la présentation ci-dessous), seront détaillées dans les différentes parties du manuel, sachant que toutes les analyses relèvent d'un même enchainement des étapes :

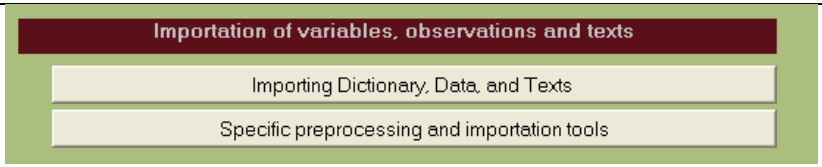
1. Sélection d'une analyse
2. Ouverture des différents fichiers de données dans le format Dtm-Vic
   - Choix des variables
   - Choix des différents paramètres spécifiques à l'analyse.
3. Création d'un fichier de commande
4. Exécution du fichier de commande
5. Visualisation des résultats.

Pour obtenir des aides sur les paramètres ou les fichiers, cliquez sur les menus Help, dans la barre du haut, qui s'affichent alors en rouge. Pour supprimer l'affichage d'une rubrique d'aide cliquez à nouveau sur le bouton correspondant. Le tutoriel (en anglais) est accessible sur cette barre.

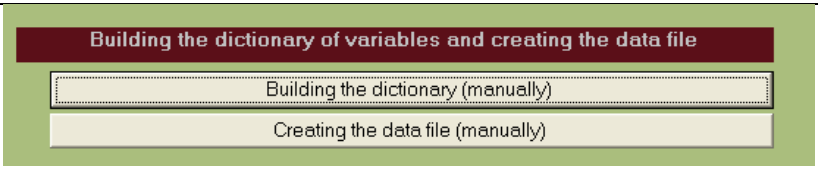
I.1 Mise en place des fichiers de données :

➢ Cliquez sur Data Importation, Preprocessing, Data Capture, Exportation dans **Data File**.

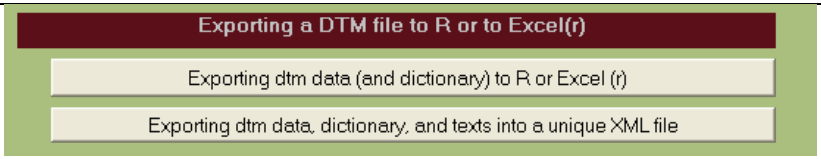
⊙ Une fenêtre suggérant différentes procédures apparaît. Voici les composants de cette fenêtre:



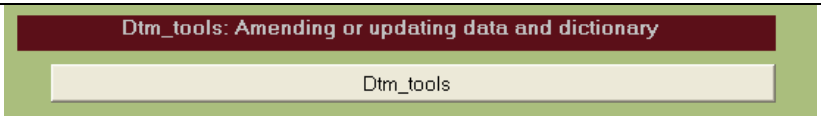
- Importation de fichiers de données numériques ou textuelles et constitution des fichiers dictionnaire, données et textes dans le format Dtm-Vic. **Voir chapitre IV**
- Quelques outils de pré-traitement.



Modules de saisie de données : construction du dictionnaire des variables et création du fichier de données. **Voir chapitre IV.**



Exportation de fichiers de données en format Excel, R ou XML.. **Voir chapitre IV**



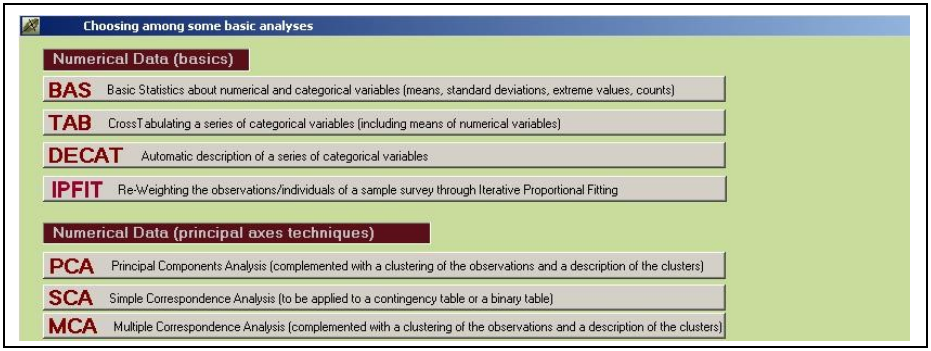
Création de nouvelles variables, sélection d'un sous-échantillon ou concaténation de plusieurs fichiers. **Voir l'accès direct à la boîte à outils *DtmVic Tools* et chapitre V**

I.2 Techniques d'analyse des données

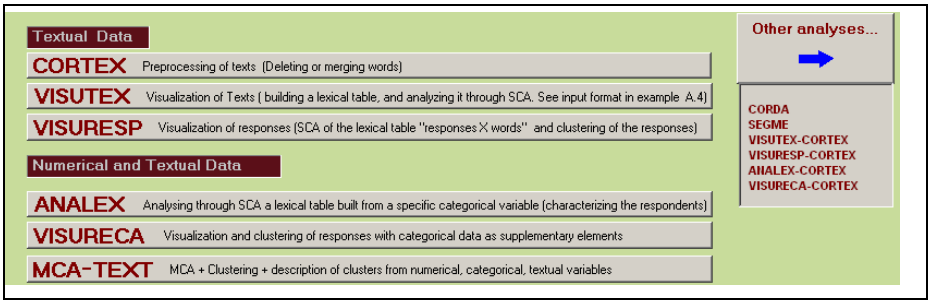
- Cliquez sur Create a command file dans la rubrique **Command File** de *Dtm – Data and Text mining*

⊙ Une fenêtre affichant différentes techniques d'analyse possibles, selon la nature numérique ou textuelle des données, apparaît :

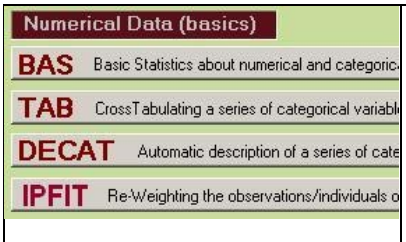
La partie supérieure de cette fenêtre traite des données numériques :



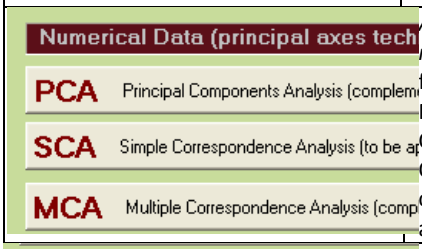
La partie inférieure de la même fenêtre traite des données textuelles :



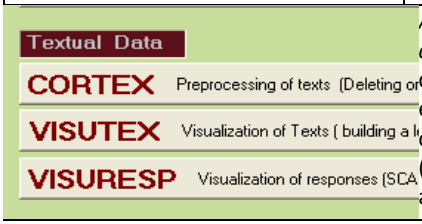
Explicitations sommaires des traitements:



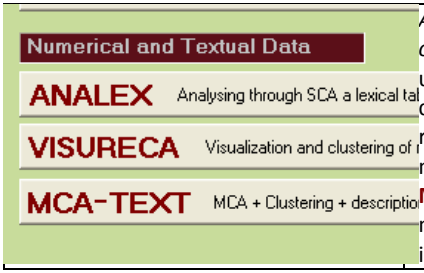
Analyse descriptive univariée, **BAS** ; *Demande de tableaux croisés* **TAB** des variables continues ou nominales;
Description automatique d'une variable par une série de variables nominales **DECAT.**
Redressement de l'échantillon, **IPFIT** (Iterative Proportional Fitting).



Analyse statistique exploratoire de données numériques : Enchainement d'une analyse factorielle (Analyse en Composantes Principales **PCA**, Analyse des Correspondances Simples **SCA**, Analyse des Correspondances Multiples **MCA**) et d'une classification (k-means et classification ascendante hiérarchique). **Voir chapitre II.**

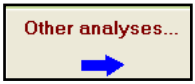


Analyse statistique exploratoire d'un corpus de textes: **CORTEXT** supprime ou regroupe des mots (lemmatisation sommaire empirique); **VISUTEXT** réalise une analyse des correspondances simples d'une table lexicale (**voir chapitre III**); **VISURESP** réalise une analyse directe de réponses ouvertes.



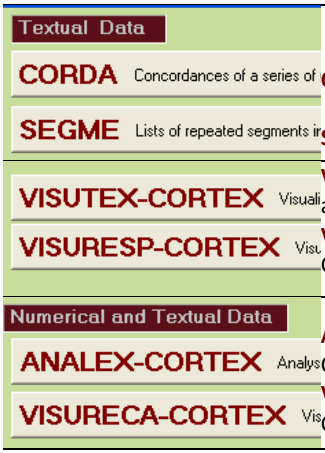
Analyse statistique exploratoire de questions ouvertes (**voir chapitre III**)*:* **ANALEX** réalise une analyse des correspondances simples d'une table lexicale agrégée; **VISURECA** réalise une analyse analogue à VISURESP, mais l'illustre avec des variables nominales **;** **MCA-TEXT :** Analyse des correspondances multiples (variables nominales), classification illustrées par les variables lexicales.

D'autres techniques d'analyse textuelle sont proposées dans le menu



➢ Si l'on clique sur ce bouton, une nouvelle fenêtre apparaît.

Les analyses CORDA et SEGME fournissent des concordances et des segments répétés, alors que les analyses suivantes incluent directement la phase CORTEX (corrections de textes) au sein des analyses VISUTEX, VISURESP, VISURECA, ANALEX.



CORDA fournit les concordances d'une liste de mots.

SEGME donne les listes de segments répétés.

VISUTEXT-CORTEXT réalise l'analyse VISUTEX précédente, après correction de textes similaire à CORTEX.

VISURESP-CORTEXT réalise l'analyse VISURESP après CORTEX.

ANALEX-CORTEXT réalise simultanément les procédures CORTEXT et ANALEX

VISURECA-CORTEXT réalise simultanément les procédures CORTEXT et VISURECA

On pourrait réaliser dans un premier temps la phase CORTEX, puis les analyses précitées. Mais CORTEX porte sur l'ensemble du fichier texte, alors que l'on peut souhaiter corriger individuellement chaque question ouverte. De plus, les réponses modales, réponses caractéristiques de chaque texte, seront les réponses originales, et non les réponses avec des mots corrigés. Mais la sélection statistique des réponses caractéristiques se fait bien, elle, sur les textes corrigés.

*
* *

Une fois le fichier de commande créé lors de la procédure Create, il est possible, toujours dans la rubrique : **Command File**, d'ouvrir directement ce fichier (bouton: Open an existing command file) pour en modifier directement certains paramètres, puis de l'exécuter (bouton: Execute). Les procédures d'analyses exploratoires de données numériques ou textuelles impliquent l'enchainement de plusieurs techniques, Analyse factorielle, Classification, Cartes de Kohonen, Validation Bootstrap. Les résultats des analyses de base peuvent être soit consultés dans la rubrique : **Result Files** (**Basic numerical results**) en navigant sur un fichier Html ou en format texte (**text format**), soit visualisés par les différents outils de la rubrique ***VIC - Visualization, Inference, Classification***.

I.3 Visualisation des résultats

Dans l'étape, ***VIC - Visualization, Inference, Classification***, une série d'outils de visualisation permettent de valider les résultats et de faciliter leur interprétation (cf. chapitres II et III).
Pour utiliser un de ces outils, Cliquer sur le menu correspondant :

– AxesView : axes factoriels.
Classements, pour chaque axe, des coordonnées des individus, des variables actives, supplémentaires, etc. pour une évaluation rapide des résultats de l'analyse factorielle.

– PlaneView : plans factoriels.
Description des plans factoriels pour tous les types d'éléments impliqués dans les analyses.

– Bootstrap : Bootstrap (BootstrapView).
Zones de confiance (ellipses ou enveloppes convexes) dans les plans factoriels pour les éléments sélectionnés.

– Seriation : sériation.
Les lignes et les colonnes de la table de contingence sont réordonnées selon le premier axe de l'analyse des correspondances de la table.

[Les techniques de Sériation sont fondées sur des permutations simples de lignes et de colonnes de la table étudiée ; elles ont l'avantage pratique et cognitif de montrer les données brutes à l'utilisateur et donc de lui éviter l'utilisation de règles de lecture complexes. Ces permutations peuvent montrer les blocs homogènes de valeurs élevées ou au contraire, de valeurs petites ou nulles. Elles peuvent également indiquer exactement une évolution continue et progressive des profils. Une propriété optimale de l'analyse de correspondance est la suivante : le premier axe d'une analyse de correspondance fournit un ordre optimal des points-ligne et des points-colonne.]

ClusterView : projection des classes de la classification sur les plans factoriels.
Représentation des positions des centres de classes dans le plan factoriel. Description des éléments caractéristiques de la classe correspondante (variables numériques, catégories, et également mots ou réponses dans le cas des questions ouvertes).

– Kohonen Map : cartes de Kohonen.
Cartes auto-organisées des individus, des variables, et simultanées des individus et des variables à partir des coordonnées factorielles (Grilles carrées de dimensions 3 x 3 à 20 x 20).

– Visualization : Outils complémentaires de visualisation.
Visualisations complémentaires des plans factoriels et de la classification. Ellipse de densité ou enveloppes convexes des classes. Tracé de l'arbre de longueur minimal, des plus proches voisins dans les plans factoriels. Visualisation pédagogique de la construction progressive des classes (cas de la procédure k-means / nuées dynamiques). Visualisation dans les plans factoriels des grilles de Kohonen et de certains graphes.

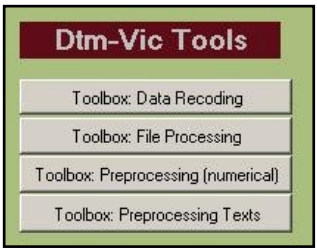
– Contiguity : analyse de contigüité.
Analyse locale, structure de graphe.

L'analyse de Contigüité relève des techniques d'analyse locale qui sont présentées au chapitre 8 de l'ouvrage précité "Statistique exploratoire multidimensionnelle". Elle considère le cas où les observations ont une structure de graphe a priori, mais aussi lorsque le graphe est intrinsèque (graphe des plus proches voisins, par exemple). Elle généralise l'analyse discriminante de Fisher (qui correspond au cas particulier du graphe associé à une partition) .

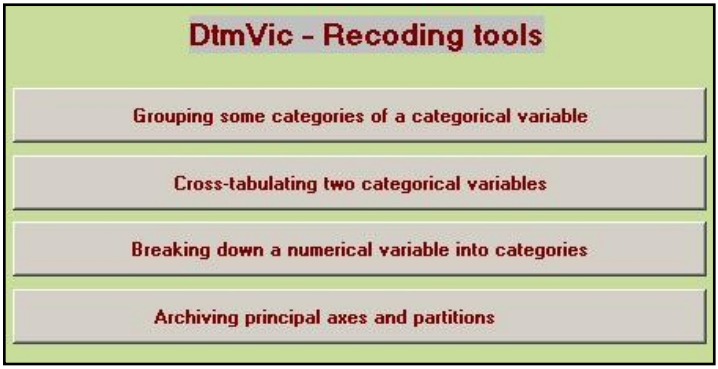
L'analyse de contiguïté est abordée dans ce manuel de prise en main dans la section VI.2 du chapitre VI.

I.4. La boîte à outils

La boîte à outils, **DtmVic Tools**, propose différents types de recodage, de stockage et de transformation des données (cf. chapitre V).



- Cliquez sur Toolbox Data Recoding
  - Le premier menu qui apparaît concerne le recodage des données et l'archivage de certains résultats.

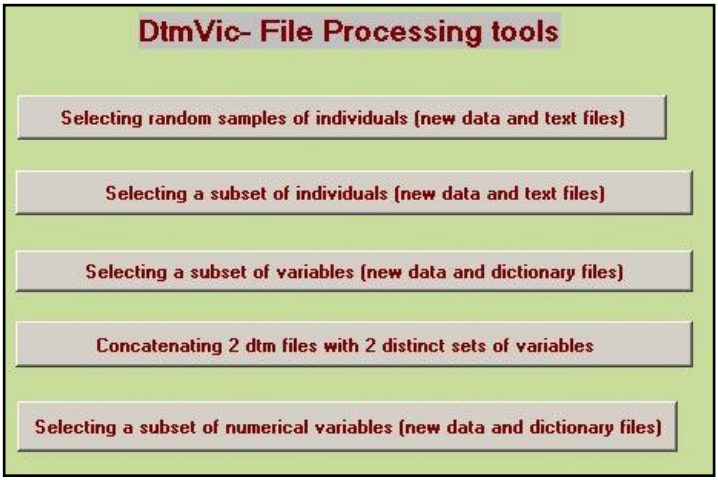


Création ou recodage de variables nominales :

i) Regroupement de modalités ;
ii) Création d'une variable nominale par croisement de deux variables nominales ;
iii) Transformation d'une variable continue en variable nominale ;

iv) Archivage des axes factoriels et des partitions.

⊙ Le second groupe d'actions concerne le menu:

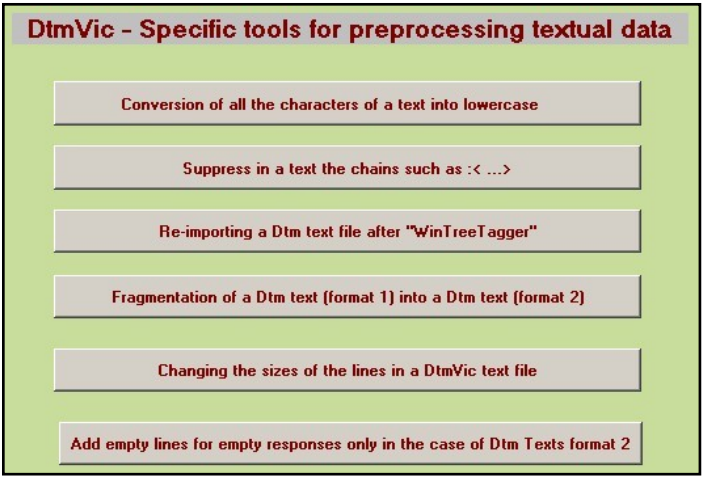


Il propose des modification de la base de données par : (Voir Chapitre V)

i) Sélection d'un sous-ensemble aléatoire d'individus (lignes) ;
ii) Sélection d'un sous-ensemble d'individus (lignes) à partir d'un filtre ;
ii) Sélection d'un sous-ensemble de variables (colonnes) ;
iii) Concaténation de deux bases de données (variables différentes).
iv) Sélection d'un sous-ensemble de variables ayant un poids maximum.

⊙ Le menu suivant propose quelques outils élémentaires de prise de contact avec les données et de prétraitements en vue de l'importation ou de l'utilisation de données numériques et textuelles.

⊙ Le dernier menu propose quelques procédures en vue de l'importation ou de l'utilisation directe des textes.



i) Conversion en minuscules des textes.

ii) Suppression des balises « < » et « > » et du texte qu'elles peuvent contenir.

iii) Ré-importation dans DtmVic d'un fichier de type Dtmic (type 1 ou 2) ayant été soumis au logiciel (gratuit) **TreeTagger**. Ceci permet de lemmatiser un texte en supprimant certaines catégories grammaticales (prépositions, articles, ...). Valable pour les textes anglais, français, espagnols, italiens.

iv) Fragmentation d'une série de textes en format 1 (textes séparés par ****) en textes de format 2, formés de une ligne, deux lignes... des textes initiaux (approximativement : fragmentation en unités de contexte). Une variable nominale est créée pour conserver l'information rattachant les unités aux textes initiaux.

v) Changement de longueur des lignes de texte. Au départ, format DtmVic (1 ou 2) sans limitation pour la longueur des lignes. A la fin : textes ayant des lignes d'une longueur choisie par l'utilisateur (mais < 200 caractères). Cette procédure permet d'importer des textes aux lignes très longues, mais aussi de formater les unités de contexte (cf. point iv ci-dessus).

vi) Cette dernière procédure limitée et spécialisée permet de faire

respecter la contrainte « une ligne vide par réponse ouverte vide » pour des fichiers qui utiliseraient deux séparateurs consécutifs.

La rubrique **DtmVic Images**, essentiellement pédagogique, montre les possibilités de compression d'images offertes par l'analyse de correspondances ou simplement par la décomposition aux valeurs singulières (section VI.4 du chapitre VI).

I.5. Format interne des données Dtm-Vic

[Version anglaise de cette section affichée par le bouton Data Format *du menu principal].*

A ce stade, il est utile de connaître le format interne des fichiers d'entrée de Dtm-Vic. Ces formats seront générés par les procédures d'importation. Trois fichiers, en format texte, constituent le format de Dtm-Vic :

Note : les noms des fichiers sont libres, mais l'extension .txt est commode pour une consultation rapide du contenu des fichiers[4].

– Exemple_dic.txt : le fichier dictionnaire fournit les noms des variables numériques et nominales. Il inclut les libellés des catégories correspondant à chaque variable nominale (cf tableau 1).

Note : les identifiants des variables et les libellés des catégories ne doivent pas contenir d'espaces vides (blancs). Ils sont par ailleurs parfois tronqués à 8 caractères dans les représentations visuelles.

– Exemple_dat.txt : le fichier de données contient les valeurs de ces variables pour un ensemble d'individus (ou : observations), ainsi que les identifiants des individus (cf tableau 2).

– Exemple_tex.txt : deux types de fichiers textes sont considérés. Un format de fichier des textes simples (type 1) peut être employé lorsqu'on traite une série de textes (cf tableau 3), sans fichier dictionnaire ni fichier de données associés. Lorsque les textes sont nombreux et qualifiés, cas des réponses à des questions ouvertes, on introduit deux niveaux de séparateurs (Fichier type 2, cf tableau 4).

Un cas d'application qui montre toutes les possibilités du logiciel est un

[4] Ces fichiers, en format texte (extension ".txt"), sont lisibles par le "bloc – notes" ou un éditeur de texte (TotalEdit, notepad, notepad++, UltraEdit, etc.), ou par l'éditeur de texte de Dtm-Vic actionné par le bouton "Open" du menu principal.

recueil de données d'enquête par sondage, comportant des réponses aux questions fermées et des réponses aux questions ouvertes. Les questions fermées peuvent donner lieu à des variables continues (ou encore quantitatives) ou à des variables nominales (ou qualitatives).

Le tableau 1 donne un exemple d'un fichier dictionnaire au format Dtm-Vic présentant quatre variables (trois nominales et une continue).

```
   2 GENDER    (nombre de catégories [2] en col. 1-4; blanc; intitulé)
MALE MALE      (identif. courts [col. 1-4]; blanc; identificateur
FEMA FEMALE    (identif. courts [col. 1-4]; blanc; identificateur
   0 AGE       (nombre de catég. [0] en col. 1-4; blanc;  var numér.)
   4 AGE_CODE  (nombre de catégories [2] en col. 1-4; blanc; intitulé)
AGE1 18_24    (identif. courts [col. 1-4]; blanc; identif. [< 20 car.]
AGE2 25_39    (identif. courts [col. 1-4]; blanc; identif. [< 20 car.]
AGE3 40_59    (identif. courts [col. 1-4]; blanc; identif. [< 20 car.]
AGE4  >60     (identif. courts [col. 1-4]; blanc; identif. [< 20 car.]
   3 EDUCATION (nbre de catégories [3] en col. 1-4; blanc; intitulé)
EDUL LOW       (identif. courts [col. 1-4]; blanc; identificateur
EDUM MEDIUM    (identif. courts [col. 1-4]; blanc; identificateur
EDUH HIGH      (identif. courts [col. 1-4]; blanc; identificateur

    [Les identificateurs ont moins de 20 caractères. Jamais de blanc
à l'intérieur d'un identificateur]
```

Tableau 1: Fichier dictionnaire en format interne Dtm-Vic pour quatre variables
Sexe (2 modalités), âge (0 modalité = variable continue), classe d'âge (4 modalités), niveau d'éducation (3 modalités). *(Les commentaires en italique donnent les explications du format fixe du fichier dictionnaire)*

Le tableau 2 donne l'exemple d'un fichier de données de Dtm-Vic correspondant aux 4 variables du fichier dictionnaire précédent pour 5 individus (sujets, observations ou répondants).

```
'n1006'   1  76  4    1   (Identificateur de l'observation : entre
'n1007'    2  20   1 2     quotes, sans blanc, < 20 caractères.
'n1008'   2  29   2  3     Separateurs entre valeurs: au moins un
'n950'     1  57  3  1      espace blanc)
'n2007'    1  21   1   2
```

Tableau 2: Fichier de données en format interne Dtm-Vic
Pour 5 individus (sujets ou observations) correspondant aux 4 variables du dictionnaire précédent : Sexe , Age, Age éclaté en 4 modalités, niveau d'éducation (cf tableau 1). Longueur maximale d'une ligne : 5000 caractères. *(commentaire du format en italique)*

Le tableau 3 donne l'exemple d'un fichier texte en format interne Dtm-Vic pour une série de trois textes (cf. exemple III.1 – poèmes).

```
****    LAMARTINE
Voilà les feuilles sans sève,
Qui tombent sur le gazon
Voilà le vent qui s'élève,
Et gémit dans le vallon
Voilà l'errante hirondelle,
Qui rase du bout de l'aile,
L'eau dormante des marais...
****    GAUTIER
L'automne va finir, au milieu du ciel terne,
Dans un cercle blafard et livide que cerne
Un nuage plombe, le soleil dort. Du fond
Des étangs remplis d'eau monte un brouillard qui Fond
Collines, champs, hameaux dans une même teinte.
.
****    VERLAINE
Les sanglots longs
Des violons
De l'automne
Blessent mon coeur
D'une langueur
Monotone.
=====
```

Tableau 3: Fichier texte en format interne (type 1) Dtm-Vic.

Les trois textes sont en format libre sur moins de 200 colonnes; les séparateurs des textes sont séparés par "****" suivis de 4 espaces puis de l'identifiant du texte comportant moins de 20 caractères; la fin du fichier est mentionné par "====". Tous les séparateurs occupent les 4 premières colonnes. Pour certaines éditions de tableaux, il est utile et important que les 4 premiers caractères de l'identifiant de texte caractérisent le texte. Si les lignes ont plus de 200 ccaractères, une procédure de Dtm-Vic-Tools permet de les reformater.

Le tableau 4 (plus bas) présente un fichier de textes concernant trois questions ouvertes pour trois répondants (cf. l'exemple III.2).

Pourquoi deux formats pour les données textuelles ? Contrairement aux données numériques, les textes peuvent poser des problèmes d'échelle, de dimensions, et donc de limites.

- Le format type 1 (séparateurs ****) permet d'accueillir des textes fort longs, par exemple les romans de la Comédie humaine de Balzac. Chaque texte peut être long, mais le nombre de texte est ici limité à 1200.
- Le format de type 2 (Séparateurs ---- [pour les observations] puis ++++ [pour les questions ouvertes, dont le nombre est limité à 12]) correspond au fichier d'enquête (le nombre de textes doit être alors inférieur à 30000, limite du nombre d'observations de Dtm-Vic dans la version actuelle). Le texte total d'un individu est alors limité à 100000 caractères.

```
---- 1006
 my sons, my kids are very important to me,
being on my own I am responsible for their education
++++
 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
++++

++++

====
```

Tableau 4: Fichier texte de questions ouvertes en format interne Dtm-Vic (type 2)

Trois individus ont répondu à trois questions ouvertes. Le format est libre sur 200 colonnes. Le séparateur entre les individus est "----" suivi par l'identifiant de l'individu (moins de 20 caractères); les questions sont séparées par "++++"; la fin du fichier est mentionné par "====". Tous les séparateurs occupent les 4 premières colonnes. Note : les lignes vides correspondent à des non-réponses (le dernier répondant n'a pas donné de réponse aux deux dernières questions ouvertes : au moins une ligne vierge est nécessaire dans ce cas). Attention : l'ordre des individus doit être celui du fichier de données numériques. Noter que la limitation est de 12 questions ouvertes par fichier texte, mais il peut y avoir plusieurs fichiers.

Notons que dans l'importation d'un fichier Excel contenant à la fois des variables numériques et textuelles, chaque réponse à une question ouverte est limitée à 8000 caractères.

Dans les exemples fournis dans Dtm-Vic, les fichiers sont déjà en format Dtm-Vic (sauf bien sur les exemples d'importation). La mise en forme dans le format de Dtm-Vic est alors inutile pour l'utilisateur.

Rappel : Dtm-Vic produit de nombreux fichiers de résultats intermédiaires liés à l'application (tous en format .txt). ***Il est, par conséquent, recommandé de créer un répertoire par application***. Au départ, un tel répertoire doit contenir les fichiers de données, dictionnaire ou texte au format Dtm-Vic.

II. Données numériques : Prise en main de Dtm-Vic à partir de trois exemples

Les exemples suivants visent à présenter Dtm-Vic à l'utilisateur d'une façon pragmatique. Ils correspondent à un dossier inclus dans le dossier **DtmVic-Examples_A_Start** qui a été téléchargé avec le logiciel Dtm-Vic. Chaque exemple rend compte d'un jeu de données adapté à une des analyses factorielles de base (Analyse en Composantes Principales, Analyse simple des Correspondances, Analyse des Correspondances Multiples) enrichie par des outils complémentaires (bootstrap, classification, cartes de Kohonen, sériation).:

1. L'exemple 1, contenu dans le dossier **EX_A01.PrinCompAnalysis**, est une analyse en composantes principales appliquée à un ensemble de variables continues : prise en compte de variables actives et supplémentaires; validation *Bootstrap* ; classification des individus et description des classes.
2. L'Exemple 2, contenu dans le dossier **EX_A02.SimpleCorAnalysis**, présente une analyse des correspondances simples adaptée à l'analyse d'un tableau de contingence : variables actives et supplémentaires ; validation *Bootstrap*.
3. L'Exemple 3, contenu dans le dossier **EX_A03.MultCorAnalysis**, porte sur l'analyse des correspondances multiples appliquée à un ensemble de variables nominales issues de données d'enquêtes : variables nominales actives, supplémentaires, variables continues; validation *Bootstrap* ; classification des individus et description des classes obtenues.

Rappel : Dtm-Vic produit de nombreux fichiers de résultats intermédiaires liés à l'application (tous en format .txt). ***Il est, par conséquent, fortement recommandé de créer un répertoire par application***. Au départ, un tel répertoire doit contenir les fichiers de données nécessaires à l'analyse au format Dtm-Vic, décrits dans le paragraphe I.5.

II.1. Analyse en Composantes Principales (ACP ou *PCA*)

Ce premier exemple (voir répertoire DtmVic-Examples_A_Start/ EX_A01.PrinCompAnalysis) vise à décrire un ensemble de variables continues par l'Analyse en Composantes Principales.

II.1.1. Les données et fichiers Dtm-Vic : Enquête "budget-temps"

Les données sont extraites d'une *Enquête Budget-temps Multimédia* effectuée par le Centre d'Étude des Supports de Publicité (www.cesp.org) en 1992 auprès de 18000 personnes. Ont été relevés le temps passé à diverses activités quotidiennes (travail, loisirs, déplacements, repas, repos, …) soit 39 activités (de V6 à V44) ainsi que le temps de fréquentation de divers médias (radio, télévision, presse) soit 5 médias (de V45 à V49). Le temps est exprimé en minutes par jour. Il est mesuré le jour précédant l'entrevue. Ont également été relevées les caractéristiques socio-économiques du répondant telles que l'âge, le sexe, l'activité, le niveau d'éducation et le lieu de résidence correspondant à 5 variables nominales (de V1 à V5). Les 18 000 répondants originaux sont groupés selon les combinaisons de cinq caractéristiques socio-économiques produisant 96 groupes qui constituent en quelque sorte des "répondants artificiels".

Le tableau de données de cet exemple dispose en ligne les 96 catégories de répondants et en colonne les 5 caractéristiques de base, le genre, l'âge, l'éducation et l'agglomération de résidence (soit 5 variables nominales), les 38 "activités" quotidiennes et 5 "fréquentation média" (soit 43 variables continues). A la croisée de la ligne i et de la colonne j est mentionné, après l'identificateur de l'individu, le cumul du temps passé (en minutes par jour) pour l'activité j par les individus de la catégorie i.

L'objectif est de définir les associations entre les différentes activités considérées comme variables actives et d'étudier le lien entre ces associations et la fréquentation des médias et aussi les caractéristiques socio-économiques (considérées comme variables supplémentaires).

A partir d'un fichier de type *Excel*, deux fichiers en format Dtm-Vic, sont importés. Ils sont contenus dans le dossier EX_A01.PrinCompAnalysis. Ils peuvent être ouverts avec un éditeur de texte (bloc note, notepad, Ultraedit, TotalEdit, Notepad++, ou l'éditeur de texte interne de Dtm-Vic).

	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	

Tableau de données "Budget-temps" (premières lignes)

1. Le fichier dictionnaire : PCA_dic.txt

Ce fichier est accessible dans le dossier en français (PCA_dic_Fr.txt) et en anglais (PCA_dic_Eng.txt). Il contient les identifiants des 44 variables et des catégories (ou modalités) des variables nominales.

```
...2.Genre_V1          ...0.Sommeil_V6          ...0.Déma_Cours_V26
 Fem Sex_Fem_1            0 Repos_V7               0 Promenad_V27
 Hom Sex_Hom_2            0 Toilette_V8            0 Courses_V28
   3 Age_V2               0 Repas_V9               0 Déplacem_V29
AMoy Age_Moy_1            0 Petit_Déj_V10          0 A_pied_V30
Ages Age_Ages_2           0 Repas_home_V11         0 En_Voitu_V31
Jeun Age_Jeun_3           0 Repas_rest_V12         0 Fréquent_V32
   2 Activité_V3          0 Travail_V13            0 Autres_a_V33
acti Act_acti_1           0 TravailR_V14           0 Total_Do_V34
inac Act_inac_2           0 Enfants_V15            0 Total_Dé_V35
   3 Education_V4         0 Ménage_V16             0 Total_ho_V36
prim Educ_prim_1          0 Relation_V17           0 Total_Me_V37
 sec Educ_sec_2           0 Visite_amis_V18        0 Radio_V38
 sup Educ_sup_3           0 Loisirs_V19            0 TV_V39
   5 agglome_V5           0 Jeux_Jar_V20           0 Presse_V40
VImp aggl_Imp_1           0 Jardinag_V21           0 Quotid_N_V41
VMoy aggl_Moy_2           0 Loisirs_ext_V22        0 Quotid_R_V42
CRur aggl_Rur_3           0 Disque_V23             0 Magazine_V43
Mixt aggl_Mixte_4         0 Lecture_V24            0 Mag_TV_V44
APar aggl_Paris_5         0 Lect_livr_V25
```

L'identifiant d'une variable nominale est précédé par le nombre N de ses modalités (colonne 5). Les N lignes suivantes sont les N modalités de réponses : un "identifiant court" en 4 caractères occupe les colonnes 1 à 5 et un "identifiant long" (<20 caractères) commence colonne 6. Conventionnellement, une variable numérique a zéro catégorie. Les espaces vides sont interdits dans les identifiants.

2. Extraits du fichier de données PCA_dat.txt

```
'1111' 1. 1. 1. 1. 1. 463.80 23.80 26.30 139.00 16.00
'1115' 1. 1. 1. 1. 5. 515.60 58.50 19.20 138.30 13.50
'1121' 1. 1. 1. 2. 1. 463.30 34.20 28.40 126.30 16.20
'1122' 1. 1. 1. 2. 2. 456.40 43.10 29.30 118.40 15.10
'1123' 1. 1. 1. 2. 3. 478.00 44.20 28.80 115.40 15.00
'1124' 1. 1. 1. 2. 4. 465.10 41.60 30.30 135.70 17.40
'1136' 1. 1. 1. 3. 5. 458.40 47.40 28.10 133.30 15.50
'1133' 1. 1. 1. 3. 3. 457.20 30.70 25.80 137.00 17.80
'1134' 1. 1. 1. 3. 4. 465.20 40.20 28.80 136.30 16.80
'1221' 1. 1. 2. 2. 1. 523.90 41.80 26.10 112.20 15.20
```

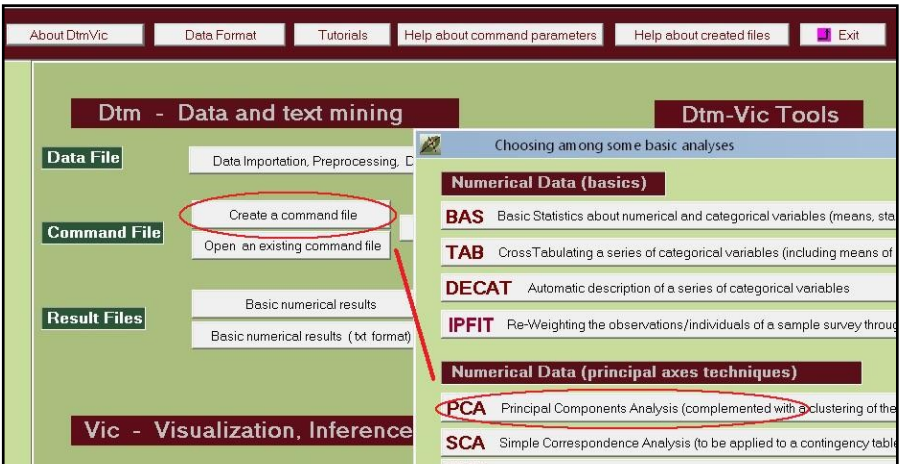
Ce fichier de données comprend 96 lignes et 45 valeurs. Pour une ligne i, la première valeur (entre *quotes*) correspond à l'identifiant de l'individu i, c'est-à-dire ici le groupe i de répondants, et les 44 autres valeurs correspondent aux réponses des 44 variables séparées par des espaces blancs : les 5 premières valeurs sont les items des 5 variables nominales (genre, âge, activité, éducation, agglomération de résidence qui sont à la base de la formation des groupes), les 32 autres valeurs correspondent aux cumuls du temps passé (minutes par jour) dans les activités par tous les individus constituant le groupe i, et les 7 dernières valeurs correspondent aux cumuls du temps passé au contact d'un média.

II.1.2. Mise en œuvre de l'analyse (*PCA*)

Le fichier paramètre est créé en 5 étapes :

Etape 1: Sélection de l'analyse

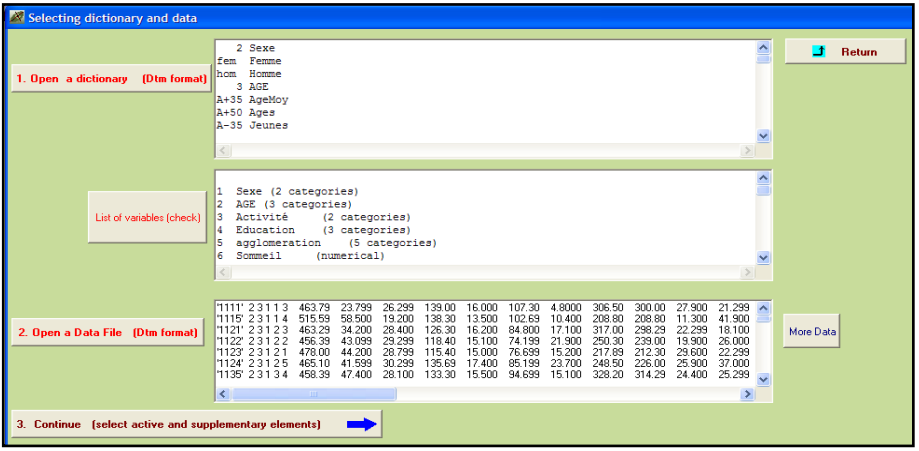
- Cliquez sur le bouton **Create a Command file** de **Command File**
  - ⊙ Une fenêtre *"Choosing among some basic analyses"* apparaît.
- Sélectionnez l'analyse : **PCA** – Principal Components Analysis dans la rubrique **numerical data (principal axes techniques)**.



- Une fenêtre "*Selecting dictionary and data*" apparaît.

Etape 2 : Sélection des fichiers dictionnaire et données

➢ Cliquez sur le bouton **Open a dictionary**. Dans le répertoire EX_A01.PrinCompAnalysis, ouvrir le fichier **PCA_dic.txt**. Il s'affiche dans une première fenêtre. Le statut (nominal [*categorical*] ou numérique) des variables est indiqué dans une deuxième fenêtre



① Cliquez sur le bouton : **Open a Data File**. Dans le répertoire DtmVic_Examples_A_Start \EX_A01.PrinCompAnalysis, ouvrir le fichier

PCA_dat.txt qui s'affiche dans une troisième fenêtre.

➢ Cliquez sur : **3. Continue** ➔ . Une fenêtre "*Selection of active et supplementary elements*" apparaît alors.

Etape 3 : Sélection des variables actives et supplémentaires

A l'intérieur de la fenêtre "*Selection of active et supplementary elements"* s'affichent trois autres fenêtres :

4. "*Variables to be selected*" où figure l'ensemble des variables
5. "*Active Variables*" qui reçoit les variables actives sélectionnées
6. "*Supplementary Variables*" qui reçoit les variables supplémentaires sélectionnées

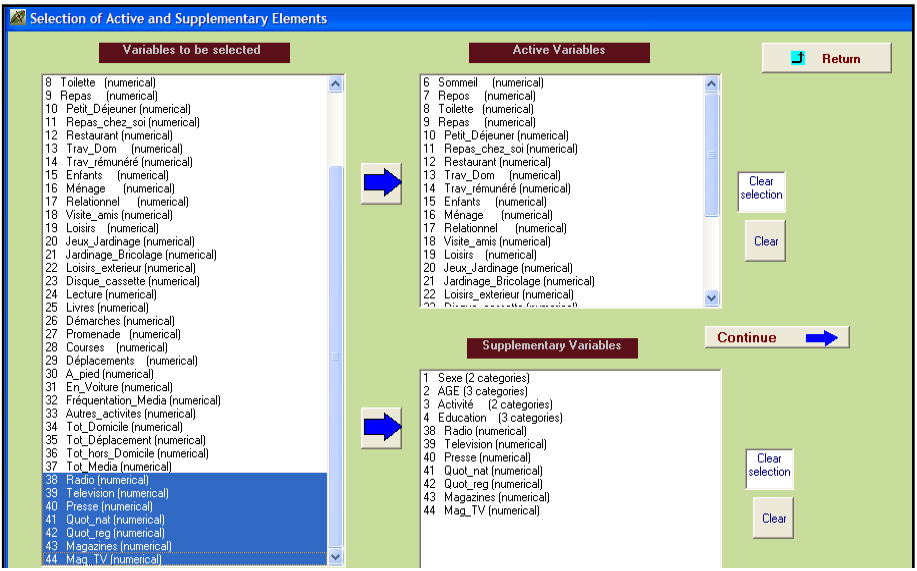
Pour l'ACP, les variables actives doivent être continues (*numerical*). Les variables supplémentaires peuvent être continues ou nominales. Nous proposons de sélectionner les variables suivantes :

- Sélection des variables continues actives : V6 à V32 à transférer dans la fenêtre intitulée "*Active Variables"* :

6. Sommeil_V6	15. Enfants_V15	24. Lecture_V24
7. Repos_V7	16. Ménage_V16	25. Lect_livr_V25
8. Toilette_V8	17. Relation_V17	26. Démarche_Course_V26
9. Repas_V9	18. Visite_amis_V18	27. Promenad_V27
10. Petit_Déj_V10	19. Loisirs_V19	28. Courses_V28
11. Repas_home_V11	20. Jeux_Jar_V20	29. Déplacem_V29
12. Repas_rest_V12	21. Jardinag_V21	30. A_pied_V30
13. Travail_V13	22. Loisirs_ext_V22	31. En_Voitu_V31
14. TravailR_V14	23. Disque_V23	32. Fréquent_V32

Sélection des variables supplémentaires à transférer dans la fenêtre "*Supplementary Variables"*

variables continues supplémentaires : V38 à V44	38. Radio 39. TV 40. Presse 41. Quotid_N	42. Quotid_R 43. Magazine 44. Mag_TV
variables nominales supplémentaires : V1 à V4	1. Sexe 2. Age	3. Activité 4. Education



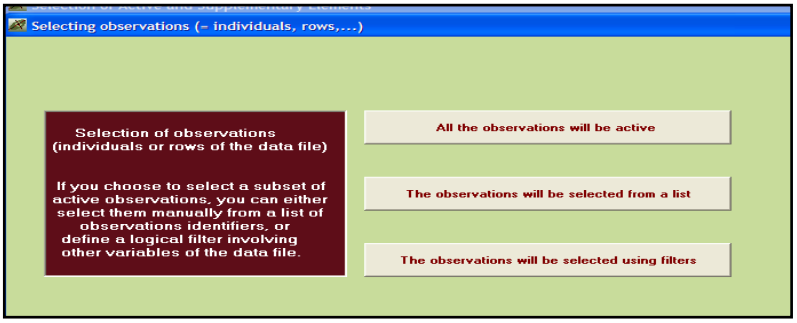
➢ Cliquez sur : **Continue** ➔

4.1. Une fenêtre "*Selecting observations*" apparaît.

Etape 4 : Sélection des observations (individus)

Trois cas de figure sont possibles :

- Considérer l'ensemble des observations
- Sélectionner les observations sur une liste
- Sélectionner les observations par un filtre



Nous prenons en compte ici l'ensemble des observations.

Cliquez sur: **All the observations will be active**

⊙ une fenêtre "*Create a starting parameter file*" apparaît.

Etape 5 : Création du fichier de commande (fichier paramètre)

A cette étape, il est possible de sélectionner, comme option, les procédures de bootstrap et/ou de classification. En effet, dans Dtm-Vic, les analyses factorielles peuvent être complétées par :

– une procédure de *bootstrap* qui permet de valider la position des variables sur le plan factoriel
– et/ou une classification avec une description automatique des classes.

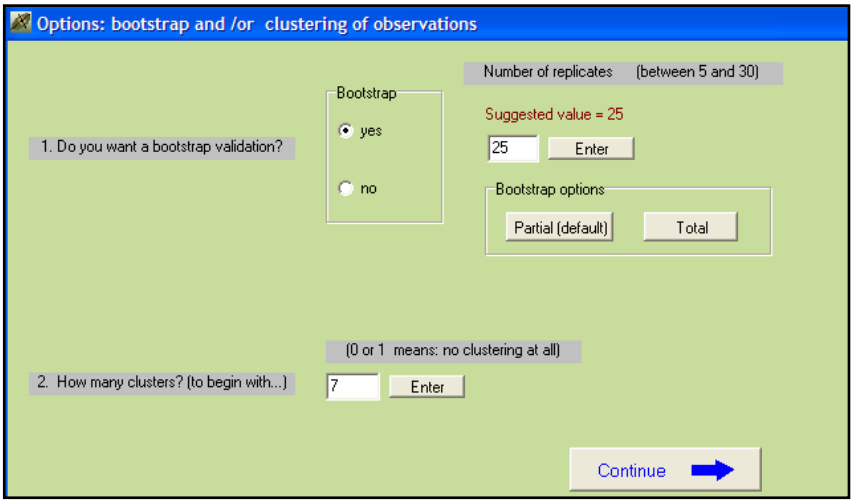
a. Sélection d'une option

- Cliquez sur **1-Select some options**

  ⊙ une fenêtre "*Options : Bootstrap and/or Clustering of observations*" apparaît.

➢ Cliquez sur : "yes" pour la procédure "bootstrap" ; indiquer le nombre de réplications (par défaut 25) puis **enter**. C'est le bootstrap partiel qui est appliqué par défaut. Si le bootstrap n'est pas adopté, cliquez sur : "no".

➢ Sélectionnez le nombre de classes souhaité (nous suggérons 7 classes) puis cliquez sur **enter**



Note technique : Les différents types de *bootstrap* pour variables non-textuelles dans Dtm-Vic.

a _ *Bootstrap* partiel pour les variables actives

Avec ce type de *bootstrap*, le plan initial sert d'espace de référence pour accueillir les réplications, qui sont projetées comme des variables supplémentaires. Le *bootstrap* partiel n'a pas pour vocation de valider la stabilité de l'espace de départ qui n'est pas remis en question. Il donne une idée de la variabilité imputable aux réplications pour chaque point-modalité pris isolément.

b _ *Bootstrap* partiel pour les variables supplémentaires

Pour les variables supplémentaires, le bootstrap ne peut être que partiel. Il s'agit d'une validation externe, et donc d'un test statistique parfaitement légitime, ces variables n'ayant pas participé à la construction du sous-espace de référence.

c _ *Bootstrap* total pour les variables actives

Rappelons que dans ce cas, chaque réplication donne lieu à une analyse en composantes principales spécifique. Il existe trois implémentations du *bootstrap* total dans Dtm-Vic.

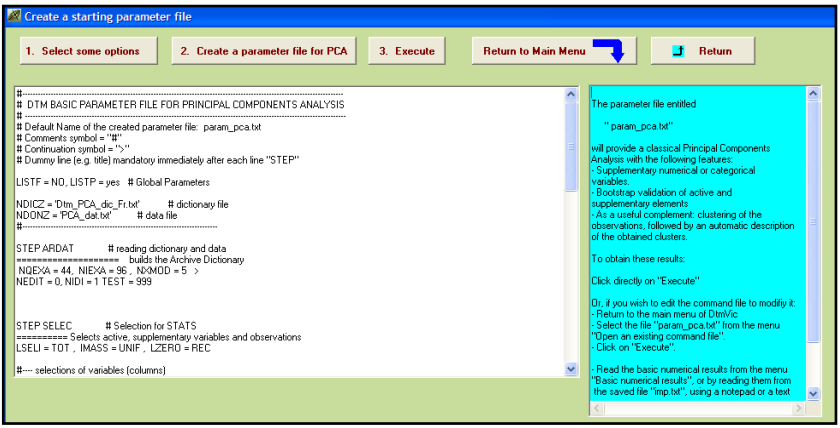
- Le *bootstrap* de type 1 (simples corrections du signe des axes pour les analyses des réplications).
- Le *bootstrap* de type 2 (corrections des interversions d'axes) est plus élaboré.
- Le *bootstrap* de type 3 (Rotations "procrustéennes" des axes répliqués de façon à les amener en correspondance avec les axes initiaux. On rejoint ainsi souvent les résultats du *bootstrap* partiel. Les options de *bootstrap* total peuvent être mises en oeuvre par les utilisateurs avancés, mais ne sont pas utilisées dans ce manuel.

➢ Cliquez sur **Continue ➔**

⊙ la fenêtre : "*Create a starting parameter file*" réapparaît.

b. Création du fichier paramètre

➢ Cliquez sur: **2-Create a parameter file for PCA.**

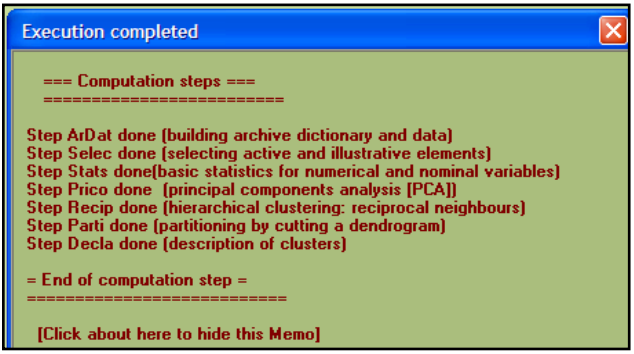


Un fichier paramètre est créé sous le nom **param_PCA.txt** dans le dossier EX_A01.PrinCompAnalysis (du dossier DtmVic_Examples_A_Start). Pour le conserver en vue d'analyses ultérieures, il sera prudent, après avoir quitté Dtm-Vic, de le renommer.

c. Exécution

➢ Cliquez sur : **3-Execute**

La séquence des procédures s'affiche en bloc après l'exécution :



Commentaires :

Ardat, (Archivage des données), Selec (Sélection des éléments actifs et supplémentaires), Stats (statistiques de base), Prico (Analyse en Composantes Principales), Recip (Classification mixte utilisant la classification ascendante hiérarchique - méthode des voisins réciproques), Parti (Coupure du dendrogramme et optimisation de la partition par la méthode des centres mobiles [*k-means*]), Decla (Description automatique des classes de la partition).

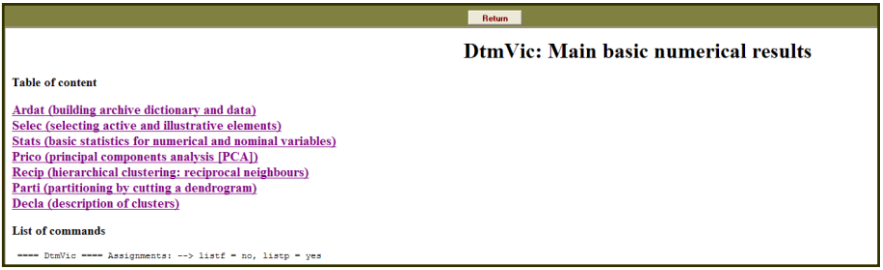
Note : Lors d'une utilisation ultérieure de Dtm-Vic, il est possible d'ouvrir le fichier paramètre param_PCA.txt dans le menu principal **Command File** avec la procédure **Open an existing command file** puis d'exécuter directement ce fichier : **Execute**.

Les utilisateurs expérimentés peuvent modifier des paramètres directement sous l'éditeur interne ou hors de Dtm-Vic avec un éditeur de texte (voir le "*Help about parameters*" disponible à partir de l'éditeur).

II.1.3 Fichier de résultats

Les résultats peuvent être consultés à partir de la rubrique : **Result Files**

- Cliquez sur : Basic numerical results pour naviguer dans le fichier de résultats, puis sur : Return pour revenir au menu principal.



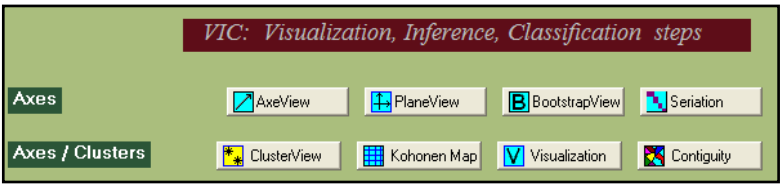
- ou cliquez sur : Basic numerical results (text format) pour ouvrir le fichier résultat en format texte.

Le fichier résultat nommé imp.txt est contenu dans le répertoire EX_A01.PrinCompAnalysis. Il est également sauvegardé sous le nom "imp" suivi de la date et l'heure de l'analyse: "imp_08.07.11_14.45.txt" signifie le 8 juillet 2011, à 14h 45. Ce fichier de sauvegarde conserve les résultats numériques principaux tandis que le fichier imp.txt est écrasé pour chaque nouvelle analyse exécutée dans le même répertoire.

Après avoir consulté les résultats numériques, revenez au menu principal. Ces résultats seront visualisés alors dans l'étape VIC de Dtm-Vic qui facilite considérablement l'interprétation (l'histogramme des valeurs propres, celui des indices de niveau et le dendrogramme doivent cependant être consultés dans l'un des fichiers imp.txt ou imp.html).

II.1.4 Visualisation des résultats

Cette deuxième phase fondamentale de Dtm-Vic fournit les outils de visualisation nécessaires à l'interprétation et la validation des résultats.



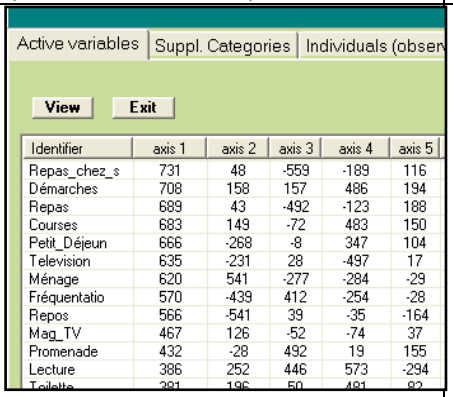
1- Axes factoriels

Cet outil fournit et classe les coordonnées sur les axes factoriels des variables actives, supplémentaires, ou des observations.

➢ Cliquez sur : AxesView.

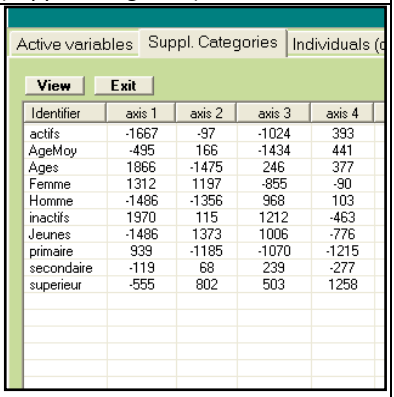
⊙ Une fenêtre propose de visualiser les coordonnées des variables actives, supplémentaires et des observations, sur les premiers axes (ces résultats sont aussi ceux de l'étape DEFAC du fichier résultat).

Coordonnées des variables continues actives et supplémentaires : (ordonnées sur l'axe 1)



Identifier	axis 1	axis 2	axis 3	axis 4	axis 5
Repas_chez_s	731	48	-559	-189	116
Démarches	708	158	157	486	194
Repas	689	43	-492	-123	188
Courses	683	149	-72	483	150
Petit_Déjeun	666	-268	-8	347	104
Television	635	-231	28	-497	17
Ménage	620	541	-277	-284	-29
Fréquentatio	570	-439	412	-254	-28
Repos	566	-541	39	-35	-164
Mag_TV	467	126	-52	-74	37
Promenade	432	-28	492	19	155
Lecture	386	252	446	573	-294
Toilette	391	196	50	481	82

Coordonnées des variables nominales supplémentaires (*Supplt categories*)



Identifier	axis 1	axis 2	axis 3	axis 4
actifs	-1667	-97	-1024	393
AgeMoy	-495	166	-1434	441
Ages	1866	-1475	246	377
Femme	1312	1197	-855	-90
Homme	-1486	-1356	968	103
inactifs	1970	115	1212	-463
Jeunes	-1486	1373	1006	-776
primaire	939	-1185	-1070	-1215
secondaire	-119	68	239	-277
superieur	-555	802	503	1258

Remarque : En cliquant sur la partie haute de l'axe 1, on identifie rapidement les oppositions visibles sur cet axe : opposition entre les activités extérieures (relation, repas au restaurant, déplacement) sur la partie positive et les activités de la maison (jardinage, repas chez soi) sur la partie négative ; sur l'axe 2, le travail rémunéré (partie positive) s'oppose au repos (partie négative)

Dans le cadre de l'analyse en composantes principales, trois éléments peuvent être examinés, les *variables continues actives* et *supplémentaires*, les *variables nominales supplémentaires* et les *observations*.

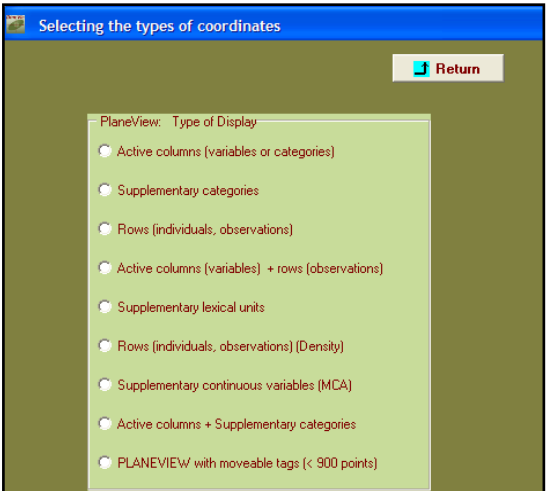
- Cliquez sur l'onglet des éléments à examiner, Active variables par exemple puis sur **View**. Il est possible d'ordonner les coordonnées sur un axe donné, en cliquant sur le libellé "axis **x**" en haut de l'axe **x**.
- Cliquez sur : **Exit** pour sortir de cet outil.

2- Plans factoriels

Cet outil fournit les plans factoriels séparés ou superposés des variables actives, supplémentaires, ou des observations.

- Cliquez sur : PlaneView

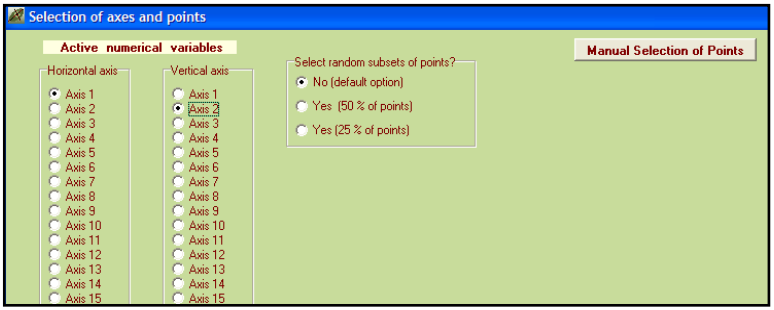
Une fenêtre propose différentes visualisations de plans factoriels.



Dans cet exemple d'analyse, six rubriques du menu sont possibles : "les colonnes actives (des variables ou des catégories)", "des catégories supplémentaires", "des lignes actives (individus, observations)", "colonnes actives + lignes actives", "individus actifs (densité)" et "colonnes actives + catégories supplémentaires". "PLANEVIEW with moveable tags" reprend certaines des rubriques précédentes et permet de déplacer les points superposés pour rendre plus lisible le graphique.

Sélectionnez la rubrique "Actives columns (variables or categories)".

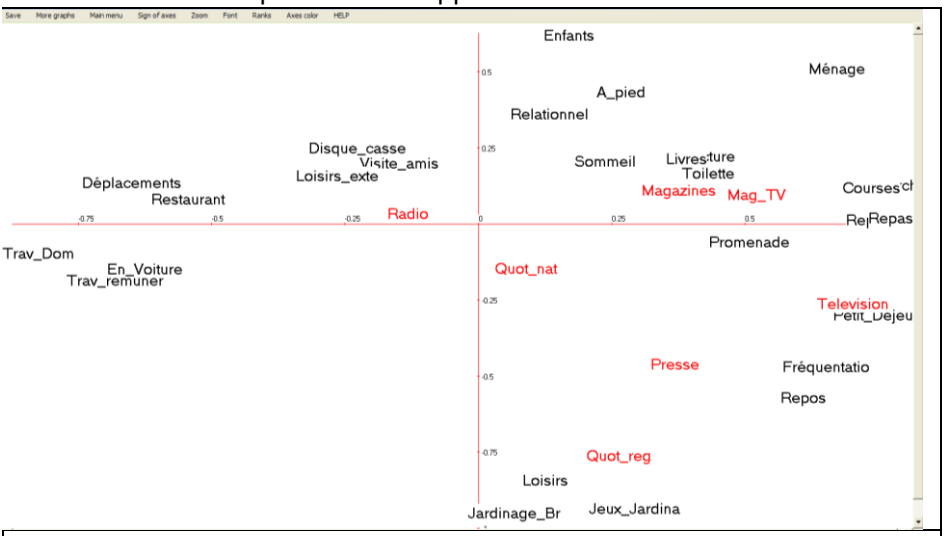
⊙ Apparaît une fenêtre pour sélectionner le plan factoriel suivant le couple d'axes souhaité.



➢ Choisir les axes 1 et 2 puis cliquez sur : **Display**. Il est possible de ne

faire figurer sur les plans que certaines variables. Cliquez alors sur : **Manual Selection of points**. Sélectionner les variables et les transférer dans la seconde fenêtre en cliquant sur : Select.

⊙ La fenêtre du plan factoriel apparaît.



Plan factoriel (1,2) – rubrique "colonnes actives (des variables ou des catégories)" : Variables continues "Activités" en actives (en noir) et variables continues "Média" en supplémentaires (en rouge)

Dans le cas de cet exemple, la première rubrique de menu "colonnes actives (variables ou catégories)" contient en fait les variables numériques actives (en noir) et des variables numériques supplémentaires (en rouge).

Note : Pour chaque graphique, le bandeau du haut contient des options :
– "Save" sauvegarde le graphique en format bmp;
– "Font" offre la possibilité de modifier la police et la couleur des caractères ;
– "More graph" permet de changer de plan factoriel
– "Sign of axes" permet d'inverser les axes ;
– "Rank", est utile seulement dans le cas des affichages très complexes, (ce qui n'est pas le cas ici) : ce bouton convertit les deux coordonnées de l'affichage courant en rangs. Par exemple, les n valeurs de l'abscisse sont converties en nombres entiers de 1 à n, ayant le même ordre que les valeurs originales. Ainsi les deux distributions sont uniformes, et les identifiants s'avèrent être beaucoup plus lisibles (au prix d'une distortion substantielle de l'affichage).

➢ Pour fermer le graphique, cliquez sur la croix en haut à droite puis sur :

Return ou directement sur la rubrique du bandeau "*Main menu*".

➢ Retournez ensuite sur : **PlaneView** pour sélectionner une autre représentation factorielle.

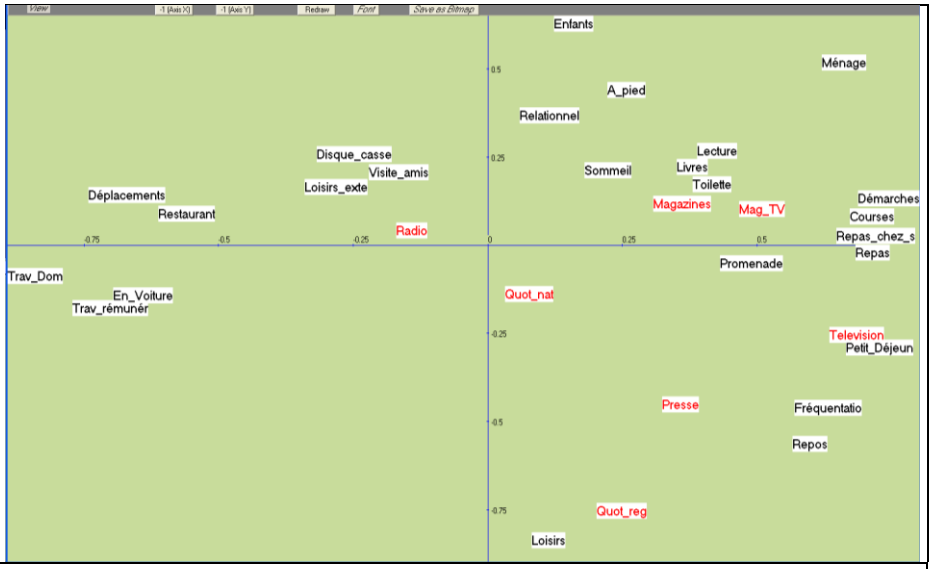
Autres graphiques :

1. rubrique : "PLANEVIEW with moveable tags" permet de déplacer les étiquettes des points du graphique.

➢ Cliquer sur "PLANEVIEW with moveable tags" puis sur Continue

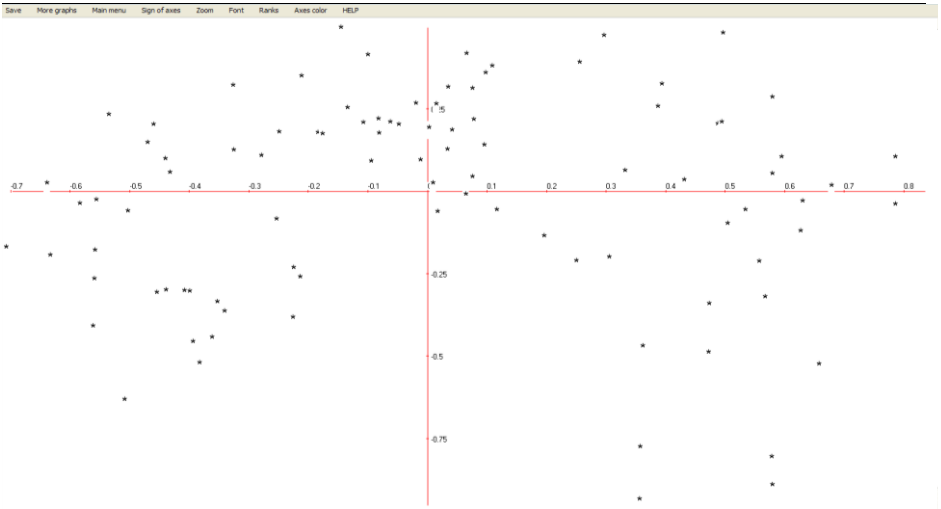
⊙ Une fenêtre apparaît.

➢ Choisir par exemple "actives columns (variables) (with continuous supplementary variables)", cliquer sur Continue et sélectionner le plan factoirel.



Plan factoriel (1,2) – rubrique "PLANEVIEW with moveable tags" puis bouton: *"actives columns (variables) (with continuous supplementary variables*"

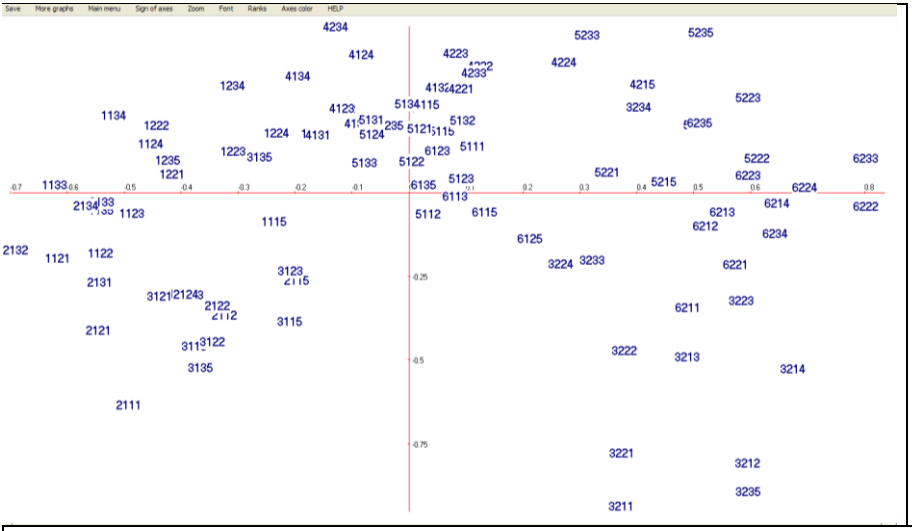
1. Rubrique "Individus actifs (densité)":



PlaneView (1,2) – Rubrique : " individus actifs (densité)"

Remarque : Les identifiants des individus sont remplacés par un caractère simple [cas de nombreux individus, plusieurs milliers par exemple]. Cet affichage montre la forme du nuage des individus et d'éventuels individus aberrants. Les identifiants d'origine peuvent s'afficher en cliquant sur le bouton droit de la souris

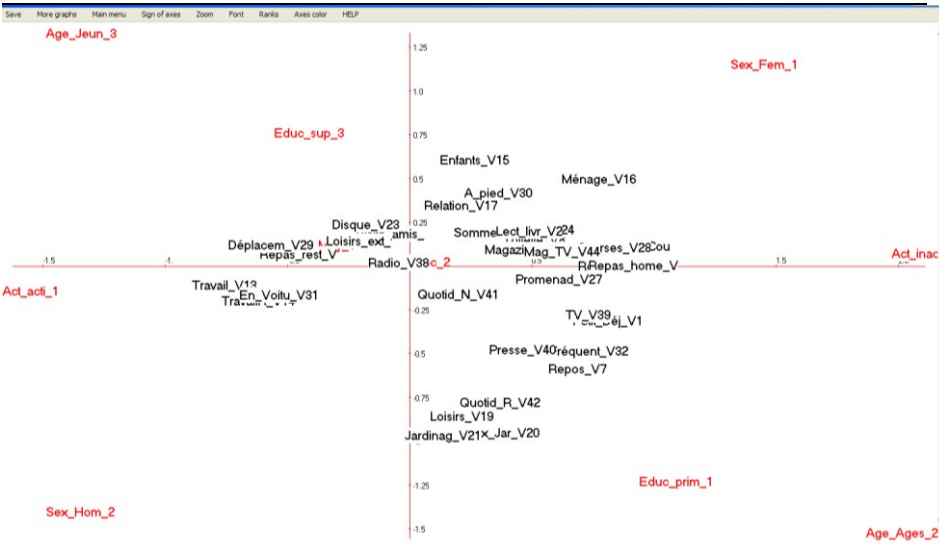
2. Rubrique " individus actifs " :



PlaneView (1,2) – rubrique "individus actifs"

Remarque : Les individus sont représentés par leur identifiants. Cet affichage est surtout intéressant lorsque les individus sont peu nombreux (< 2000).

3. rubrique "colonnes actives + catégories supplémentaires" :



Résultat – PlaneView – rubrique "colonnes actives + catégories supplémentaires"
Remarque : Sont présentes les variables continues et nominales supplémentaires)

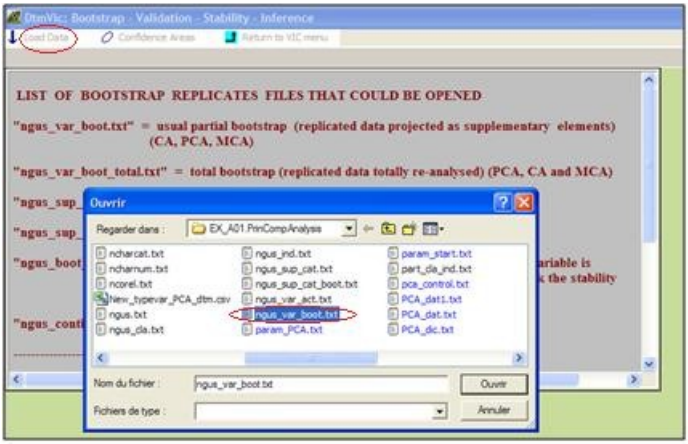
3- Validation Bootstrap

Cet outil permet de valider la position des variables sur le plan factoriel.

1. Cliquez sur : B Bootstrap

   ⊙ Une fenêtre *"DtmVic – Bootstrap – Validation – Stability - Inférence*" apparaît.

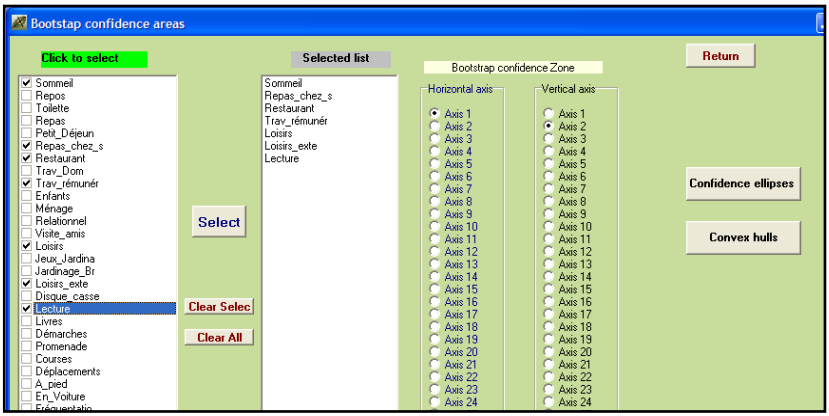
➢ Cliquer sur **Load Data** puis ouvrir dans le répertoire le fichier des réplications selon le bootstrap choisi. Sélectionnez le fichier **ngus_var_boot.txt** pour un bootstrap partiel. Répondre **OK** à la fenêtre "*Set of principal coordinates loaded*" qui s'affiche.



➢ Puis cliquez sur **Confidence Areas**.

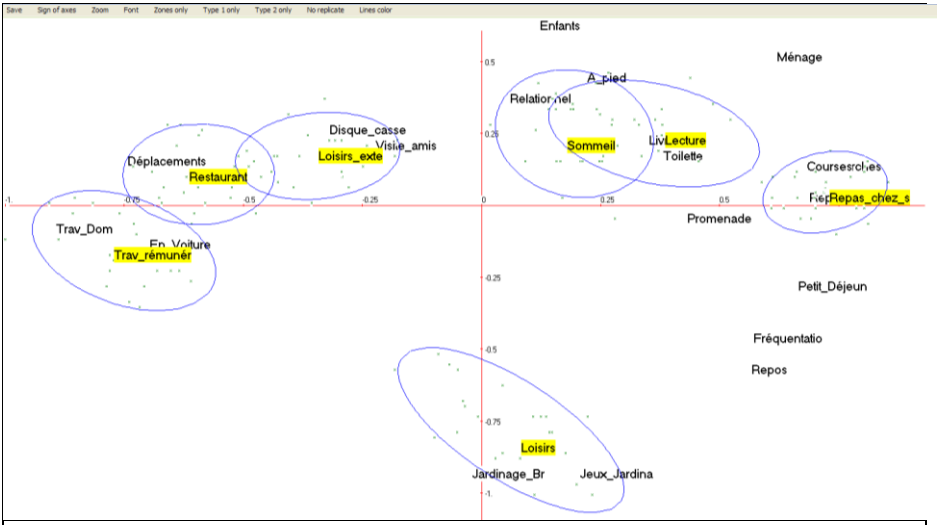
1. une fenêtre "*Bootstrap confidence areas*" s'affiche

➢ sélectionnez dans la rubrique "*Click to Select"* les variables dont on veut visualiser les ellipses. Les transférer avec **Select**, dans la fenêtre "*selected list*".



➢ Choisir ensuite le plan factoriel puis cliquez sur **Confidence ellipses** pour obtenir l'affichage graphique des variables actives (si le fichier ngus_var_boot.txt a été chargé), ou des catégories supplémentaires (si le fichier ngus_sup_cat_boot.txt a été chargé).

⊙ une fenêtre des zones de confiance bootstrap s'affiche



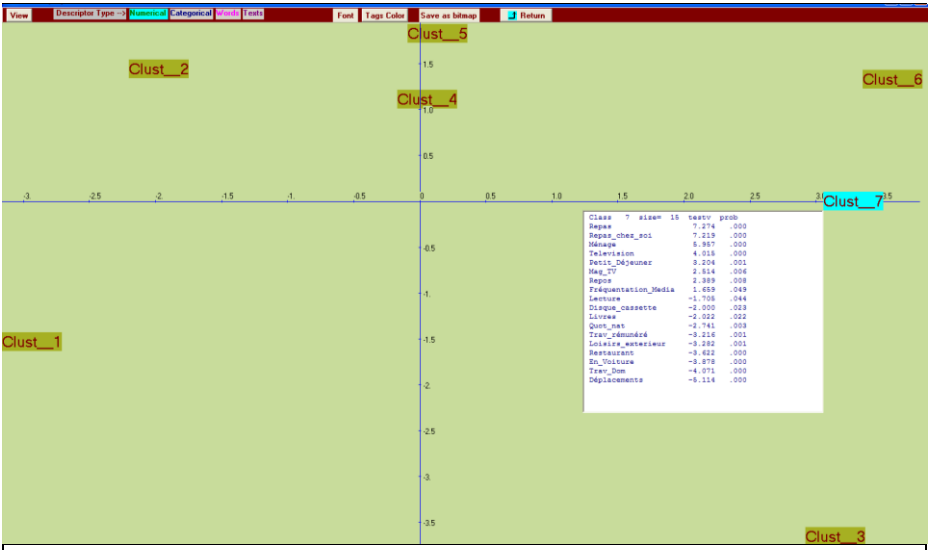
Commentaires : Les ellipses sont assez grandes en raison du faible nombre de groupes d'individus. L'utilisation du bootstrap, dans ce cas, donne des zones de confiance pessimistes pour les points. Dans une application réelle, le fichier individuel original (comportant des milliers d'individus) donnerait lieu à des ellipses de confiance beaucoup plus petites.

- Fermez la fenêtre et choisissez maintenant le bouton : **convex Hulls**. Les ellipses sont remplacées par les enveloppes convexes des réplications bootstrap pour chaque point. Les enveloppes convexes prennent en considération les points périphériques, tandis que les ellipses sont dessinées en utilisant la densité des nuages des réplications. Les deux informations sont complémentaires.
- Pour revenir au menu principal de Dtm-Vic, cliquez, selon la fenêtre, soit sur la croix en haut à droite, soit sur **return**.

4- Classification

Cette option permet de visualiser les centres des classes, qui sont projetés sur le plan factoriel.

- Cliquez sur **ClusterView**. Choisissez les axes (1 et 2 pour commencer), et **Continue**.
  - La fenêtre "*DTM-Display of clusters*" apparaît.



Commentaire : En actionnant ce bouton "numérique", nous observons le lien entre les variables numériques (variables actives et supplémentaires) du fichier de données et les 7 classes. En raison du petit nombre d'individus de l'exemple, certaines classes ne produisent pas des résultats significatifs. Dans le cadre de cet exemple, les autres rubriques du menu principal ne sont pas appropriées.

- Cliquez sur **View**. Les centres des 7 classes apparaissent sur le plan factoriel. Cliquez ensuite sur la rubrique **Numerical** du bandeau. Cette rubrique est désormais activée. Puis en cliquant (bouton **droit** de la souris) sur une classe, les variables les plus descriptives de la classe apparaissent.

L'ensemble des résultats figure dans la procédure DECLA du fichier sortie ("Basic numerical results"). ClusterView nous permet d'apprécier la forme du nuage des centres de classes et d'interroger interactivement leurs caractéristiques.

Nous pouvons facilement imaginer l'intérêt de l'outil pour une visualisation relative à des centaines de variables, des milliers d'individus regroupés, par exemple, en une vingtaine de classes.

II.2. Analyse des correspondances (AC ou *SCA*)

Ce deuxième exemple vise à décrire un petit tableau de contingence par l'analyse des correspondances (les données sont dans le répertoire : **DtmVic-Examples_A_Start/ EX_A02. SimpleCorAnalysis**).

II.2.1. Les données et fichiers Dtm-Vic :

(Fréquentation multimédia)

Les données proviennent d'une enquête multimédia par échantillonnage (effectuée par le CESP en 1992) pour laquelle on retient ici deux variables nominales : une variable : "média" à 6 modalités (radio, télévision, presses nationales et régionales, magazines, magazines de TV) et une variable : "statut d'activité" à 8 modalités (agriculteur, petit patron, cadre supérieur, profession intermédiaire, employé, ouvrier qualifié, ouvrier non qualifié, inactif). Le tableau de contingence considéré est obtenu par croisement de ces deux variables.

Les 6 modalités "médias" sont représentées en colonne et les 8 modalités "statuts d'activité" sont les lignes de la table de contingence. La cellule (i, j) de la table contient le nombre de contacts (le jour précédent l'enquête) entre les répondants appartenant au statut i avec le média j. Rappelons que les lignes et les colonnes représentent deux variables et jouent un rôle identique (contrairement au cas de l'analyse en composantes principales qui distingue variables et observations).

Identifiers	Radio	TV	Quot_Nat	Quot_Reg	Magazine	Mag_TV
Agriculteur	96	118	2	71	50	17
Petit_patron	122	136	11	76	49	41
Aff._Cadre_sup	193	184	74	63	103	79
Prof._interm	360	365	63	145	141	184
Employé	511	593	57	217	172	306
Ouvrier_qualif	385	457	42	174	104	220
Ouvr_non_qualif	156	185	8	69	42	85
Inactif	1474	1931	181	852	642	782

Tableau de contingence croisant les médias et les statuts d'activité

L'objectif est de décrire les relations entre les différents médias et les statuts d'activité pour la population considérée.

Nous considérons également, en ligne, trois autres caractéristiques socio-économiques, le sexe, l'âge et le niveau d'étude comme variables supplémentaires. Les tableaux de contingence croisant ces variables avec la variable "média" sont ainsi juxtaposés au tableau précédent.

Le dossier **EX_A02.SimpleCorAnalysis** contient le fichier de données et le fichier dictionnaire qui peuvent être importés à partir d'un fichier de données de type *Excel*).

- **fichier de données :** SCA_dat.txt

```
'Agriculteur'          96    118      2     71     50     17
'Petit patron'        122    136     11     76     49     41
'Aff. Cadre sup'      193    184     74     63    103     79
'Prof. interm'        360    365     63    145    141    184
'Employ,'             511    593     57    217    172    306
'Ouvrier qualif'      385    457     42    174    104    220
'Ouvrier non qual'    156    185      8     69     42     85
'Inactif'            1474   1931    181    852    642    782
'Homme'              1630   1900    285    854    621    776
'Femme'              1667   2069    152    815    683    938
'15-24 ans'           660    713     69    216    234    360
'25-34 ans'           640    719     84    230    212    380
'35-49 ans'           888   1000    130    429    345    466
'50-64 ans'           617    774     84    391    262    263
'65 ans ou +'         491    761     70    402    251    245
'Primaire'            908   1307     73    642    360    435
'Secondaire'          869   1008    107    408    336    494
'Techn. prof.'        901   1035     80    140    311    504
'Superieur'           619    612    177    209    298    281
```

Ce fichier de données comporte 20 lignes (dont 8 seront actives) et 7 colonnes. Chaque ligne contient l'identifiant des catégories socio-économiques (entouré du symbole "*quote*") suivi des 6 valeurs correspondant aux fréquences absolues de 6 médias, séparées par au moins un espace vide.

- **fichier dictionnaire :** SCA_dic.txt

```
Radio
Television
Quot_Nat
Quot_Reg
Magazine
Mag_TV
```

Dans ce format interne de Dtm-Vic, les libellés des catégories commencent à la colonne 6, [une police à intervalle fixe telle que le "courier" peut être employée pour faciliter l'utilisation de ce genre de format].

Rappel : les espaces vides dans les identifiants (individus et variables) ne sont pas permis.

II.2.2. Mise en œuvre de l'analyse (*SCA*)

Comme dans l'exemple 1, le fichier paramètre est créé en 5 étapes :

Etape 1 : Sélection de l'analyse

Dans la fenêtre du menu principal, cliquez sur : **Create** (**Command File**).

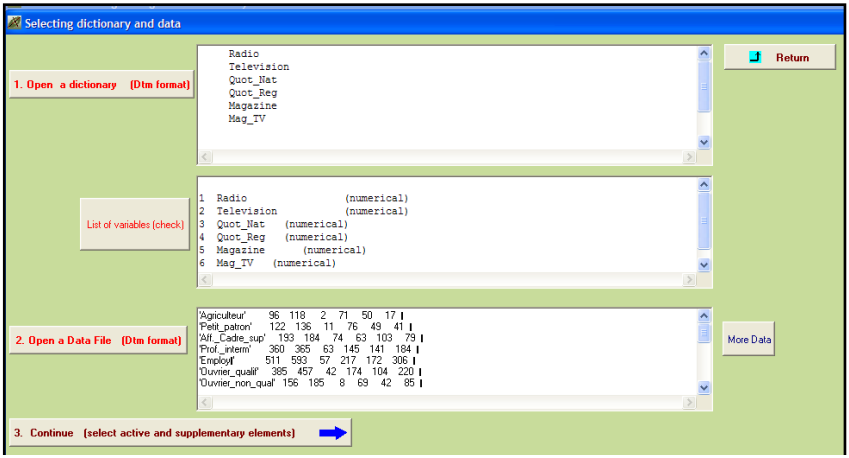
- Une fenêtre *"Choosing among some basic analyses"*

➢ Sélectionnez l'analyse : **SCA** – Simple Correspondence Analysis dans la rubrique : **Numerical data (principal axes techniques)**.

- Une fenêtre d'ouverture des "*fichiers dictionnaires et de données*" apparaît.

Etape 2 : Sélection des fichiers dictionnaires et de données

➢ Cliquez sur le bouton **Open a dictionary**. Dans le dossier EX_A02.SimpleCorAnalysis du jeu d'exemples de Dtm-Vic, ouvrir le fichier **SCA_dic.txt**. Il s'affiche dans une première fenêtre. La liste et le statut (numérique par défaut dans cet exemple) des variables sont indiqués dans une deuxième fenêtre.



Les colonnes de fréquences, pour une variable nominale donnée, sont considérées ici comme des variables numériques. Nous verrons que pour l'analyse des correspondances multiples (section II.3 ci après), les variables nominales ont le statut de "categorical variable", comme nous l'avons vu à propos de certaines variables supplémentaires en ACP.

4. Cliquez sur le bouton **Open a Data File**. Dans le même dossier EX_A02.SimpleCorAnalysis, ouvrir le fichier **SCA_dat.txt** qui s'affiche dans une troisième fenêtre.

Note : il est possible qu'une boîte de message annonce l'existence d'une dernière ligne vide". Cliquer alors sur OK deux fois.

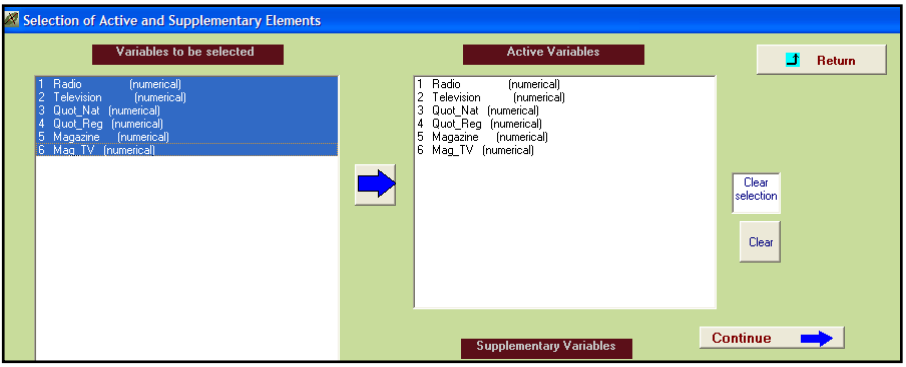
Cliquez sur : **3. Continue**

⊙ une fenêtre "*Selection of active et supplementary elements*" apparaît.

Etape 3 : Sélection des variables actives et supplémentaires

Dans le cas d'une table de contingence, les variables sont en fait les modalités de la variable considérée en colonne c'est-à-dire ici les médias. Le jeu de données présente ici peu de variables (types de médias) qui sont toutes considérées comme actives.

➢ Sélection des variables continues actives : V1 à V6 à transférer dans la fenêtre "*Active Variables*"



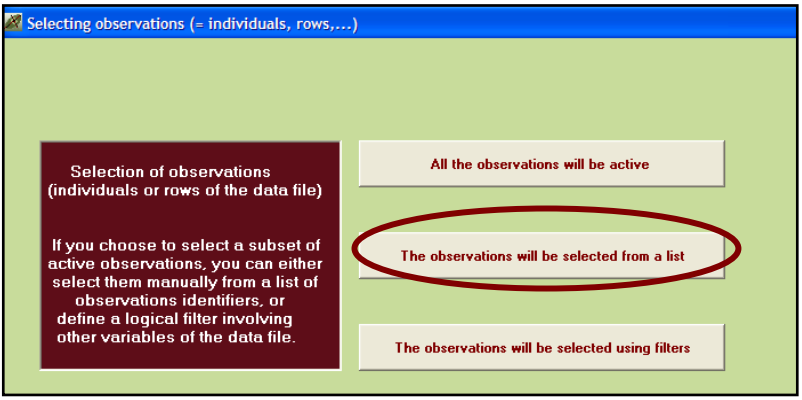
➢ Cliquez sur : **Continue**

⊙ Une fenêtre "*Selecting observations*" apparaît

Etape 4 : Sélection des observations (individus)

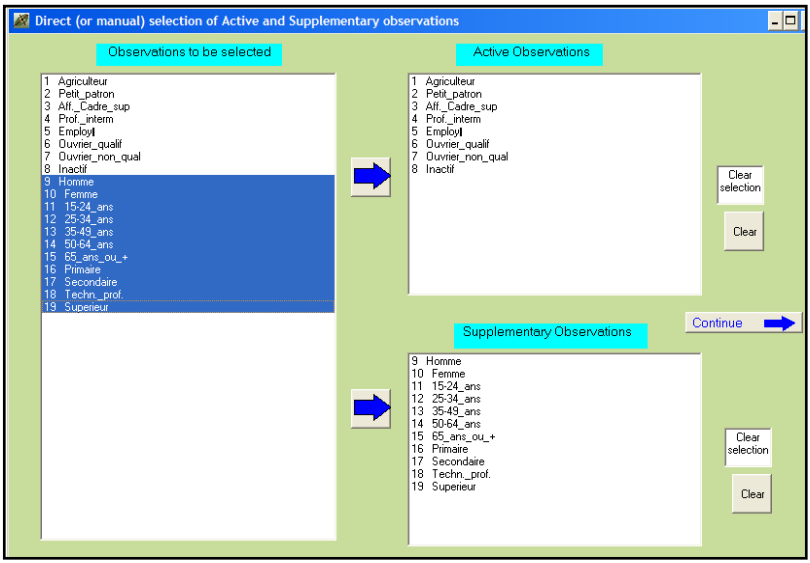
Les lignes ne représentent pas ici des observations ou individus comme pour l'ACP ou l'Analyse des Correspondances Multiples (plus loin) mais des modalités de variables. Aussi de la même manière que l'on considère des variables actives et/ou supplémentaires, on procède à la sélection des modalités actives et/ou supplémentaires représentées en ligne. Nous retenons ici l'ensemble des 8 statuts d'activité comme variables actives, et le sexe, l'âge et le niveau d'étude comme variables supplémentaires.

➢ Cliquez sur: **The observations will be selected from a list**



⊙ La fenêtre "*selection of Active and Supplementary observations*" apparaît.

- Sélectionnez les modalités de la variable "statut d'activité" comme éléments actifs. Puis sélectionnez les modalités des variables "sexe", "âge", "niveau d'étude" comme éléments supplémentaires.



- Cliquez sur Continue

  ⊙ une fenêtre : "*Create a starting parameter file*" apparaît.

Etape 5 : Création du fichier paramètre

Nous faisons ici le choix d'une procédure *bootstrap*. (Si elle n'est pas retenue, cliquez directement sur : **2-Create a parameter file for SCA**).

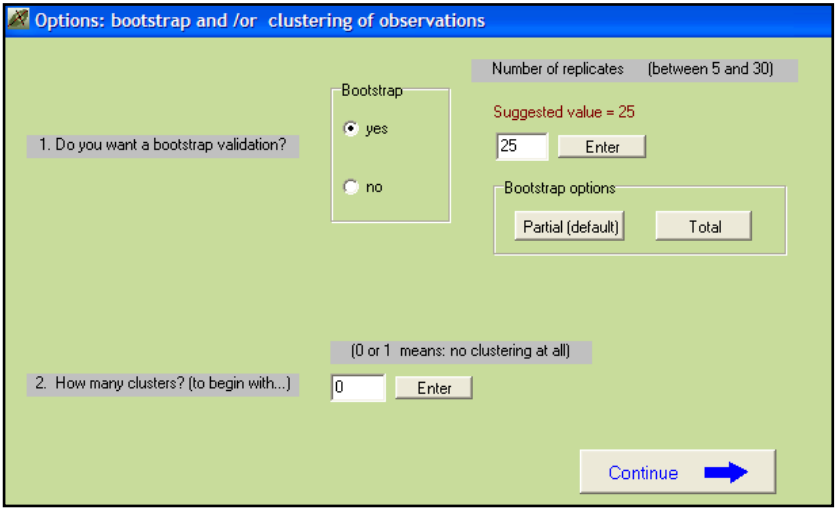
- Cliquez sur **1-Select some options**

  ⊙ une fenêtre "*Options : Bootstrap and/or Clustering of observations*" apparaît.

Compte tenu du petit nombre d'individus, aucune classification n'est nécessaire : nous ne considérons ici que la procédure du *bootstrap*.

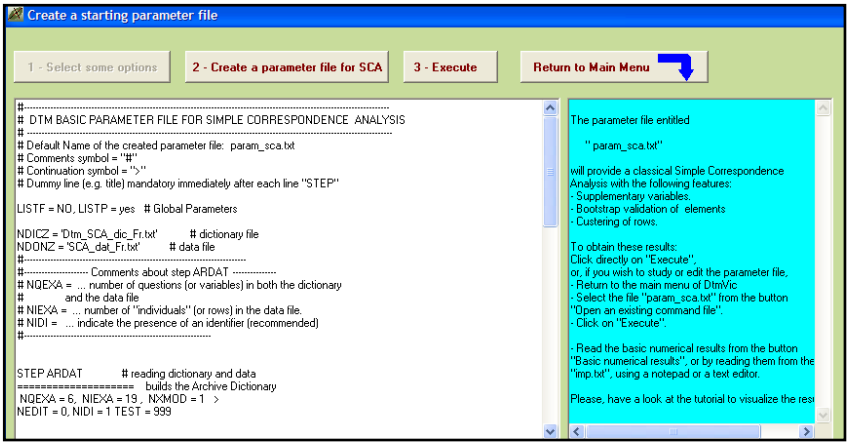
- Cliquez sur "yes" pour la procédure *bootstrap* ; indiquer le nombre de réplications (par défaut 25) puis : **Enter**. C'est le *bootstrap* partiel qui est appliqué par défaut. (cf. encadré technique section II.1.2 Etape 5 à propos de l'ACP).

- Choisir 0 ou 1 classe puis cliquez sur : **Enter**. Nous ne voulons pas effectuer de classification.
- Cliquez sur : Continue ➔

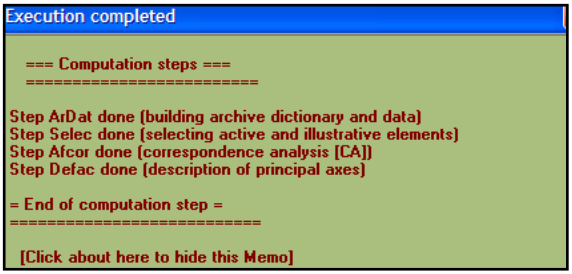


  - ⊙ la fenêtre : "*Create a starting parameter file*" réapparaît.

- Cliquez sur: **2-Create a parameter file for SCA**. Un fichier paramètre vient d'être créé sous le nom **param_SCA.txt** et stocké dans le dossier EX_A02.SimpleCorAnalysis du répertoire DtmVic_Examples_A_Start. (Pour le conserver en vue de réitérer directement la même analyse plus tard, il faudra le renommer après l'analyse).



➢ Cliquez sur: **3-Execute**



Les procédures s'affichent en bloc à la fin de l'exécution : ArDat (Archivage des données), Selec (Sélection des éléments actifs et supplémentaires), Afcor (Analyse des correspondances) et Defac (Description des axes factoriels).

Note : Lors d'une utilisation ultérieure de Dtm-Vic, il est possible d'ouvrir le fichier paramètre param_SCA.txt dans le menu principal **Command File** avec la procédure **Open an existing command file** puis d'exécuter ce fichier **Execute**.

II.2.3 Fichier de résultats

Les résultats peuvent être consultés dans l'étape **Result Files**

➢ Cliquez sur: Basic numerical results pour ouvrir le fichier en format html ou sur: Basic numerical results (text format) pour ouvrir le fichier résultat en format texte puis cliquer sur: **Return** pour en sortir et revenir au menu principal.

Le fichier résultat est nommé selon les mêmes principes que pour l'analyse en composantes principales.

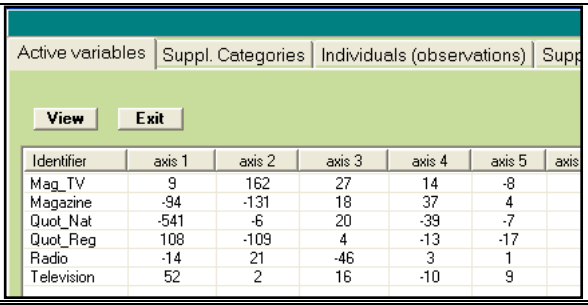
II.2.4 Visualisation des résultats

Nous renvoyons le lecteur au paragraphe II.1.4 pour la présentation de la deuxième phase de Dtm-Vic et le détail des différents outils de visualisation. Nous considérons ici comme outils : AxesView, PlaneView et Bootstrap.

1- Axes factoriels

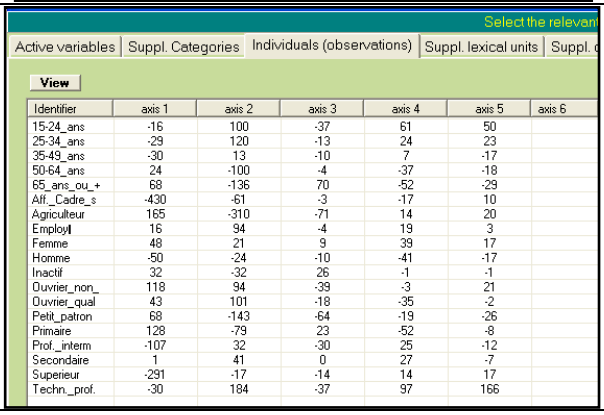
- Cliquez sur: AxesView. Une fenêtre propose de visualiser les coordonnées des variables actives, supplémentaires et des observations, sur les premiers axes (résultats correspondant à l'étape DEFAC du fichier résultat).
- Cliquez sur: Active variables puis sur: **View** pour obtenir les coordonnées des modalités "média". Cliquez ensuite sur: Individuals (observations) puis sur: **View** pour obtenir les coordonnées des modalités actives "statut d'activité" et des modalités supplémentaires.

Coordonnées des modalités de la variable "média"



Identifier	axis 1	axis 2	axis 3	axis 4	axis 5	axis
Mag_TV	9	162	27	14	-8	
Magazine	-94	-131	18	37	4	
Quot_Nat	-541	-6	20	-39	-7	
Quot_Reg	108	-109	4	-13	-17	
Radio	-14	21	-46	3	1	
Television	52	2	16	-10	9	

Coordonnées des modalités de la variable "statut d'activité". (Cette variable est positionnée en ligne et considérée ici comme individus)



Identifier	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6
15-24_ans	-16	100	-37	61	50	
25-34_ans	-29	120	-13	24	23	
35-49_ans	-30	13	-10	7	-17	
50-64_ans	24	-100	-4	-37	-18	
65_ans_ou_+	68	-136	70	-52	-29	
Aff._Cadre_s	-430	-61	-3	-17	10	
Agriculteur	165	-310	-71	14	20	
Employé	16	94	-4	19	3	
Femme	48	21	9	39	17	
Homme	-50	-24	-10	-41	-17	
Inactif	32	-32	26	-1	-1	
Ouvrier_non_	118	94	-39	-3	21	
Ouvrier_qual	43	101	-18	-35	-2	
Petit_patron	68	-143	-64	-19	-26	
Primaire	128	-79	23	-52	-8	
Prof._interm	-107	32	-30	25	-12	
Secondaire	1	41	0	27	-7	
Superieur	-291	-17	-14	14	17	
Techn._prof.	-30	184	-37	97	166	

L'axe 1 oppose la presse quotidienne nationale aux autres médias et les cadres aux autres catégories
L'axe 2 oppose la presse régionale et magazine à la presse TV, et les agriculteurs et indépendants aux employés et ouvriers

- Cliquez sur : **exit** pour sortir de cet outil.

2- Plans factoriels

➢ Cliquez sur : PlaneView.

⊙ Une fenêtre s'affiche proposant différentes visualisations de plans factoriels.

Cette option fournit les plans factoriels séparés ou superposés des variables actives, supplémentaires, ou des observations. Là encore, variables et observations représentent les modalités des deux variables de la table de contingence. Dans ce cas, le sous-menu *"Actives columns + Active rows"* est approprié pour le tableau de contingence.

➢ Cliquez sur la rubrique : "**Actives columns + Active rows"** puis sélectionnez les axes principaux désirés (ici les axes 1 et 2). Cliquez ensuite sur : **display**.

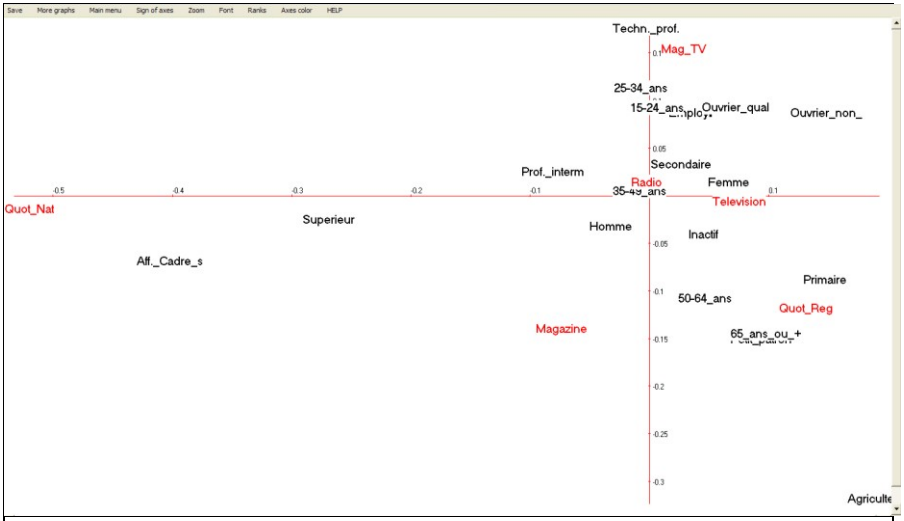
⊙ Apparaît une fenêtre pour sélectionner le plan factoriel suivant la paire d'axes souhaitée.

➢ Choisir les axes 1 er 2 (choix par défaut) puis cliquez sur : **display**. Il est possible de ne faire figurer sur les plans que certaines variables. Cliquez alors sur : **Manual Selection of points**. Sélectionner les variables et les transférer dans la seconde fenêtre en cliquant sur : select.

Rappel : Pour chaque graphique, le bandeau du haut contient des options :
– "Save" sauvegarde le graphique en format bmp;
– "Font" offre la possibilité de modifier la police et la couleur des caractères ;
– "More graph" permet de changer de plan factoriel ;
– "Sign of axes" permet d'inverser les axes ;
– "Rank", est utile seulement dans le cas des affichages très complexes, (ce qui n'est pas le cas ici): ce bouton convertit les deux coordonnées de l'affichage courant en rangs (voir note de la section précédente).

⊙ La fenêtre du plan factoriel apparaît. Choisir une option puis cliquez sur : ***View***

⊙ Retournez ensuite sur : "PlaneView" pour sélectionner une autre représentation factorielle. Pour fermer le graphique, cliquez sur : **Return** ou sur la croix en haut à droite, puis sur : **Return** dans la fenêtre de sélection des axes principaux.



Commentaire : On relève également, sur le plan factoriel principal, l'opposition entre Presse quotidienne Nationale et Régionale, et aussi entre Cadres et les autres catégories. Puis, sur le second axe, l'opposition entre les magazines TV et les autres supports de presse.

- Retourner ensuite sur : "PlaneView" pour sélectionner une autre représentation factorielle. Pour fermer le graphique, cliquez sur : **Return** ou sur la croix en haut à droite, puis sur : **return** dans la fenêtre de sélection des axes principaux.

- Pour revenir au menu principal de VIC, cliquez soit sur la croix en haut à droite, soit sur "Main menu" soit sur : **return**.

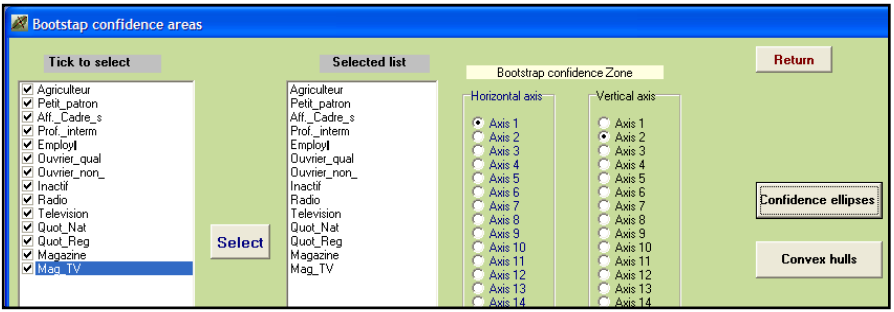
3- Validation Bootstrap

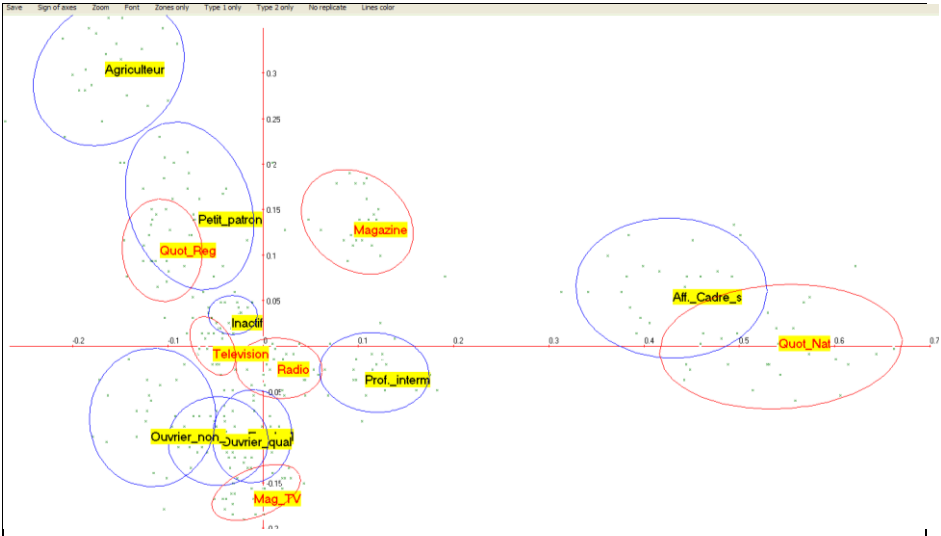
- Cliquez sur : B Bootstrap pour valider la position des variables dans les plans factoriels.

  - Une fenêtre : *"DtmVic – Bootstrap – Validation – Stability - Inférence*" apparaît.

- Cliquez ensuite sur : **Load Data** puis ouvrir dans le répertoire le fichier des réplications selon le bootstrap choisi. On sélectionne ici le fichier **ngus_var_boot.txt** pour un bootstrap partiel. Répondre : **OK** à la boîte de message : "*Set of principal coordinates loaded*" qui s'affiche.

- sélectionnez ("*Tick to select*") les variables dont on veut visualiser les ellipses. Les transférer avec **Select**, dans la fenêtre "*selected list*". Choisir ensuite le plan factoriel puis cliquez sur **Confidence ellipses** pour l'affichage graphique des variables actives (fichier ngus_var_boot.txt).





Commentaire : Les zones de confiance des points-médias (en rouge) montrent que ceux-ci ont des positions distinctes, donc des profils d'activités typés. En revanche, certains points-activité (en noir) ont des zones de confiance qui empiètent. Ainsi, on ne peut conclure que les Ouvriers non qualifiés, les Ouvriers qualifiés et les Employés occupent des positions distinctes, et donc que ces catégories ont des profils de fréquentation média distincts.

- Pour fermer le graphique, cliquez sur : **return**.

II.3. Analyse des Correspondances Multiples (ACM ou *MCA*)

Le troisième exemple (voir répertoire : **DtmVic-Examples_A_Start/ EX_A03.MultCorAnalysis)** décrit un ensemble de variables nominales par l'Analyse des Correspondances Multiples.

II.3.1. Les données : Extraits de l'enquête : *"Conditions de vie et Aspirations des Français"*

Les données sont extraites d'une enquête par sondage effectuée par le CREDOC en 1986 sur "les conditions et aspirations des Français"[5]. Elles traitent des réponses d'un petit sous-échantillon de 315 individus et 49 questions. Une première série de questions concerne les caractéristiques objectives du répondant ou de son ménage (âge, statut, genre, équipements,…). D'autres séries de questions se rapportent à l'attitude ou aux opinions des enquêtés sur la perception du niveau de vie, la famille, l'environnement physique et technologique, la santé, la justice, la société.

Dans le dossier EX_A03.MultCorAnalysis du répertoire **DtmVic-Examples_A_Start**, sont contenus les fichiers dictionnaire et des données en format Dtm-Vic :

1. le fichier dictionnaire : MCA_dic.txt

```
   8 region
AA01 region paris
AA02 bassin parisien
AA03 nord
AA04 est
AA05 ouest
AA06 sud-ouest
AA07 centre-est
AA08 mediterranee
   9 taille d'agglomeration
AB01 <2000
AB02 2001-5000
AB03 5001-10000
AB04 10001-20000
AB05 20001-50000
AB06 50001-100000
BC01 satisf.log:tres
BC02 satisf.log:assez
BC03 satisf.log:peu
BC04 satisf.log:pas du tout
   4 satisfaction cadre de vie
BD01 cdv:tres
BD02 cdv:assez
BD03 cdv:peu
BD04 cdv:pas du tout
   5 statut d'occupat logement
BE01 accedant
BE02 proprietaire
BE03 locataire
BE04 loge gratuit
BE05 autre
   6 depenses de logement
```

[5] Cf. Lebart L. (1987) - Conditions de vie et aspirations des Français. Evolution et structure des opinions de 1978 à 1984. *Futuribles*, 1, p 25-56. Cf. aussi: Lebart L. (1986) - Qui pense quoi ? Evolution et structure des opinions en France de 1978 à 1984. *Consommation Revue de Socio-Economie*, Dunod, 4, p 3-22.

```
AB07 100001-200000
AB08 >200000
AB09 paris.agglo.paris
   2 sexe
AC01 masculin
AC02 feminin
   0 age
   7 situation
AD01 actif
AD02 etudiant
AD03 menagere s.prof.
AD04 malade invalide
AD05 retraite
AD06 militaire
AD07 chomeur
   5 A1-statut matrimonial
AG01 celibataire
AG02 marie(e)
AG03 concubinage
AG04 separe(e) divorce
AG05 veuf(ve)
    3 la famille est le seul end
AI01 famille:-oui-
AI02 famille:-non-
AI03 famille:nsp-nr
   4 opinion sur le mariage
AJ01 mariage:indissoluble
AJ02 mariage:dissout si pb grave
AJ03 mariage:dissout si accord
AJ04 mariage:ne sait pas
   4 travaux/menage/enfants
AK01 la femme seule
AK02 plutot la femme
AK03 homme et femme
AK04 tr.femmes:ne sait pas
   4 satisfaction_logement
```

```
BF01 negligeable
BF02 sans gros probleme
BF03 une lourde charge
BF04 tres lourde charge
BF05 ne fait pas face
BF06 ne sait pas
……..
   4 activite professionnelle
GB01 plein temps
GB02 temps partiel
GB03 non activite
GB04 n'a jamais travail.
   2 conflits travail-vie person
GC01 conflits oui
GC02 conflits non
   2 chomage douze derniers mois
GD01 chomage oui
GD02 chomage non
   2 maux de tete
HA01 maux de tete oui
HA02 maux de tete non
   2 mal au dos
HB01 mal au dos oui
HB02 mal au dos non
   2 nervosite
HC01 nervosite oui
HC02 nervosite non
   2 etat depressif
HD01 etat depressif oui
HD02 etat depressif non
   4 satisfaction sante
HG01 satisfaction sante:tres
HG02 satisfaction sante:satisf
HG03 satisfaction sante:peu
HG04 satisfaction sante:pas du t
   0 nombre de personnes logt
```

Le dictionnaire MCA_dic.txt contient les identifiants de 49 variables (39 nominales et 10 continues).

Rappel : L'identifiant d'une variable nominale est précédé par le nombre N de ses catégories (en colonne 5). Les N lignes suivantes identifient les N catégories des réponses : un identifiant en 4 caractères occupe les colonnes 1 à 4 et un identifiant long (20 caractères maximum) commence à la colonne 6 [utiliser une police à intervalle fixe]. Une variable numérique telle que l'âge ou le nombre d'enfants, a, conventionnellement, zéro catégorie.*Les espaces vides dans les identifiants ne sont pas permis.*

2. **fichier de données** (extraits) : **MCA_dat.txt**

```
'0005' 8. 1. 2. 27. 3. 2. 7. 1. 2. 3. 1. 1. 2. 2. 2. 2. 2. 3. 0. 0. 1. 1..... 4. 7. 7. 6. 6. 6. 3. 3. 2. 4.  1  3
'0011' 8. 1. 2. 32. 3. 2. 2. 1. 3. 3. 1. 2. 3. 3. 2. 2. 2. 4. 0. 0. 2. 1..... 1. 7. 5. 4. 7. 7. 1. 5. 3. 4.  2  1
'0018' 8. 8. 1. 21. 2. 1. 8. 2. 1. 3. 2. 3. 1. 4. 2. 2. 1. 4. 0. 0. 2. 1..... 4. 7. 7. 7. 5. 7. 3. 7. 2. 4.  1  3
'0024' 5. 1. 2. 42. 1. 2. 3. 1. 2. 3. 1. 2. 1. 3. 2. 2. 2. 2. 1. 2. 2. 1..... 1. 7. 6. 7. 5. 5. 7. 5. 2. 4.  3  1
'0030' 5. 1. 1. 29. 1. 2. 2. 1. 2. 3. 1. 2. 1. 2. 2. 2. 2. 2. 2. 1. 1. 2..... 3. 7. 7. 4. 4. 7. 4. 3. 4. 4.  1  1
'0036' 2. 4. 2. 35. 1. 2. 7. 1. 2. 2. 1. 1. 2. 2. 1. 1. 2. 1. 1. 2. 1. 1..... 4. 7. 7. 5. 6. 7. 5. 5. 2. 4.  2  3
```

```
'0042' 2. 4. 1. 71. 5. 2. 8. 1. 3. 3. 4. 2. 3. 2. 2. 2. 1. 3. 0. 0. 2. 2..... 2. 5. 7. 7. 5. 5. 1. 3. 4. 4.  4  3
'0048' 5. 1. 1. 62. 1. 2. 1. 1. 3. 2. 2. 2. 2. 3. 2. 2. 2. 1. 1. 2. 1. 1..... 3. 6. 6. 6. 6. 6. 3. 3. 3. 1.  4  1
'0054' 5. 5. 1. 24. 1. 3. 3. 1. 3. 2. 2. 2. 3. 2. 2. 2. 2. 1. 2. 2. 2. 2..... 4. 7. 4. 7. 5. 7. 4. 3. 3. 3.  1  1
'0060' 4. 1. 1. 52. 1. 2. 3. 1. 2. 3. 2. 2. 2. 2. 2. 2. 2. 1. 2. 2. 1. 1..... 2. 7. 7. 5. 4. 5. 7. 3. 3. 2.  3  1
```

Le fichier de données comporte 315 lignes correspondant aux individus enquêtés et 50 valeurs. Pour une ligne i, la première valeur (entre quotes) correspond à l'identifiant de l'individu i, et les 49 autres valeurs correspondent aux réponses des 49 variables numériques ou aux valeurs codant les items de réponse aux variables nominales, séparées par des espaces blancs.

II.3.2. Mise en œuvre de l'ACM

Selon le même principe de mise en œuvre de l'analyse en composantes principales (cf § II.1.2), le fichier paramètre est créé en 5 étapes :

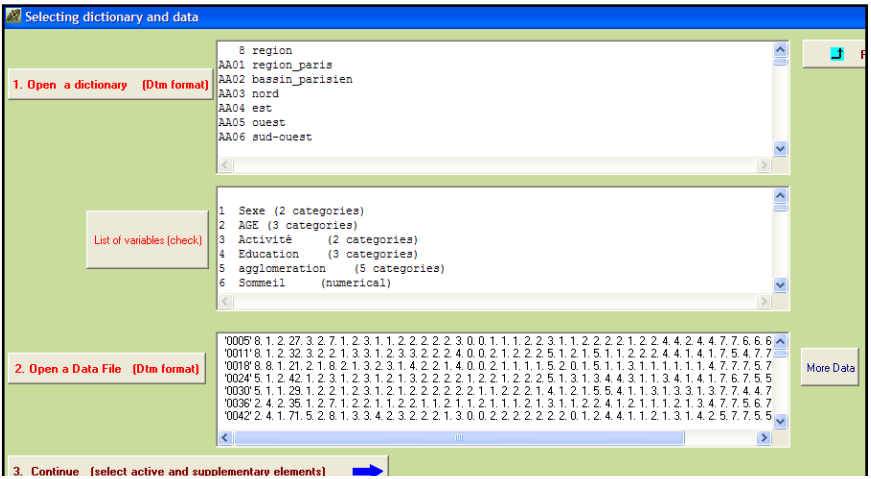
Etape 1 : Sélection de l'analyse

- Cliquez sur le bouton : **Create a command file** , ligne : **Command File**
  - Une fenêtre: *"Choosing among some basic analyses"* apparaît.
- Sélectionnez l'analyse : **MCA** – Multiple Correspondances Analysis dans la rubrique **Numerical Data (principal axes techniques)**.
  - Une fenêtre d'ouverture des "*fichiers dictionnaires et de données*" apparaît.

Etape 2 : Sélection des fichiers dictionnaires et de données

3. Cliquez sur le bouton : **Open a dictionary**. Dans le répertoire : DtmVic-Examples_A_Start/EX_A03.MultCorAnalysis, ouvrir : **MCA_dic.txt**. Ce fichier s'affiche dans une première fenêtre. Le statut (*categorical* ou *numerical*) des variables est indiqué dans une deuxième fenêtre.

- Cliquez sur le bouton : **Open a Data File**. Dans le répertoire DtmVic-Examples_A_Start /EX_A03.MultCorAnalysis, ouvrir le fichier **MCA_dat.txt** qui s'affiche dans une troisième fenêtre.



➢ Cliquez sur **3. Continue ➔**

⊙ une fenêtre " *Selection of active and supplementary elements* " apparaît.

Etape 3 : Sélection des variables actives et supplémentaires

A l'intérieur de la fenêtre "*Selection of active and supplementary elements"* s'affichent trois autres fenêtres :

– "*Variables to be selected*" où figurent l'ensemble des variables
– "*Active Variables*" qui reçoit les variables actives sélectionnées
– "*Supplementary Variables*" pour les variables supplémentaires sélectionnées

Dans le cadre de l'analyse des correspondances multiples, les variables actives doivent être nominales (catégorielles). Les variables supplémentaires peuvent être continues ou nominales.

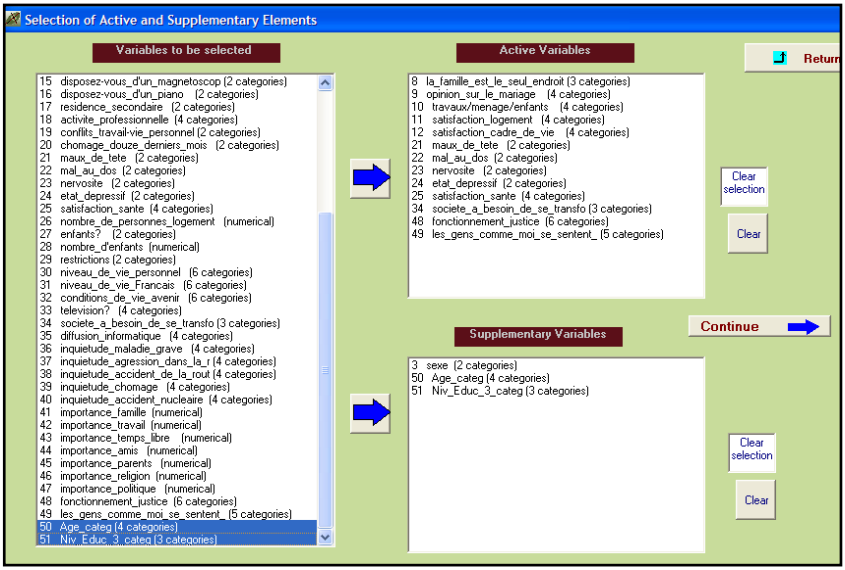
Nous suggérons de sélectionner les variables suivantes comme variables actives et supplémentaires :

➢ Variables actives à transférer dans la fenêtre "*Active Variables*"

8 . la_famille_est_le_seul_endroit_où …	23 . nervosite
9 . opinion_sur_le_mariage	24 . etat_depressif
10 . travaux/menage/enfants	25 . satisfaction_sante
11 . satisfaction_logement	34 . societe_a_besoin_de_se_transf
12 . satisfaction_cadre_de_vie	48 . fonctionnement_justice
21 . maux_de_tete	49 . les_gens_comme_moi_se_sentent_seuls
22 . mal_au_dos	

➢ Sélection des variables supplémentaires à transférer dans la fenêtre "*Supplementary Variables*"

variables nominales supplémentaires :	3 . sexe 50 . Age_categ 51 . Niv_Educ_3_categ



➢ Cliquez sur : **Continue** ➔

⊙ Une fenêtre : "*Selecting observations*" apparaît

Etape 4 : Sélection des observations (individus)

Trois cas de figure sont possibles :

1. Prendre en compte l'ensemble des observations
2. Sélectionner les observations sur une liste
3. Sélectionner les observations par un filtre

Nous prenons en considération ici l'ensemble des observations.

➢ Cliquez sur: **All the observations will be active**

⊙ une fenêtre : "*Create a starting parameter file*" apparaît.

Etape 5 : Création du fichier paramètre

A cette étape, il est possible de sélectionner, comme option, les procédures de *bootstrap* et/ou de classification. Rappelons que dans Dtm-Vic les analyses factorielles sont systématiquement complétées par :

- un *bootstrap* qui permet de valider les positions des variables .
- une classification avec une description automatique des classes.

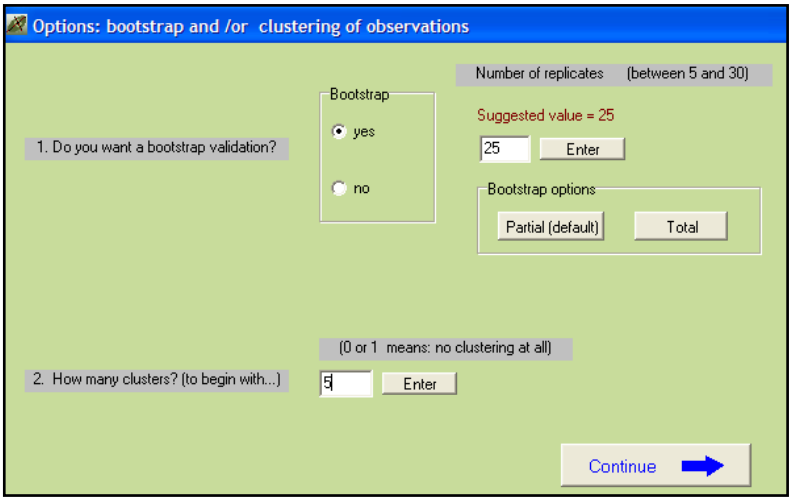
➢ Cliquez sur : **1-Select some options**

⊙ une fenêtre "*Options : Bootstrap and/or Clustering of observations*" apparaît.

Pour un rappel sur les différents types de bootstrap dans Dtm-Vic, voir l'encadré technique à propos de l'ACP, section II.1.2, Etape 5.

➢ Cliquez sur : "yes" pour la procédure "bootstrap" ; indiquer le nombre de réplications (par défaut 25) puis : **Enter**. C'est le bootstrap partiel qui est appliqué par défaut.

Si le bootstrap n'est pas adopté, cliquez sur "no" et passer directement à l'option de classification.



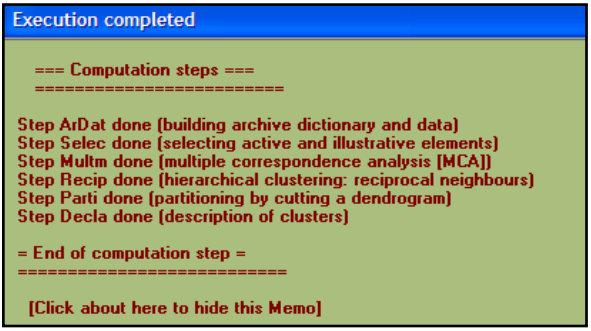
- Sélectionnez le nombre de classes souhaité (nous suggérons 5 classes) puis cliquez sur : **Enter**.

- Cliquez sur **Continue ➔**

  - la fenêtre "*Create a starting parameter file*" réapparaît.

- Cliquez sur **2-Create a parameter file for MCA.** Un fichier paramètre vient d'être créé sous le nom **param_MCA.txt** et stocké dans le dossier EX_A03.MultCorAnalysis du répertoire DtmVic-Examples_A_Start. Pour le conserver en vue de répéter l'analyse ultérieurement, il faudra le renommer.

- Cliquez sur **3-Execute**



Les procédures s'affichent en bloc à la fin de l'exécution.

Commentaires sur les procédures :

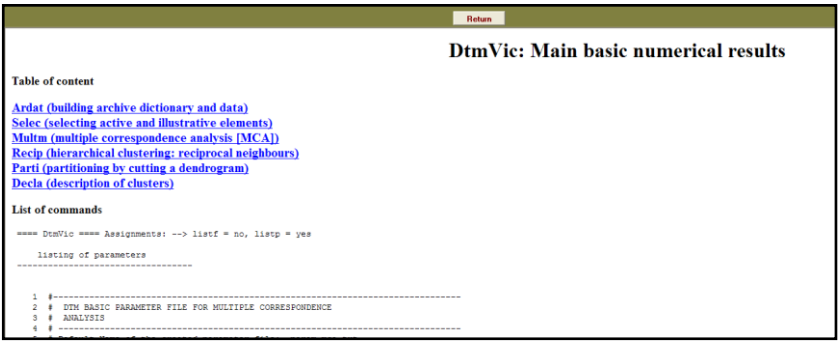
ArDaT (Archivage des données), Selec (Sélection des éléments actifs et supplémentaires), Multm (Analyse des correspondances multiples), Recip (Classification mixte utilisant la classification ascendante hiérarchique, méthode des voisins réciproques), Parti (Coupure du dendrogramme et optimisation de la partition par la méthode des centres mobiles [*k-means*]), Decla (Description automatique des classes).

Note : Une fois créé, il est possible, lors d'une utilisation ultérieure de Dtm-Vic d'ouvrir le fichier paramètre param_MCA.txt dans le menu principal avec la procédure **Open an existing command file** puis d'exécuter à nouveau ce fichier **Execute**. Les utilisateurs expérimentés peuvent modifier des paramètres directement, ou avec n'importe quel autre éditeur de textes après avoir quitté Dtm-Vic..

II.3.3 Fichier de résultats

Les résultats peuvent être consultés dans la rubrique : **Result Files**

- Cliquez sur Basic numerical results pour naviguer dans le fichier en format html puis sur Return pour en sortir et revenir au menu principal.



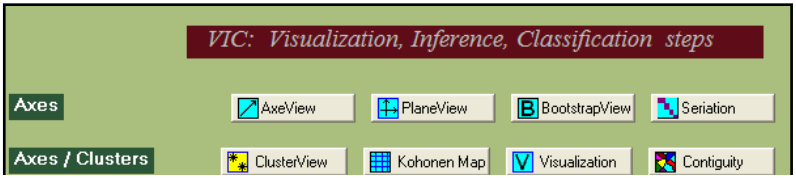
- ou encore : cliquez sur Basic numerical results (.txt format) pour ouvrir le fichier de résultats en format texte.

Les deux fichiers "imp.txt" et "imp.html" sont contenus dans le répertoire EX_A03.MultCorAnalysis. Ils sont également sauvegardés sous le nom "imp" suivi de la date et l'heure de l'analyse. Ces fichiers de sauvegarde archivent les résultats numériques principaux tandis que les fichiers "imp.txt/html" sont écrasés pour chaque nouvelle analyse exécutée dans le même répertoire.

Après avoir parcouru les résultats numériques, revenez au menu principal. Ces résultats sont visualisés alors dans l'étape VIC de Dtm-Vic. Cette visualisation va faciliter les interprétations.

II.3.4 Visualisation des résultats

Cette deuxième phase de Dtm-Vic fournit les outils de visualisation nécessaires à l'interprétation et la validation des résultats.



1- Axes factoriels

- Cliquez sur AxesView. Une fenêtre propose de visualiser les coordonnées des variables actives, supplémentaires et des observations, sur les premiers axes [cf. aussi l'étape DEFAC du fichier résultats].

  Dans le cadre d'une ACM, trois éléments peuvent être examinés, les ***variables nominales actives*** et ***supplémentaires***, les ***variables continues supplémentaires*** et les ***observations***.

- Cliquez sur l'onglet des éléments à examiner, **Active variables** par exemple puis sur : **View**. Il est possible d'ordonner les coordonnées d'un axe donné, par exemple l'axe 2, en cliquant sur "Axis 2".

Active variables | Suppl. Categories | Individuals (observ

View | Exit

Identifier	axis 1	axis 2	axis 3
satisfaction_sante:p	-2256	-300	-1250
satisfaction_sante:p	-1370	122	898
etat_depressif_oui	-1350	-317	-569
justice:ne_sait_pas	-1001	935	-716
mariage:ne_sait_pas	-906	1282	-698
la_femme_seule	-879	1442	626
transf-soc:ne_sait_p	-865	1383	-307
maux_de_tete_oui	-785	-145	-61
solitude:assez_d'acc	-694	-363	17
solitude:tres_d'acco	-651	-848	995
nervosite_oui	-640	-160	-160
mal_au_dos_oui	-570	150	54
satisf.log:peu	-358	-680	1883
justice:refus/repond	-144	1110	-17
satisf.log:assez	-129	-358	-19
plutot_la_femme	-119	280	54
mariage:dissout_si_p	-83	50	-165
cdv:assez	-70	-123	231
famille:-oui-	-63	184	289

Coordonnées (x 1000) des variables nominales actives

Active variables | Suppl. Categories | Individuals (obs

View | Exit

Identifier	axis 1	axis 2	axis 3
Age_super_60	-333	374	363
feminin	-204	-54	-101
Niv_Educ_bas	-203	59	142
Age_inf.60	-85	104	87
Niv_Educ_moyen	14	-64	-224
Age_inf.40	82	-14	-264
Age_inf.30	248	-347	-133
masculin	261	70	129
Niv_Educ_haut	335	-65	-115

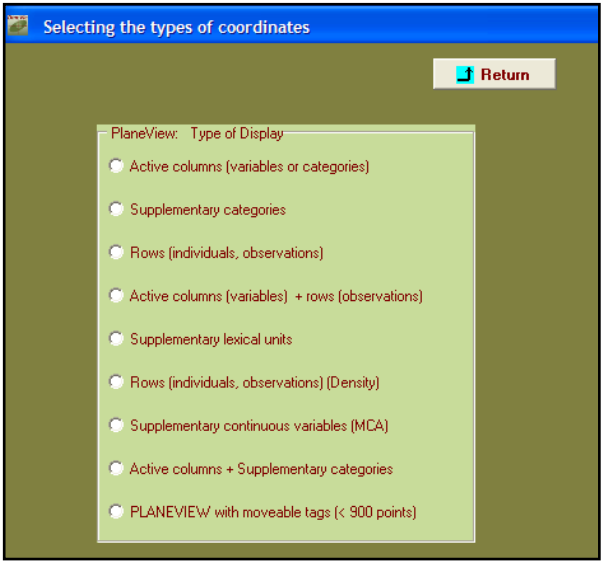
Coordonnées (x 1000) des var. nominales supplémentaires

2- Plans factoriels

Cet outil fournit les plans factoriels séparés ou superposés des variables actives, supplémentaires, ou des observations.

➢ Cliquez sur : PlaneView

⊙ Une fenêtre s'affiche proposant différentes visualisations.



Dans cet exemple d'analyse, six rubriques sont possibles : "colonnes actives (variables, catégories)", "catégories supplémentaires", "lignes actives (individus, observations)", "colonnes actives + lignes actives", "individus actifs (densité)" et "colonnes actives + catégories supplémentaires". L'item : "PLANEVIEW with moveable tags" reprend certaines des rubriques précédentes et permet de déplacer les points superposés pour rendre plus lisible le graphique.

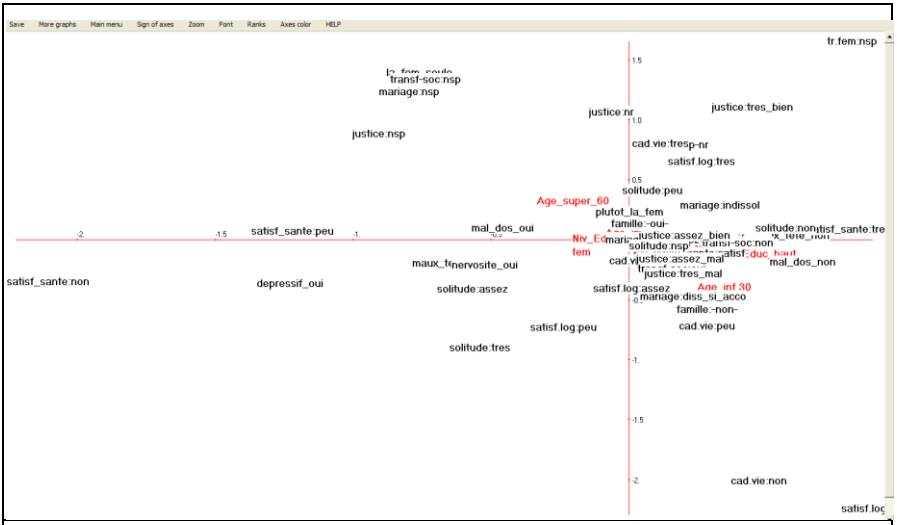
➢ Sélectionnez : "colonnes actives + catégories supplémentaires".

1. Apparaît une fenêtre pour sélectionner le couple d'axes souhaités.

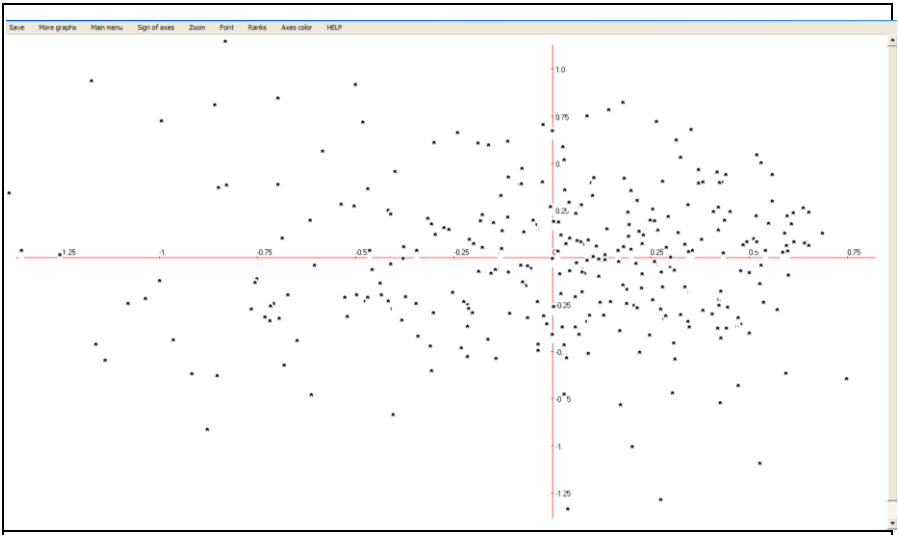
➢ Laisser les axes 1 er 2 (option par défaut) puis cliquez sur : **display**. Il est possible de ne faire figurer sur les plans que certaines variables.

➢ Cliquez alors sur : **Manual Selection of points**. Sélectionner les variables et les transférer dans la seconde fenêtre en cliquant sur : select.

⊙ La fenêtre du plan factoriel apparaît



PlaneView (1,2) – rubrique "colonnes actives + catégories supplémentaires"



PlaneView – rubrique "individus actifs (densité)"

Commentaires : Dans "les individus actifs (densité)", les identifiants des individus sont remplacés par un caractère simple [cas d'un ensemble d'individus très grand]. Cet affichage montre principalement la forme du nuage des individus, mais les identifiants d'origine peuvent s'afficher en cliquant sur le bouton droit de la souris.

Rappel : Pour chaque graphique, le bandeau du haut contient des options :
– *Font* offre la possibilité de modifier la police et la couleur des caractères ;
– *Sign of axes* permet d'inverser les axes ;
– *Save* sauvegarde le graphique en format *bmp*;
– *Rank*, est utile seulement dans le cas des affichages très complexes: ce bouton convertit les deux coordonnées de l'affichage courant en rangs.

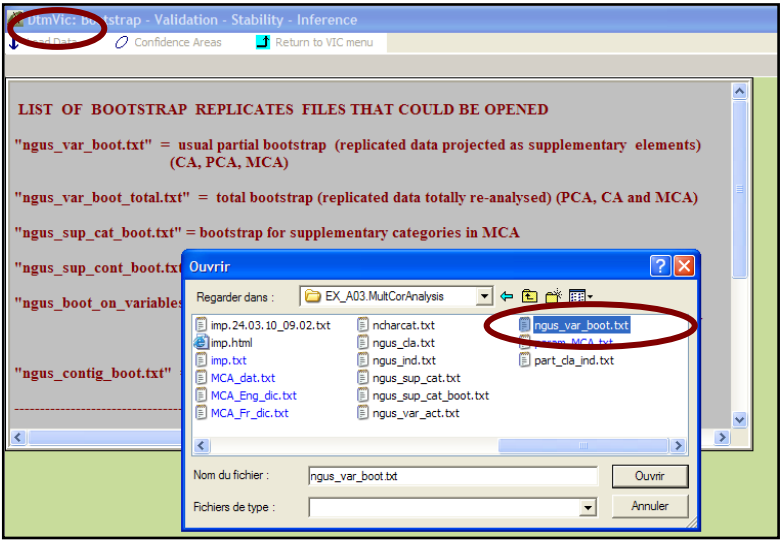
➢ Pour revenir au menu principal de Dtm-Vic, cliquez, selon la fenêtre, soit sur la croix en haut à droite, soit sur **Return**.

3- Validation Bootstrap

Cet outil permet de valider la position des variables sur le plan factoriel.

➢ Cliquez sur B Bootstrap

⊙ Une fenêtre *"DtmVic – Bootstrap – Validation – Stability - Inférence*" apparaît.

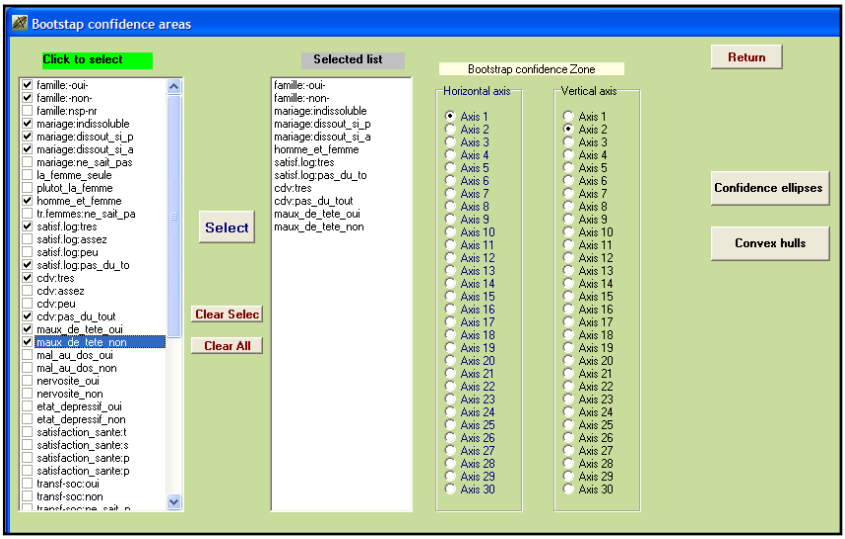


➢ Cliquer sur **Load Data** puis ouvrir dans le répertoire le fichier des réplications selon le *bootstrap* choisi.

➢ Sélectionnez le fichier **ngus_var_boot.txt** pour un bootstrap partiel.

Répondre **OK** à la fenêtre "*Set of principal coordinates loaded*" qui s'affiche.

- Puis cliquez sur **Confidence Ellipse**.
  - ⊙ une fenêtre "*Bootstrap confidence areas*" s'affiche
- sélectionnez dans la rubrique "*CLick to select"* les variables dont on veut visualiser les ellipses.
- Les transférer avec **Select**, dans la fenêtre "*Selected list*".
- Choisir ensuite le plan factoriel puis cliquez sur **Confidence ellipses** ou sur **Convex Hulls** pour obtenir l'affichage graphique des variables actives (si le fichier ngus_var_boot.txt a été chargé), ou de la catégorie supplémentaire (si le fichier ngus_sup_cat_boot.txt a été chargé).

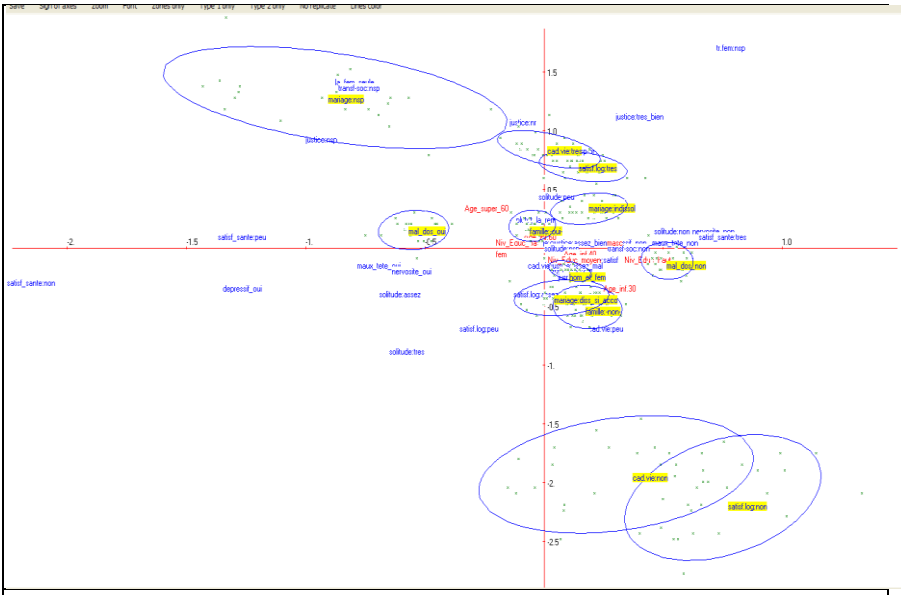


Les ellipses de confiance prennent en compte la densité du nuage de points-réplications, mais peuvent laisser quelques points à l'extérieur. Chaque ellipse de confiance est calculée à partir d'une analyse en composantes principales spécifique de l'ensemble des réplications.

Les enveloppes convexes (*Convex hulls*) enveloppent toutes les réplications, mais donnent du poids aux points périphériques sans aucune

considération de densité. Cf. par exemple le chapitre 7 de : *Multiple Correspondence Analysis and Related Techniques* (M. Greenacre and J. Blasius, eds) : Validation Techniques in Multiple Correspondence Analysis (L. Lebart). Chapman and Hall, 2006.

➢ Pour revenir au menu principal VIC, cliquez, selon la fenêtre, soit sur la croix en haut à droite, soit sur **Return**.



Exemples de zones de confiance bootstrap

4- Classification

Cette option positionne les classes obtenues sur le plan factoriel.

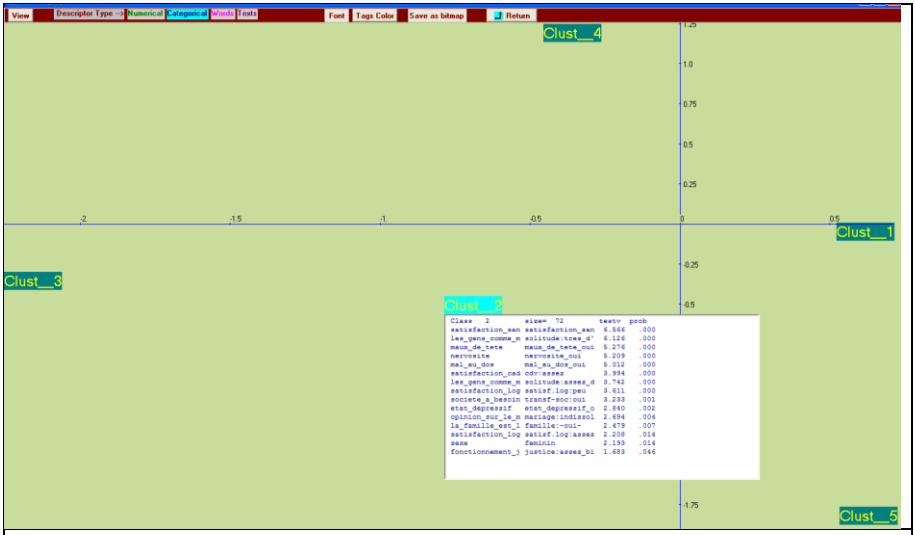
➢ Cliquez sur **ClusterView**. Choisissez les axes (1 et 2 pour commencer), et **Continuer**.

⊙ La fenêtre "*DTM-Display of clusters*" apparaît.

➢ Cliquez sur **View**. Les centroïdes des 5 classes apparaissent sur le plan factoriel.

➢ Actionnez le bouton **Categorical** du bandeau. Puis en cliquant (droit)

sur une classe, les variables descriptives de la classe apparaissent. L'ensemble des résultats figure dans la procédure DECLA du fichier de résultats.



Un clic droit sur l'étiquette d'une classe provoque l'affichage des éléments les plus caractéristiques de la classe. L'activation des éléments (*numerical, categorical*) se fait sur le bandeau supérieur du graphique.

On verra à propos des analyses textuelles que la même procédure ClusterView permet d'afficher aussi les mots caractéristiques des classes (pour la réponse des individus à une question ouverte) et les réponses caractéristiques (sous forme de texte) des classes.

Rappel : Dtm-Vic produit de nombreux fichiers de résultats intermédiaires liés à l'application (tous en format .txt). ***Il est, par conséquent, recommandé de créer un répertoire par application***. Au départ, un tel répertoire doit contenir les fichiers de données, dictionnaire et texte au format Dtm-Vic.

III. Données textuelles et mixtes : Prise en main de Dtm-Vic à partir de trois exemples

Ce chapitre présente un exemple d'analyse textuelle simple et deux exemples d'analyses élaborées utilisant à la fois des données numériques et textuelles (Dossier : **DtmVic_Examples_A_Start** de **DtmVic_Examples**)

- L'Exemple 4, contenu dans le sous-dossier **EX_A04.Text-Poems**, réalise une analyse lexicale à partir d'une série de textes (poèmes) : codage numérique des réponses ; application de l'analyse des correspondances au tableau lexical croisant les mots et les poèmes ; validation *Bootstrap* ; description des poèmes par leurs mots et vers caractéristiques ; carte de Kohonen des mots et poèmes ; sériation.
- L'Exemple 5, contenu dans le sous-dossier **EX_A05.Text-Responses_1**, porte sur l'analyse d'un jeu de données numériques et textuelles correspondant à des questions fermées et ouvertes d'une enquête : traitement des réponses à une question ouverte utilisant une variable nominale spécifique pour regrouper les réponses; codage numérique des réponses; analyse des correspondances de la table lexicale croisant les mots et les catégories d'individus; validation *Bootstrap* ; description des catégories par leurs mots et réponses ; carte de Kohonen simultanée des mots et des catégories.
- L'Exemple 6 utilise les mêmes données et dictionnaire que l'exemple 5. Il est contenu dans **EX_A06.Text-Responses_2** toujours dans le dossier **DtmVic_Examples_A_Start.** Il procède à une analyse directe des réponses à une question ouverte, sans regroupement préalable, avec classification des réponses et description des classes à partir des mots, des réponses caractéristiques et des caractéristiques des répondants.

Rappel : Dtm-Vic produit de nombreux fichiers de résultats intermédiaires liés à l'application (tous en format .txt). ***Il est, par conséquent, recommandé de créer un répertoire par application***. Au départ, un tel répertoire doit contenir les fichiers de données, dictionnaire ou texte au format Dtm-Vic.

III.1 Simples textes : Série de poèmes

Cet exemple élémentaire traite la forme la plus simple d'analyse des textes. Les données correspondent à une série de textes composée ici des 20 premiers sonnets de Shakespeare[6]. Dans ce format simple, Dtm-Vic peut traiter jusqu'à 1000 textes sans limitation de taille pour chaque texte. Cette portion de corpus, prise comme exemple, est ainsi un "modèle réduit", soulignant seulement les fonctionnalités (mais pas la puissance) de Dtm-Vic.

III.1.1 Le fichier DtmVic : "Série de poèmes"

Dans le cadre d'une analyse de texte, un seul fichier Dtm-Vic contenant l'ensemble des textes suffit. Celui de notre exemple est nommé Sonnet_LowerCase.txt et est contenu dans le répertoire **DtmVic-Examples_A_Start/EX_A04.Text-Poems** .

```
****    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,
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.

****    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,
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,
```

[6] Pour un ensemble plus important de sonnets et les commentaires attenants, se reporter au site : http://www.shakespeare-online.com/sonnets/.

```
if thou couldst answer 'this fair child of mine

****    S 20
a woman's face with nature's own hand painted
hast thou, the master-mistress of my passion;
a woman's gentle heart, but not acquainted
with shifting change, as is false women's fashion;
an eye more bright than theirs, less false in rolling,
gilding the object whereupon it gazeth;
a man in hue, all 'hues' in his controlling,
much steals men's eyes and women's souls amazeth.
and for a woman wert thou first created;
till nature, as she wrought thee, fell a-doting,
and by addition me of thee defeated,
by adding one thing to my purpose nothing.
but since she prick'd thee out for women's pleasure,
mine be thy love and thy love's use their treasure.

====
```

Les textes pouvant avoir des longueurs très différentes, une ligne spécifique sépare un sonnet d'un autre. Elle est caractérisée par des séparateurs "****" suivis de 4 espaces blancs et du nom du texte. Le symbole "====" indique la fin du fichier. Comme tous les fichiers de données en format Dtm-Vic, celui-ci est en format "txt". La conversion en minuscules permet ici de ne pas traiter différemment le premier mot de chaque vers.

L'objectif est de décrire les textes à partir de la table de contingence lexicale croisant les textes avec les mots les plus fréquents. La méthodologie générale à la base du traitement est présentée dans les livres : "*Statistique textuelle*" (L. Lebart, A. Salem, Dunod, 1994) et "*Exploring Textual Data*" (L. Lebart, A. Salem, L. Berry ; Kluwer, 1998, Dordrecht). L'ouvrage "*Statistique textuelle*" peut être librement téléchargé à partir du site : www.dtmvic.com.

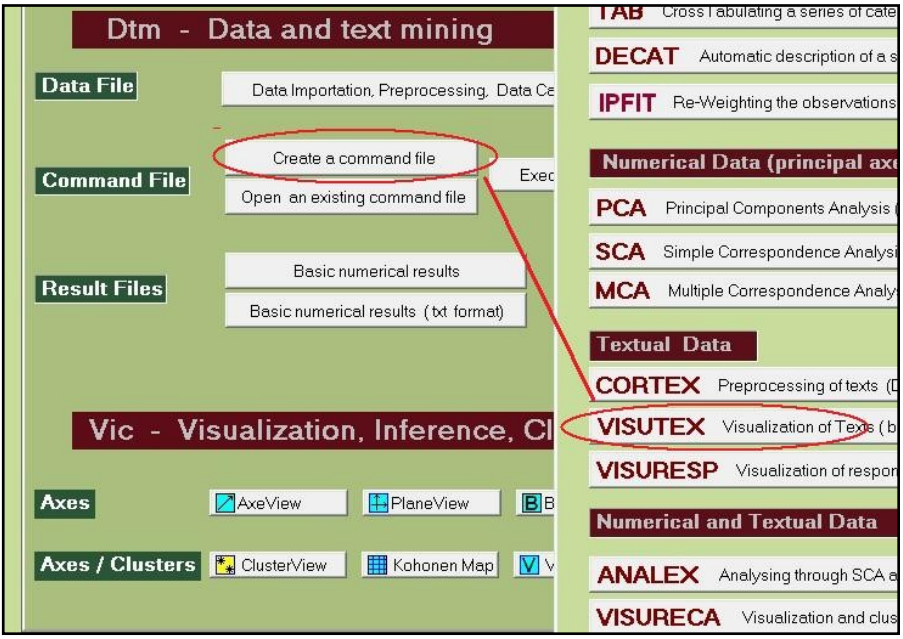
III.1.2. Mise en œuvre de l'analyse textuelle : "VISUTEXT"

Le fichier de commande, ou fichier paramètre, est créé en 4 étapes :

Etape 1 : Sélection de l'analyse

- Dans la fenêtre du menu principal, cliquez sur le bouton : **Create a ccommand file** de **Command File**

➢ Une fenêtre *"Choosing among some basic analyses"* apparaît.



➢ Sélectionnez l'analyse : **VISUTEX** – Visualization of texts de la rubrique : **Textual Data**.

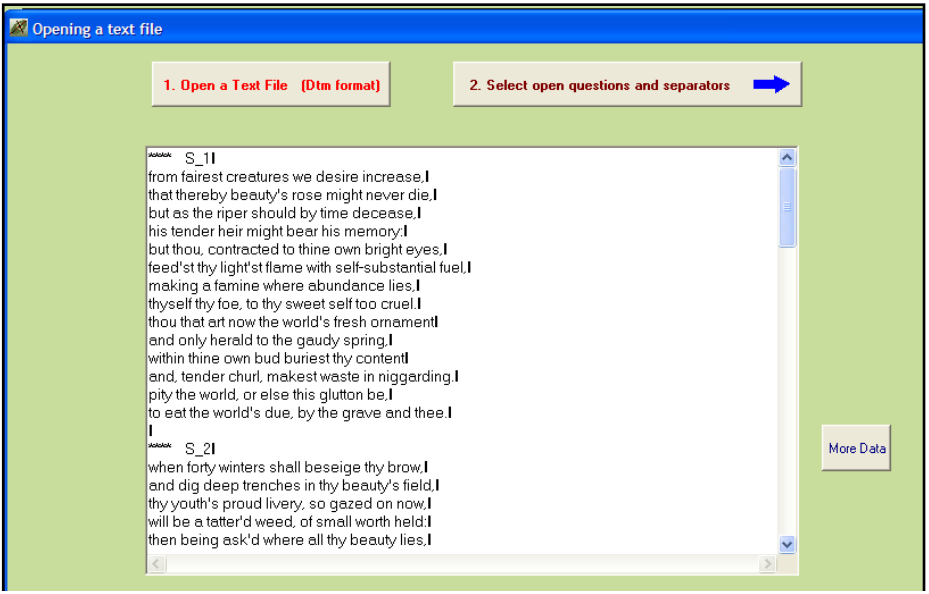
⊙ Une fenêtre : "*Opening a text file*" apparaît.

Etape 2 : Sélection du fichier texte

➢ Cliquez sur le bouton : **1. Open a text File**. Dans le répertoire **EX_A04.Text-Poems**, ouvrir le fichier Sonnet_LowerCase.txt.

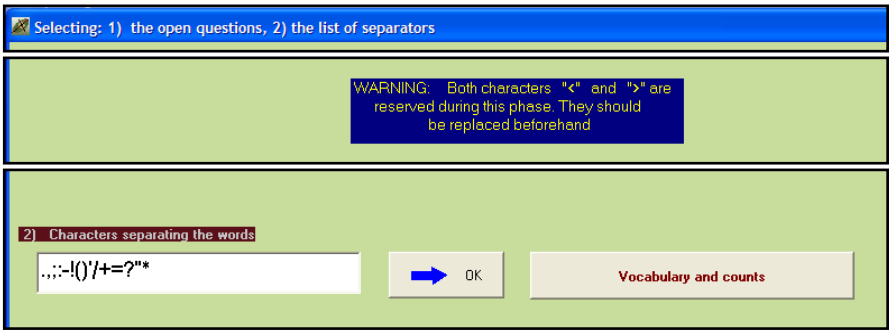
Après avoir cliqué sur : **OK** sur la boite de message donnant le nombre de lignes et de textes, le fichier s'affiche dans une première fenêtre.

➢ Cliquez ensuite sur : **2. Select Open questions and separators** ➔.



Etape 3 : Sélection des questions, mots et vocabulaire

La fenêtre suivante permet de sélectionner soit les questions ouvertes (ce qui n'est pas le cas ici), soit de compléter la liste des séparateurs des mots.

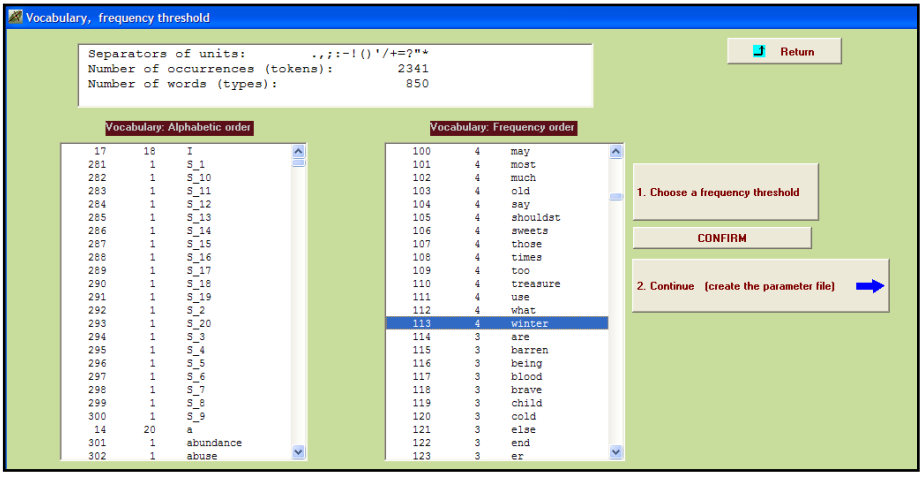


➢ Cliquez directement sur : **vocabulary and counts**

La fenêtre suivante présente le vocabulaire (ordre alphabétique et ordre de fréquence).

Nous devons choisir un seuil de fréquence en choisissant une ligne dans la rubrique : **Vocabulary : Frequency order**. La ligne 113 correspond à la

fréquence 4 (c'est une petite fréquence, adaptée à un petit corpus. Il s'agit ici simplement d'explorer l'éventail des commandes, sans interprétation linguistique pertinente…).

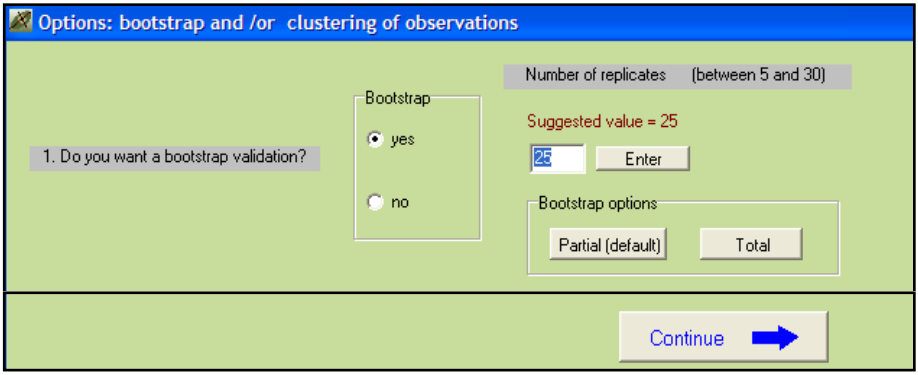


- Sélectionnez cette ligne 113 puis cliquez sur **CONFIRM**. La fréquence apparaît. Répondre **OK** à la boite de message.
- Cliquez sur **2. continue (create a parameter file)**.

Etape 4 : Création du fichier paramètre

C'est à cette étape de constitution du fichier paramètre qu'est proposée l'option *bootstrap* (cf. les trois exemples précédents).

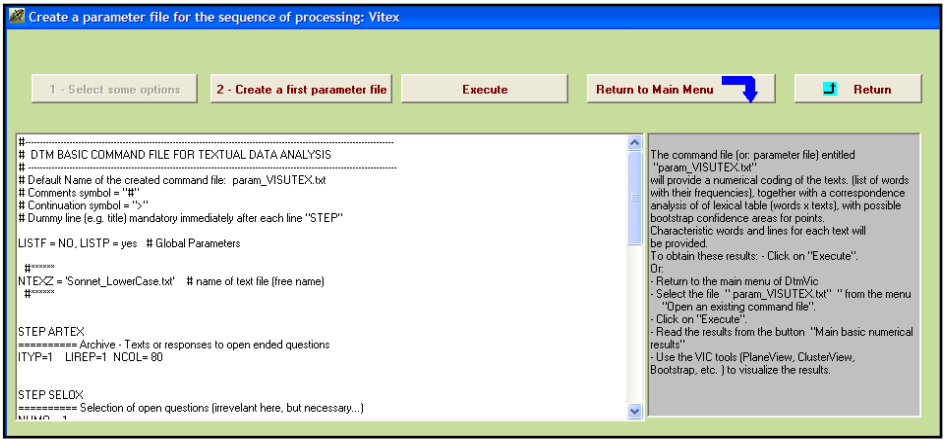
- Cliquez sur **1-Select some options**
  - une fenêtre "*Options : Bootstrap and/or Clustering of observations*" apparaît.



- Cliquez sur "yes" pour la procédure "bootstrap" ; indiquez le nombre de réplications (par défaut 25) puis **Enter**. Si le bootstrap n'est pas adopté, cliquez sur "no".
- Cliquez sur Continue ➔

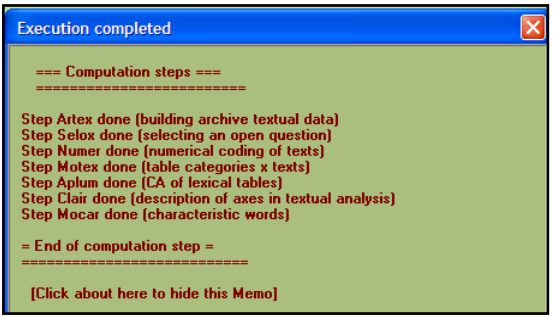
  ⊙ la fenêtre "*Create a parameter file*" apparaît de nouveau.

- Cliquez sur **2-Create a first parameter file.** Un fichier de commande (*parameter file*) vient d'être créé sous le nom param_VISUTEX.txt et stocké dans le dossier **EX_A04.Text-Poems** du répertoire **DtmVic-Examples_A_Start**. (Pour le conserver en vue d'analyses ultérieures, il faudra le renommer).



- Cliquez sur **3-Execute**

Les procédures s'affichent en bloc après l'exécution : Artex (Archivage des textes), Selox (Sélection des questions ouvertes), Numer (Numérisation du texte), Motex (table de contingence Mots-textes), Aplum (analyse des correspondances pour ce type de tables), Clair (brève description des axes factoriels), Mocar (mots et lignes caractéristiques).



Note : Une fois le fichier de commande créé (fichier paramètre : param_VISUTEXT.txt), il est possible de l'ouvrir, lors d'une utilisation ultérieure de DtmVic , dans le menu principal **Command File** avec le bouton : **Open an existing command file** puis d'exécuter ce fichier : **Execute**. Les utilisateurs expérimentés peuvent aussi modifier les paramètres directement sous l'éditeur proposé par Open (avec l'aide du bouton "Help about parameters" disponible dans l'éditeur) ou avec un autre éditeur de texte hors de Dtm-Vic.

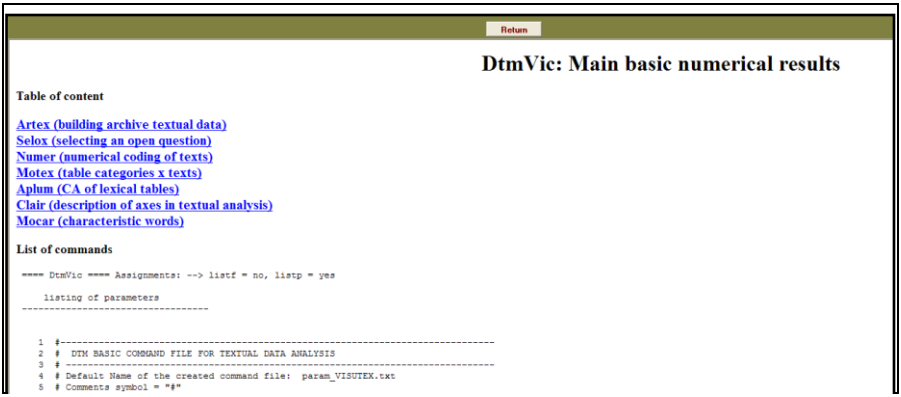
III.1.3 Fichier de résultats

Les résultats peuvent être consultés dans la rubrique : **Result Files**

Cliquez sur : Basic numerical results pour naviguer dans le fichier de résultats en format html puis sur : Return pour en sortir et revenir au menu principal, ou cliquez sur Basic numerical results (text format) pour ouvrir le fichier de résultats en format texte.

Les fichiers de résultats sont dans le répertoire **EX_A04.Text-Poems**.

Rappel : Le fichier résultat "imp.txt" (comme son homologue "imp.html") est également sauvegardé sous le nom "imp" suivi de la date et l'heure de l'analyse : "imp_18.07.11_14.45.txt" signifie le 18 juillet 2011, à 14h 45. Ce fichier de sauvegarde garde comme archives les résultats numériques principaux tandis que les dossiers "imp.txt" et "imp.html" sont écrasés à chaque nouvelle analyse exécutée dans le même répertoire.


Return

DtmVic: Main basic numerical results

Table of content

Artex (building archive textual data)
Selox (selecting an open question)
Numer (numerical coding of texts)
Motex (table categories x texts)
Aplum (CA of lexical tables)
Clair (description of axes in textual analysis)
Mocar (characteristic words)

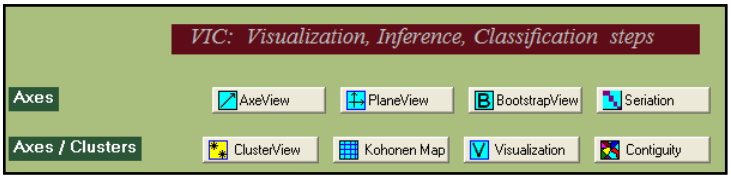
List of commands

```
==== DtmVic ==== Assignments: --> listf = no, listp = yes

    listing of parameters
 ---------------------------------


  1  #-------------------------------------------------------------------------
  2  #  DTM BASIC COMMAND FILE FOR TEXTUAL DATA ANALYSIS
  3  #-------------------------------------------------------------------------
  4  # Default Name of the created command file:  param_VISUTEX.txt
  5  # Comments symbol = "#"
```

La lecture de ce fichier est utile pour prendre connaissance de certains résultats qui ne peuvent être visualisés. La procédure NUMER, nous apprend, par exemple, que la table lexicale se présente sous la forme de 280 réponses (lignes), avec un nombre total de mots (occurrences) de 2321, impliquant 830 mots distincts. Utilisant un seuil de fréquence de 4, ce qui signifie que l'on conserve les mots de fréquence supérieure à trois) le nombre de mots conservés se réduit à 1384, tandis que le nombre de mots distincts est ramené à 114.

III.1.4 Visualisation des résultats et interprétation

Cette deuxième phase fondamentale de Dtm-Vic fournit les outils de visualisation nécessaires à la validation et l'interprétation des résultats.

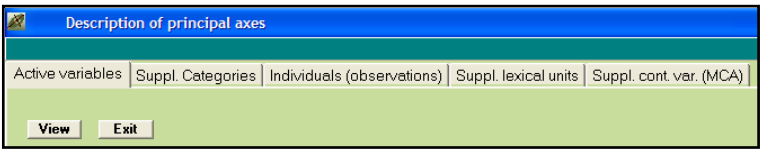


1- Axes factoriels

Cet outil fournit les coordonnées sur les axes factoriels des variables actives, supplémentaires, ou des observations.

- Cliquez sur : AxesView.

Dans le contexte de cette analyse textuelle, seulement deux options sont envisageables : "active variables" (qui correspondent ici aux poèmes) et les "observations" (qui correspondent ici aux mots).



- Cliquez sur l'onglet des éléments à examiner, Active variables ou Individuals (observations) puis sur **View**. Il est possible d'ordonner les coordonnées d'un axe donné, en cliquant sur cet axe.
- Cliquez : **Exit** pour sortir de cet outil.

Active variables | Suppl. Categories | Individuals (observatio

View | Exit

Identifier	axis 1	axis 2	axis 3	axis 4	axis 5
S_1	-263	237	4	-214	580
S_10	-340	-360	273	-9	634
S_11	-321	-158	246	-296	-136
S_12	68	744	331	370	-583
S_13	1402	-799	50	-298	-46
S_14	-61	535	442	465	-17
S_15	574	337	25	104	-239
S_16	1156	-236	247	-81	119
S_17	583	-98	-172	108	137
S_18	-64	370	20	540	-59
S_19	25	319	354	74	4
S_2	-136	202	-196	381	197
S_20	-135	-10	-195	50	211
S_3	-307	34	70	-208	381
S_4	-741	-612	-237	-750	-683
S_5	104	9	-1052	167	-837

Coordonnées des sonnets (variables actives)

Active variables | Suppl. Categories | Individuals (observations

View

Identifier	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	ax
a	316	14	-504	-114	157	321	-
age	83	582	-442	-776	221	-1047	1(
all	-8	483	301	393	-256	64	1
an	-17	172	-910	-75	783	172	-6
and	87	328	90	158	-156	-35	1
another	-713	-177	414	-470	212	-686	1
art	-601	-370	221	578	736	123	1
as	34	418	289	259	-39	-2	
be	-648	-774	-222	143	279	239	2
bear	565	-505	832	-615	402	104	-4
beauty	-149	68	-423	90	-266	216	9
but	250	104	-174	-182	43	187	-2
by	-61	293	365	-100	223	270	-4
can	386	740	495	933	-314	319	3
change	-114	-203	218	86	810	634	-3
d	-72	-35	-486	246	-193	-188	2
day	691	686	-59	-28	-488	-391	4
death	-4	-1304	-82	1006	-179	-683	2

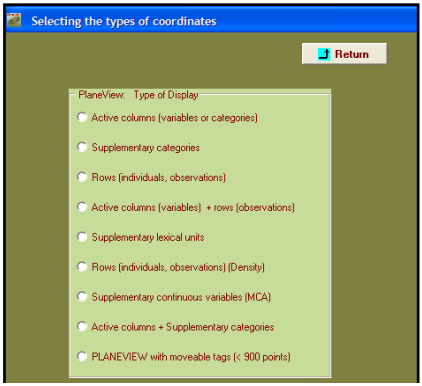
Coordonnées des mots (observations)

2- Plans factoriels

Cette option fournit les plans factoriels séparés ou superposés des sonnets (variables actives) et des mots (observations).

- Cliquez sur PlaneView
  - ⊙ Une fenêtre s'affiche proposant différents plans factoriels.

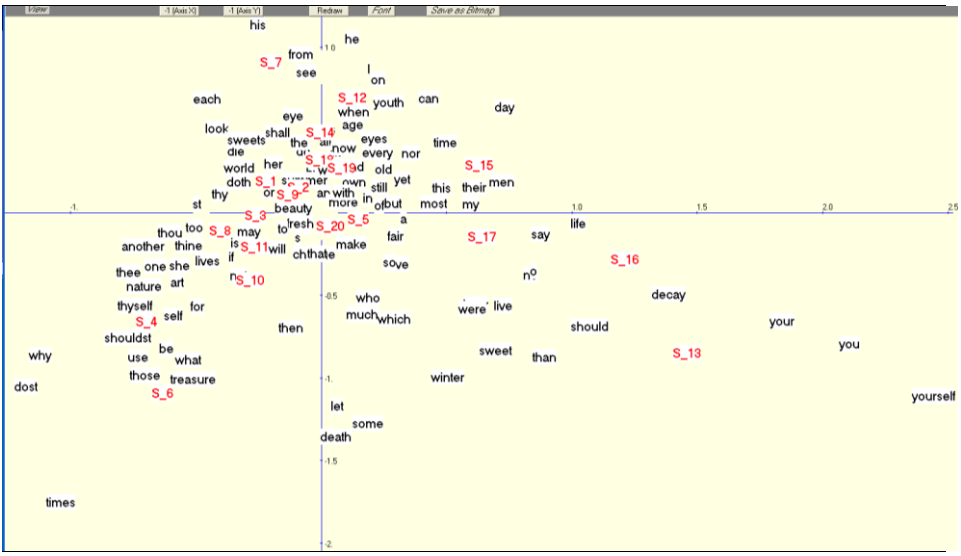
Parmi les configurations de plans factoriels proposées, l'option "active columns + actives rows" est adaptée à cette analyse.



- Sélectionnez la rubrique "Actives columns (variables) + rows (observations)".
  - ⊙ Une fenêtre pour sélectionner le plan factoriel suivant la paire d'axes souhaitée apparaît.
- Choisir les axes 1 er 2 puis cliquez sur : **display**. Il est possible de ne faire figurer sur les plans que certaines variables. Cliquez alors sur : **Manual Selection of points**. Sélectionner les variables et les transférer dans la seconde fenêtre en cliquant sur : select.
  - ⊙ La fenêtre du plan factoriel apparaît.

On peut également choisir ce menu par l'intermédiaire de "PLANEVIEW with moveable tags" qui reprend certaines des rubriques précédentes et permet de déplacer les points superposés pour rendre plus lisible le graphique.

Rappel : Pour chaque graphique, le bandeau du haut contient des options : *Font* offre la possibilité de modifier la police et la couleur des caractères ; "Sign of axes" permet d'inverser l'orientation des axes ; "Save" sauvegarde le graphique en format bmp; "Rank", est utile seulement dans le cas des affichages très complexes : ce bouton convertit les deux coordonnées de l'affichage courant en rangs. Par exemple, les n valeurs de l'abscisse sont converties en nombres entiers de 1 à n, ayant le même ordre que les valeurs originales. Ainsi les deux distributions sont uniformes, et les identifiants se recouvrent moins au prix d'une déformation substantielle de l'affichage).



Positionnement des sonnets et des mots dans le plan factoriel principal.

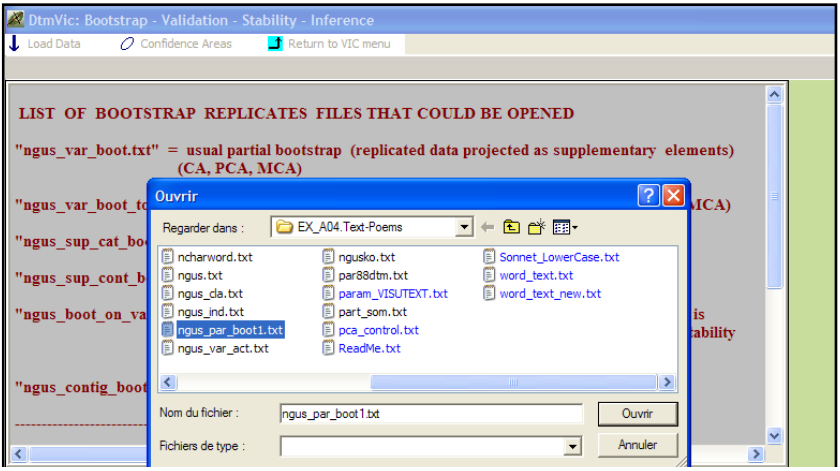
Choisir une option puis cliquez sur : ***View***

- Pour revenir au menu principal de Dtm-Vic, cliquez sur : **return**.

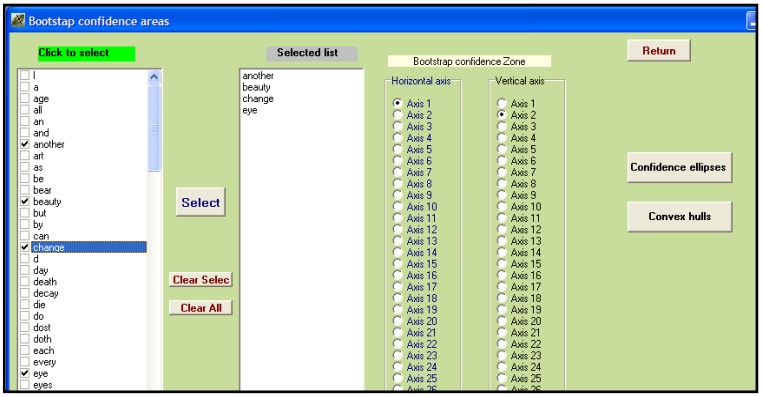
3- Validation Bootstrap

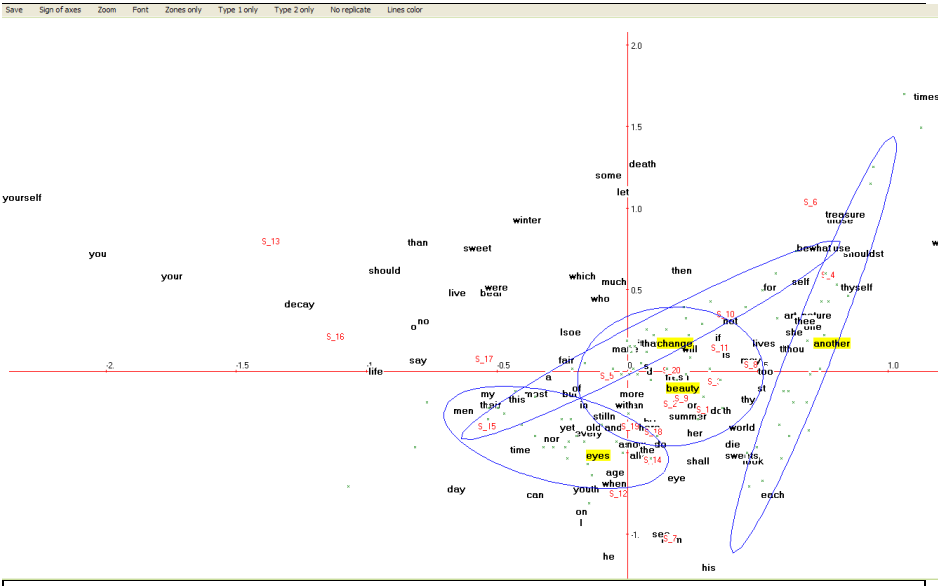
[Voir l'encadré technique sur le bootstrap, chap. II, section II.1.2, Etape 5]

- Cliquez sur : B Bootstrap pour valider la position des variables sur les plans factoriels.
  - ⊙ Une fenêtre : *"DtmVic – Bootstrap – Validation – Stability – Inférence*" apparaît.
- Cliquez sur : **Load Data** puis ouvrir dans le répertoire le fichier des réplications selon le bootstrap choisi. Sélectionnez le fichier ngus_par_boot1.txt pour un bootstrap partiel dans le cas textuel.
- Répondre : **OK** à la fenêtre : "*Set of principal coordinates loaded*" qui s'affiche.



- Puis cliquez sur : **Confidence Ellipse**.
  - ⊙ une fenêtre : "*Bootstrap confidence areas*" s'affiche
- sélectionnez dans la rubrique : "*Click to select"* les variables dont on veut visualiser les ellipses. Les transférer avec : **Select**, dans la fenêtre "*selected list*". Choisir ensuite le plan factoriel puis cliquer sur : **Confidence ellipses** ou sur sur : **Convex Hulls** (cf § II.1.4.3-Bootstrap) pour obtenir l'affichage graphique des éléments actifs (si le dossier ngus_par_boot1.txt a été chargé).



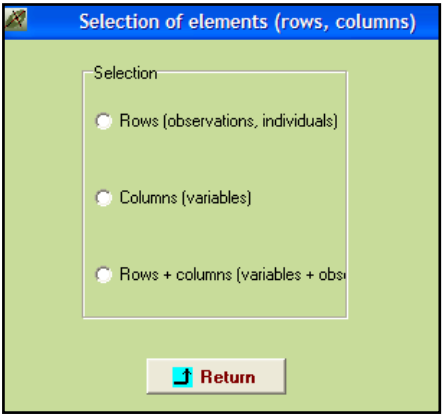


Commentaires : Les ellipses correspondant aux points "change" et "beauty" contiennent l'origine des axes : on ne peut rejeter l'hypothèse selon laquelle la distribution des ces points est indifférenciée dans les 20 textes. En revanche, le mot "another" a une position typée sur le premier axe (et neutre sur le second). Le mot "eye" a une position significative sur le second axe.

4- Cartes auto-organisées de Kohonen

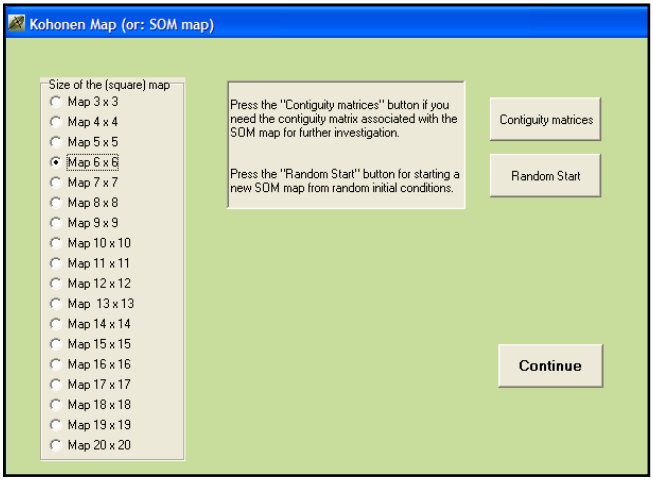
➢ Cliquez sur Kohonen Map.

⊙ Une fenêtre *"Selection of elements*" apparaît.



Les colonnes c'est-à-dire les variables actives sont les mots, et les lignes c'est-à-dire les observations, sont les poèmes. On souhaite représenter sur une même carte les mots et les poèmes.

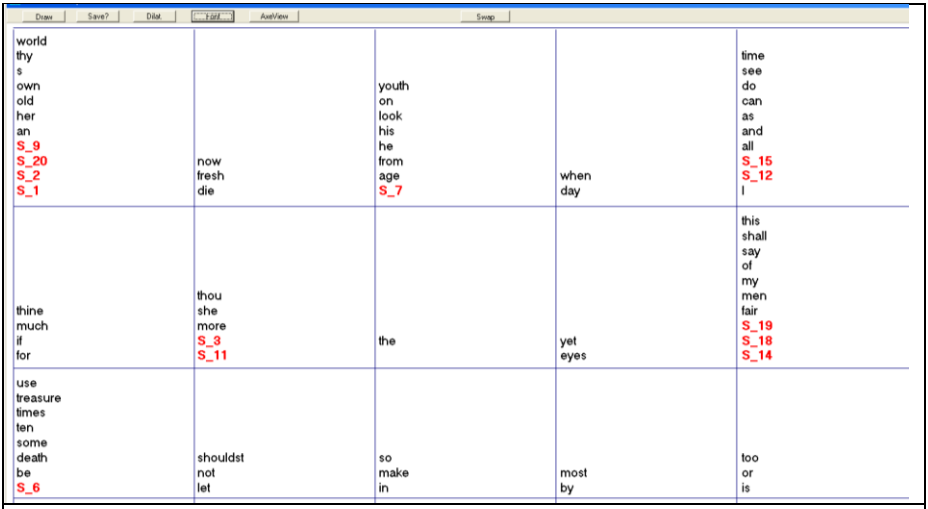
- Cliquez sur "Rows + columns"
  - ⊙ Une fenêtre *"Kohonen map*" apparaît.



- Choisir la carte "map 5x5" puis **Continue** et répondre OK à la boite de message : "SOM map completed".
  - ⊙ Une nouvelle fenêtre s'affiche.
- Actionnez **Draw**. La Carte de Kohonen apparaît.

Nous avons obtenu une représentation simultanée des lignes et des colonnes, due à l'utilisation, comme fichier d'entrée, des coordonnées de l'analyse de correspondance de la table lexicale. Dans le cadre de cet exemple, les autres articles du menu principal ne sont pas appropriés.

Notons que, pour toute l'analyse présentée, aucune transformation préalable n'a été opérée sur le vocabulaire. La procédure CORTEX aurait pu précéder la procédure VISUTEXT pour fusionner des mots (formes graphiques relatives à un même lemme) ou pour supprimer certains mots (mots outils par exemple). Toutefois, une analyse préalable des matériaux bruts est toujours conseillée.



world thy s own old her an S_9 S_20 S_2 S_1	now fresh die	youth on look his he from age S_7	when day	time see do can as and all S_15 S_12 I
thine much if for	thou she more S_3 S_11	the	yet eyes	this shall say of my men fair S_19 S_18 S_14
use treasure times ten some death be S_6	shouldst not let	so make in	most by	too or is

Extraits de la carte de Kohonen représentant simultanément les sonnets et les mots.

Remarque : Il est possible de changer de taille de police ("Font") et de dilater la carte de Kohonen obtenue ("Dilat") pour rendre le graphique plus lisible.
Les mots apparaissant dans la même cellule sont souvent associés aux mêmes réponses (sonnets). Cette propriété tient, à un moindre degré, pour les cellules contiguës.

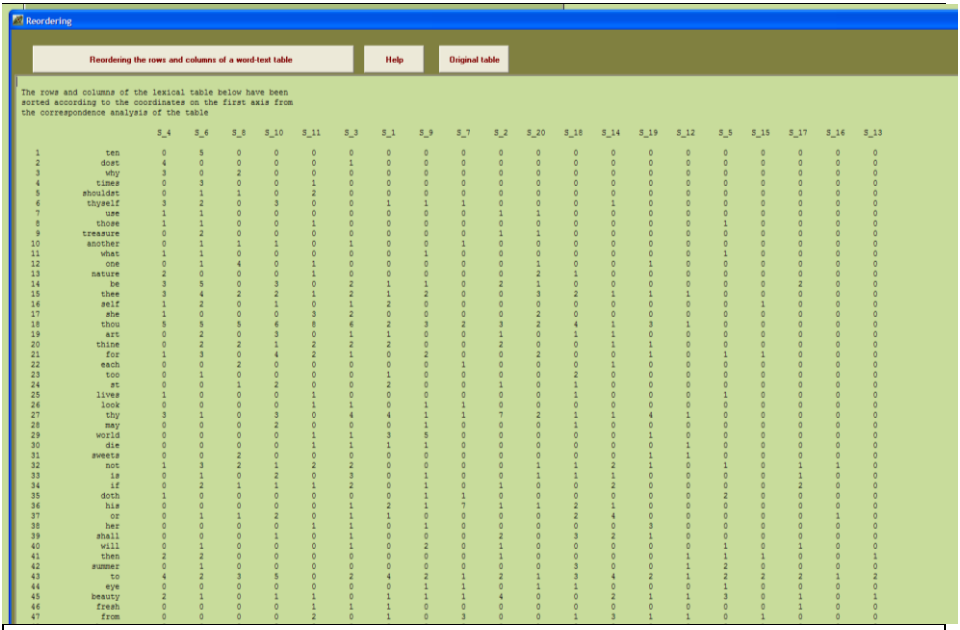
5- Sériation

(Voir l'encadré du paragraphe I.3 du chapitre 1)

La sériation est appliquée ici à la table lexicale croisant les 20 sonnets et les mots choisis (mots apparaissant au moins 4 fois dans le corpus).

- Cliquez sur Seriation.
  - ⊙ La fenêtre "*Reordering*" apparaît.
- Cliquez sur **Reordering the rows and the columns of a word-text table**.
- Répondre OK à "Seriation of rows and columns of the lexical table completed".

La table réordonnée en lignes et en colonnes croisant les 20 sonnets et les mots retenus est alors constituée.



		S_4	S_6	S_8	S_10	S_11	S_3	S_1	S_9	S_7	S_2	S_20	S_18	S_14	S_19	S_12	S_5	S_15	S_17	S_16	S_13
1	ten	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	dost	4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	why	3	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	times	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	shouldst	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	thyself	3	2	0	3	0	0	1	1	1	0	0	0	1	0	0	0	0	0	0	0
7	use	1	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
8	those	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
9	treasure	0	2	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
10	another	0	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
11	what	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
12	one	0	1	4	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
13	nature	2	0	0	0	1	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0
14	be	3	5	0	3	0	2	1	1	0	2	1	0	0	0	0	0	0	2	0	0
15	thee	3	4	2	2	1	2	1	2	0	0	3	2	1	1	1	0	0	0	0	0
16	self	1	2	0	1	0	1	2	0	0	0	0	0	0	0	0	0	1	0	0	0
17	she	1	0	0	0	3	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0
18	thou	5	5	5	6	8	6	2	3	2	3	2	4	1	3	1	0	0	0	0	0
19	art	0	2	0	3	0	1	1	0	0	1	0	1	1	0	0	0	0	0	0	0
20	thine	0	2	2	1	2	2	2	0	0	2	0	0	1	1	0	0	0	0	0	0
21	for	1	3	0	4	2	1	0	2	0	0	2	0	0	1	0	1	1	0	0	0
22	each	0	0	2	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
23	too	0	1	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0
24	st	0	0	1	2	0	0	2	0	0	1	0	1	0	0	0	0	0	0	0	0
25	lives	1	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
26	look	0	0	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0
27	thy	3	1	0	3	0	4	4	1	1	7	2	1	1	4	1	0	0	0	0	0
28	may	0	0	0	2	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
29	world	0	0	0	0	1	1	3	5	0	0	0	0	0	1	0	0	0	0	0	0
30	die	0	0	0	0	1	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0
31	sweets	0	0	2	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
32	not	1	3	2	1	2	2	0	0	0	0	1	1	2	1	0	1	0	1	1	0
33	is	0	1	0	2	0	3	0	1	0	0	1	1	1	0	0	0	0	1	0	0
34	if	0	2	1	1	1	2	0	1	0	1	0	0	2	0	0	0	0	2	0	0
35	doth	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	2	0	0	0	0
36	his	0	0	0	0	0	1	2	1	7	1	1	2	1	0	0	0	0	0	0	0
37	or	0	1	1	2	0	1	1	0	0	0	0	2	4	0	0	0	0	0	1	0
38	her	0	0	0	0	1	1	0	1	0	0	0	0	0	3	0	0	0	0	0	0
39	shall	0	0	0	1	0	1	0	0	0	2	0	3	2	1	0	0	0	0	0	0
40	will	0	1	0	0	0	1	0	2	0	1	0	0	0	0	0	1	0	1	0	0
41	then	2	2	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	1
42	summer	0	1	0	0	0	0	0	0	0	0	0	3	0	0	1	2	0	0	0	0
43	to	4	2	3	5	0	2	4	2	1	2	1	3	4	2	1	2	2	2	1	2
44	eye	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	0	0
45	beauty	2	1	0	1	1	1	1	1	1	4	0	0	2	1	1	3	0	1	0	1
46	fresh	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0
47	from	0	0	0	0	2	0	1	0	3	0	0	1	3	1	1	0	1	0	0	0

Commentaire : On peut voir (ou deviner... si les caractères sont trop petits) que les premiers mots de la liste des mots réordonnée caractérisent (parfois exclusivement) les premiers sonnets dans la liste elle-même réordonnée de sonnets. Les derniers mots de la même liste ordonnée sont absents ou rarement observés parmi ces sonnets. Cependant, ils sont fréquents parmi les derniers sonnets (côté droit de la table).

Le bouton : **Original table** permet d'inspecter la table lexicale pour laquelle les lignes et les colonnes ont leur disposition initiale.

III.2. Analyse textuelle de questions ouvertes

Cet exemple vise à décrire les réponses à une question ouverte dans une enquête par sondage en relation avec des réponses à des questions fermées. Il s'agit de confronter les profils lexicaux des réponses de certaines catégories de répondants choisies *a priori*.

III.2.1 Les données et fichiers Dtm-Vic :

"Enquête internationale sur les attitudes et valeurs"

L'enquête qui va nous servir d'exemple a été menée dans sept pays (Japon, France, Allemagne, Royaume-Uni, Etats-Unis, Pays Bas, Italie) vers la fin des années 80[7]. Nous présentons ici le volet britannique de cette enquête, que nous désignerons par "Enquête *Life*", qui traite les réponses de 1043 individus à 14 questions fermées et à 3 questions ouvertes. Les questions fermées concernent à la fois les caractéristiques objectives du répondant ou de son ménage (âge, statut, genre, équipements) et des questions sur les attitudes et les valeurs des personnes interrogées, dont la plupart furent extraites du questionnaire de l'enquête "Aspiration" (exemple de la section II.3, ACM).

Trois questions ouvertes ont été posées :

- *"Qu'est ce qui est le plus important pour vous dans la vie ?"*
- "*Quelles sont les autres choses très importantes pour vous ?"* (relance de la première question)
- *"Que pensez vous de la culture de votre pays ?"*

Nous nous intéressons ici aux deux premières questions que nous voulons par la suite mettre en relation avec l'âge et le niveau d'instruction du répondant. Une variable nominale à 9 catégories est créée combinant les trois niveaux d'âge avec trois degrés d'instruction.

Cet exemple est disponible dans le dossier **EX_A05.Text-Responses_1**

[7] Cf. Hayashi C., Suzuki T., Sasaki M. (1992): *Data Analysis for Social Comparative research: International Perspective*, North-Holland, Amsterdam. Le Professeur Chikio Hayashi, ancien Directeur de *l'Institute of Statistical Mathematics* (Tokyo) et maître d'œuvre de ces enquêtes, fût aussi un de premiers « découvreur » de l'analyse des correspondances.

inclus dans le répertoire **DtmVic-Examples_A_Start**. On y trouve 3 fichiers d'entrée Dtm-Vic : Dictionnaire, Données numériques, Données textuelles.

Ces fichiers en format Dtm-Vic peuvent être générés par une procédure d'importation à partir d'un fichier Excel unique (cf. chapitre IV).

1 - fichier de données pour les questions fermées : TDA_dat.txt **(extrait)**

```
'___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
```

Ce fichier comprend 1043 lignes (les individus) et 15 colonnes séparées par des espaces blancs. La première colonne correspond à l'identifiant de l'individu, les 14 autres sont les valeurs des réponses aux questions fermées représentées par des variables nominales ou numériques continues.

2. Fichier dictionnaire des questions fermées : TDA_dic.txt **(extraits)**

```
   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
```

Le fichier dictionnaire contient les identifiants des 14 variables.

Rappel 1 : L'identifiant d'une variable nominale est précédé par le nombre N de ses catégories (en colonne 5). Les N lignes suivantes identifient les N catégories des réponses : un "identifiant court" en 4 caractères occupe les colonnes 1 à 5 et un "identifiant long" (20 caractères maximum) commence à la colonne 6. Une variable numérique telle que l'âge ou le nombre d'enfants, a 0 catégorie.

Rappel 2 : les espaces vides dans les identifiants ne sont pas permis.

3. Fichier des textes des questions ouvertes : TDA_tex.txt (extraits)

```
----'___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

..

----1042
 to see my daughter settled in a job
++++
 health,  healthy enough to keep them secure,  that I get
  on well with my neighbours,  a life outside my family circle,
++++
 folk music,  architecture, particularly religious
  architecture,
----1043
 contentment
++++
 my children's health and happiness
++++

====
```

Ce fichier contient les réponses libres de 1043 individus aux trois questions ouvertes citées précédemment. Le format du fichier des textes est assez spécifique, mais transparent pour l'utilisateur (format .txt).

Rappel sur le format interne Dtm-Vic : Puisque les réponses peuvent avoir des longueurs très différentes, des séparateurs sont utilisés pour distinguer les questions des individus (ou répondants). Les individus [qui doivent impérativement être dans le même ordre que dans le fichier de données numériques] sont séparés par la chaîne de caractères "----" (commençant à la colonne 1) suivie éventuellement de l'identifiant de l'individu.

Puis à la ligne suivante, viennent les réponses aux questions ouvertes, séparées par "++++" (commençant à la colonne 1). Le symbole "====" indique la fin du fichier. Comme tous les fichiers de données Dtm-Vic, ce fichier est un dossier de texte brut (.txt). Si le dossier des textes vient d'une phase de traitement de textes, il doit être sauvé en ".txt".

Après archivage des fichiers dictionnaire, des données et des textes, le codage numérique du texte nous permet de construire une table lexicale croisant les mots avec une variable nominale sélectionnée. Une analyse de correspondance est alors exécutée sur cette table lexicale[8]. Des zones de confiance *bootstrap* pourront être dessinées autour des mots et des catégories d'individus.

III.2.2. Mise en œuvre de l'analyse textuelle sur tableau lexical agrégé – ANALEX

Le fichier paramètre est créé en 5 étapes :

Etape 1 : Sélection de l'analyse

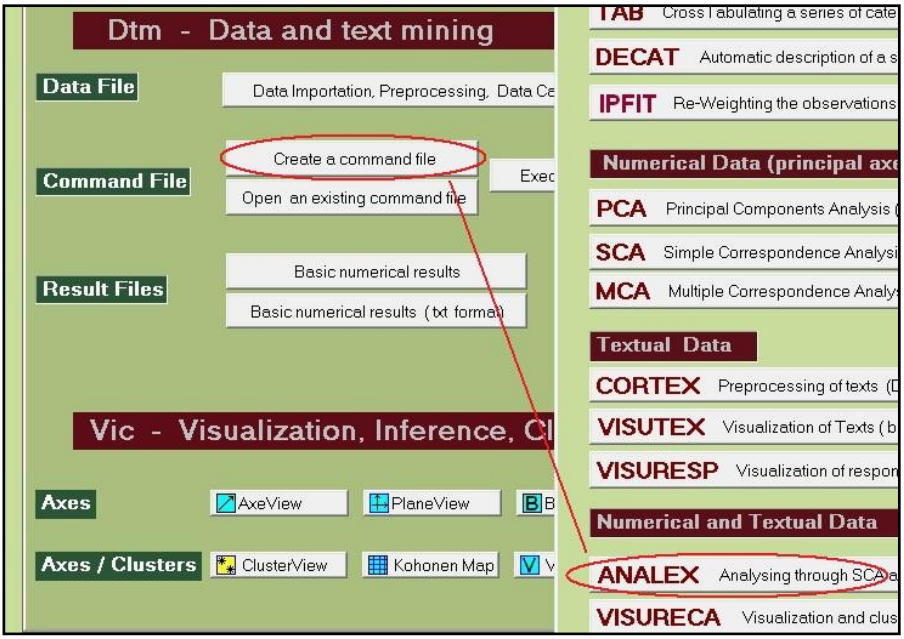
➢ Dans le *menu principal*, cliquez sur : **Create** de **Command File**.

⊙ Une fenêtre: *"Choosing among some basic analysis"* apparaît.

➢ Sélectionnez l'analyse **ANALEX** – Analysing through SCA of a lexical table built from a specific categorical variable dans la rubrique **Numerical and Textual Data**.

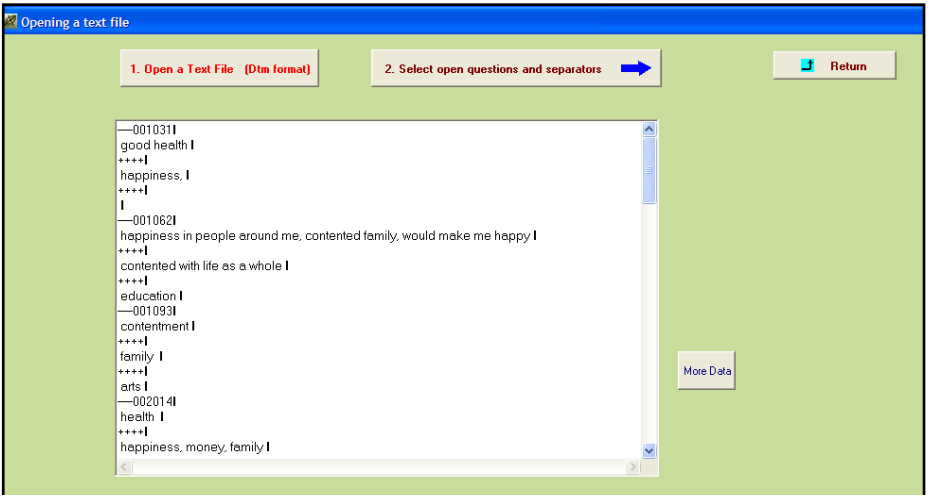
⊙ Une fenêtre : "*Opening a text file*" apparaît.

[8] De plus amples explications à propos de cet exemple particulier et de la méthodologie correspondante peuvent être trouvées dans le livre : « *Exploring Textual Data*» (L. Lebart, A. Salem, L. Berry ; Kluwer AcademicPublisher, 1998).



Etape 2 : Sélection du fichier texte

- Cliquez sur le bouton : Open a text File. Dans le répertoire **EX_A05.Text-Responses**, ouvrir le fichier : TDA _tex.txt.
- Une boite de message récapitule les informations de ce fichier : 7329 lignes (correspondant à l'ensemble des réponses aux trois questions), 1043 observations (les répondants) et 3 questions ouvertes.
- Cliquez sur : **OK**, le fichier texte en format Dtm-Vic de type 2 s'affiche dans une première fenêtre.

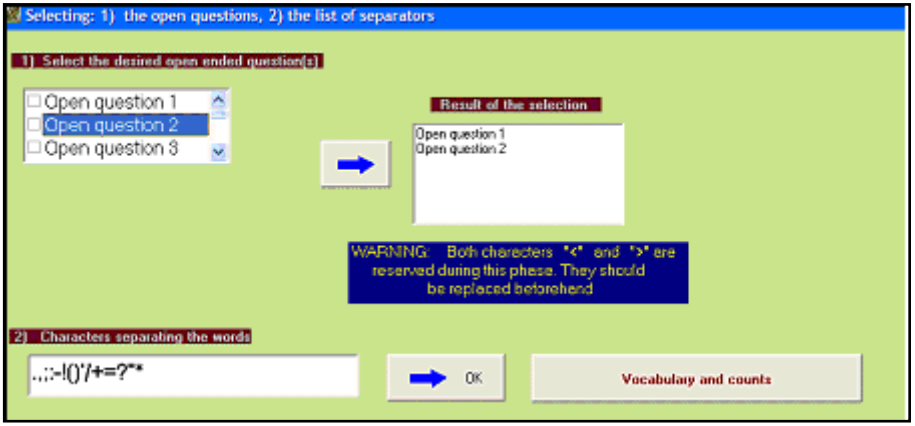


➢ Cliquez sur : **2.Select Open questions and separators**

⊙ Une nouvelle fenêtre ayant pour titre : "*Selecting : 1) the open questions, 2) the list of separators*" apparaît.

Etape 3 : Sélection des questions ouvertes

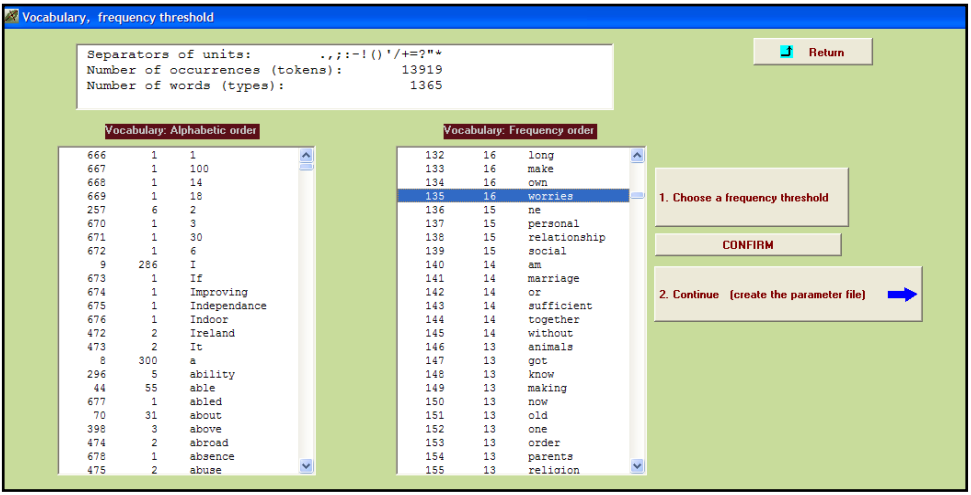
➢ Sélectionnez les questions ouvertes 1 et 2 et les transférer dans "*Result of the selection*". Puis choisir les séparateurs. Ici, nous adoptons ceux proposés par défaut. Cliquez alors sur **Vocabulary and counts.**



➢ La fenêtre suivante présente le vocabulaire (alphabétique et par ordre de fréquence).

Nous devons choisir un seuil de la fréquence en choisissant une ligne dans la rubrique "*Vocabulary (frequency order)*". La ligne 135 correspond à la fréquence 16.

➢ Sélectionnez cette ligne puis : CONFIRM. La fréquence apparaît.
Répondre OK



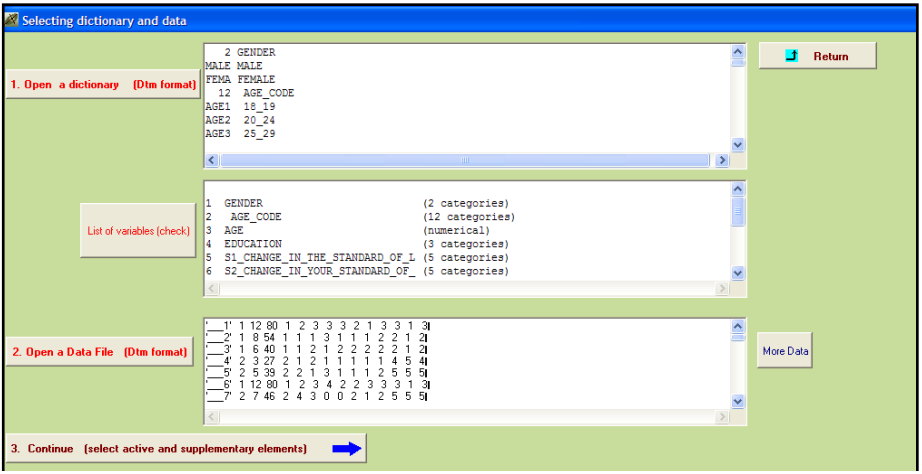
➢ Cliquez sur **2. Continue (create the parameter file)**.

⊙ Une fenêtre d'ouverture "*fichiers dictionnaires et données*" apparaît

Etape 4 : Sélection des fichiers dictionnaire et de données

➢ Cliquez sur le bouton : **Open a dictionary**. Dans le répertoire **EX_A05.Text-Responses**, ouvrir le fichier TDA_dic.txt. Il s'affiche dans une première fenêtre. Le statut (nominal ou numérique) des variables est indiqué dans une deuxième fenêtre

➢ Cliquez sur le bouton : **Open a Data File**. Dans le répertoire **EX_A05.Text-Responses**, ouvrir le fichier TDA_dat.txt qui s'affiche dans une troisième fenêtre.

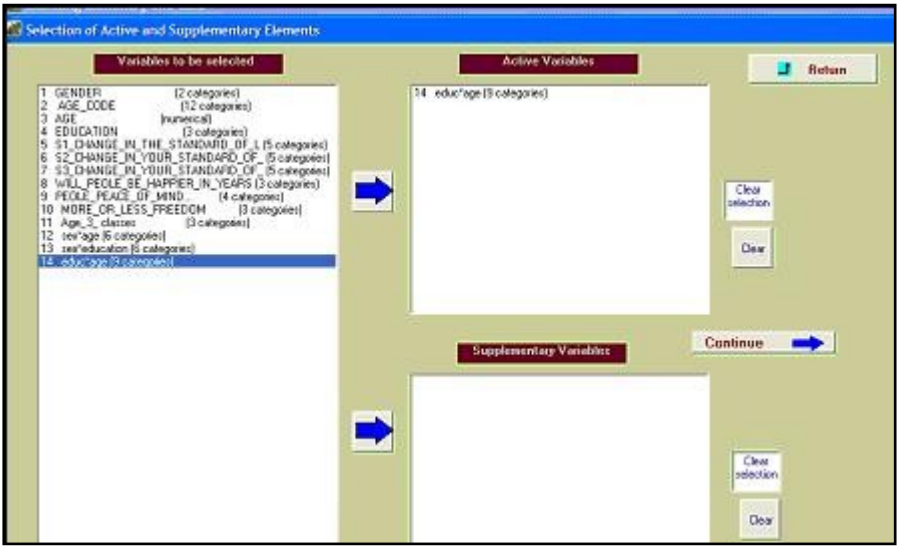


➢ Cliquez sur : **3. Continue ➔**

⊙ une fenêtre : " *Selection of active et supplementary elements* " apparaît.

Etape 5 : Sélection des variables actives et supplémentaires

A l'intérieur de la fenêtre "*Selection of active et supplementary elements*" s'affichent trois autres fenêtres :



- "*Active Variables*" qui reçoit les variables actives sélectionnées
- "*Supplementary Variables*" qui reçoit les variables supplémentaires.

Pour ce type d'analyse, la variable active, unique, est celle dont les modalités vont servir à regrouper les réponses aux questions ouvertes. Nous suggérons de sélectionner la variable nominale numéro 14 "*Educ*age*" comme variable active et nous ignorons les variables supplémentaires. Dans ce cas, les variables supplémentaires pourraient servir à décrire la variable active, pour compléter l'étape "ClusterView".

➢ Cliquez sur : **Continue ➔**

⊙ Une fenêtre : "*Selecting observations*" apparaît.

Etape 6 : Sélection des observations (individus)

Trois cas de figure sont possibles :

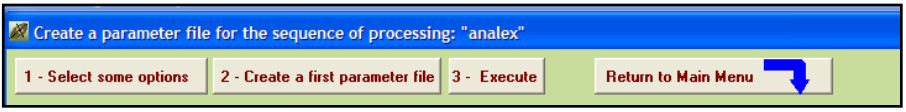
1. Considérer l'ensemble des observations.
2. Sélectionner les observations sur une liste.
3. Sélectionner les observations par un filtre.

Nous considérons ici l'ensemble des observations.

➢ Cliquez sur: **All the observations will be active**

⊙ une fenêtre : "*Create a starting parameter file*" apparaît.

Etape 7 : Création du fichier paramètre

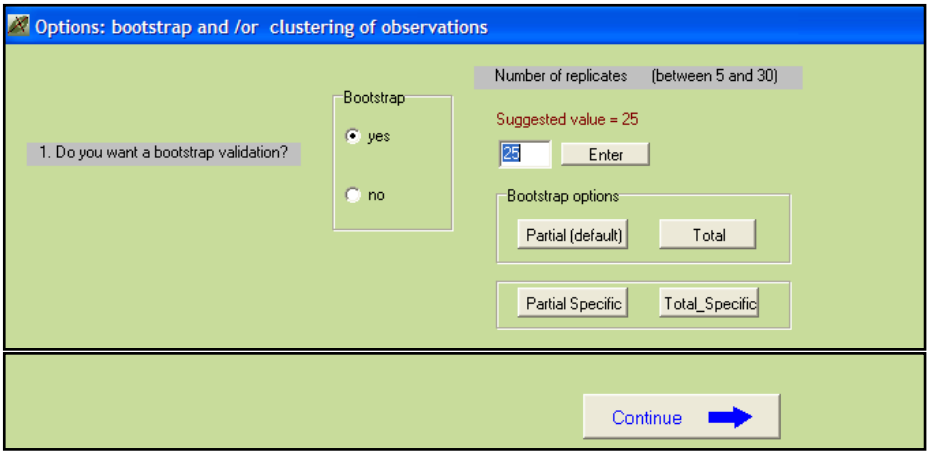


A cette étape, il est possible de sélectionner, comme option, les procédures de bootstrap. Rappelons que dans Dtm-Vic, les analyses factorielles peuvent être complétées par un *bootstrap* qui permet de valider la position des variables sur le plan factoriel

➢ Cliquez sur **1-Select some options**

⊙ une fenêtre: "*Options : Bootstrap and/or Clustering of observations*" apparaît.

➢ Cliquez sur "yes" pour la procédure "bootstrap" ; indiquer le nombre de réplications (par défaut 25) puis : **Enter**. C'est le bootstrap partiel qui est appliqué par défaut. Si le bootstrap n'est pas souhaité, cliquez sur "no" et continuer.



➢ Cliquez sur : **Continue ➔**

⊙ la fenêtre : "*Create a starting parameter file*" réapparaît.

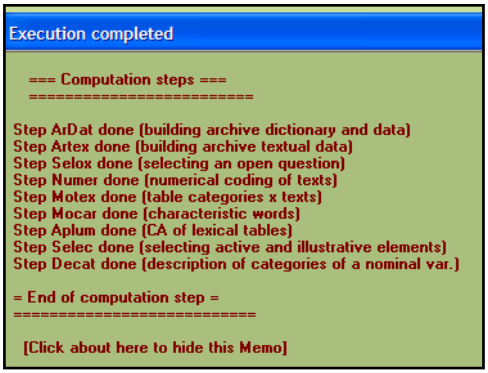
➢ Cliquez sur : **2-Create a first parameter file**.

Un fichier paramètre vient d'être créé sous le nom param_ANALEX.txt et stocké dans le répertoire **EX_A05.Text-Responses**, du répertoire **DtmVic-Examples_A_Start**.

➢ Cliquez sur **3-Execute**

La liste des procédures s'affiche en bloc à la fin de l'exécution: Ardat (Archivage des données), Artex (Archivage des textes), Selox (sélection des questions ouvertes), (Sélection des éléments actifs et supplémentaires), Numer (Numérisation du texte), Motex (table de contingence Mots-textes – les textes étant ici les regroupement de réponses selon la variable active sélectionnée), Mocar (mots et réponses caractéristiques), Aplum (analyse des correspondances pour ce type de tables), Selec (Selection des variables en vue de la description de la variable active), Decat (description

automatique des modalités de la variable active à partir des variables supplémentaires).



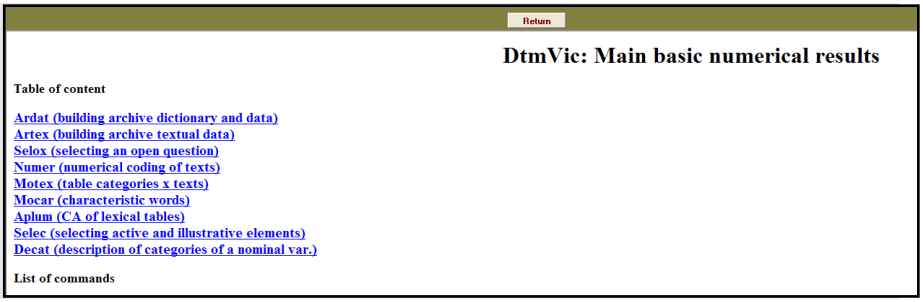
Note : Une fois le fichier paramètre param_ANALEX.txt créé, il est possible, après avoir quitté Dtm-Vic, de l'ouvrir à nouveau dans le menu principal **Command File** avec la procédure **Open an existing command file** puis d'exécuter ce fichier **Execute**. Les utilisateurs expérimentés peuvent modifier les paramètres directement sous l'éditeur proposé par **Open an existing command file** ou avec un autre éditeur de texte hors de Dtm-Vic (voir le bouton "Help about parameters", menu principal et menu de l'éditeur de texte interne).

III.2.3 Fichier de résultats

Les résultats peuvent être consultés dans la rubrique **Result Files** du menu principal.

- Cliquez sur Basic numerical results pour naviguer dans le fichier en format html puis sur Return pour en sortir et revenir au menu principal.

Rappel : Le fichier résultat "imp.txt" (comme son homologue "imp.html") est également sauvé sous le nom "imp" suivi de la date et l'heure de l'analyse. Ce fichier de sauvegarde garde comme archives les résultats numériques principaux tandis que les dossiers "imp.txt" et "imp.html" sont écrasés à chaque nouvelle analyse exécutée dans le même répertoire.


Return

DtmVic: Main basic numerical results

Table of content

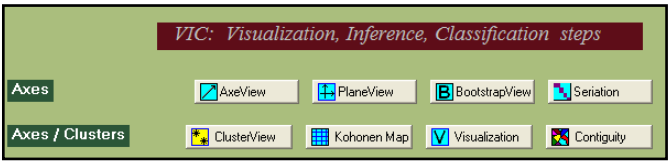
Ardat (building archive dictionary and data)
Artex (building archive textual data)
Selox (selecting an open question)
Numer (numerical coding of texts)
Motex (table categories x texts)
Mocar (characteristic words)
Aplum (CA of lexical tables)
Selec (selecting active and illustrative elements)
Decat (description of categories of a nominal var.)

List of commands

La lecture de ce fichier est nécessaire pour prendre connaissance de certains résultats qui ne peuvent être visualisés. Ainsi la procédure NUMER nous dit que nous avons 1043 individus et 13 919 mots dont 1365 mots distincts. Utilisant un seuil de fréquence de 16 (ce qui signifie que l'on conserve les mots de fréquence supérieure à 16), le nombre de mots conservés se réduit à 10738, tandis que le nombre de mots distincts est ramené à 136. Le livre "*Exploring Textual Data*" (op. cit.) traite les détails de ce prétraitement et tous les résultats qui suivent.

III.2.4 Visualisation des résultats et interprétation

Cette deuxième phase fondamentale de Dtm-Vic fournit les outils de visualisation nécessaires à la validation et l'interprétation des résultats.



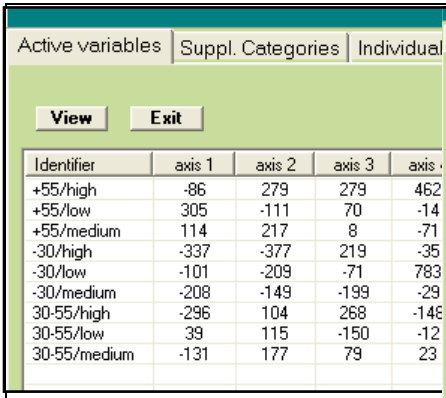
1- Axes factoriels

➢ Cliquez sur AxesView.

Une fenêtre propose de visualiser les coordonnées des variables actives, supplémentaires et des observations sur les premiers axes. Dans le contexte de l'analyse textuelle, seulement deux options sont envisageables: "actives variables" (qui correspondent aux catégories) et les "observations" (qui correspondent aux mots).

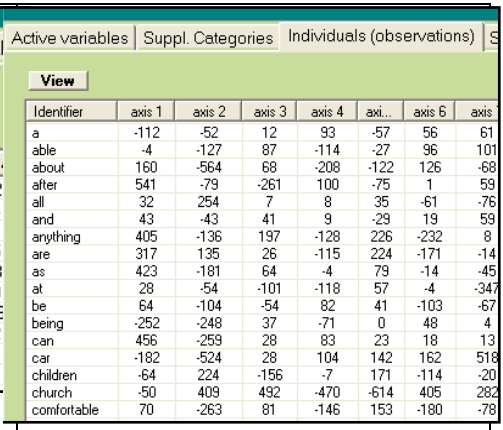
➢ Cliquez sur l'onglet des éléments à examiner, Active variables ou

Individuals (observations) puis sur **View**. Il est possible d'ordonner les coordonnées d'un axe donné, en cliquant sur cet axe. Cliquez sur **Exit** pour sortir de cet outil.



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

Coordonnées des variables nominales actives



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

Coordonnées des mots (observations)

2- Plans factoriels

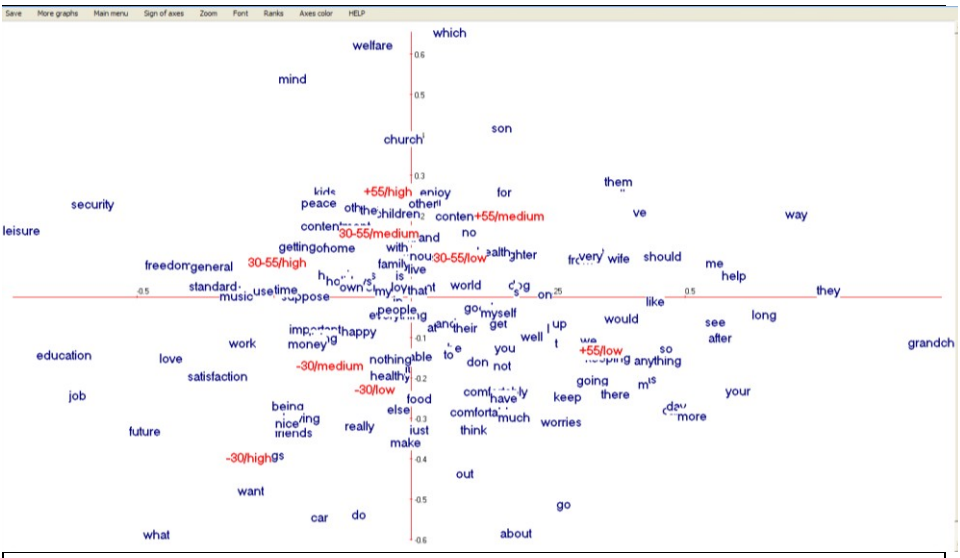
➢ Cliquez sur PlaneView.

⊙ Une fenêtre s'affiche proposant différentes visualisations de plans factoriels.

➢ Choisir la rubrique "Actives columns (variables) + rows (observations)", adaptée à cette analyse. En effet, elle concerne des lignes et des colonnes de la table lexicale.

⊙ Apparaît alors une fenêtre pour sélectionner le plan factoriel suivant la paire d'axes souhaitée. Choisir les axes 1 er 2 puis cliquez sur **display**. Le plan factoriel apparaît.

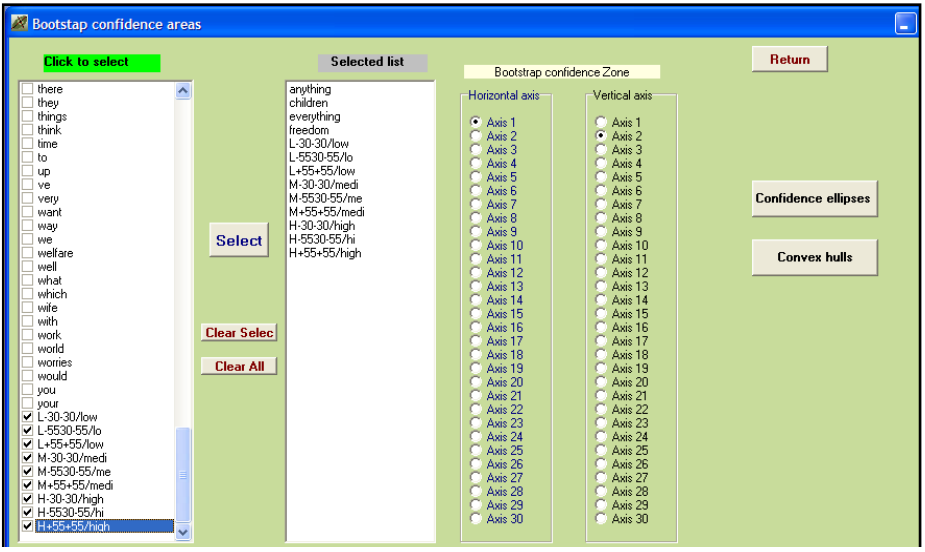
On peut également choisir ce menu par l'intermédiaire de "PLANEVIEW with moveable tags" qui reprend certaines des rubriques précédentes et permet de déplacer les points superposés pour rendre plus lisible le graphique.

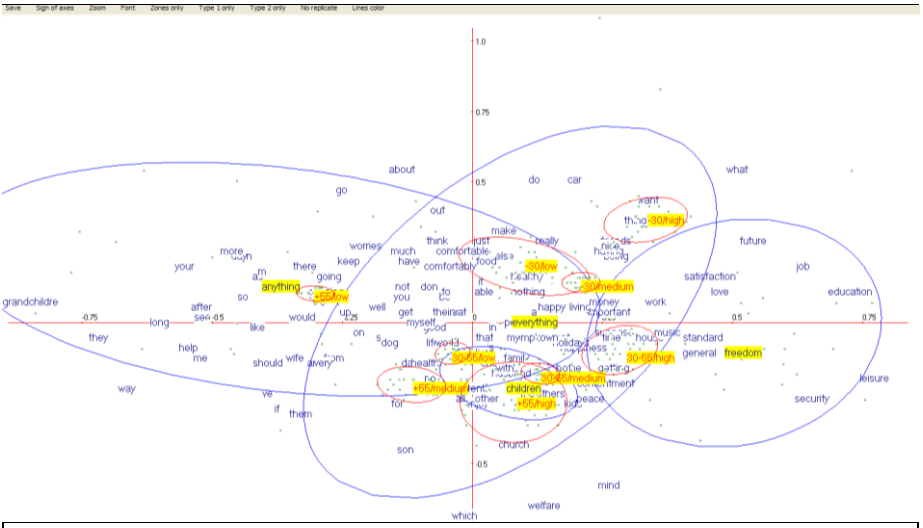


Remarque : Les catégories actives "Age x Education" (colonnes de la table lexicale) sont imprimées en rouge, alors que les mots actifs (lignes) sont imprimés en bleu. Les rôles des différents boutons sont décrits précédemment, notamment dans les exemples A.1 et A.2).

3- Validation Bootstrap

- Cliquez sur : **B** Bootstrap pour valider la position des variables sur le plan factoriel.
  - ⊙ Une fenêtre : *"DtmVic – Bootstrap – Validation – Stability - Inférence*" apparaît.
- Cliquez sur : **Load Data** , puis ouvrir dans le répertoire le fichier des réplications selon le *bootstrap* choisi. Sélectionnez le fichier : ngus_par_boot1.txt pour un *bootstrap* textuel partiel. Répondre **OK** à la boîte : "*Set of principal coordinates loaded*" qui s'affiche.
- Puis cliquez sur : **Confidence Ellipse**.
  - ⊙ une fenêtre : "*Bootstrap confidence areas*" s'affiche
- sélectionnez dans la rubrique "*Click to select*" les variables dont on veut visualiser les ellipses. Les transférer avec **Select**, dans la fenêtre "*selected list*". Choisir ensuite le plan factoriel puis cliquez sur : **Confidence ellipses** ou sur : **convex Hulls** (cf § II.1.4) pour obtenir l'affichage graphique des variables actives.





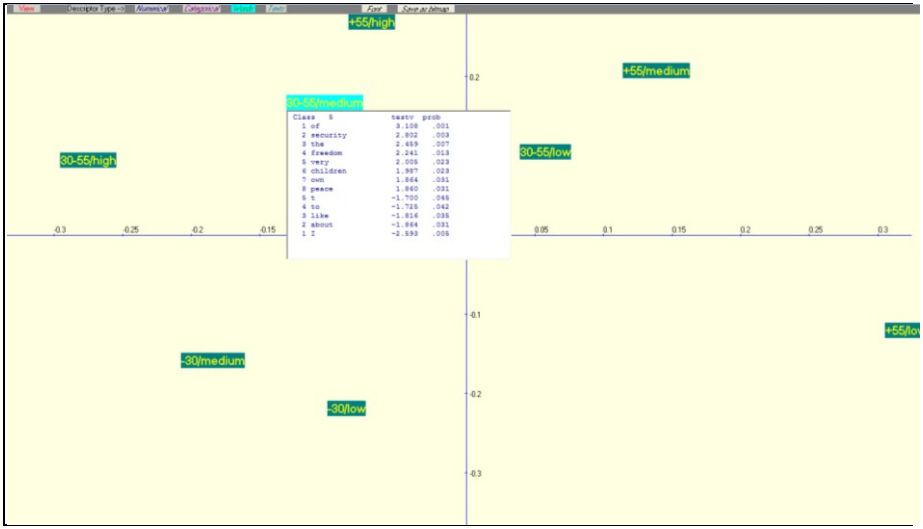
Zones de confiance pour quelques points-mots et points-catégories :

Commentaires : Nous pouvons voir que, individuellement, quelques mots n'ont aucune position significative ("everything" et "anything" par exemple). Dans cet affichage, nous apprenons par exemple que presque tous les groupes d'âge-éducation (points -colonne) ont des "profils lexicaux" distincts, si l'on excepte les catégories "- 30-low" [moins de 30 ans, de bas niveau de l'éducation] et "- 30-medium" [moins de 30 ans, niveau moyen d'éducation] dont les zones de confiance se recouvrent en grande partie.

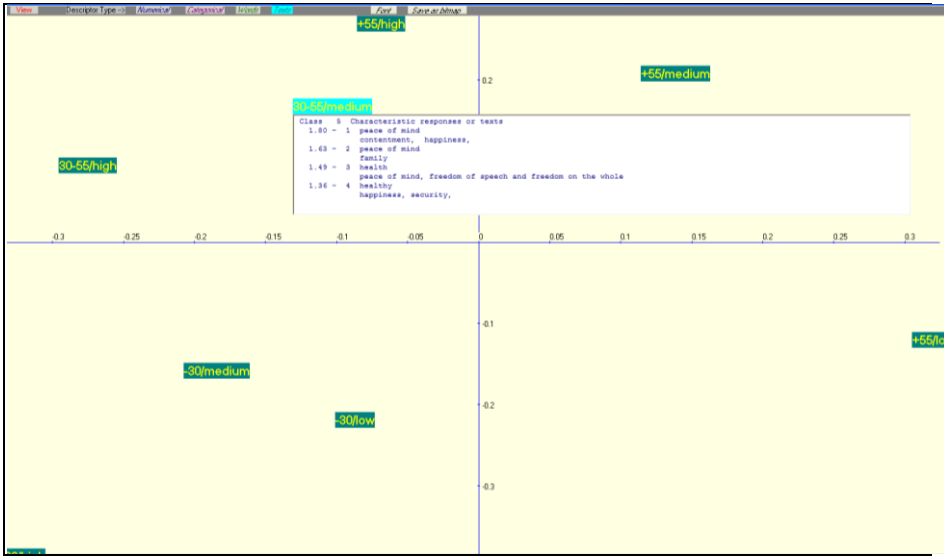
4- ClusterView

Dans le cas d'ANALEX, il ne s'agit pas des résultats d'une classification, mais des catégories de la variable active. Cette option positionne les 9 catégories de la variable "*14_educ*age*" sur le plan factoriel et fournit les mots et textes caractéristiques pour chacune de ces catégories.

- Cliquez sur : ClusterView. Choisissez les axes (1 et 2 pour commencer), et : **Continue**.
  - La fenêtre du plan factoriel s'affiche. Cliquez sur ***View***. La localisation des 9 classes apparaissent sur le plan factoriel.
- Actionnez dans un premier temps le bouton ***Words*** du bandeau. Puis en cliquant (droit) sur une catégorie, les mots descriptifs de la catégorie apparaissent.



- Actionnez ensuite le bouton *Texts* du bandeau. Puis en cliquant (droit) sur une catégorie, les textes descriptifs (réponses caractéristiques ou réponses modales) de la catégorie apparaissent.



5- Carte auto-organisée : Kohonen map

- Cliquez sur Kohonen Map.
  - Une fenêtre *"Selection of elements*" apparaît.
- Cliquez sur "Rows + columns"
  - Une fenêtre *"Kohonen map or SOM map*" apparaît.
- Choisir la carte "map 5x5" puis **continue** et répondre OK à la boite de message : "SOM map completed"
  - Une nouvelle fenêtre *"Kohonen map*" s'affiche
- Actionnez **Draw**. La Carte de Kohonen apparaît.

Les variables actives sont les mots (en noir) et les observations représentent les catégories de la variable (en rouge).

what want think things satisfaction nice having future friends do being about **-30/high**	really nothing else **-30/medium**	work money kids house happy happiness a	time job important	suppose security others music love leisure general freedom education **30-55/high**
out just go comfortable car able	to it healthy comfortably be and	their that my in family everything at **30-55/low**	with the living is home holidays getting enjoy children **30-55/medium**	standard of contentment
not more make m keep have employment **-30/low**	worries up t s myself get don	world son no life health good dog daughter	which live husband for enough content all	welfare peace own other mind
so can	see long after	wife we keeping going are	very them on	people if from **+55/high**
you food church **+55/medium**	well way ve should our i	your there me like grandchildre as anything **+55/low**	they	would much help day

Remarque ; Il est possible de changer de taille de police ("Font") et de dilater la carte de Kohonen obtenue ("Dilat") pour rendre la graphique plus lisible.

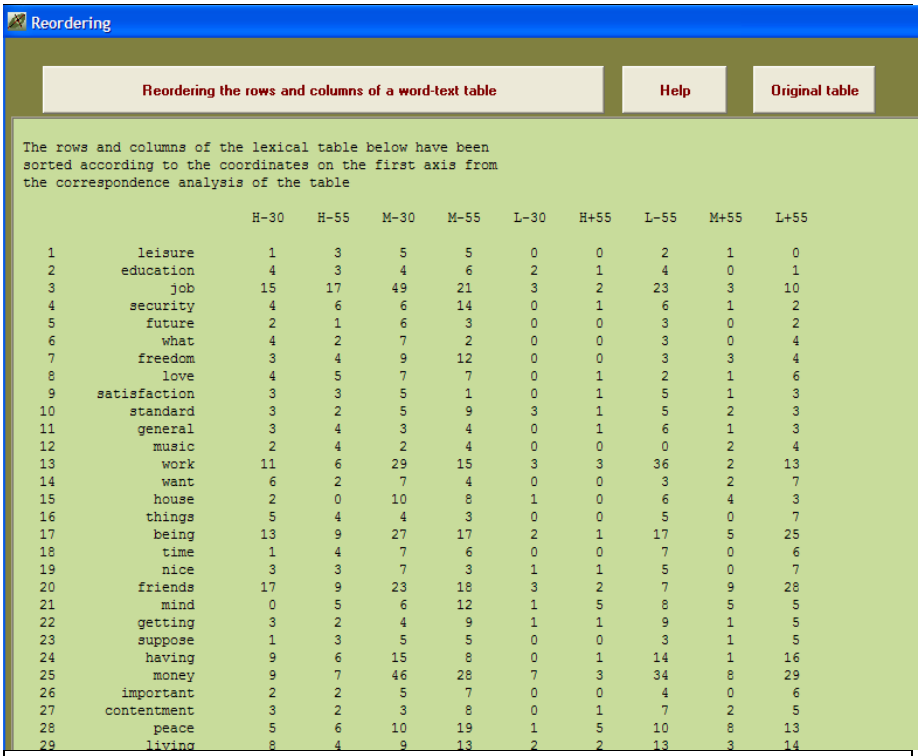
6- Sériation

(Voir l'encadré du paragraphe I.3 du chapitre 1)

La sériation est appliquée ici à la table lexicale croisant les 9 catégories de répondants et les mots choisis (mots apparaissant au moins 16 fois dans le corpus). Dans cette version de Dtm-Vic, la sériation peut être obtenue seulement après les deux types d'analyse : VISUTEX et ANALEX. Ces deux approches impliquent l'analyse de correspondance des tables lexicales.

- Cliquez sur Seriation.
  - La fenêtre "*reordering*" apparaît.
- Cliquez sur **Reordering the rows and the columns of a word-text table**. Et répondre OK à "Seriation of rows and columns of the lexical table completed".

La table lexicale réordonnée croisant les 9 catégories des répondants et les mots choisis est alors constituée.



Reordering

Reordering the rows and columns of a word-text table | Help | Original table

The rows and columns of the lexical table below have been sorted according to the coordinates on the first axis from the correspondence analysis of the table

		H-30	H-55	M-30	M-55	L-30	H+55	L-55	M+55	L+55
1	leisure	1	3	5	5	0	0	2	1	0
2	education	4	3	4	6	2	1	4	0	1
3	job	15	17	49	21	3	2	23	3	10
4	security	4	6	6	14	0	1	6	1	2
5	future	2	1	6	3	0	0	3	0	2
6	what	4	2	7	2	0	0	3	0	4
7	freedom	3	4	9	12	0	0	3	3	4
8	love	4	5	7	7	0	1	2	1	6
9	satisfaction	3	3	5	1	0	1	5	1	3
10	standard	3	2	5	9	3	1	5	2	3
11	general	3	4	3	4	0	1	6	1	3
12	music	2	4	2	4	0	0	0	2	4
13	work	11	6	29	15	3	3	36	2	13
14	want	6	2	7	4	0	0	3	2	7
15	house	2	0	10	8	1	0	6	4	3
16	things	5	4	4	3	0	0	5	0	7
17	being	13	9	27	17	2	1	17	5	25
18	time	1	4	7	6	0	0	7	0	6
19	nice	3	3	7	3	1	1	5	0	7
20	friends	17	9	23	18	3	2	7	9	28
21	mind	0	5	6	12	1	5	8	5	5
22	getting	3	2	4	9	1	1	9	1	5
23	suppose	1	3	5	5	0	0	3	1	5
24	having	9	6	15	8	0	1	14	1	16
25	money	9	7	46	28	7	3	34	8	29
26	important	2	2	5	7	0	0	4	0	6
27	contentment	3	2	3	8	0	1	7	2	5
28	peace	5	6	10	19	1	5	10	8	13
29	living	8	4	9	13	2	2	13	3	14

On peut voir que les premiers mots de la liste réordonnée caractérisent les catégories plutôt jeunes et instruites. Les derniers mots de la même liste réordonnée sont absents ou rarement observés parmi ces catégories. Cependant, ils sont fréquents parmi les dernières catégories (côté droit de la table).

Rappel : Dtm-Vic produit de nombreux fichiers de résultats intermédiaires liés à l'application (tous en format .txt). ***Il est, par conséquent, recommandé de créer un répertoire par application***. Au départ, un tel répertoire doit contenir les fichiers de données, dictionnaire et/ou texte au format Dtm-Vic.

III.3. Analyse directe de réponses libres

Cet exemple reprend l'exemple précédent et procède à une analyse directe des réponses à une question ouverte, sans regroupement préalable.

III.3.1 Les données et fichiers Dtm-Vic :

"Enquête internationale sur les attitudes et valeurs".

Il s'agit encore de l' "Enquête *Life*", volet britannique de l'enquête internationale sur les attitudes et valeurs (voir section précédente III.2.1). Nous nous intéressons ici aux deux premières questions que nous voulons analyser directement, sans regroupement préalable :

- *"Qu'est ce qui est le plus important pour vous dans la vie ?"*
- "*Quelles sont les autres choses très importantes pour vous ?"*

Nous voulons détecter quelles sont les variables nominales les plus liées aux réponses, pour éventuellement les utiliser pour procéder aux regroupements de réponses (procédure ANALEX de la section précédente).

La section III.2 donne toutes les informations nécessaires sur les trois fichiers Dtm-Vic de base qui vont être utilisés :

- Fichier de données pour les questions fermées : TDA_dat.txt
- Fichier dictionnaire des questions fermées : TDA_dic.txt
- Fichier des textes des questions ouvertes : TDA_tex.txt

III.3.2. Mise en œuvre de l'analyse textuelle directe des réponses – "VISURECA"

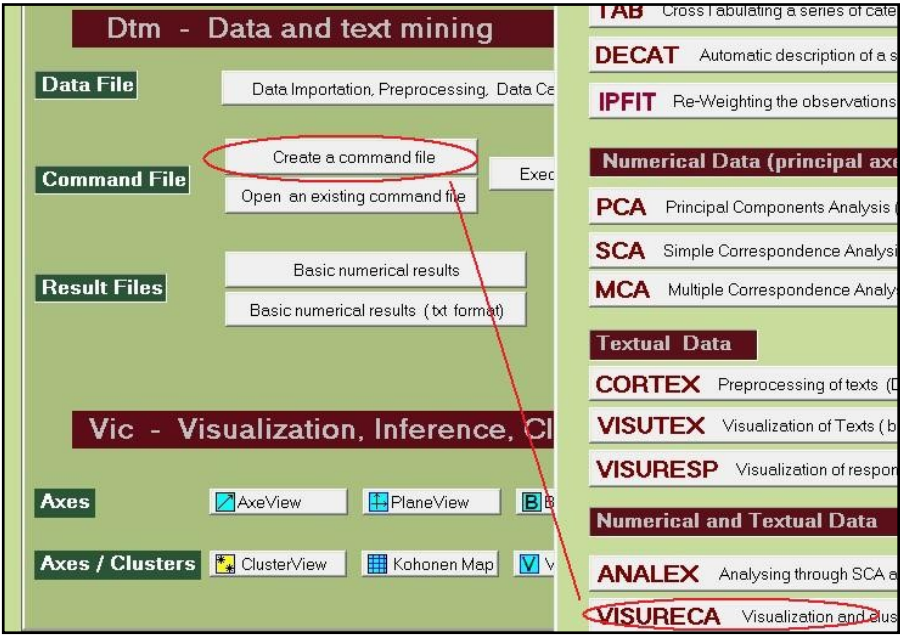
Le fichier paramètre est créé en 5 étapes :

Etape 1 : Sélection de l'analyse

➢ Dans le *menu principal*, cliquez sur : **Create** de **Command File**.

⊙ Une fenêtre: *"Choosing among some basic analysis"* apparaît.

➢ Sélectionnez l'analyse **VISURECA – Visualization and Clustering of responses with categorical data as suplementary elements** dans la rubrique **Numerical and Textual Data**.



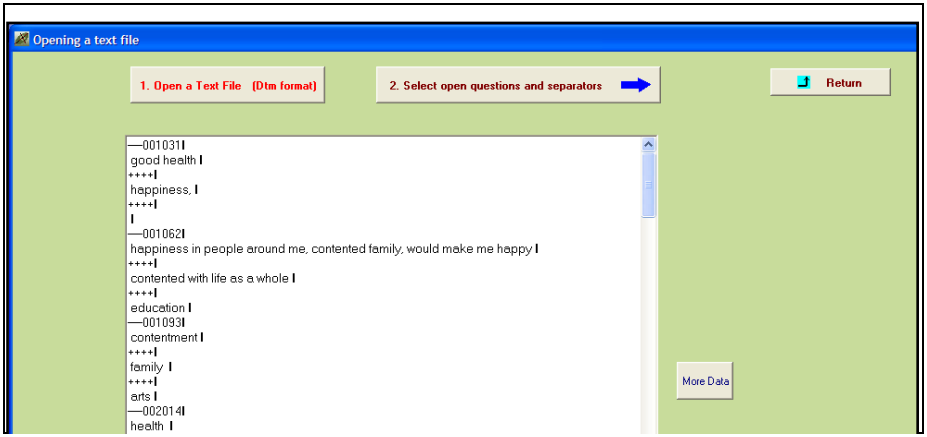
⊙ Une fenêtre : "*Opening a text file*" apparaît.

Etape 2 : Sélection du fichier texte

➢ Cliquez sur le bouton : **Open a text File**. Dans le répertoire **EX_A06.Text-Responses_2**, lui-même inclus dans le dossier **DtmVic_Examples_A_Start** ouvrir le fichier : TDA _tex.txt.

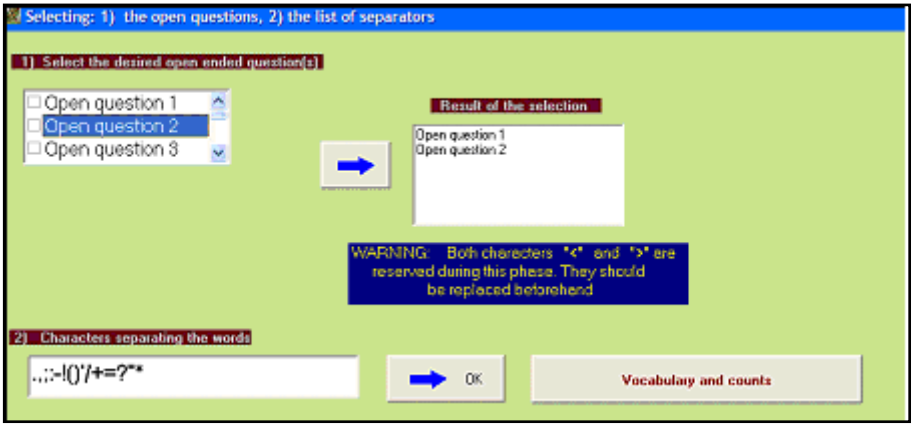
➢ Une boite de message récapitule les informations de ce fichier : 7329 lignes (correspondant à l'ensemble des réponses aux trois questions), 1043 observations (les répondants) et 3 questions ouvertes.

➢ Cliquez sur : **OK**, le fichier s'affiche dans une première fenêtre.



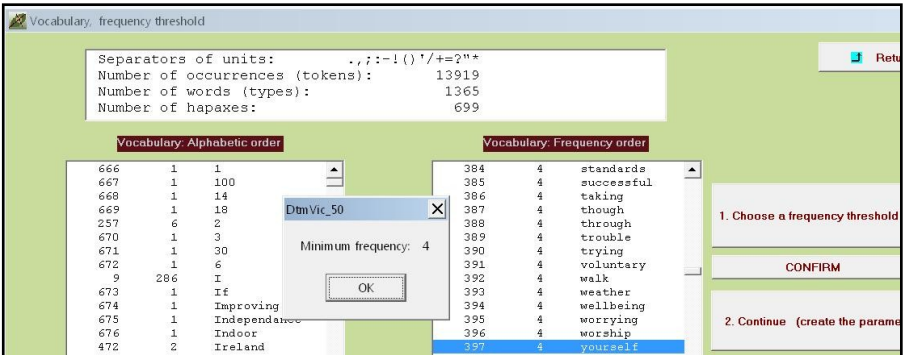
➢ Un deuxième bouton : **2.Select Open questions and separators** apparaît. Cliquez sur ce bouton.

⊙ Une nouvelle fenêtre: "*Selecting : 1) the open questions, 2) the list of separators*" se présente.



Etape 3 : Sélection des questions ouvertes

➢ Sélectionnez les questions ouvertes 1 et 2 et les transférer dans "*Result of the selection*". Puis choisir les séparateurs. Ici, nous adoptons ceux proposés par défaut. Cliquez alors sur **Vocabulary and counts.** La fenêtre suivante présente le vocabulaire (alphabétique et par ordre de fréquence).



Nous devons choisir un seuil de la fréquence en choisissant une ligne dans la rubrique "*Vocabulary (frequency order)*". La ligne 397 correspond à la fréquence 4. (nous avions pris un seuil de 16 précédemment : pour des réponses individuelles, très pauvres lexicalement, il faut plus de mots, pour ne pas générer trop de réponses vides après le choix du seuil). Nous allons donc garder les 397 mots les plus fréquents.

➢ Sélectionnez cette ligne puis : CONFIRM. La fréquence apparaît. Répondre OK.

➢ Cliquez sur **2. Continue (create the parameter file)**.

➢ Une fenêtre d'ouverture des "*fichiers dictionnaires et de données*" apparaît.

Etape 4 : Sélection des fichiers dictionnaire et données

➢ Cliquez sur le bouton : **Open a dictionary**. Dans le répertoire **EX_A06.Text-Responses_2**, ouvrir le fichier TDA_dic.txt. Il s'affiche dans une première fenêtre.

Le statut (nominal ou numérique) des variables est indiqué dans une deuxième fenêtre

➢ Cliquez sur le bouton : **Open a Data File**. Dans le répertoire **EX_A06.Text-Responses_2**, ouvrir le fichier TDA_dat.txt qui s'affiche dans une troisième fenêtre. L'image de l'écran correspondant est la même que pour l'exemple II.2.

➢ Cliquez sur : **3. Continue ➔**

⊙ une fenêtre : "*Selection of active et supplementary elements*" apparaît.

Etape 5 : Sélection des variables actives et supplémentaires

A l'intérieur de la fenêtre "*Selection of active et supplementary elements"* s'affichent trois autres fenêtres :

- "*Variables to be selected*" où figure l'ensemble des variables
- "*Active Variables*" : Il n'y a pas de variable active, puisque c'est le texte des réponses qui est actif ici. Nous avons en fait choisi des variables actives en sélectionnant plus haut les réponses aux questions ouvertes 1 et 2.
- "*Supplementary Variables*" reçoit les variables supplémentaires sélectionnées. Nous pouvons toutes les sélectionner : Elles nous serviront à décrire nos axes et nos classes.

➢ Cliquez sur : **Continue ➔**

⊙ Une fenêtre : "*Selecting observations*" apparaît.

Etape 6 : Sélection des observations (individus)

Nous considérons ici l'ensemble des observations.

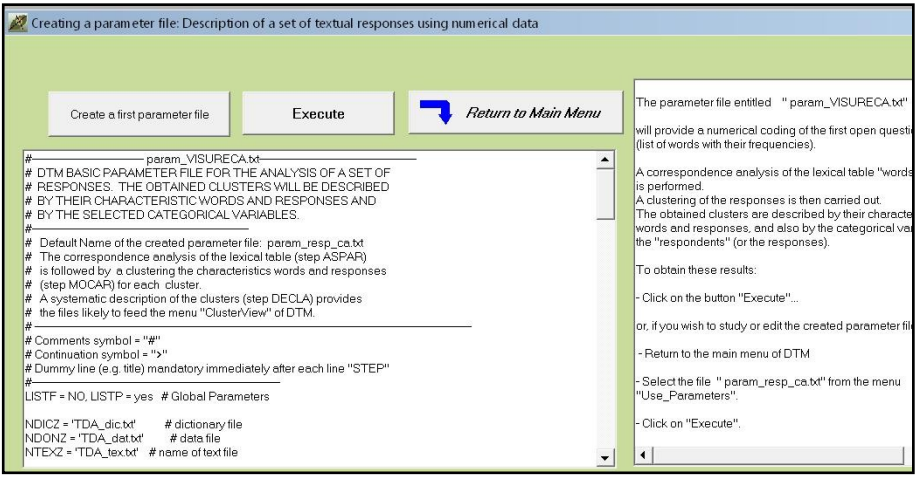
➢ Cliquez sur: **All the observations will be active**

⊙ une fenêtre : "*Create a starting parameter file*" apparaît.

Etape 7 : Création du fichier paramètre

➢ Cliquez sur : **2-Create a first parameter file.**

Un fichier paramètre vient d'être créé sous le nom param_VISURECA.txt et stocké dans le répertoire **EX_A06.Text-Responses_2**, du répertoire **DtmVic-Examples_A_Start**.



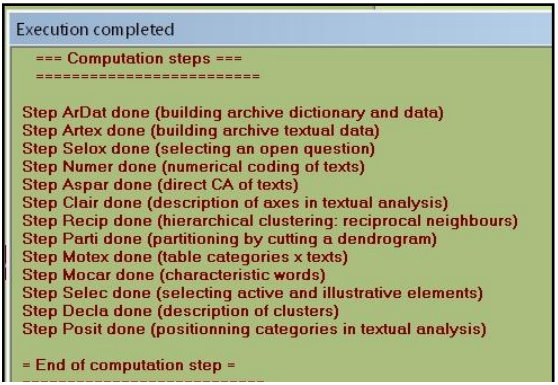
Pour ce type d'analyse, il n'y a pas (encore) de validation *bootstrap*. La classification est automatique, et le nombre de classes est choisi (par défaut) en fonction du nombre de réponses (ici 30 classes). [Ce nombre de classe peut être modifié en éditant le fichier de commande (ou fichier paramètre) avant l'exécution, paramètres des étapes (STEP) "PARTI" et "DECLA"].

➢ Cliquez sur **Execute**

La liste des procédures s'affiche en bloc à la fin de l'exécution.

Commentaires sur les étapes de calcul :

Ardat (Archivage des données), Artex (Archivage des textes), Selox (sélection des questions ouvertes), (Sélection des éléments actifs et supplémentaires), Numer (Numérisation du texte), Aspar (analyse des correspondances directe de la table clairsemée (*sparse*) individus x mots), Recip (classification hiérarchique des réponses par la méthode des voisins réciproques), Parti (coupure de l'arbre et optimisation de la partition obtenue), Motex (table de contingence Mots-textes – les textes étant ici les regroupement de réponses selon les classes de la partition), Mocar (mots et réponses caractéristiques pour chacune des classes), Selec (Selection des variables en vue de la description des classes de la partition des individus), Decla (description automatique des classes à partir des variables supplémentaires nominales et continues), enfin Posit (positionnement des variables nominales supplémentaires dans les plans factoriels construits, rappelons-le, avec les mots des réponses aux questions ouvertes actives).



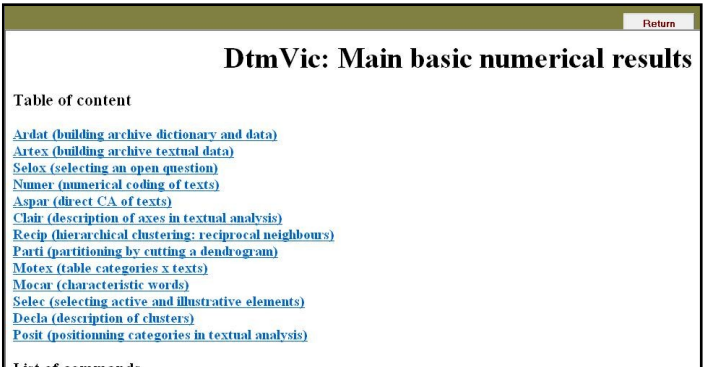
Affichage des étapes de calcul après l'exécution

Note **:** Une fois créé, il est possible, après avoir quitté Dtm-Vic, d'ouvrir à nouveau le fichier paramètre param_VISURECA.txt dans le menu principal **Command File** avec la procédure **Open an existing command file** puis d'exécuter ce fichier **Execute**. Les utilisateurs expérimentés peuvent modifier les paramètres directement sous l'éditeur proposé par **Open an existing command file** ou avec un autre éditeur de texte hors de Dtm-Vic (voir le bouton "Help about parameters", menu principal).

III.3.3 Fichier de résultats

Les résultats peuvent être consultés dans la rubrique **Result Files** du menu principal (MP).

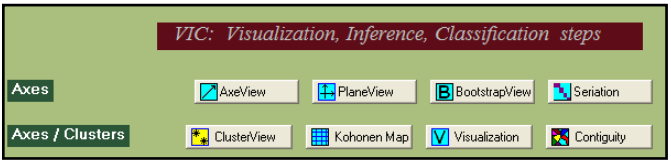
- Cliquez sur Basic numerical results pour naviguer dans le fichier en format html puis sur Return pour en sortir et revenir au MP.



Rappel : Le fichier résultat "imp.txt" (comme son homologue "imp.html") est également sauvé sous le nom "imp" suivi de la date et l'heure de l'analyse. Ce fichier de sauvegarde garde comme archives les résultats numériques principaux tandis que le dossier "imp.txt" (resp. "imp.html") est écrasé à chaque nouvelle analyse exécutée dans le même répertoire.

III.3.4 Visualisation des résultats et interprétation

Cette deuxième phase fondamentale de Dtm-Vic fournit les outils de visualisation nécessaires à la validation et l'interprétation des résultats.



1- Axes factoriels

➢ Cliquez sur AxesView.

L'utilisation de AxesView est parfaitement similaire à celle des analyses précédentes. Les consulter pour naviguer dans cet outil.

2- Plans factoriels

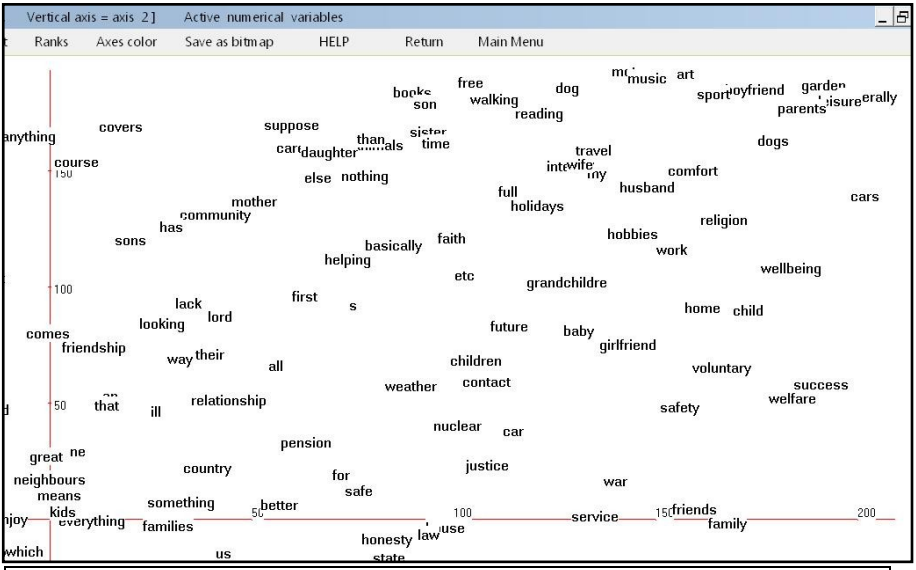
➢ Cliquez sur PlaneView.

⊙ Une fenêtre s'affiche proposant différentes visualisations de plans

factoriels.

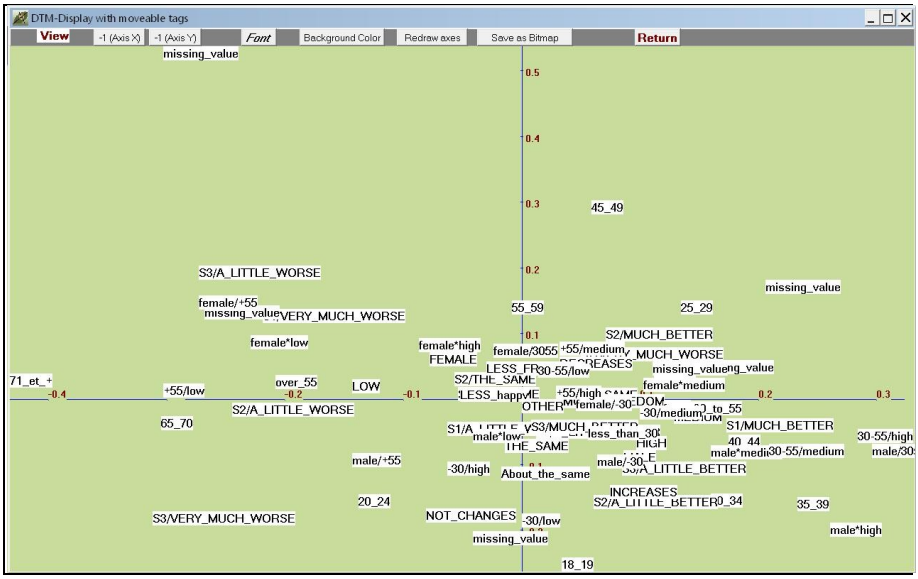
- Choisir la rubrique "Actives columns (variables)", adaptée à cette analyse. En effet, elle concerne les mot utilisés. Les proximités entre mots signifient que ces mots sont utilisés dans les mêmes réponses, donc souvent dans les mêmes phrases. Il y a une composante syntaxique plus prononcée dans les associations que lors de l'analyse précédente qui rapprochait les mots utilisés par les mêmes catégories de répondant, et donc à l'intérieur de textes beaucoup plus importants.

  - Apparaît alors une fenêtre pour sélectionner le plan factoriel suivant la paire d'axes souhaitée. Choisir les axes 1 er 2 puis cliquez sur **display**. Le plan factoriel apparaît.



Ici, compte tenu de la présence de 398 mots, nous avons choisi l'option "RANK" pour déformer les coordonnées (transformation en rangs) sans modifier leur ordre sur les axes. Nous avons également demandé un "Zoom" de façon à détacher un peu plus les mots, mais nous n'avons sur la copie d'écran ci-dessus que le quadrant supérieur droit du plan factoriel. La police (FONT) a également été augmentée.

On peut également choisir l'option : "PLANEVIEW with moveable tags" qui reprend certaines de des rubriques précédentes et permet de déplacer les points superposés pour rendre plus lisible le graphique.



Catégories supplémentaires avec l'option « Etiquettes déplaçables »

Dans le sous-menu proposé par "PLANEVIEW with moveable tags", nous avons sélectionné les catégories supplémentaires, qui constituent le principal intérêt de ce type d'analyse directe des réponses. Le graphique ci-dessus nous montre que l'âge est une des variables très importantes dans la dispersion des réponses ouvertes, ainsi que le niveau d'instruction et le genre (sexe).

C'est à la suite de ce type d'analyse réalisée sans "a priori" que l'on peut choisir les critères de regroupement des réponses les plus pertinents.

Les autres outils (ClusterView, Kohonen) peuvent être utilisés selon les préconisations des sections précédentes.

IV. Importation (création, exportation) des fichiers au format Dtm-Vic

Les fichiers en format interne de Dtm-Vic sont les fichiers dictionnaire, les fichiers de données numériques et les fichiers de textes, présentés au paragraphe I.3. Ils sont nécessaires pour procéder à une analyse de données numériques ou à une analyse de données textuelles. Le cas le plus complet qui met en oeuvre ces trois types de fichiers est celui d'une enquête comportant des réponses à la fois à des questions fermées (fichiers dictionnaire et données) et à des questions ouvertes (fichier texte).

Les fichiers internes sont des fichiers en format ".txt" et s'obtiennent soit de façon manuelle à partir d'un mode de saisie d'importation intégré à Dtm-Vic soit, le plus souvent, à partir de fichiers préexistants en format ".doc" pour certaines données textuelles ou en format ".csv" issu d'Excel pour les données numériques et textuelles, ou encore simplement en format texte (codes ASCII).

La procédure d'importation ne s'opère qu'une fois, au début du processus de l'analyse.

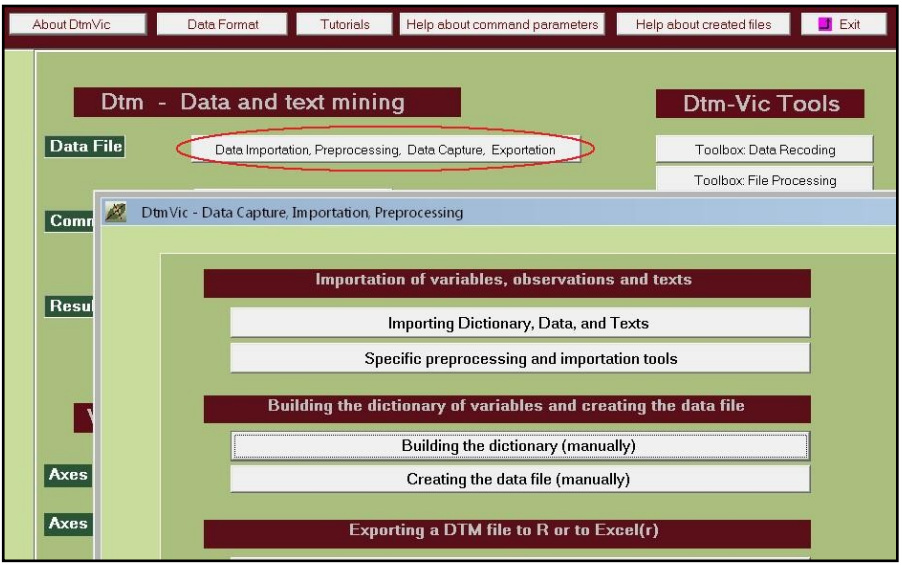
Nous approfondirons ici l'importation standard, en format "Excel", de données numériques et textuelles, telles que les données d'enquêtes composées de questions fermées et ouvertes, puis, dans une seconde partie, nous présenterons la procédure de saisie directe des données.

D'autres procédures sont présentées dans le Tutoriel (en Anglais) intégré à Dtm-Vic. Les textes simples (format interne type 1 décrit en section I.5, et illustré par l'exemple III.1 du chapitre III) ne donnent pas lieu à une procédure d'importation particulière : il suffit d'insérer les séparateurs entre des textes aux formats usuels.

- Cliquez sur le bouton Data Importation, Preprocessing, Data Capture, Exportation

Une fenêtre s'affiche et offre différentes possibilités pour constituer un jeu de données numériques ou textuelles en format Dtm :

- **Importation of variables, observations and texts** : importer des données numériques ou textuelles en format Excel, libre ou fixe; des données textuelles en format libre; ou encore des fichiers XML contenant des données numériques ou textuelles.
- **Building the dictionnary of variables and creating the data file** : créer les fichiers dictionnaires et les fichiers de données numériques ou textuelles manuellement à partir d'un mode de saisie d'importation intégré à Dtm-Vic. Les deux autres procédures, **Exporting a DTM file to R or to Excel(r)** et **Dtm_tools**, concernent l'exportation, le recodage, et l'archivage des données.



IV.1. Importation de fichiers Excel ®

IV.1.1. Présentation du fichier Excel

Nous considérons le tableau de données de l' "enquête "*Life*" présentée dans les deux derniers exemples du chapitre III précédent. Le fichier correspondant dispose en ligne de 1043 individus et en colonnes de 17 variables : 9 variables nominales (le genre, l'âge recodé, le niveau d'éducation et 6 variables d'opinion), 1 variable continue (l'âge), 3 variables textuelles correspondant aux 3 questions ouvertes, enfin 4 autres variables nominales qui correspondent à des variables signalétiques recodées (l'âge en 3 classes, les croisements du genre avec l'âge en 3 classes, le niveau d'éducation, le croisement de l'âge en 3 classes avec le niveau d'éducation).

ident	gender	age_code	age	education	important_life	important_probe	change_last_years	change_your_last_yrs	change_your_next_yrs	peole_be_happier?	peole_peace_of_mind.	more_or_less_freedom	culture	...
1	1	80	12	1	good health	happiness,	2	3	3	3	2	1		...
2	1	54	8	1	happiness in peop	contented with life as	1	1	3	1	1	1	education	
3	1	40	6	1	contentment	family	1	2	1	2	2	2	arts	
4	2	27	3	2	health	happiness, money, fa	1	2	1	1	1	1	the way british people	
5	2	39	5	2	to be happy	healthy, have enough	2	1	3	1	1	1		
6	1	80	12	1	my wife	music, holidays, I like	2	3	4	2	2	3	not much it's very imp	
7	2	46	7	2	health	happiness	4	3	0	0	2	1		
8	2	33	4	1	to be healthy	just to live long enou	3	4	1	2	3	1		
9	2	64	10	1	health,	keeping going, family	4	3	3	2	1	2	culture is good,	
10	2	65	11	1	husband	new baby grand dau	2	1	0	2	2	1	goodwill,	
11	1	58	9	3	companionship	job, good life, money	1	2	5	2	2	3	It's important, has exi	
12	2	74	12	1	good health	happiness, togethern	2	3	0	2	3	3	heritage, concerts, dr	
13	2	29	3	2	family	friends, pets,	2	2	2	3	2	1	theatre, national trust	
14	1	82	12	3	togetherness	peace of mind, good	3	3	0	2	2	2	music, poetry, ballet,	
15	2	68	11	1	my family really	health, walking	2	2	4	3	3	3	the beauty of our cou	
16	2	37	5	2	my children	my husband, my fam	1	2	1	3	0	1	can't think of anything	
17	1	34	4	2	my own time, not	my friends, plants, fo	2	4	3	0	2	2	the music of henry pu	
18	1	30	4	2	freedom of choice	sport, work, parents	2	1	2	1	2	1	literature, the theatre,	
19	1	27	3	3	I suppose work	family, friends, gener	2	1	2	3	1	0	sausages, beefeaters	
20	1	85	12	1	health	family	0	3	3	2	1	2		
....														

La première ligne et la première colonne contiennent les identifiants respectivement des individus et des variables. Toutes les valeurs alphanumériques, celles par exemple des identifiants ou encore des catégories des variables nominales, doivent être composées de moins de

20 caractères et de préférence de moins de 10 et ne doivent pas contenir d'espace vide. Les réponses aux questions ouvertes sont des textes de moins de 8000 caractères. Par contre les données manquantes sont exprimées par des espaces vides. Pour un tableau de données à n individus et p variables, quelque soit leur nature, le tableau "Excel" dispose donc de n+1 lignes et de p+1 colonnes.

Le fichier est sauvegardé en format ".csv" dont les séparateurs sont des points-virgules (version française d'Excel).

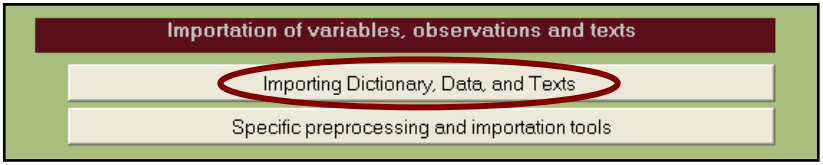
Ce fichier qui va nous servir d'exemple a pour nom : datbase_global.csv il se trouve dans le répertoire (dossier) :

DtmVic_Examples_D_Import\EX_D01.Importation.Num_Text.

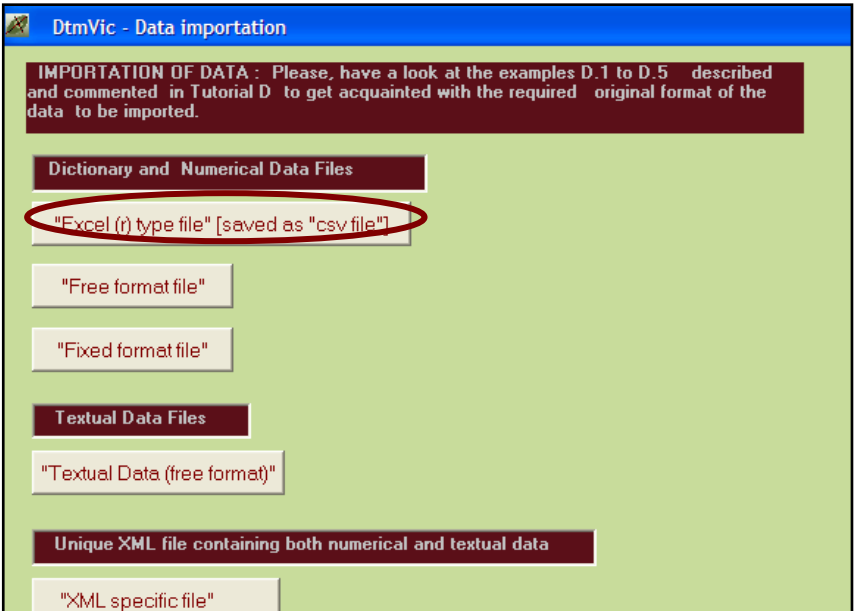
lui-même dans le dossier **DtmVic-Examples** téléchargeable avec Dtm-Vic

Dans certaines versions d'Excel, notamment les versions anglophones, le séparateur, pour le format ".csv", n'est pas le point virgule, mais la virgule. La procédure d'importation de DtmVic prévoit une possibilité de changement des séparateurs. De fait, tout comme les espaces vides, les points-virgules et les apostrophes dans l'expression des valeurs alphanumériques ne sont pas autorisés et doivent être remplacés par un autre symbole. De même les valeurs numériques, notamment les nombres à plus de 3 chiffres ne doivent pas contenir de blancs (écriture des francophones laissant un demi-espace pour séparer les milliers). Enfin, dans la version française et dans quelques versions européennes d'Excel, "les virgules décimales" doivent être remplacées par les points décimaux habituels dans les notations anglo-saxonnes et dans les langages de programmation.

IV.1.2. Procédure d'importation

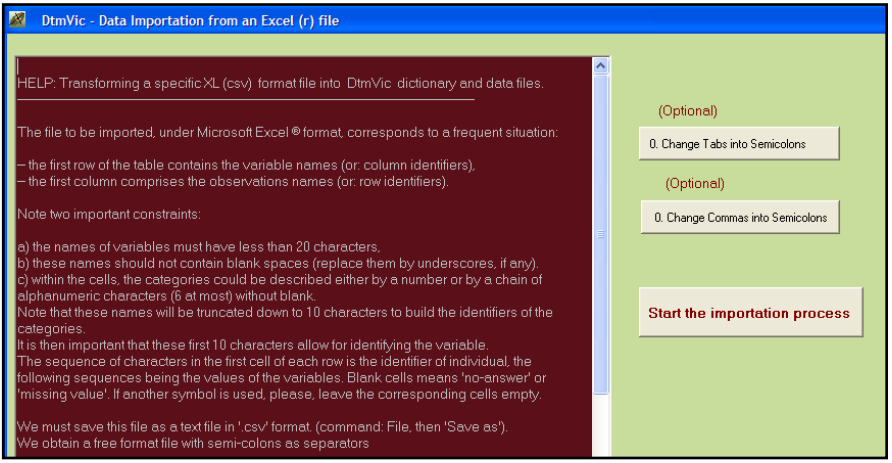


- Sélectionnez, dans le menu principal, Data Importation, Preprocessing, Data Capture, Exportation puis Importing Dictionnary, Data and Texts dans Importation of variables, observations and texts. Une fenêtre apparaît.



➢ Cliquez ensuite sur Excel (r) type file [saved as "csv file"].

⊙ Une fenêtre "*Data Importation from an Excel ® file*" apparaît proposant plusieurs options.



Si le fichier Excel a été sauvegardé en utilisant des "tabulations" ou des "virgules" comme séparateurs, cliquez sur un des boutons optionnels :

- Change Tabs into Semicolons change les tabulations en points-virgules [après avoir vérifié que le fichier original ne contenait pas de points-virgules, et remplacé ceux-ci le cas échéant].
- Change Commas into Semicolons change les virgules en points-virgules. [après avoir vérifié que le fichier original ne contenait pas de virgules, et remplacé celles-ci le cas échéant].

Dans ce cas, sélectionnez le fichier Excel sauvegardé avec des tabulations ou des virgules, et convertissez-le. Un nouveau nom est donné au fichier créé. Le procédé d'importation continuera d'employer ce nouveau fichier.

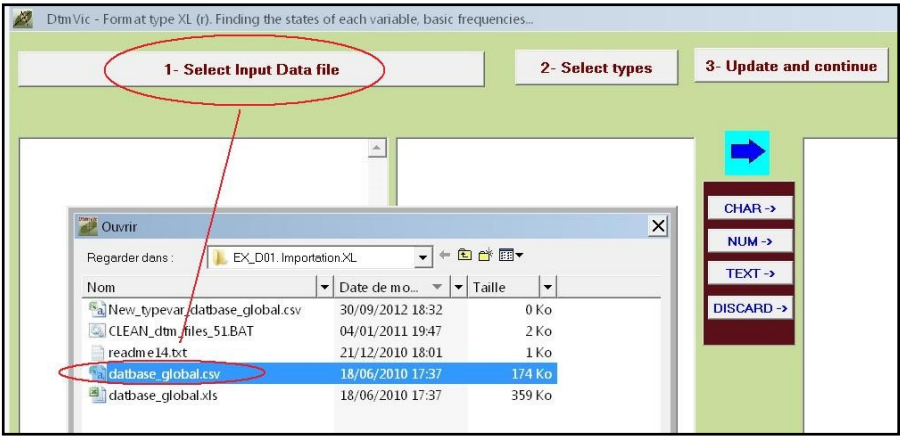
Dans tous les cas :

➢ Cliquez sur le bouton **Start the importation process**.

⊙ Une nouvelle fenêtre "*Format type XL®, Finding the states of each categorical variable, basic frequencies…*" apparaît.

➢ Cliquez sur **1.Select Input Data file** et ouvrez le fichier XL en format ".csv". Pour l'exemple, on choisit le fichier datbase_global.csv dans le répertoire :
DtmVic_Examples_D_Import\EX_D01.Importation.Num_Text.

➢ Répondre **OK** à la boîte de message.



Le descriptif des variables s'affiche dans la fenêtre de gauche. Dans la fenêtre centrale, nous pouvons lire entre crochets le nombre de valeurs

distinctes observées dans le fichier et entre parenthèses une lettre A ou N.

La lettre (A) signifie que l'on a observé des valeurs non numériques; la lettre (N) indique que ce sont uniquement des valeurs numériques. Il est alors plus facile de choisir le statut des variables correspondant à la deuxième étape de cette procédure. Pour cela :

- **2. Select types** : Sélectionnez une ou plusieurs variables dans la liste de la fenêtre centrale puis spécifiez leur statut en cliquant sur :

  CHAR -> pour une variable nominale (ou catégorielle, ici les variables signalétiques (1,2,4) et d'opinion (7 à 12)

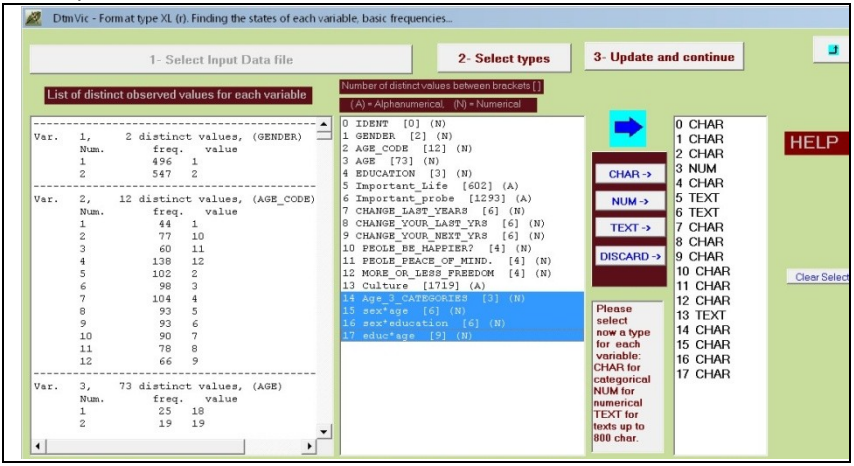
  NUM -> pour variable numérique (ou continue, ici la variable 3-Age)

  TEXT -> pour les variables textuelles, les réponses aux questions ouvertes (variables 5,6,13).

  DISCARD -> pour abandonner des variables.

- Une fois l'attribution du statut accompli, cliquez sur le bouton **3.Updating and continue** puis répondez **OK** sur le "number of observations".

[Cette procédure crée un nouveau fichier d'importation, nommé automatiquement New_typevar_datbase_global.csv, dont la deuxième ligne contient les types des variables. Mais l'utilisateur n'a pas à se préoccuper de ce fichier.]



Précisions sur la nature de l'importation :

Le procédé d'importation consiste en la construction d'un dictionnaire et d'un fichier de données de DtmVic à partir du fichier original de données. Les noms des variables seront extraits à partir des identificateurs des variables dans le fichier de départ. Le nombre de catégories pour chaque variable nominale et les noms de ces catégories seront établis à partir de ce fichier.

Pour chaque variable, toutes les différentes séquences des caractères observées dans le fichier de données sont détectées et comptées. Les catégories des variables nominales sont rangées selon l'ordre alphabétique de leurs identifiants.

Les lignes du fichier de données de DtmVic commenceront par l' identifiant figurant dans la première colonne « identifiant » du fichier Excel.

Les modalités des variables nominales seront des nombres entiers consécutifs commençant par la valeur "1", au lieu d'un symbole alphanumérique (l'ordre des modalités sera l'ordre alphabétique de leurs symboles dans le fichier d'origine). Les valeurs manquantes (cases vides dans le fichier de départ) donnent lieu à une modalité particulière, identifiée dans le dictionnaire Dtm-Vic par la lettre « b » (comme « blanc »).

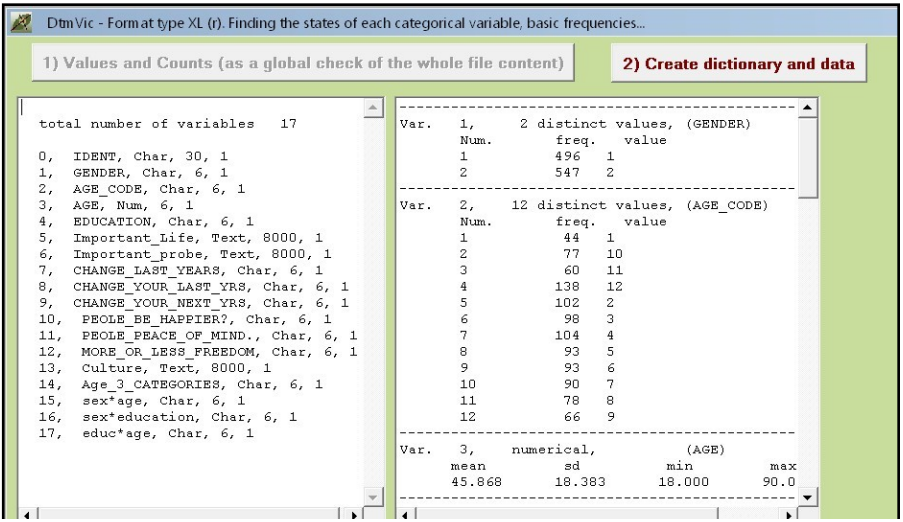
Les valeurs des variables numériques seront identiques à celles du fichier de données original, les valeurs manquantes (cases vides dans le fichier de départ) sont remplacées, dans cette version de DtmVic, par la valeur conventionnelle "999".

Les variables textuelles (réponses aux questions ouvertes) donnent lieu à un fichier textuel séparé (format textuel de type 2, cf. chapitre I, section I.5).

- ⊙ Une seconde fenêtre "*Format type XL . Finding the states of each categorical variable, basic frequencies…*" apparaît.

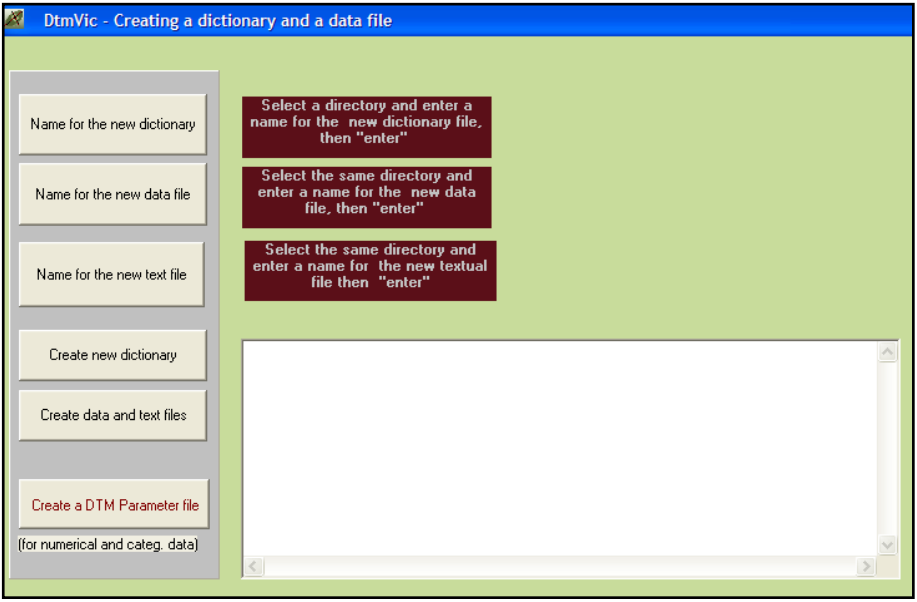
➢ Cliquez sur **Values and counts**.

Le nom des variables s'affiche dans la fenêtre de gauche. La fenêtre de droite présente les statistiques élémentaires de ces variables. Il s'agit seulement de permettre à l'utilisateur de vérifier que les statuts qu'il a choisis pour les variables sont corrects.



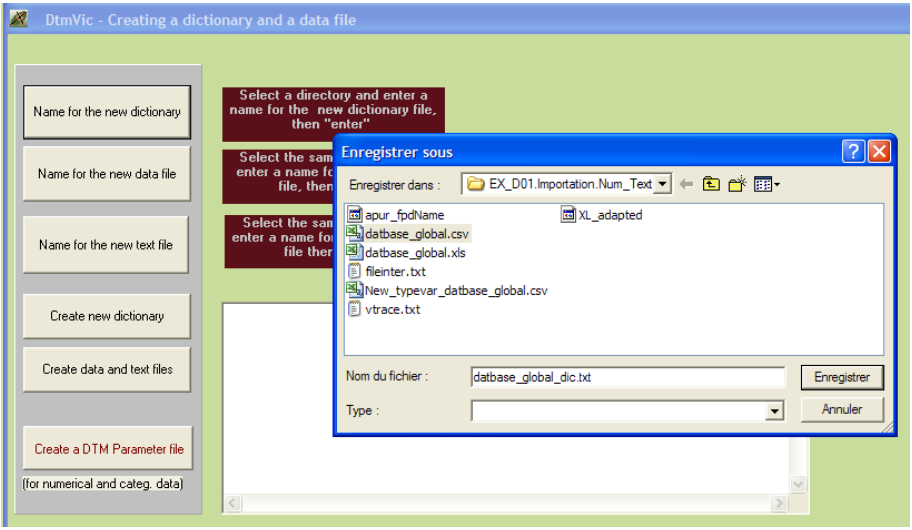
➢ Cliquez sur **Create dictionary and data**.

⊙ Une fenêtre "*creating a dictionary and a data file*" apparaît sur l'écran.

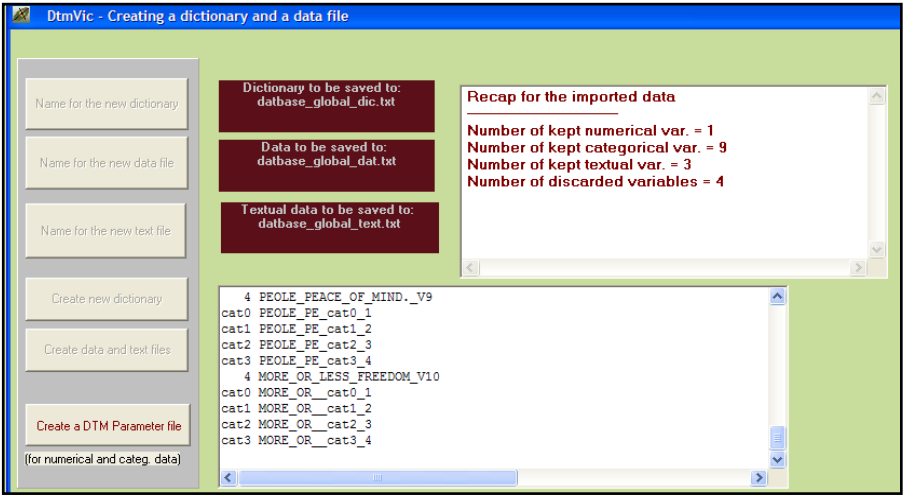


➢ Cliquez sur **Name for the new dictionary**. Entrez le nom du fichier

dictionnaire Datbase_global_dic.txt (par exemple) et enregistrez.



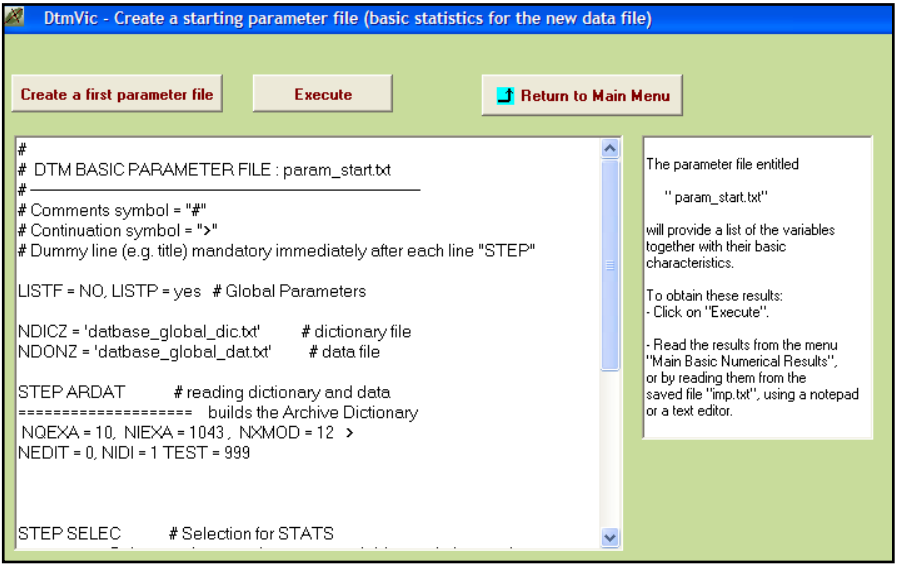
- Cliquez ensuite sur **Name for the new data file**. Entrez le nom du fichier de données Datbase_global_dat.txt et enregistrez.
- Cliquez sur **Name for the new text file**. Entrez le nom du fichier dictionnaire Datbase_global_text.txt (par exemple) et enregistrez. S'il n'y a pas de données textuelles, passez à l'étape suivante.



- Cliquez sur **Create new dictionary**. Le fichier dictionnaire de DtmVic est créé automatiquement et s'affiche dans la fenêtre. Répondre **OK** à "New Dictionary completed". De la même façon en cliquant sur **Create new data file**, le fichier de données de DtmVic est crée. Une boîte de message donne le nombre d'individus. Répondre **OK**. En cas de présence de questions ouvertes, cliquez sur Create new text file.

Un récapitulatif des données importées apparaît dans une nouvelle fenêtre.

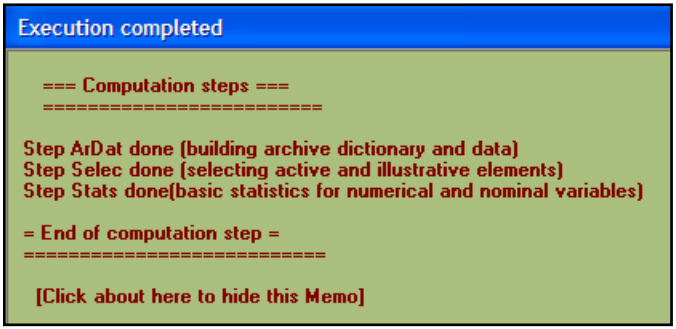
- Cliquez enfin sur le bouton **Create a DTM Parameter file**.
  - ⊙ Une fenêtre "*create a first parameter file*" apparaît sur l'écran.
- Cliquez alors sur **Create a first parameter file**. Un fichier de commande de DtmVic est affiché dans la fenêtre inférieure (dans DtmVic, les expressions "fichier de paramètre" et "fichier de commande" sont équivalentes). Les opérations et les commentaires restent identiques à ceux de l'introduction.



- Cliquez enfin sur **Execute** pour créer le fichier paramètre. Le fichier paramètre est automatiquement sauvegardé sous le nom de param_start.txt dans le dossier de travail.

Le fichier paramètre n'inclut aucune commande d'analyse statistique élaborée. Il se limite au calcul des statistiques de base des variables. Il sert simplement de contrôle à l'importation des *données numériques*.

⊙ La fenêtre d'exécution, identique à toutes procédures d'analyse, apparaît dans la fenêtre du menu principal.



Les procédures s'affichent en bloc à la fin de l'exécution : l'étape Ardat archive les données et le dictionnaire. L'étape Selec choisit les variables pour le traitement suivant ; dans ce cas-ci, toutes les variables disponibles sont choisies. L'étape Stats calcule les statistiques générales.

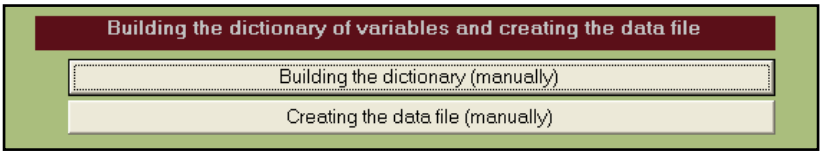
Les résultats peuvent être consultés dans l'étape **Result Files**

➢ Cliquez sur Basic numerical results pour ouvrir le fichier en format html puis sur Return pour en sortir et revenir au menu principal.

➢ ou cliquez sur Basic numerical results (text format) pour ouvrir le fichier résultat en format texte. L'importation est terminée.

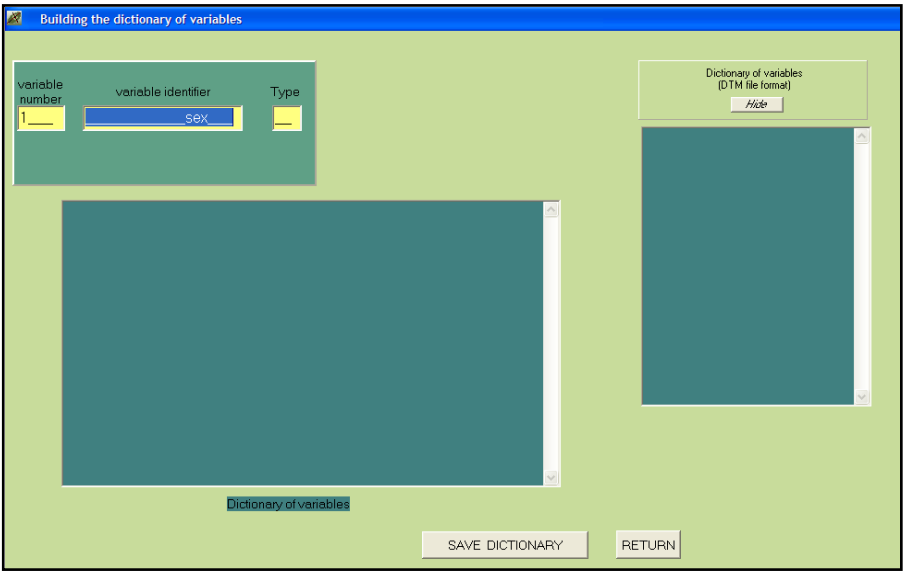
IV.2. Saisie manuelle

DtmVic propose un module de collecte de **données numériques**. Il est surtout utilisable dans un contexte pédagogique, pour saisir de petits jeux de données numériques. Ce module ne permet cependant pas de saisir des questions ouvertes. Le passage par un fichier "Excel" est souhaitable.



IV.2.1. Le fichier dictionnaire

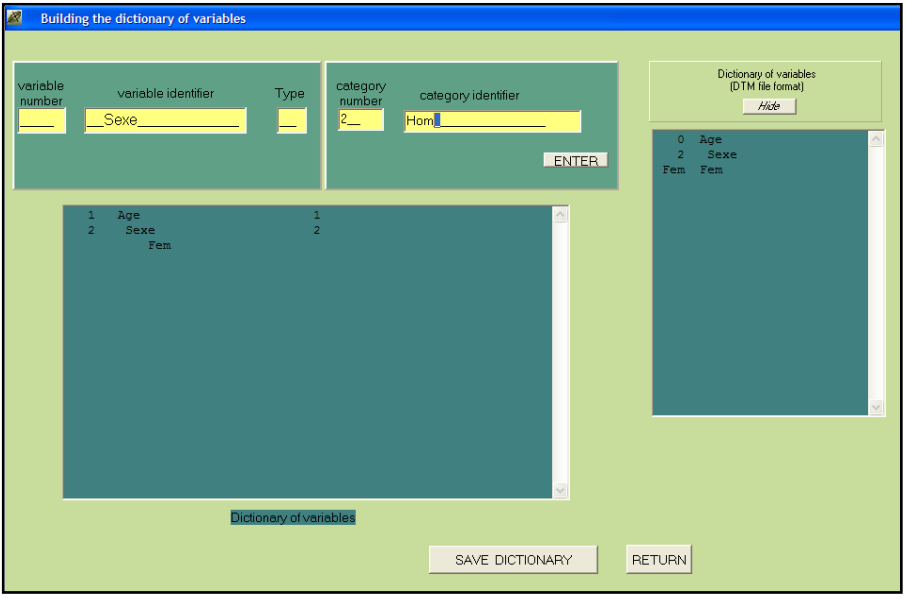
- Sélectionnez, dans le menu principal, Data Importation, Preprocessing, Data Capture, Exportation puis Building the dictionary dans Building the dictionary of variables and creating the data file.
  - Une fenêtre dédiée à la construction du dictionnaire apparaît.



La première sous-fenêtre, en haut à gauche, permet de saisir le numéro, le nom et le type de chacune des variables.

- La 1ère fenêtre jaune affiche : "1", le numéro de la 1ère variable à saisir. Dans la deuxième fenêtre, tapez le nom de la variable puis dans la 3ème fenêtre donnez le "Type" de la variable c'est-à-dire le nombre de modalités si la variable est nominale ou tapez "0" si la variable est continue. Un bouton ENTER s'affiche à l'issu de la saisie du type de la variable. Si celle-ci est continue, continuez la saisie. Si elle est nominale, une fenêtre apparaît pour saisir les numéros et les modalités de la variable nominale. Une fois les modalités enregistrées, cliquez sur ENTER (ou appuyez sur la touche "entrée"). Continuez de saisir l'ensemble des variables.

Le résultat de la capture du dictionnaire des variables apparaît dans la fenêtre inférieure ainsi que dans celle de droite, dans laquelle elle apparaît dans le format interne de DtmVic.



Par exemple, une première variable "Age" a été saisie. Etant une variable continue le type est "0". Une seconde variable " Sexe" est saisie. Ayant deux modalités, le type "2" est saisi. Il fait alors apparaître une fenêtre contigüe dans laquelle sont saisis les libellés des deux modalités.

Cliquez sur ENTER (ou pressez la touche "Entrée") après chaque saisie.

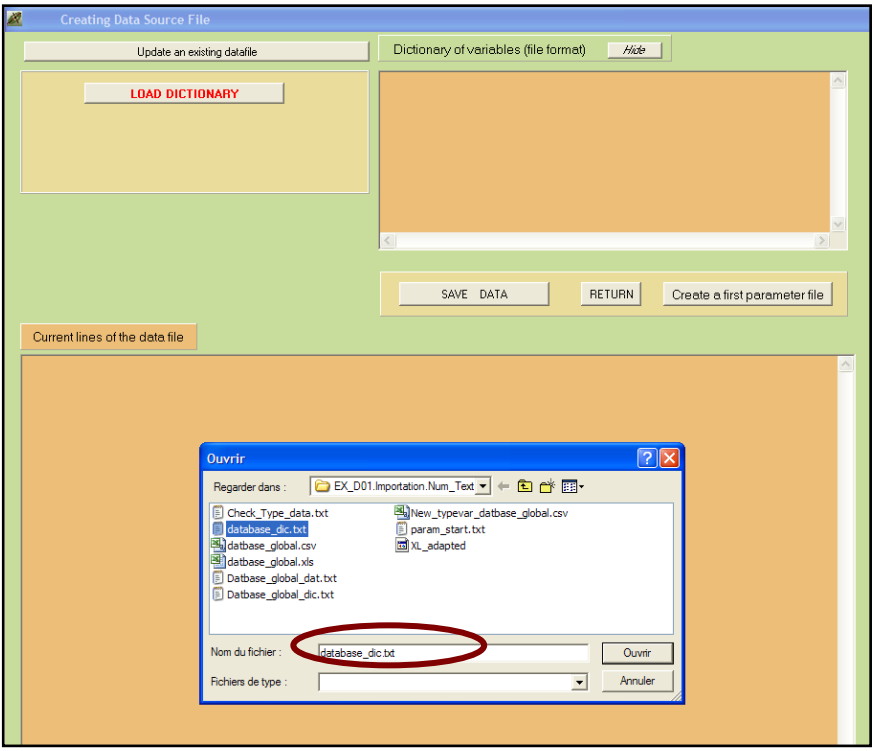
- Une fois l'ensemble des variables capturées, cliquez sur SAVE DICTIONARY et enregistrez un nom pour le fichier du dictionnaire.

On peut le nommer : Database_dic.txt. Cliquez ensuite sur RETURN.

IV.2.2. Le fichier des données

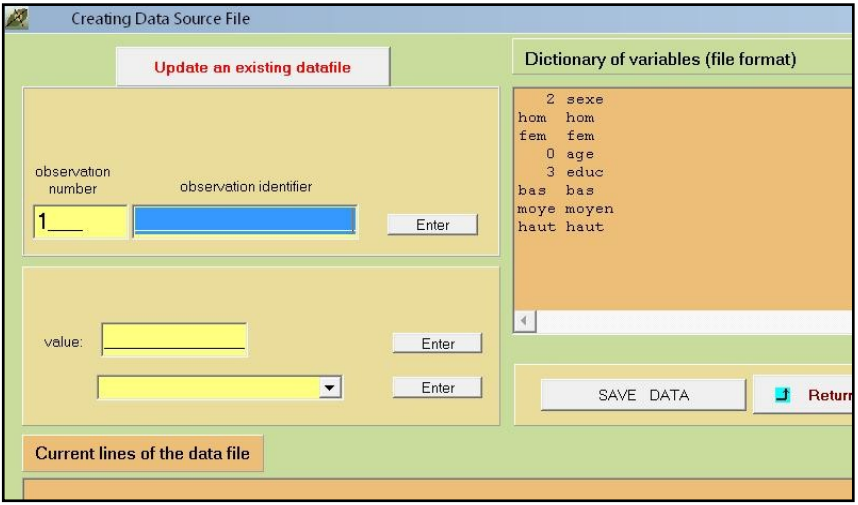
Une fois le fichier dictionnaire créé :

- Sélectionnez, Creating the data file dans Building the dictionary of variables and creating the data file.
  - ⊙ Une fenêtre pour la construction du fichier de données apparaît.

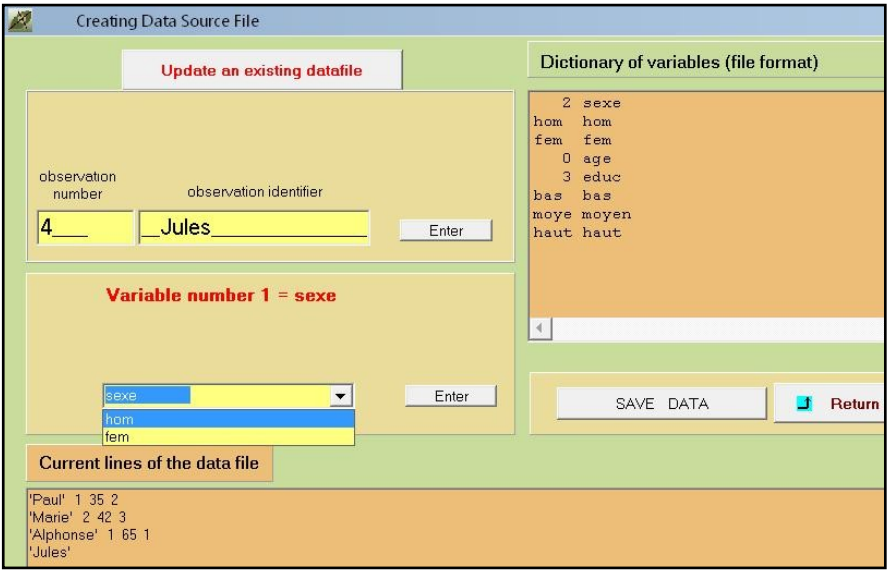


➢ Cliquez sur LOAD DICTIONARY et ouvrez le fichier dictionnaire créé précédemment Database_dic.txt.

⊙ Une fenêtre pour la capture de données apparaît. Le dictionnaire des variables s'affiche dans la fenêtre de droite.



➢ Saisir l'identifiant de l'individu et cliquer sur Enter (ou appuyer sur "Entrée" sur le clavier). La 1ère variable s'affiche dans la fenêtre.



- Sélectionnez la modalité correspondant à l'individu avec le menu déroulant puis cliquez sur Enter (ou appuyez sur "Entrée" sur le clavier).

La 2ème variable s'affiche. Il s'agit de la saisir de la même façon. Une fois les variables capturées pour l'individu, l'individu suivant apparaît.

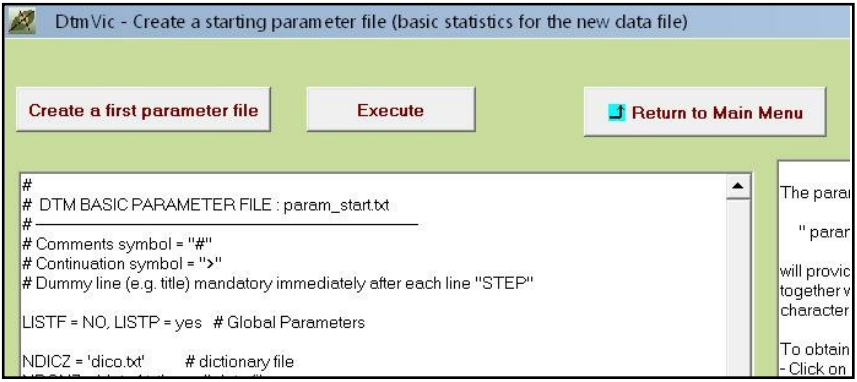
Le dictionnaire s'affiche dans la fenêtre en haut et droite et le fichier des données dans la fenêtre en bas.

IV.2.3. Création des fichiers DtmVic

Une fois la saisie achevée :

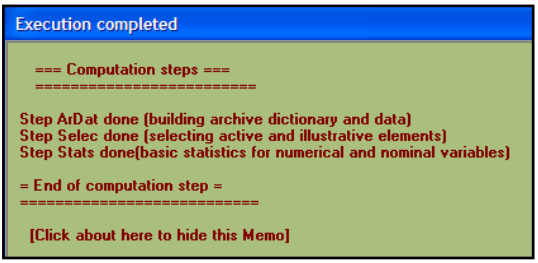
- sauvegardez le fichier en cliquant sur SAVE DATA et enregistrer le nom du fichier de données : Database_dat.txt (par exemple) relatif au fichier dictionnaire créé précédemment puis :

- Cliquer sur, Creating a first parameter file.
  - ⊙ Une fenêtre pour la création du fichier paramètre apparaît.
- Cliquer sur le nouveau bouton: Create a first parameter file. Le fichier paramètre apparaît dans la fenêtre du bas



- Cliquer sur Execute.

- La fenêtre d'exécution apparaît, identique à celle de la procédure d'importation (simple vérification et statistiques de base pour les données enregistrées).



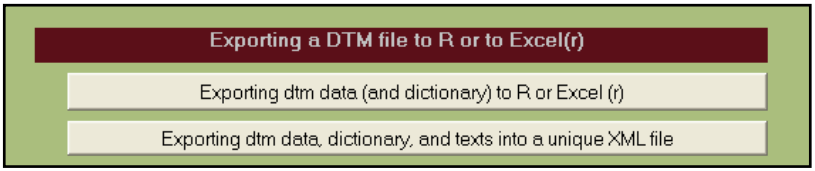
Les fichiers saisis (dictionnaire et données) sont maintenant prêts pour les analyses.

IV.4. Exportation de fichiers de données en format "Excel ®" (ou : *XL*)

La procédure d'exportation présente principalement l'intérêt d'exporter des variables recodées et surtout des coordonnées factorielles archivées ou une partition calculée et archivée (les procédures d'archivage sont traitées au chapitre V).

On propose ici d'exporter le fichier de données issu de l'exemple de l'analyse des correspondances multiples du chapitre II. L'exportation peut se faire vers un format Excel ou vers un format voisin acceptable par la procédure "read.table" du langage R (fichier dont le format est identique au format Excel, à l'exception de la première ligne).

IV.4.1. Procédure d'exportation



- Cliquer sur Exportation dtm data dans Exporting a DTM file to R or to Excel.
  - ⊙ Une fenêtre apparaît.
- Cliquez sur Open a dictionary. Ouvrir alors, à titre d'exemple, le fichier MCA _dic.txt" dans "EX_A03.MultCorAnalysis
  - ⊙ Une première fenêtre affiche le libellé des variables et des modalités.
- Cliquez ensuite sur Open a Data file et ouvrez le fichier "MCA_dat.txt" dans "EX_A03.MultCorAnalysis". Puis cliquez sur List of variables.

Il est possible d'exporter soit en format Excel ® soit en format R. Ici, nous faisons le choix d'un fichier Excel.

- Sélectionner Create new data file for Excel et répondre OK à la boite de message: "New data file created".

Un nouveau fichier MCA_d_dtm_XL.csv est créé dans le répertoire EX_A03.MultCorAnalysis.

Un extrait de ce fichier Excel (14 individus, 4 variables) figure ci-dessous.

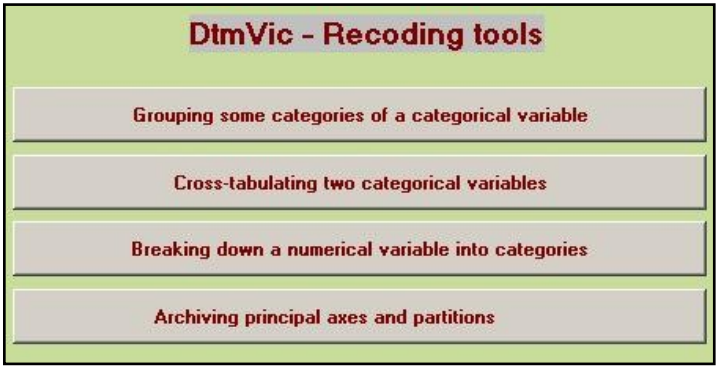
Identifiers	region	size_of_town	gender	age
5	mediterranee	<2000	female	27.000000
11	mediterranee	<2000	female	32.000000
18	mediterranee	>200000	male	21.000000
24	ouest	<2000	female	42.000000
30	ouest	<2000	male	29.000000
36	bassin_parisien	10001-20000	female	35.000000
42	bassin_parisien	10001-20000	male	71.000000
48	ouest	<2000	male	62.000000
54	ouest	20001-50000	male	24.000000
60	est	<2000	male	52.000000
66	est	10001-20000	female	42.000000

V. Recodage, archivage, outils divers

L'exploitation des données statistiques est un processus interactif nécessitant souvent plusieurs itérations. Parmi les opérations les plus courantes, le regroupement des modalités d'une variable nominale, le croisement de deux variables nominales, la division en classes d'une variable continue sont fréquemment suscités par les résultats d'une analyse antérieure. L'archivage des partitions ou des axes factoriels est également utile pour avancer dans la compréhension des données en permettant de réaliser des analyses qui les prennent en compte. Ces étapes de recodage sont en fait assez fondamentales. Bien que Dtm-Vic ne soit pas un logiciel de gestion de données, il a paru nécessaire de rendre ces opérations accessibles à partir de la boîte à outils *(Toolbox)*.

V.1. Recodage

- Cliquez sur Toolbox Data Recoding
  - Le menu qui apparaît concerne le recodage des données et l'archivage de certains résultats.



Création ou recodage de variables nominales :

i) Regroupement de modalités ;
ii) Création d'une variable nominale par croisement de deux variables nominales ;
iii) Transformation d'une variable continue en variable nominale ;
iv) Archivage des axes factoriels et des partitions.

Que ce soit pour le regroupement de modalités d'une variable nominale, pour la création d'une variable par croisement de deux variables nominales ou pour la transformation d'une variable continue en une variable nominale, la première étape consiste à :

– ouvrir le fichier dictionnaire : 1. Open a dictionary
– puis celui des données : 2. Open a data file
– à lister les variables : 3. List of variables
– puis, cliquer sur : 4. Continue

Les opérations suivantes sont effectuées à partir du jeu de données de l'exemple **EX_A03.MultCorAnalysis** dans le dossier **DtmVic_A_Start**.

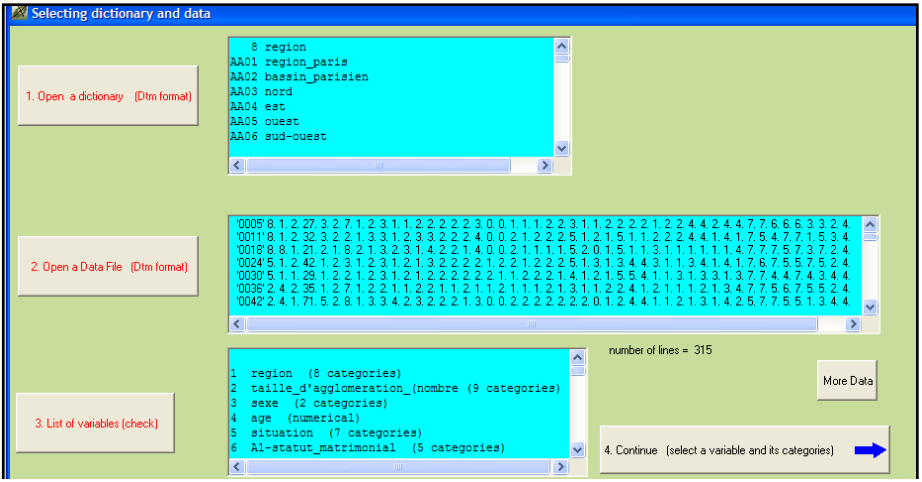
V.1.1 Regroupement de modalités d'une variable nominale

Lors du dépouillement de données d'enquête et à l'occasion de tris à plat effectués sur les variables nominales, on doit parfois regrouper certaines modalités d'une variable nominale pour satisfaire, dans la mesure du possible, certaines règles de recodage : éviter des modalités à faible effectif, équilibrer le nombre de modalités des variables nominales, regrouper des catégories similaires ou trop fines.

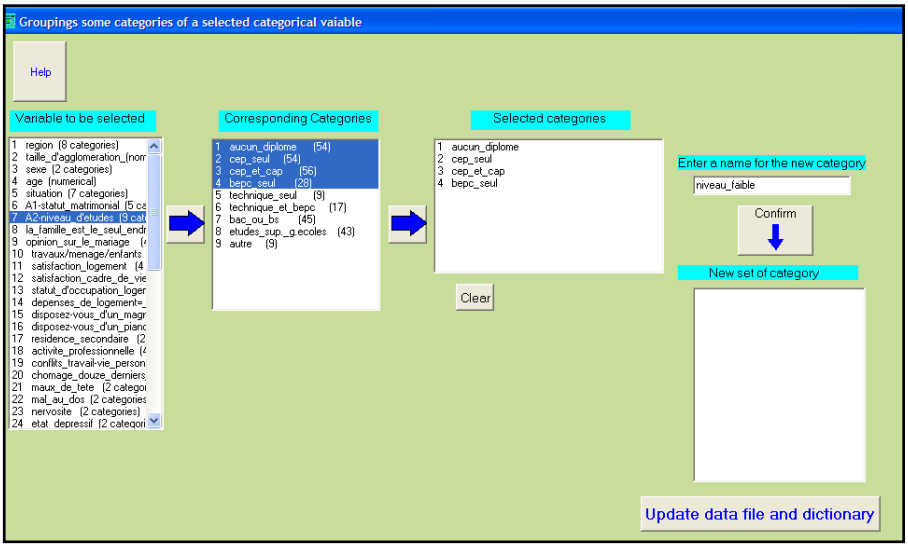
➢ Cliquez sur **Grouping some categories of a categorical variable**.

⊙ La fenêtre de sélection des fichiers dictionnaire et des données apparaît.

➢ Ouvrir les fichiers MCA_dic.txt et MCA_dat.txt dans le dossier **EX_A03.MultCorAnalysis**, lister les variables et cliquer sur 4. Continue.



⊙ Une nouvelle fenêtre apparaît.



➢ Sélectionnez la variable à recoder. Ici nous choisissons, dans la 1ère fenêtre, la variable "7-niveau d'étude" en 9 catégories. Les catégories (modalités) de cette variable s'affichent dans une 2ème fenêtre. Sélectionnez l'ensemble des modalités à regrouper qui apparaissent dans une 3ème fenêtre. Entrez le nom de la nouvelle modalité dans la

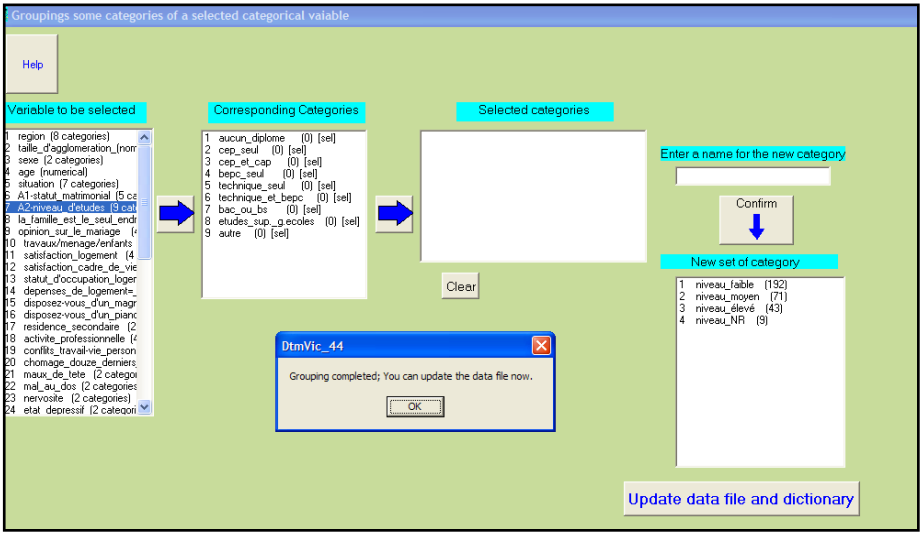
4ème fenêtre puis confirmez. La nouvelle modalité apparaît dans la 5ème fenêtre.

- Recommencez la procédure pour toutes les modalités de la variable. Si une modalité n'est pas à regrouper, la sélectionner et lui attribuer une étiquette.

Dans l'exemple, nous avons regroupé les 4 premières modalités en "niveau_faible", les 3 autres en "niveau_moyen", la 8ème modalité en "niveau_élevé" et la 9ème en "niveau_NR" (Non-réponse).

Les modalités de la nouvelle variable apparaissent dans la 5ème fenêtre. Cette variable est positionnée à la fin du fichier et se nomme "var7-4cat".

- Une fois les regroupements terminés, répondre : OK puis cliquez sur : **Update data file and dictionary**.



Deux nouveaux fichiers dictionnaire et de données sont créés dtm_dic_newG7.txt et dtm_dat_newG7.txt, toujours dans le même dossier **EX_A03.MultCorAnalysis**.

- ⊙ Une fenêtre s'affiche pour présenter ces nouveaux fichiers (pour lesquels l'utilisateur pourra choisir de nouveaux noms, s'il le juge utile).

➢ Cliquez sur : **Return**. L'opération de regroupement des modalités est terminée.

V.1.2. Croisement de deux variables nominales

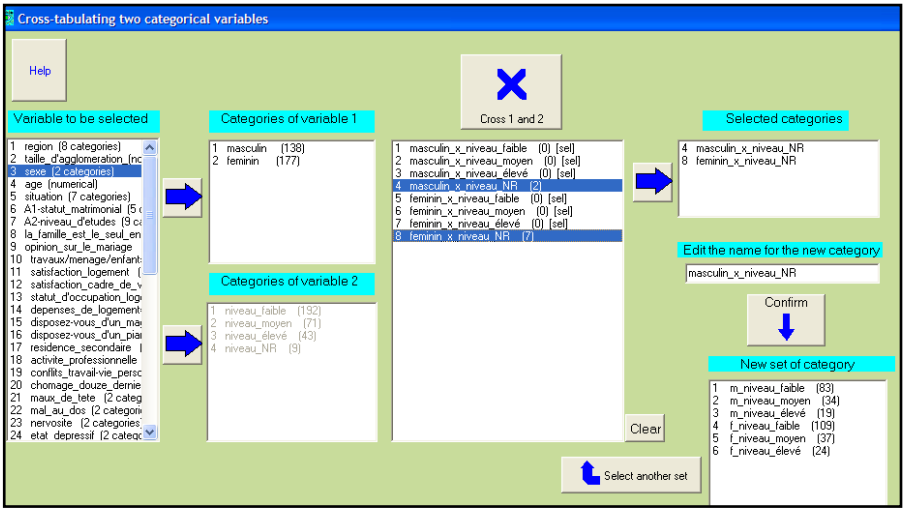
On souhaite dans ce cas augmenter les possibilités d'analyse et d'interprétation en créant une nouvelle variable nominale à partir du croisement de deux variables nominales (Exemple : sexe X âge).

➢ Cliquez sur: **Cross-tabulating two categorical variables**.

⊙ La fenêtre de sélection des fichiers dictionnaires et des données apparaît.

➢ Ouvrir les fichiers dictionnaire et de données concernés (pour l'exercice, on pourra ouvrir les fichiers précédemment créés dans le dossier **EX_A03.MultCorAnalysis** : dtm_dic_newG7.txt et dtm_dat_newG7.txt), lister les variables, puis : Continuer.

⊙ Une fenêtre apparaît. (cf. ci-dessous)



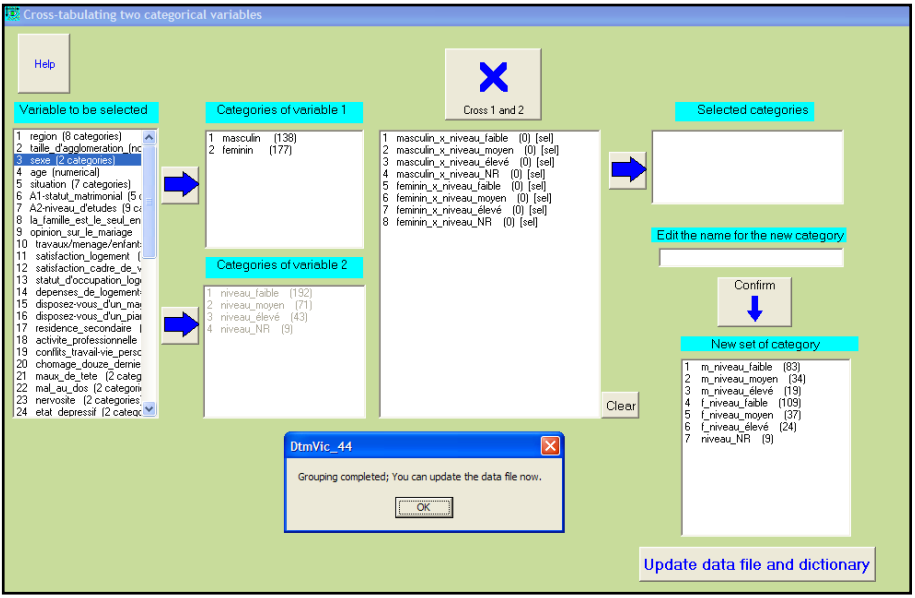
➢ Sélectionnez les modalités à regrouper ou à valider qui apparaissent dans une 3ème fenêtre.

- Entrez l'étiquette de la nouvelle modalité dans la 4ème fenêtre puis confirmez. La nouvelle modalité apparaît dans la 5ème fenêtre.

- Recommencez la procédure d'étiquetage pour toutes les nouvelles modalités. Si une modalité n'est pas à regrouper, la sélectionner et lui attribuer une étiquette.

- Une fois les regroupements terminés, répondre : OK à la boîte de message, puis cliquez sur **Update data file and dictionary**.

Deux nouveaux fichiers dictionnaire et de données sont créés : dtm_dic_newCr3x52.txt et dtm_dat_newCr3x52.txt dans le dossier **EX_A03.MultCorAnalysis**. Une fenêtre s'affiche pour présenter ces nouveaux fichiers.



- Cliquez sur **Return**. Une fois l'opération terminée, vous pouvez modifier les noms des fichiers par défaut si ceux-ci ne conviennent pas.

V.1.3. Transformation d'une variable continue en variable nominale

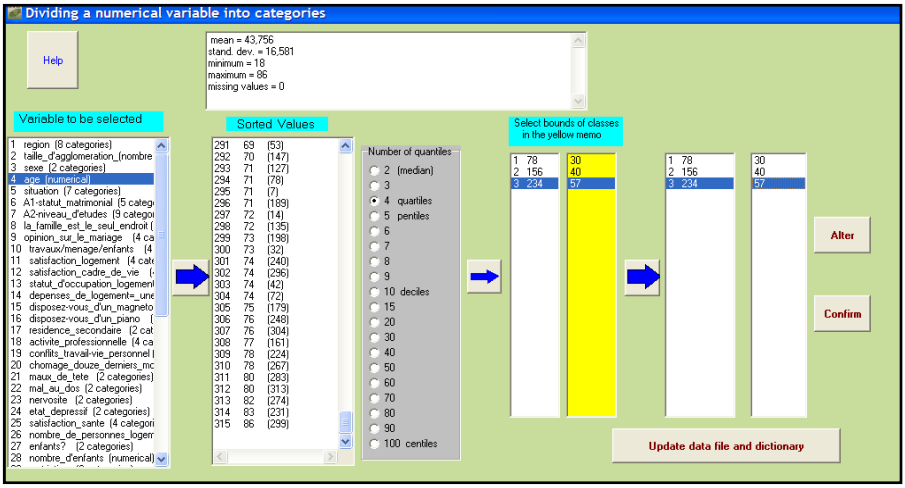
Cette procédure permet de transformer une variable continue en une variable nominale, en regroupant les valeurs numériques en classes. Ce regroupement en k classes se fait à partir d'un découpage préalable en n quantiles (n classes d'effectifs égaux), n étant beaucoup plus grand que k. Ce découpage est utile car il "délinéarise" le rôle de la variable dans les calculs (des liaisons non linéaires peuvent alors être prises en compte).

Cliquez sur **Breaking down a numerical variable into categories**.

⊙ La fenêtre de sélection des dictionnaires et des données apparaît.

➢ Ouvrir, dans le dossier **EX_A03.MultCorAnalysis,** les fichiers dictionnaire et de données MCA_Fr_dic.txt et MCA_dat.txt.

⊙ Une fenêtre apparaît.



➢ Sélectionnez la variable continue (V4_age) et transférez la dans la 2ème fenêtre Sorted Values. Choisir le nombre de quantiles (5 par exemple, on peut aussi choisir 20 (ou 100) quantiles pour mieux maîtriser les limites de classes).

➢ Transférez en cliquant sur ➔. Confirmer et répondre OK lors de l'affichage du nombre de modalités.

➢ Une fois les regroupements terminés, répondre OK puis cliquez sur **Update data file and dictionary**. Deux nouveaux fichiers dictionnaire et de données sont créés : dtm_dic_newD4.txt et dtm_dat_newD4.txt ainsi qu'un fichier "Dissecting_Check" qui présente les détails de l'opération. Cliquez sur **Return** pour revenir au menu principal.

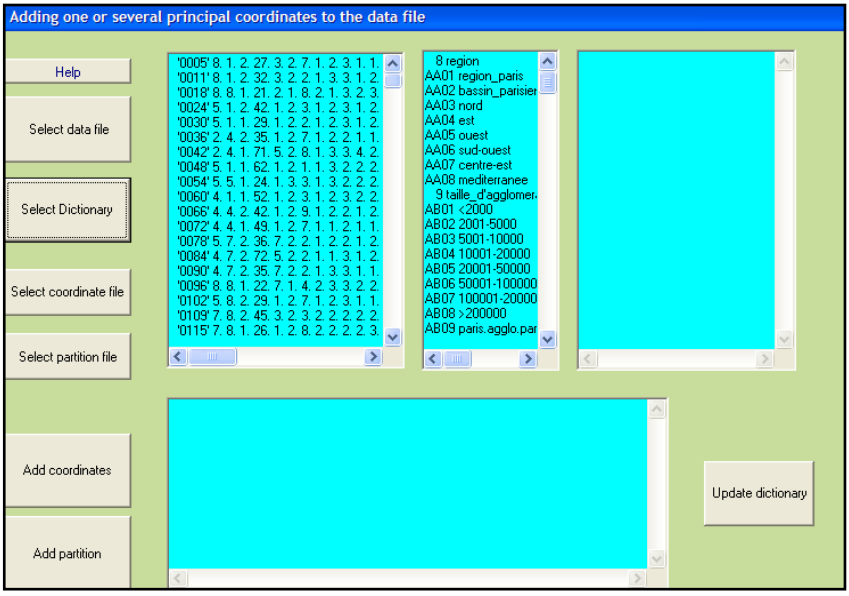
V.1.4. Archiver des facteurs ou des partitions

On peut vouloir enrichir le fichier de données initial par les résultats d'une analyse factorielle ou d'une classification. Les facteurs ou partitions sont alors considérés comme de nouvelles variables.

Attention : On ne peut archiver des facteurs ou des partitions si l'analyse qui les a produits a utilisé un filtre interne sur les individus (lors de la création du fichier de commande). En revanche, on peut utiliser un filtre externe (avant toute analyse) tel que défini au paragraphe V.2.1 ci-après.

➢ Cliquez sur **Archiving principal axes and partitions**.

⊙ Une fenêtre apparaît.

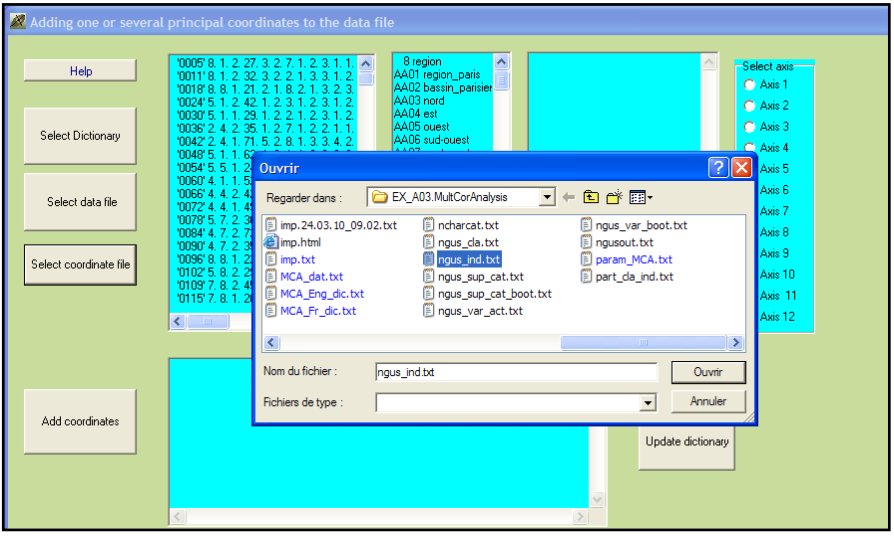


➢ Ouvrir le fichier dictionnaire (MCA_dic.txt) puis celui de données (MCA_dat.txt) et sélectionner l'archivage d'un facteur : Select coordinate file ou d'une partition : Select partition file.

a. Archiver un facteur

➢ Cliquer sur Select coordinate file

⊙ Une fenêtre apparaît affichant le dossier **EX_A03.MultCorAnalysis** où figure le fichier ngus_ind.txt des coordonnées factorielles créé lors de la procédure : **MCA – Multiple Correspondence Analysis**



➢ ouvrez le fichier ngus_ind.txt, puis sélectionnez l'axe à archiver.

⊙ Les coordonnées factorielles apparaissent dans la 3ème fenêtre.

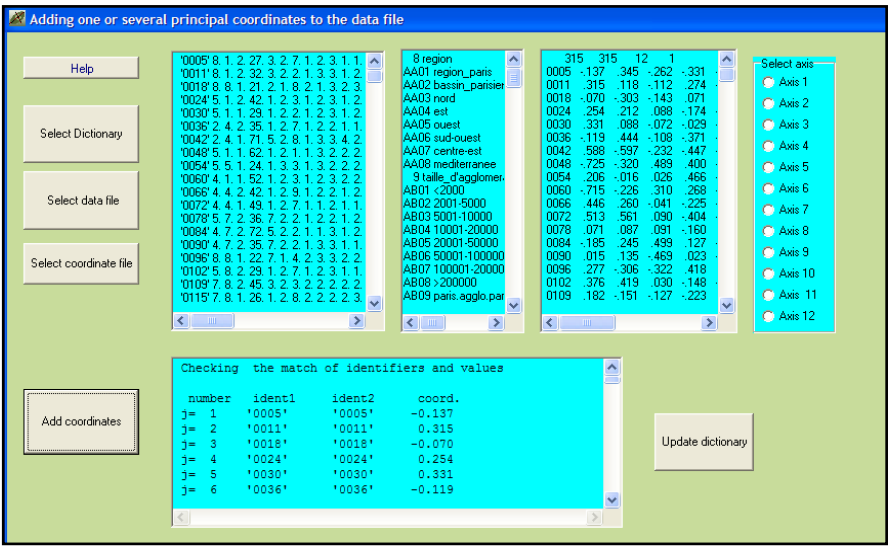
➢ Cliquez sur Add coordinates.

⊙ Une boite de message : "Coordinate added. Please, update the dictionary" apparaît. Répondre OK. L'archivage des coordonnées s'affiche dans la fenêtre du bas.

➢ Cliquez sur Update dictionary et répondre OK dans la boite de message "Dictionary updated" qui s'affiche.

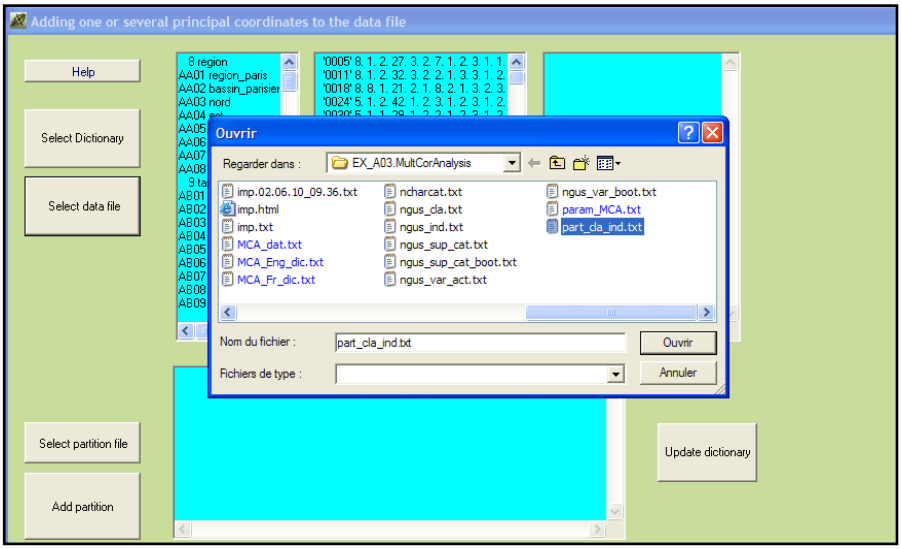
Les fichiers dictionnaire et des données sont créés dans le dossier **EX_A03.MultCorAnalysis** et sont nommés : dtm_dico_newA1.txt et dtm_data_newA1.txt.

Pour archiver un deuxième facteur recommencer la procédure en sélectionnant les **nouveaux** fichiers dictionnaire et données : dtm_dico_newA1.txt et dtm_data_newA1.txt. Même procédure pour archiver une partition à la suite.

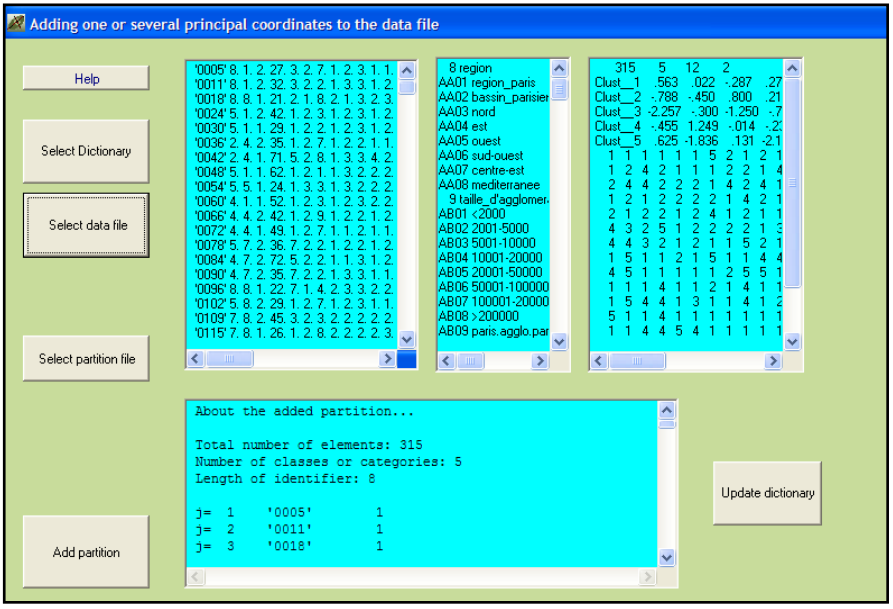


b. Archiver une partition

- Cliquez sur Select partition file
  - Une fenêtre du dossier : **EX_A03.MultCorAnalysis** s'affiche où figure le fichier : part_cla_ind.txt du stockage de la partition créée lors de la procédure : **MCA – Multiple Correspondances Analysis** et dont le nombre de classes a été spécifié lors du paramétrage de l'analyse.
- Ouvrez, dans le dossier : **EX_A03.MultCorAnalysis**, le fichier : part_cla_ind.txt (fichier de la partition, voir les noms des divers fichiers texte créés par Dtm-Vic dans le "Help about files" du menu principal.
- Cliquez sur Add partition.



⊙ Une fenêtre: "Partition added. Please, update the dictionary" apparaît. Répondre : OK.



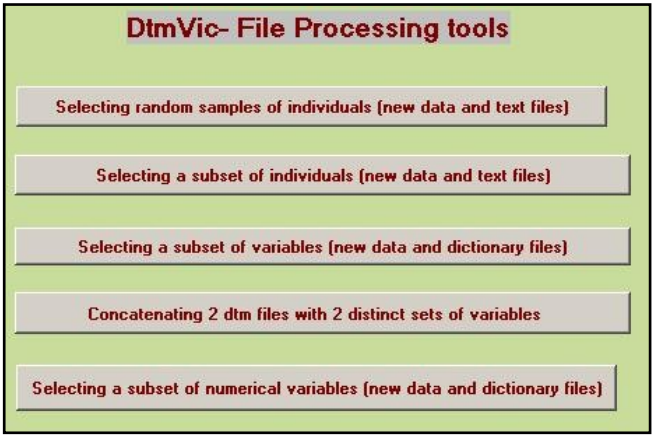
⊙ L'archivage de la partition s'affiche dans la fenêtre inférieure.

➢ Cliquez sur : Update dictionary et répondre : OK dans la fenêtre : "Dictionary update" qui s'affiche.

Les fichiers dictionnaire et des données sont créés dans le dossier **EX_A03.MultCorAnalysis** et sont nommés : dtm_dico_newP1.txt et dtm_data_newP1.txt.

V.2. Interventions élémentaires sur la base de données

⊙ Le second groupe d'actions est obtenu en cliquant sur : ToolBox : File Processing



i) Sélection d'un sous-ensemble aléatoire d'individus (lignes) ;
ii) Sélection d'un sous-ensemble d'individus (lignes) à partir d'un filtre ;
iii) Sélection d'un sous-ensemble de variables (colonnes) ;
iv) Concaténation de deux bases de données (variables différentes).
v) Sélection d'un sous-ensemble de variables ayant un poids maximum.

Les sections i) et v) ne seront pas traitées de façon détaillées ici. Elles comportent des rubriques « HELP » qui devraient faciliter la tâche des utilisateurs.

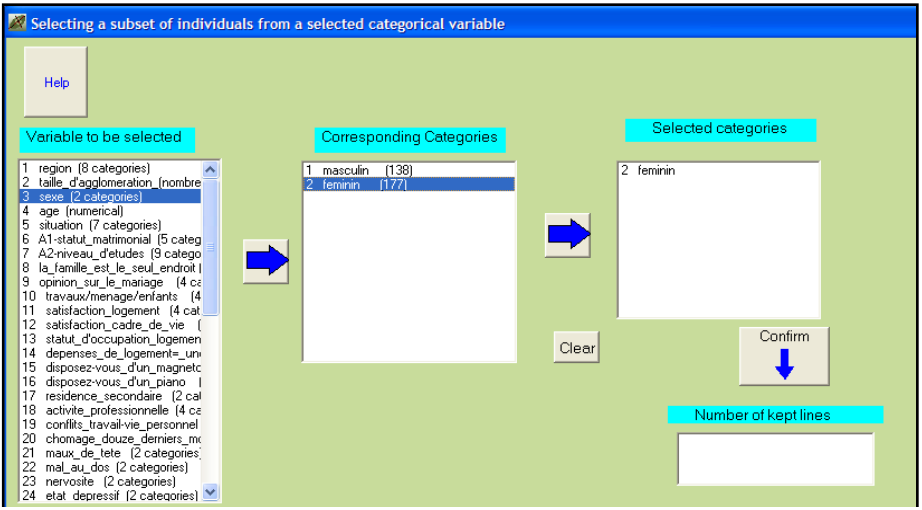
La section i) permet de diviser par 2 ou 4 la taille de l'échantillon de départ (formé de la réunion des 2 ou 4 groupes). Ceci permet de tester des analyses de façon plus économique, mais aussi de valider des structures observées.

La section v) est très particulière et répond à la situation pratique suivante : Si les données comportent un grand ensemble homogènre de n variables numériques dont la somme sur les individus a un sens, alors on peut sélectionner les p variables ($p < n$) de plus fortes sommes. Exemple : on a pour 10 000 individus 1200 variables (nombre de visites pour 1200 sites webs). On peut sélectionner les 400 sites les plus visités, pour travailler sur ce seul sous ensemble.

V.2.1 Sélection d'un sous-ensemble d'individus par filtrage

Il est fréquent d'avoir à travailler de façon approfondie sur une sous-population, par exemple les femmes, les personnes ayant accès à internet à leur domicile, etc.. Il est alors commode de sélectionner un sous-fichier Dtm-Vic, sans avoir à re-importer les données à partir de la base initiale.

- Cliquez sur **Selecting a subset of individuals**.
  - ⊙ Une fenêtre apparaît.
- Ouvrir les fichiers dictionnaire (par exemple MCA_dic.txt), de données (par exemple MCA_dat.txt), lister les variables, ouvrir le fichier texte des questions ouvertes s'il existe, puis continuer.
  - ⊙ Une nouvelle fenêtre apparaît.



- Sélectionnez la variable nominale dans la 1ère fenêtre (par exemple 3-Sexe), la transférer dans la 2ème fenêtre.
- Sélectionnez la modalité de filtrage (par exemple "féminin").
- Cliquez sur Confirm. Le nombre de lignes (individus) conservées s'affichent dans la fenêtre "Number of kept lines" et correspond au nombre d'individus de la catégorie affiché dans la fenêtre "Corresponding Categories", catégorie qui ne s'affiche plus après la procédure de confirmation.
- Cliquez sur **Update data file and text file**.

Un fichier dont le nom par défaut est : dtm_data_Subset.txt est crée dans le dossier **EX_A03.MultCorAnalysis**. Le fichier dictionnaire MCA_dic.txt reste inchangé. L'opération est terminée.

V.2.2 Sélection d'un sous-ensemble de variables

- Cliquez sur **Selecting a subset of variables**. Une fenêtre apparaît.
- Ouvrir les fichiers dictionnaire et de données de la base concernée, lister les variables puis continuer. Une nouvelle fenêtre apparaît.
- Sélectionner dans la 1ère fenêtre l'ensemble des variables à conserver dans la nouvelle base, les transférer dans la 2ème fenêtre.

➢ Cliquer sur **Update data file and dictionary**.

Deux fichiers dtm_dic_SELVAR.txt et dtm_dat_SELVAR.txt sont créés dans le dossier **EX_A03.MultCorAnalysis**.

V.2.3 Concaténation d'ensembles de variables

Cette option permet de concaténer deux bases de données de Dtm-Vic pour créer une nouvelle base de données réunissant deux ensembles de variables (opération utile lorsque les fichiers livrés sont segmentés, comme dans le cas des versions d'Excel pour lesquelles le nombre de colonnes est limité). **Attention ! Les deux bases doivent contenir les mêmes individus en lignes, triés dans le même ordre.**

➢ Cliquez sur **Concatenating 2 dtm files with 2 distinct sets of variables**.

⊙ Une fenêtre apparaît.

➢ Ouvrir les deux fichiers des données puis des dictionnaires à concaténer. Ils s'affichent dans chacune des quatre fenêtres.

➢ Cliquez sur Merge Sorted Files.

⊙ Une série de fenêtres s'affichent successivement. Les deux premières précisent l'intégration des deux fichiers de données In file, 0 individuals have no counterparts : répondre OK. Une troisième fenêtre donne le nombre d'individus du nouveau fichier : Répondre OK.

Enfin, une quatrième fenêtre indique que la procédure "merge" des deux fichiers de données est effectuée : répondre OK. Les identifiants des deux fichiers apparaissent dans la fenêtre du bas.

⊙ Cliquez sur Merge dictionaries.

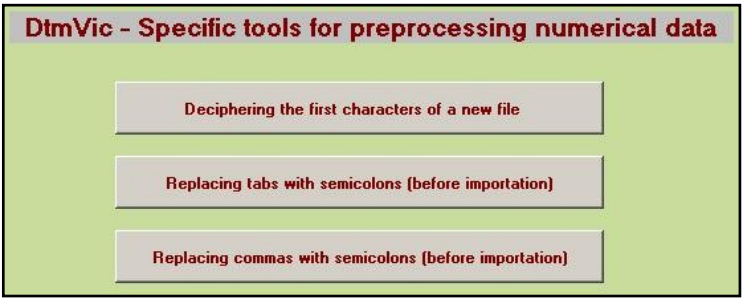
⊙ Une fenêtre indique que la procédure "merge" des dictionnaires est effectuée : répondre OK, et cliquez sur Exit.

Deux fichiers dtm_dico_new et dtm_data_new sont alors créés.

V.3. Outils spécifiques de pré-traitement

V.3.1 Données numériques et textuelles

⊙ Le bouton ToolBox : Preprocessing (numerical) propose des outils élémentaires de prise de contact et de prétraitements en vue de l'importation ou de l'utilisation de données numériques et textuelles.



Lorsque l'on reçoit un fichier de données (internet, clé USB, DVD), il est utile de vérifier la nature des caractères présents (numériques, alphanumériques, séparateurs, ponctuation, éventuelles tabulations, etc.).

Le premier bouton "Deciphering the characters of a new file" nous donne le code ASCII correspondant aux 6000 premiers caractères d'un fichier, opération aussi utile (parfois) qu'élémentaire.

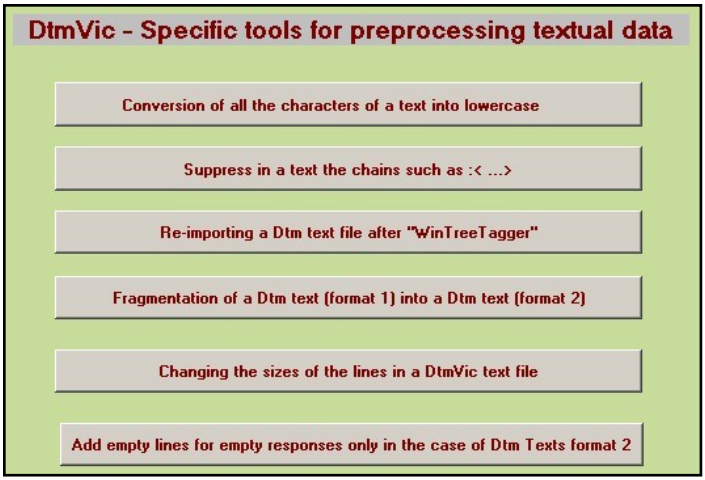
Le second bouton, Replacing Tabs with semicolons, est utile lors de l'importation d'un fichier Excel®. Dans certaines versions d'Excel, le séparateur du format ".csv" est une virgule (*comma*) (cas fréquent des pays pour lesquels la notation décimale utilise des points à la place des virgules, la virgule pouvant alors jouer un rôle de séparateur d'enregistrement). Le passage par la sauvegarde avec les tabulations comme séparateurs est alors plus pratique. Il faut ensuite utiliser ce bouton. *Attention ! Si un tel fichier contient déjà des points-virgules, la transformation ne pourra avoir lieu.*

Le troisième bouton, Replacing commas with semicolons, est utile lorsque le fichier fourni a déjà été sauvegardé avec des virgules comme séparateur. Comme précédemment, si le fichier contient déjà des points-

virgules, la transformation ne pourra avoir lieu. Il convient donc de les remplacer par un autre symbole avant d'actionner le bouton.

V.3.2 Données textuelles uniquement

⊙ Le dernier bouton ToolBox : Preprocessing texts propose quelques procédures en vue de l'importation ou de l'utilisation directe des textes.



i) Conversion des textes en minuscules.

Le bouton "Conversion of the characters of a text into lowercase" transforme tous les caractères en minuscules. Ceci fait gagner de l'information en termes de fréquences pour le vocabulaire banal, mais des traitements préliminaires peuvent s'imposer, pour traiter, par exemple, l'homonymie entre certains noms propres (noms de lieu par exemple) et noms communs (Tour, Paris, Pierre, Constant). L'étape CORTEX (après le bouton "Create" du menu principal) doit en général intervenir avant ce type de transformation.

ii) Suppression des balises XML ouvertes et fermées « < » et « > » et du texte qu'elles peuvent contenir.

Le second bouton "Suppress in a text the chains such as <....>" est utile si le texte transmis contient des balises dont on ne veut pas tenir compte (textes formatés pour le logiciel Lexico3 par exemple). Toutefois, ce type de transformation doit intervenir après que le texte ait été segmenté à partir de certaines balises.

iii) Ré-importation dans DtmVic d'un fichier de type Dtmic (type 1 ou 2) ayant été soumis au logiciel (gratuit) TreeTagger.

Le bouton : Re-importing a Dtm text file after WinTreeTagger permet de lemmatiser un texte (remplacer les formes graphiques par le lemme correspondant). Il permet également de supprimer certaines catégories grammaticales (prépositions, articles, etc..). Quatre options sont disponibles respectivement pour les textes anglais, français, espagnols, italiens. Ceci suppose l'installation du logiciel (gratuit) WinTreeTagger.

TreeTagger : Auteur: Helmut Schmid, IMS, University of Stuttgart, TreeTagger est un analyseur morpho-syntaxique indépendant des langues dans son principe. Les informations et le téléchargement se font à partir du site web:
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
On notera que TreeTagger n'a pas d'interface graphique. (Il fonctionne avec ligne de commande). Comme suggéré par Helmut Schmid, on peut utiliser l'interface Windows plus conviviale WinTreeTagger réalisée par Ciarn O'Duibhin.:
http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm

Notez que le fichier alimentant WinTreetagger doit impérativement être un fichier texte au format Dtm-Vic: Le nouveau fichier à importer issu de WinTreetagger contient trois colonnes séparées par des tabulations. - *Première colonne:* occurrence - *Deuxième colonne* : Etiquette grammaticale - *Troisième colonne:* Lemme. Un tel fichier contient autant de lignes qu'il y a d'occurrences et de signes de ponctuation. (voir le « Help » de Dtm-Vic). C'est ce fichier que la procédure remet en format Dtm-Vic.

iv) Fragmentation d'un texte en format 1 (textes séparés par *) en textes de format 2.***

Le bouton :

Fragmentation of a Dtm text (format 1) into a Dtm Text (format 2) permet de fragmenter les textes importants en petites unités de

longueurs variables. Ces unités sont formés de une ligne, deux lignes... des textes initiaux (il s'agit approximativement d'une fragmentation en unités de contexte). On verra ci-dessous que la longueur des lignes peut être modifiée dans certaines limites. Une variable nominale est créée pour conserver l'information rattachant les unités aux textes initiaux. (voir le « Help » *in situ*).

v) Changement de longueur des lignes de texte.

Le bouton Changing the size of the lines in a DtmVic text file permet une importation ou un reformatage des fichiers textes. Au départ, on dispose de textes en format DtmVic (1 ou 2) sans limitation pour la longueur des lignes. A la fin : textes ayant des lignes d'une longueur choisie par l'utilisateur, (mais < 200 caractères). Cette procédure permet d'importer des textes aux lignes très longues, mais aussi de formater les unités de contexte (cf. point iv ci-dessus).

vi) Enfin le dernier bouton déclenche une procédure limitée et spécialisée qui permet de faire respecter la contrainte « une ligne vide par réponse ouverte vide » pour des fichiers qui utiliseraient deux séparateurs consécutifs. Elle est parfois utile après la ré-importation après TreeTagger d'un fichier de type 2.

VI. Autres analyses avec Dtm-Vic

Visualisations élaborées, Contiguïté, Graphes, Images

L'orientation principale de Dtm-Vic est l'analyse exploratoire multi-dimensionnelle des données numériques et textuelles, avec validation systématique des résultats (par la complémentarité d'approches différentes et par les méthodes de *Bootstrap*). D'autres applications et d'autres outils qui permettent d'envisager des analyses plus élaborées sont présentés dans ce chapitre.

Dans le dossier : **DtmVic-Examples/DtmVic-Examples_C_NumData**, une série d'exemples reprend les techniques d'analyses de base sur données numériques. Cette série va nous donner l'occasion d'approfondir les outils Visualization et Contiguity du volet VIC de Dtm-Vic : *VIC steps*. Nous étudierons ensuite l'application des analyses en axes principaux aux visualisations de graphes et aux compressions d'images

1. L'exemple 1, dans le dossier **EX_C01.PCA_Semio**, vise à décrire un ensemble de variables numériques (un extrait de données semiométriques) par analyse en composantes principales. Les axes principaux sont complétés par une classification et une description automatique des classes (un fichier de commande tout préparé nous permet d'accéder directement à la phase "VIC"). On ne présentera ici que le sous-menu "Visualisation" de la phase "VIC": visualisation des classes (ou catégories) en utilisant des symboles ou des couleurs, des enveloppes convexes ou ellipses de densité pour les classes, le tracé de l'arbre de longueur minimale (Minimum Spanning Tree), les visualisations des graphes des plus proches voisins, classifications de type k-means "à la volée", etc ...
2. L'exemple 2, dans le dossier **EX_C02.PCA_Contiguity**, analyse un ensemble classique de variables numériques (les données IRIS d'Anderson et Fisher, bien connues des statisticiens) par l'analyse

en composantes principales, la classification, l'analyse de contiguïté et l'analyse discriminante. Cet exemple reprend les procédures de base de l'exemple 1 précédent : Analyse en composantes principales et classification (clustering) d'un ensemble de données numériques, avec différents outils de visualisation, impliquant aussi une variable nominale spécifique (la variable identifiant les 3 espèces d'iris). L'exemple présente ensuite les améliorations apportées par l'analyse de contiguïté, dont l'analyse linéaire discriminante et un cas particulier.

3. L'exemple 3, dans le dossier **EX_C03-Graphs** vise à décrire trois types simples de graphes planaires symétriques, principalement au moyen de l'analyse des correspondances. Contrairement aux exemples précédents, le répertoire contient plusieurs jeux de données : un graphe en forme de damier, un cycle, et des graphes empiriques représentant des régions du Japon et de France. Ces exemples veulent jeter un pont entre les différentes possibilités du logiciel Dtm-Vic : un même graphe peut provenir de données d'entrée différentes : données numériques, données textuelles, et aussi dans ce cas un "format externe" spécifique pour les graphes.

4. L'exemple 4, dans le dossier **EX_C04.Images** a une vocation plutôt pédagogique : montrer les propriétés de compression numériques des méthodes en axes principaux (et des séries de Fourier discrètes, à titre de comparaison). Les images nécessitant un format spécifique, cette application ne s'insère pas dans les chaînes de traitement les plus usuelles de Dtm-Vic. Une interface spécialisée est obtenue par le bouton SVD and CA of Images de la rubrique "DtmVic Images" du menu principal.

Les analyses de base auxquelles les exemples 1 à 3 ont recours sont celles présentées au chapitre II. Nous ne revenons donc pas sur la mise en place interactive du *fichier de commande* (ou : *fichier paramètre*) et des analyses. Nous présentons ici directement ces analyses à partir du *fichier de commande* déjà préparé et fourni avec chaque exemple.

VI.1. Données numériques : "Sémiométrie"

L'exemple 1, dans le dossier **EX_C01.PCA_Semio**, vise à analyser un ensemble de variables numériques (extrait de "données sémiométriques") par analyse en composantes principales. Les principaux axes de visualisation sont complétés par une classification, avec une description automatique des classes. La procédure "Vizualisation" propose différents outils de visualisation des enveloppes convexes ou des ellipses de densité pour les classes, le tracé de l'arbre de longueur minimale (*Minimum Spanning Tree*) et la visualisation des graphes des plus proches voisins. Une nouvelle classification des variables (ou des observations ou individus) à travers une méthode de type k-means peut être obtenue et visualisée, itération après itération, à partir du sous-menu "Visualisation".

VI.1.1. Les données sémiométriques

Dans la plupart des enquêtes en marketing, il est courant d'inclure des informations sur les modes de vie et des valeurs des personnes interrogées. Ces informations sont généralement obtenues par une série de questions décrivant les attitudes et les opinions.

La "Sémiométrie" est une technique introduite par Jean-François Steiner[9]. L'idée de base consiste à insérer dans le questionnaire, une série de questions composées uniquement de mots (une liste de 210 mots est actuellement utilisée, mais il va être question ici d'une liste abrégée contenant un sous-ensemble de 70 mots). Les personnes interrogées doivent noter ces mots selon une échelle comportant sept niveaux, le niveau le plus bas (1), est relatif à un sentiment "plus désagréable (ou déplaisant) vis-à-vis du mot présenté", le plus haut niveau (7), relatif à une sensation "plus agréable (ou plaisante) "au sujet de ce mot.

Le traitement des questionnaires par l'Analyse en Composantes Principales met en évidence une structure stable (la stabilité concerne l'espace des 8 premiers axes principaux). Des propriétés très similaires sont observées dans dix pays différents, malgré les problèmes posés par la

[9] Pour de plus amples informations, se référer à l'ouvrage : "La sémiométrie" par L. Lebart, M. Piron, JF Steiner; Editeur: Dunod, Paris, 2003. Ce livre peut être téléchargé à partir du site: www.dtmvic.com (rubrique "Publications").

traduction de la liste des mots. Comme pour les études "styles de vie", les espaces obtenus permettent de positionner des produits, des marques ou des services dans le cadre d'études de recherche marketing.

Les trois fichiers qui composent cet exemple se trouvent dans le répertoire **DtmVic-examples/DtmVic-Examples_C_NumData/EX_C01. PCA_Semio**.

1. **le fichier de données :** PCA_semio.dat.txt
   Cet exemple est de taille réduite et comprend 300 répondants (au lieu de 1000 ou 2000 qui sont les tailles usuelles des échantillons d'enquête sémiométrique) et 76 variables: 70 mots (les notes attribuées à ces mots sont considérées ici comme des variables numériques) et 6 variables nominales décrivant les caractéristiques des répondants.

2. **le fichier de dictionnaire :** PCA_semio.dic.txt
   Le fichier dictionnaire contient les identifiants des 76 variables. Dans le dictionnaire interne de DtmVic, les identificateurs de catégories doivent commencer : "colonne 6" [une police à intervalle fixe telle que "courrier" peut être utile pour faciliter ce genre de format].

3. **le fichier de commandes :** EX_C01_Param.txt
   La phase de calcul de l'analyse est décomposée en "étapes". Chaque étape nécessite quelques paramètres décrits brièvement dans le menu principal de DtmVic (bouton: Help about command parameters).
   Notons qu'un fichier de commande similaire au "fichier de commande EX_C01_Param.txt peut également être généré en cliquant sur le bouton : Create du menu principal (étapes de base), comme indiqué au chapitre 2 de ce manuel. Une fenêtre "Select a basic analysis" s'affiche. Cliquez ensuite sur : Principal Components analysis situé dans la rubrique "Numerical Data", et suivez les instructions.

VI.1.2. Calculs de base (PCA et classification)

(Exécution de l'exemple C.01 "sémiométrie" et lecture des résultats)

a. Ouverture du fichier paramètre

➢ Cliquez sur le bouton : **Open an existing command file** de la rubrique Command File (menu principal).

Ensuite, recherchez le dossier **DtmVic-Examples_C_NumData** dans **DtmVic-examples**. Dans ce répertoire (ou dossier), ouvrez le répertoire **EX_C01. PCA_Semio**.

Ouvrez le fichier de paramètres: EX_C01_Param.txt.

⊙ Le fichier paramètre s'affiche dans la fenêtre de l'éditeur de texte :

```
#-------------------------------------------------
LISTP = yes, LISTF = no, LERFA = yes  # global parameters
#
NDICZ = 'PCA_semio.dic.txt'          # Dictionary file
NDONZ = 'PCA_semio.dat.txt'           # Data file

STEP ARDAT
========== Reading data and dictionary
NIDI = 1,  NIEXA = 300 NQEXA = 76


STEP SELEC
========== Selecting active and supplementary variables
LSELI = TOT,  IMASS = UNIF, LZERO = NOREC, LEDIT = short
CONT ACT 1--70
NOMI ILL 71--76
END

STEP STATS
==========  Basic descriptions
LHIST=no

STEP PRICO
========== Principal component analysis
LCORR = 2, ……..
```

Vérifier que les fichiers de données et dictionnaires inscrits dans le fichier paramètre sont cohérents avec ceux du répertoire.

Dix "étapes" sont effectuées:

- ARDAT (Archivage des données),
- SELEC (Sélection des éléments actifs et supplémentaires),
- PRICO (analyse en composantes principales),
- DEFAC (brève description des axes factoriels),
- RECIP (Classification ascendante hiérarchique – méthode des voisins réciproques),
- PARTI (Coupure du dendrogramme produit par l'étape précédente, et optimisation de la partition obtenue),
- DECLA (Description automatique des classes de la partition),
- SELEC (Sélection d'une variable spécifique),
- EXCAT (Extraction de la variable spécifique, sélectionnés par l'étape SELEC qui précède, pour être utilisée dans la suite).

Dans ce fichier de commandes, l'étape SELEC joue comme toujours un rôle fondamental pour décider quelles variables seront actives ou supplémentaires. L'étape RECIP effectue une classification hiérarchique des observations en utilisant l'algorithme "de la recherche en chaîne de voisins réciproques" et l'étape PARTI coupe l'arbre obtenu selon le nombre de classes fixé *a priori*, puis optimise la partition par des itérations de type "k-means" (RECIP et PARTI exécutent un algorithme "hybride" de classification[10]).

L'éditeur de texte interne de Dtm-Vic contient aussi un bouton Help about command parameters qui donne brièvement (en Anglais) la signification de chacun des paramètres.

Nous ne modifierons pas le fichier de commande.

- Cliquez sur Return to execute dans le bandeau pour revenir au menu principal.

b. Exécution du fichier de commande (fichier paramètre)

- Cliquez sur : Execute de Command File

Les étapes de calcul de base présentes dans le fichier de commande sont exécutées : archivage de données et le dictionnaire, choix des éléments actifs et supplémentaires, statistiques élémentaires, analyse en composantes principales de la table sélectionnée, réplications "bootstrap" de la table, brève description des axes, classification, description approfondie des classes. Les 9 étapes décrites ci-dessus s'affichent à la fin de l'exécution. Pour examiner les résultats numériques, comme précédemment :

- Cliquez sur : Basic numerical results de Result Files

  Les résultats numériques sont du même type que ceux présentés en section II.1.3 (Analyse en composantes principales, chapitre II).

VI.1.3. Visualisation et lecture des résultats

Nous procédons tout d'abord comme dans le chapitre II à propos de la

[10] "*Statistique Exploratoire Multidimensionnelle*" (4ème édition, L. Lebart, M. Piron, A. Morineau, Dunod, Paris, 2006).

visualisation des résultats en utilisant les possibilités offertes par la seconde phase : VIC *: Visualization, Inference, Classification steps*.

L'analyse réalisée permet d'examiner les axes et les plans factoriels : boutons AxesView et PlaneView, la validation des positions des points sur les graphiques par *Bootstrap,* avec : BootstrapView, la classification avec le bouton : ClusterView et les cartes auto-organisées avec : Kohonen Map.

Les fonctionnalités de ces quatre premiers boutons ont été décrites à propos des exemples des chapitres II et III. Nous allons dans cette section nous focaliser sur les fonctionnalités du bouton Visualization.

Cette option propose des outils de visualisations complémentaires des plans factoriels et de la classification : ellipse de densité ou enveloppes convexes des classes ; tracé de l'arbre de longueur minimale, tracé des plus proches voisins dans les plans factoriels ; visualisation pédagogique de la construction progressive des classes (cas de la procédure k-means / nuées dynamiques) ; visualisation dans les plans factoriels des cartes de Kohonen et de certains graphes.

a. Visualisation utilisant la partition demandée dans le fichier de commande (étapes RECIP et PARTI)

- Cliquez sur le bouton Visualization
  - ⊙ Une fenêtre intitulée "DTM-visualization: loading files, selecting axes" apparaît.
- Cliquez sur Load coordinates. Dans le sous-menu correspondant, choisir, dans un premier temps, le fichier: ngus_ind.txt. Les principales coordonnées des individus (lignes) sont sélectionnées.
  - ⊙ Une sous-fenêtre donne les caractéristiques du fichier.
- Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, sélectionnez la partition obtenue précédemment à l'étape de calcul. Choisir alors Load partition File et ouvrir le fichier part_cla_ind.txt (*classes de la partition pour les individus*).
- - Cliquez sur Graphics puis, dans la fenêtre "Sélection des axes", choisir les axes 2 et 3 (qui constituent le premier "plan sémio-

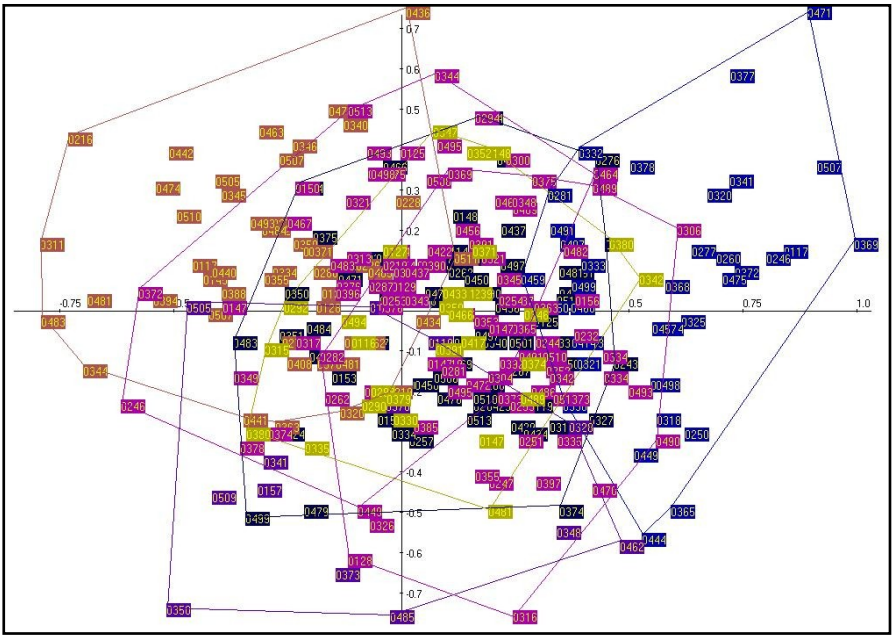
métrique", car l'axe 1 est un "axe de notation").

➢ Cliquez ensuite sur Continue puis sur DISPLAY.

⊙ Le Plan factoriel (2, 3) s'affiche.

Dans le bandeau vertical de gauche de la fenêtre "Graphics" figure une série de boutons : On appuie sur un bouton pour l'activer (couleur rouge), et on appuie de nouveau pour le désactiver (couleur noire).

- Le bouton C.Hull (*Convex Hull* = Enveloppe convexe) trace l'enveloppe convexe de chaque classe. Pressez ce bouton : La figure ci-dessous représente les 300 individus dans le plan (2, 3), avec une couleur par classe et une enveloppe convexe par classe.



Enveloppes convexes (Convex Hulls) des 7 classes dans le plan (2 , 3) après activation du bouton : "C.Hull" puis du bouton : "Colours".

b. Visualisation à partir d'une variable nominale

La visualisation précédente va être reprise, mais au lieu d'utiliser une partition fournie par un algorithme de classification, nous allons utiliser la partition induite par les catégories d'une variable nominale spécifique. Il s'agit de la variable numéro 76 (sexe), sélectionnée et extraite à travers les

deux étapes SELEC et EXCAT (à la fin du fichier de commande).

- Cliquez à nouveau sur V Visualization
- Dans la fenêtre intitulée "DTM-visualization: Loading files, Selecting axes", cliquez sur Load coordinates

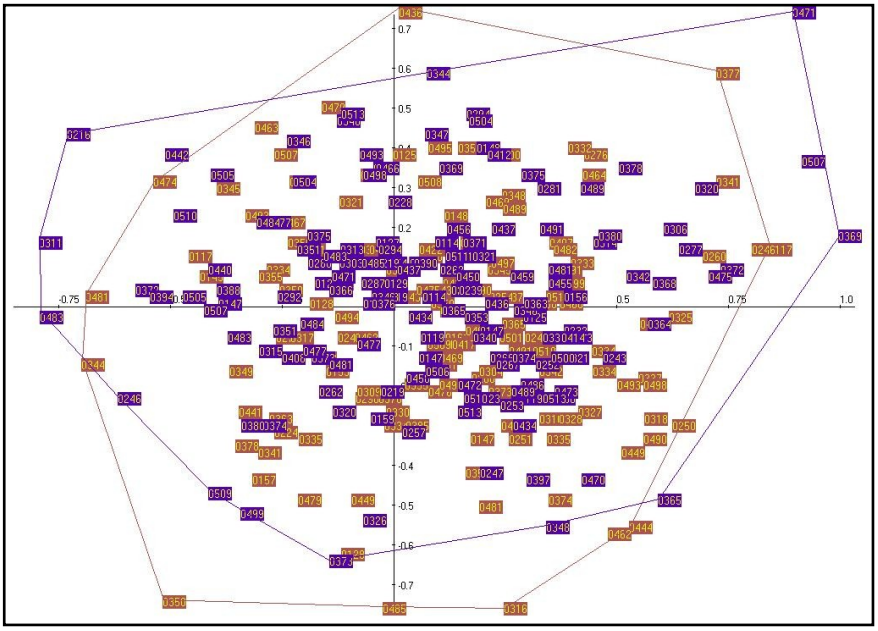
  Dans le sous-menu correspondant, choisir à nouveau le fichier: "ngus_ind.txt". Les coordonnées des individus (lignes) sont sélectionnées.

- Cliquez ensuite sur Load or create a partition

  Dans le sous-menu correspondant, choisissez le fichier "part_cat.txt". La partition induite par les catégories de la variable 76 (sexe) est chargée.

- Cliquez sur Graphics puis choisissez encore les axes 2 et cliquez sur Continue puis sur DISPLAY. Le Plan factoriel (2, 3) s'affiche.

- Cliquez sur le bouton C.Hull (*Convex Hull* = Enveloppe convexe). La figure ci-dessous représente alors les 300 individus dans le plan (2, 3), avec une couleur par classe et une enveloppe convexe par classe.



Enveloppes convexes des deux sous-nuages hommes/femmes dans le plan sémiométrique (2, 3) (après usage du bouton "Colours" de façon à contraster les deux sous-populations..

Commentaire:
Les deux catégories "Homme" [violet] et "Femme" [marron] sont en fait étroitement liées à l'axe vertical 3 (on peut le vérifier à partir des zones de confiance *bootstrap*). Mais ce lien est à peine visible quand on regarde directement les enveloppes convexes des deux sous-nuages correspondant à ces deux catégories de répondants. Ce résultat (presque) paradoxal illustre la différence entre "statistiquement significatif" (qui est le cas ici) et "nettement distinct" (qui n'est pas le cas ici).

c. Arbre de longueur minimum et plus proches voisins dans l'espace des variables (mots)

- Cliquez sur Visualization
  - ⊙ Une fenêtre intitulée "DTM-visualization: loading files, selecting axes" apparaît.
- Cliquez sur Load coordinates. Dans le sous-menu correspondant, choisissez le fichier: ngus_var_act.txt pour une classification de **variables** ; les coordonnées principales des variables actives sont sélectionnées.
  - ⊙ Une sous-fenêtre donne les caractéristiques du fichier.
- Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, sélectionnez la partition obtenue précédemment à l'étape de calcul. Choisissez alors No partition.
- 1 - Cliquez sur Min. Span. Tree (Minimum Spanning Tree). Choisissez le nombre d'axes qui serviront à calculer l'arbre de longueur minimale; par exemple ici les 3 premiers axes. Confirmer en cliquant OK sur le nombre d'axes conservés.
- 2- Cliquez sur N.N (recherche de plus proches voisins [*Nearest Neighbours*] limité à 20 NN). Répondre OK à la recherche des plus proches voisins.
- 3- Cliquez sur Graphics puis choisissez encore les axes 2 et 3 (qui constituent le premier "plan sémiométrique", car l'axe 1 est une "axe de notation") dans la fenêtre "Sélection des axes", et cliquez

sur Continue puis sur DISPLAY.

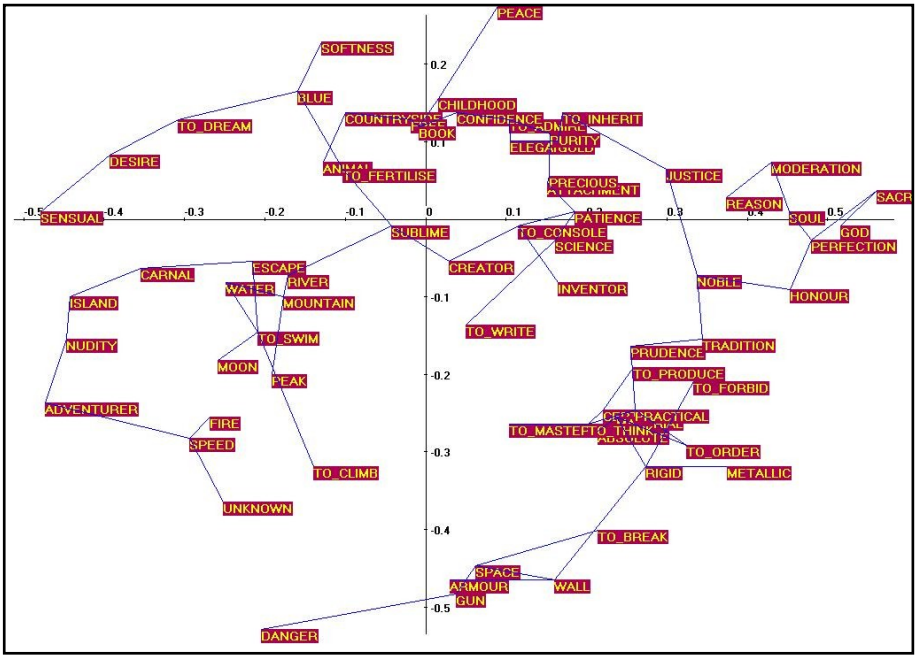
⊙ Le Plan factoriel (2, 3) s'affiche.

Dans le bandeau de gauche de la fenêtre "Graphics" figurent quatre familles de boutons :

Sur la barre d'outils verticale gauche, on appuie sur un bouton pour l'activer (couleur rouge), et on appuie de nouveau pour le désactiver (couleur noire)

- Le bouton MST (*Minimum Spanning Tree*) trace l'arbre de longueur minimale.
- Le bouton N.N (*Nearest Neighbours* = plus proches voisins) joint chaque point à ses voisins les plus proches. Le bouton N.N.up permet d'incrémenter le nombre de plus proches voisins (<= 20).

La figure ci-dessous montre l'espace des mots (plan (2, 3) avec le tracé de l'arbre de longueur minimum. Cet arbre étant calculé dans l'espace des trois premiers axes, il apporte un complément par rapport au plan. Les figures obtenues à partir des plus proches voisins sont analogues.

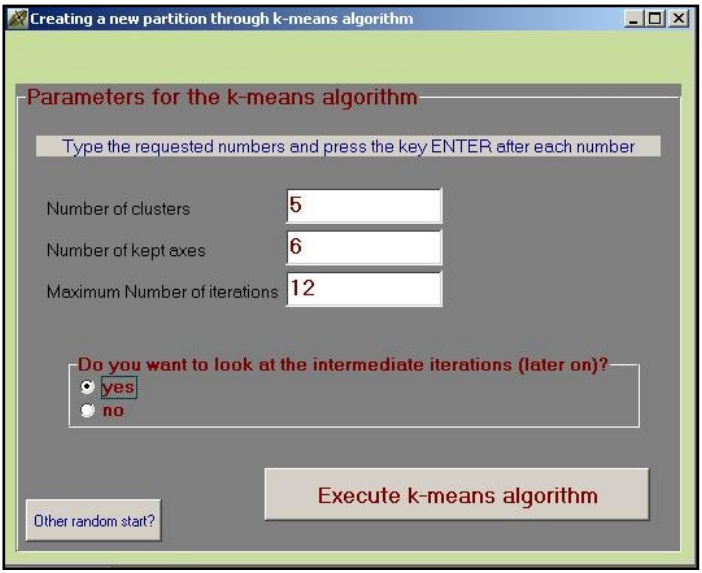


Tracé de l'Arbre de longueur minimale dans le plan sémiométrique (2, 3)
(après avoir actionné le bouton "Colours").

d. Calcul direct d'une partition dans le menu "Visualisation"

Dtm-Vic permet de construire "à la volée" (c'est-à-dire en dehors du "fichier de commande") une "partition k-means" de variables (ou des individus).

- Cliquez sur V Visualization
  - ⊙ Une fenêtre intitulée "DTM-visualization: Loading files, Selecting axes" apparaît.
- Cliquez sur Load coordinates. Dans le sous-menu correspondant, choisissez le fichier: ngus_var_act.txt pour une classification des variables actives ; Pour un regroupement d'individus, sélectionnez le fichier: ngus_ind.txt.
- Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, sélectionnez l'option "Create a new k-means partition". Vous devez ensuite sélectionner (figure ci-dessous) le nombre de classes désirées, le nombre de coordonnées principales pour les calculs de distances, le nombre maximum d'itérations (généralement < 12) et vous devez cocher "yes" si vous désirez visualiser les itérations.

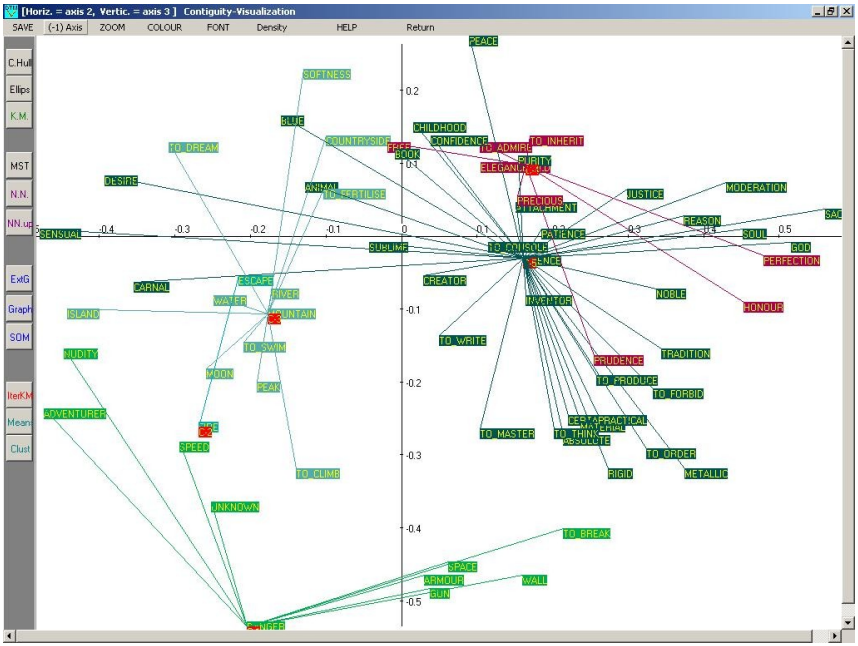


Exemple du choix de 5 classes, calculées avec 6 axes, en 12 itérations au maximum.

A titre pédagogique, on peut visualiser les différentes étapes de construction de la partition dans la fenêtre, après avoir cliqué sur Graphics. Il faut ensuite sélectionner les axes 2 et 3, puis cliquer sur Continue puis enfin cliquer sur : DISPLAY.

Dans la barre verticale gauche, il faut alors cliquer sur IterKM, puis cliquer alternativement sur Means (calcul des centres des classes) et sur Clust (affectation des éléments aux nouveaux centres de classes) jusqu'à ce que la convergence soit atteinte. Notez que la partition obtenue par cet algorithme classique des k-moyennes ne coïncidera pas en général avec la partition induite par les paramètres du fichier de commande.

Voir l'encadré de la section VI.1.2 précédente à propos des calculs réalisés par les instructions du fichier de commande (étapes RECIP et PARTI).



Exemple de visualisation de la première itération de la construction de la partition en 5 classes. Les variables (ici : les mots) sont reliées par des segments de droites aux centres provisoires de classes auxquels elles sont affectées (les 5 mots qui servent de centres provisoires de classes sont repérables par un carré rouge).

VI.2. Données numériques et contiguïté : Iris

Cette section concerne l'analyse exploratoire d'un ensemble de variables numériques (Les données "Iris" de Anderson et Fisher, jeu de données classique pour les statisticiens) par l'analyse en composantes principales et la classification (avec une description automatique des classes obtenues). Elle ajoute à ces approches de base, l'analyse de contiguïté et l'analyse discriminante.

La première partie de cet exemple est très semblable à l'exemple VI.1 de la section précédente: analyse en composantes principales et classification (clustering) d'un ensemble de données numériques, avec divers outils de visualisation, impliquant également la présence de données nominales.

Les paragraphes qui suivent présentent les améliorations apportées par l'analyse de contiguïté.

VI.2.1 Rappel sur l'Analyse de Contiguïté

Dans l'analyse de la contiguïté, nous considérons le cas d'un ensemble d'observations multidimensionnelles (n objets décrits par p variables, conduisant à une matrice X (n, p)). Les observations ont *a priori* une structure de graphe. Les n observations sont ainsi les n sommets d'un graphe symétrique G, dont la matrice associée symétrique (n, n) est la matrice M ($m_{ii'}$= 1 si les sommets i et i' sont reliés par une arête, $m_{ii'}$= 0 sinon).

Une telle situation se produit lorsque les sommets représentent les points d'une série chronologique ou des zones géographiques. L'Analyse de contiguïté, confronte les variances locales et globales, et généralise ainsi l'analyse discriminante, qui confronte les variances internes et globales (ou, de façon équivalente les variances internes et externes). Elle permet de mettre en évidence les niveaux responsables des patterns observés (locaux ou globaux). Le graphe constitue donc une information externe.

Dans cet exemple, nous allons traiter la situation dans laquelle la matrice M et la structure du graphe ne sont pas externes, mais proviennent de la matrice des données X elle-même, G étant par exemple le graphe symétrisé des k plus proches voisins provenant d'une distance entre les observations.

(Le cas d'un graphe externe fait partie des fonctionnalités du logiciel Dtm-Vic, mais n'est pas présenté dans ce manuel de prise en main).

Il s'agit donc ici d'une analyse de contiguïté "intrinsèque", ouvrant des possibilités intéressantes d'exploration de données. L'idée de déduire des données une métrique susceptible de mettre en évidence l'existence de classes a été suggérée par Art *et al.* (1982) et Gnanadesikan *et al.* (1982).

Quelques références pour la section VI.2.1
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 Nerworks*, 19, 847-854.

VI.2.2 Les données "Iris" de Fisher / Anderson :

Pour les données numériques en format texte de Dtm-Vic, cherchez le répertoire **DtmVic_Examples**. Dans ce répertoire, ouvrez le dossier : **DtmVic_Examples_C_NumData**. Puis ouvrez le dossier de l'exemple C.2, nommé **EX_C02. PCA_Contiguity** .

Comme d'habitude, il est recommandé d'utiliser un répertoire pour chaque application, car Dtm-Vic produit beaucoup de fichiers-textes intermédiaires liés à l'application.

Au départ, le répertoire doit contenir 3 fichiers:

a) le fichier de données,
b) le fichier dictionnaire,
c) le fichier de commandes.

a) Fichier de données: iris_dat.txt

L'exemple comporte 150 observations et 5 variables: 4 mesures (ces variables numériques sont les longueurs des différents constituants des fleurs: Longueur des sépales, Largeur des sépales, Longueur des pétales, largeur des pétales) et une variable nominale décrivant l'appartenance aux espèces (trois espèces d'iris : *setosa, versicolor, virginica*). Référence: Anderson, E. (1935). The irises of the Gaspe Peninsula, *Bulletin of the American Iris Society*, **59**, 2–5.

Le fichier de données iris_dat.txt comprend donc 150 lignes et 6 colonnes (l'identificateur de lignes [entre quotes] suivi de 5 valeurs [correspondant à 4 variables numériques et une variable nominale, séparées par au moins un espace).

b) Dictionnaire: iris_dic.txt

Le fichier-dictionnaire iris_dic.txt contient les identificateurs de ces 5 variables. Dans cette version du dictionnaire interne Dtm-Vic, les identifiants des catégories doivent commencer en colonne 6 [une police à intervalles fixe – *courrier*, par exemple - est nécessaire pour représenter clairement ce genre de format].

c) Fichier de commandes: EX_C02_Param.txt

La phase de calcul de l'analyse est décomposée en "étapes". Chaque étape nécessite quelques paramètres décrits brièvement dans le menu principal de Dtm-Vic (bouton: Help about parameters).

Notons qu'un autre fichier de commande similaire (mais pas forcément identique) au fichier de commande : EX_C02_Param.txt peut également être généré en cliquant sur le bouton Create, rubrique Command File du menu principal ("Basic Steps"). Procéder alors comme le montre le premier exemple de la section II.1 dévolu à l'analyse en composantes principales.

VI.2.3 Calculs de base (ACP et classification)

(Exécution de l'exemple C.2 "Iris" et lecture des résultats)

a. Ouverture du fichier paramètre

- Cliquez sur le bouton : **Open an existing command file** de la rubrique Command File (menu principal). Recherchez dans **DtmVic_Examples** le sous-répertoire **DtmVic_Examples_C_NumData**. Dans ce répertoire, ouvrir le répertoire de l'exemple C.2 nommé **EX_C02. PCA_Contiguity** .

➢ Ouvrir alors le fichier de commande: EX_C02_Param.txt

⊙ Le fichier paramètre s'affiche dans une fenêtre (qui est aussi un éditeur de texte).

Dans ce fichier de commandes, on peut lire, après avoir identifié les deux fichiers (données et dictionnaire), que 9 "étapes" sont effectuées :

- ARDAT (Archivage des données),
- SELEC (sélection des éléments actifs et supplémentaires),
- PRICO (analyse en composantes principales),
- DEFAC (Brève description des axes factoriels),
- RECIP (classification hiérarchique),
- PARTI (coupure du dendrogramme produit par l'étape précédente, et l'optimisation de la partition obtenue),
- DECLA (description automatique des classes de la partition),
- SELEC (sélection d'une variable nominale, dans ce cas),
- EXCAT (extraction d'une variable nominale (3 espèces d'iris) sélectionnée par l'étape SELEC)

Notez que le bouton: Help about parameters est accessible à partir de cet éditeur de texte pour expliciter (en Anglais) les paramètres de chaque étape.

b. Exécution du fichier de commande (fichier paramètre)

Revenir au menu principal et exécuter les étapes de calcul de base.

➢ Cliquez sur Return to execute dans le bandeau pour revenir au menu principal.

➢ Cliquer sur le bouton : Execute de : Command File .

Cette opération exécute les étapes de calcul du fichier de commandes.

c. Lecture des résultats

➢ Cliquer sur le bouton : Basic numerical results de : Result Files

Le browser ouvre le fichier HTML nommé "imp.html" qui contient les principaux résultats des étapes précédentes de calcul de base. Après lecture de ces résultats numériques, retour au menu principal.

VI.2.4. Visualisation et lecture des résultats

Comme pour l'exemple C.1 précédent portant sur la sémiométrie, nous allons maintenant utiliser les fonctionnalités du bouton V Visualization.

a. Visualisation à partir d'une partition induite par une variable nominale (espèce d'iris)

Nous allons visualiser les différentes espèces de fleurs (variable n° 5) dans le plan engendré par les premiers axes principaux de l'ACP.

- Cliquez sur V Visualization

  ⊙ Une fenêtre intitulée "DTM-visualization ... " apparaît.

- Cliquez sur Load coordinates. Dans le sous-menu correspondant, choisir, dans un premier temps, le fichier: ngus_ind.txt. Les principales coordonnées des individus (lignes) sont sélectionnées.

- Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, choisissez alors "Load partition File" et ouvrir le fichier part_cat.txt, la partition induite par les 4 catégories de la variable 5 (les 4 espèces d'iris). Cette partition a été choisie et extraite à travers les 2 dernières étapes SELEC et EXCAT du fichier de commande ci-dessus.

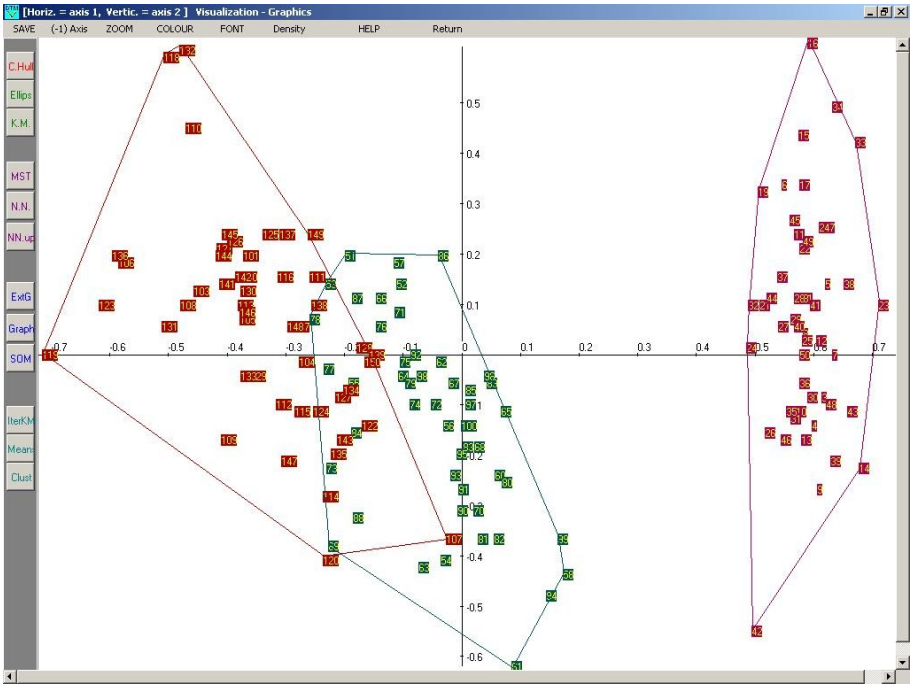
- Cliquez sur Graphics puis choisissez les axes 1 et 2 (par défaut) dans la petite fenêtre "Sélection des axes" et cliquez sur "Continue" puis sur DISPLAY.

Dans la nouvelle fenêtre intitulée "Visualization - Graphics" sont affichés les individus dans le plan engendré par les axes sélectionnés. Une couleur aléatoire est attribuée à chaque catégorie. Le bouton Colour permet d'essayer un nouveau jeu de couleurs.

Sur la barre d'outils verticale gauche, on appuie sur un bouton pour l'activer (couleur rouge), et on appuie de nouveau pour le désactiver (couleur noire)

- Le bouton Density, par souci de clarté, permet de remplacer les identifiants des individus par un seul caractère rappelant sa classe (l'identifiant et le numéro de la classe s'obtiennent en cliquant sur le bouton gauche de la souris au voisinage des points).

- Pressez le bouton C.Hull (*Convex Hull* = enveloppe convexe) qui trace l'enveloppe convexe de chaque classe. Le tracé apparaît ci-dessous.



Plan principal de l'ACP des 4 variables continues (mesures) avec tracé des enveloppes convexes correspondant aux trois espèces d'iris. L'identification des trois espèces par des couleurs différentes est réalisée *a posteriori*, après l'analyse en composantes principales. On voit que deux espèces se chevauchent sur ce plan principal.

À cette étape, nous avons obtenu un affichage des 150 individus, avec les enveloppes convexes correspondant aux trois espèces. C'est l'affichage classique dans le plan principal de l'ACP, montrant que sur la droite, la première espèce *setosa* (nombre = 50) est bien séparée des espèces deux et trois qui, elles, se chevauchent.

b. Visualisation d'une partition en trois classes non supervisée

Nous allons maintenant revenir au menu principal et refaire la visualisation précédente, mais au lieu de charger la partition induite par les 4 catégories de la variable 5 (4 espèces d'iris), nous allons charger une partition en trois classes produite par l'algorithme de classification

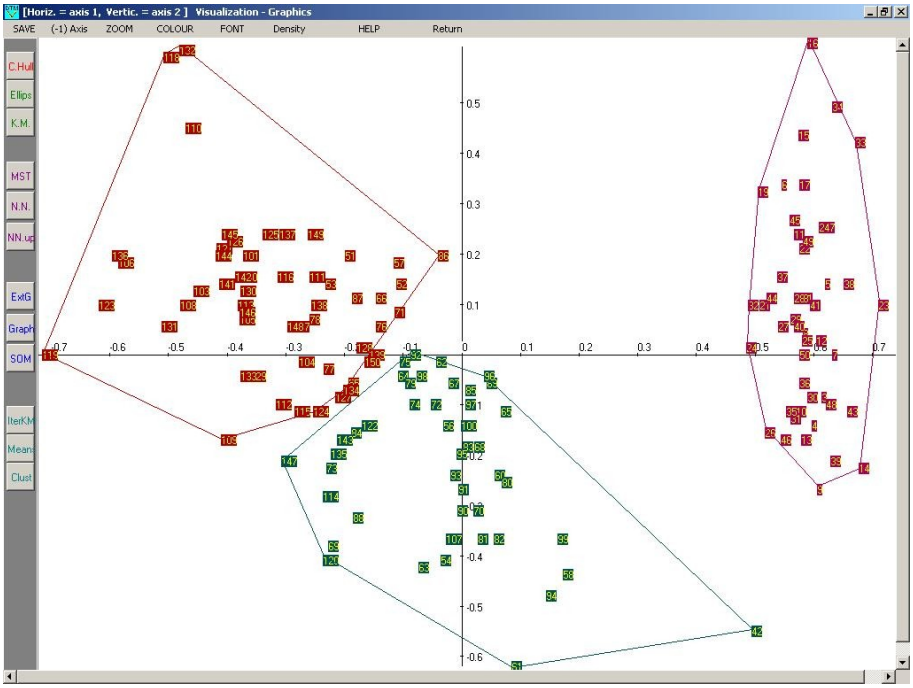
contenu dans les étapes de base : cette partition correspond aux étapes RECIP et PARTI (voir le fichier de commande). Elle ne suppose pas connue la division en espèces, d'où la dénomination de partition non-supervisée.

➢ Cliquez sur V Visualization

⊙ La fenêtre intitulée "DTM-visualization..." apparaît.

➢ Cliquez sur Load coordinates. Dans le sous-menu correspondant, choisir le fichier: ngus_ind.txt. Les principales coordonnées des individus (lignes) sont sélectionnées.

Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, choisissez alors Load partition File et ouvrir le fichier part_cla_ind.txt (partition en 3 classes issue des phases RECIP et PARTI).



Même plan principal que la figure précédente. **Attention !** Les couleurs différencient les classes (issues de l'algorithme de classification non supervisée) et non plus les espèces. La classification non supervisée en trois classes ne réussit à isoler que la classe de droite. Les deux autres espèces sont mélangées au sein des deux classes restantes.

Comme on le soupçonnait, la partition obtenue directement à partir des

mesures numériques, en ignorant l'espèce, n'est pas en mesure de séparer les trois espèces. Seule l'espèce "*setosa*", bien séparée des deux autres espèces, coïncide avec une des classes (*cluster*) de la partition.

Retour vers : VIC *: Visualization, Inference, Classification steps*

VI.2.5. Analyse de contiguïté

Deux analyses de contiguïté vont être exécutées. La première, non supervisée, utilise le graphe des plus proches voisins. C'est l'analyse de contiguïté intrinsèque. La seconde, supervisée, utilise le graphe formé de trois cliques disjointes correspondant aux trois espèces d'iris (tous les couples d'individus appartenant à une même espèce sont voisins, deux couples appartenant à deux espèces différentes ne sont jamais voisins). Dans ce cas pour lequel l'appartenance à une espèce est connue *a priori*, l'analyse de contiguïté coïncide avec l'analyse discriminante linéaire.

a. Graphes des plus proches voisins

Nous allons effectuer une analyse de contiguïté utilisant un "graphe des plus proches voisins" provenant des mesures. La partition en trois espèces n'est pas prise en compte. Il s'agit donc d'une approche non-supervisée.

- Cliquez sur le bouton :  Contiguity.
- Cliquez sur Parameter/Edit. Choisissez l'élément Create

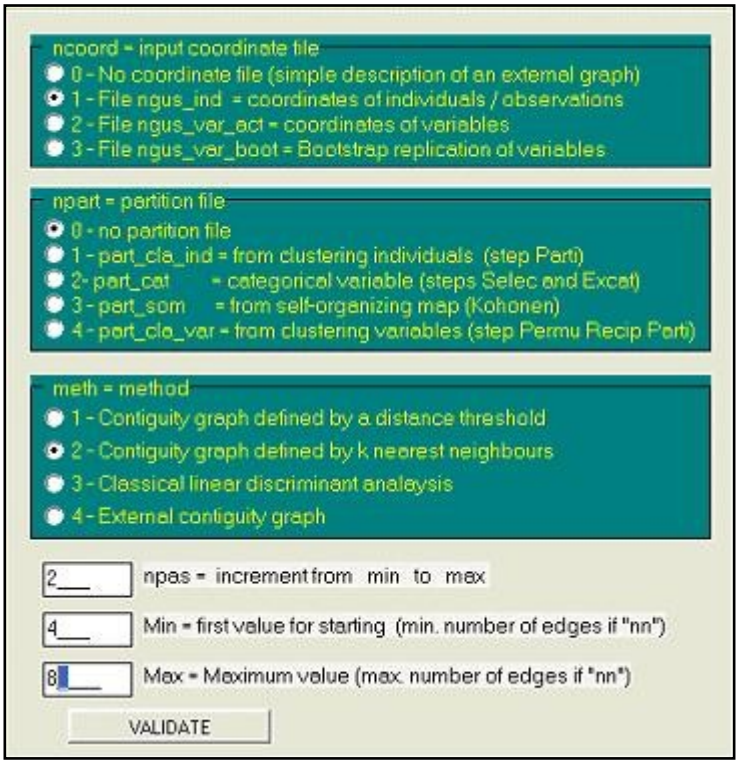
La fenêtre suivante apparaît : *(page suivante*)

Nous allons établir les paramètres nécessaires à une analyse de contiguïté:

- Dans le premier bloc intitulé "ncoord = Input coordinate file", cochez "1" (*File ngus_ind: coordinates of individuals/observations*). L'analyse de contiguïté utilisera les coordonnées des individus ou observations comme données d'entrée.

- Dans le deuxième bloc intitulé "npart = partition file" cochez "0" *(no partition)*

- Dans le troisième bloc intitulé "meth = method" cochez " 2 "*(Contiguity graph defined by k nearest neighbours).*



- Ensuite, nous aurons à entrer les valeurs numériques suivantes :
- npas = 2 (incrémentation du nombre de plus proches voisins)
- Min = 4 (nombre minimal de plus proches voisins)
- Max = 8 (nombre maximum de plus proches voisins)

Trois analyses de contiguïté seront alors effectuées pour les trois graphes correspondant respectivement à 4, 6, 8 plus proches voisins (de Min=4 jusqu'à Max=8, avec un incrément de npas =2).

➢ Cliquez sur VALIDATE.

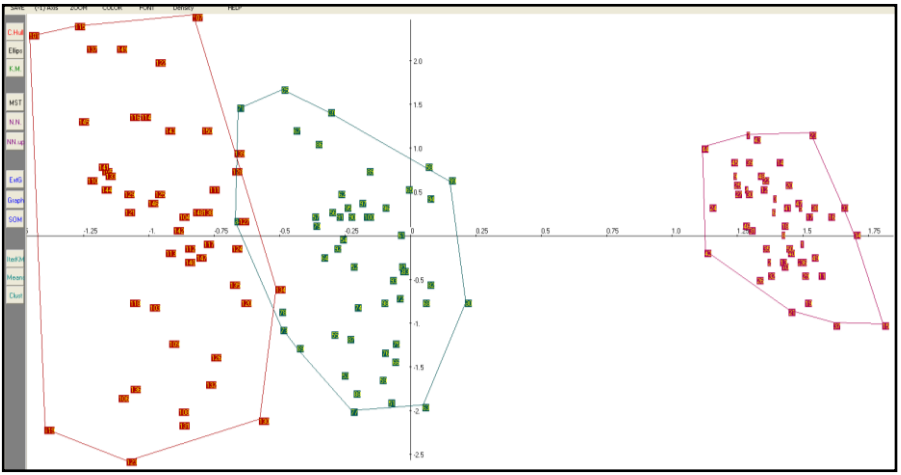
⊙ Un résumé des paramètres apparaît.

➢ Dans la barre supérieure de la fenêtre, cliquez sur Execute. Les calculs sont effectués.

- La rubrique Results permet de consulter les détails techniques des calculs impliqués dans l'analyse de contiguïté.
- Cliquez ensuite sur Contiguïty View.
  - La fenêtre "Visualization : loading files, selecting axes" qui correspondait au bouton V Visualization apparaît.
- Dans le menu Load coordinates de la nouvelle fenêtre, ouvrez le fichier ngus_contig.txt. Au lieu d'utiliser les coordonnées principales de l'ACP (ngus_ind.txt comme précédemment), nous utilisons maintenant le résultat de l'analyse de contiguïté : ngus_contig.txt.
- Cliquez ensuite sur Load or create a partition. Dans le sous-menu Load partition File, sélectionnez le fichier: part_cat.txt. (Avec ce fichier, nous allons identifier les espèces). Nous ne pouvons pas calculer l'arbre de longueur minimale ("minimum Spanning Tree"), ni les plus proches voisins à partir du fichier : ngus_contig.txt.
- Cliquez sur Graphics. Choisissez ensuite les axes 1 et 2 (qui sont d'ailleurs les valeurs par défaut)
- Choisissez (cochez) le numéro du niveau de contiguïté, par exemple 2, qui correspond à 6 plus proches voisins. (Le niveau 1 correspond à 4 plus proches voisins, et le niveau 3 à 8 plus proches voisins).
- Cliquez sur DISPLAY. Changer les couleurs, si nécessaire.
- Cliquez sur : C.Hull. Les trois espèces sont maintenant mieux séparées.

Cela signifie que le graphe (symétrisé) des 6 plus proches voisins permet de calculer une matrice des covariances "locale" qui peut jouer le rôle d'une matrice des covariances "interne". Dans cet exemple, le plan principal d'une analyse de la contiguïté est similaire au plan principal d'une analyse linéaire discriminante de Fisher.

Nous devons garder à l'esprit que l'analyse de contiguïté n'utilise pas la connaissance *a priori* des espèces. C'est une méthode non supervisée, contrairement à l'analyse discriminante, qui, elle, tente de séparer au mieux les espèces.



L'analyse de contiguïté réussit à séparer assez correctement les trois variétés d'Iris. La matrice des covariances "locale" calculée à partir des plus proches voisins fournit ici l'estimation d'une matrice des covariances "interne". Les excellents résultats sont dûs au fait que les plus proches voisins sont calculés dans un espace ayant plus de 2 dimensions, et, pour cet exemple, au fait que les 3 classes sont assez bien séparées dans cet espace.

b. Analyse discriminante

Nous allons maintenant effectuer une "analyse de contiguïté" qui coïncide exactement avec une analyse discriminante linéaire classique.

L'Analyse discriminante linéaire en k classes est en effet un cas particulier de l'analyse de contiguïté. Dans un tel cas, le graphe impliqué dans l'analyse de contiguïté est fait de k cliques (graphes complets) correspondant aux k classes de l'analyse discriminante. Dans notre cas particulier, k = 3. Tous les couples d'observations appartenant à une même espèce sont reliés par une arête. Aucune arête ne relie deux observations appartenant à deux espèces différentes.

- Revenir au menu principal et cliquez sur Contiguity.
- Cliquez sur Parameter/Edit. Choisissez l'élément "Create".
- Cochez :
  - "1" (*File ngus_ind: coordinates of individuals/observations*) dans le premier bloc "ncoord = Input coordinate file"
  - "**2**" *(part_cat.txt, nominales)* dans le deuxième bloc "npart = partition file" (partition utilisée pour construire le graphe).

- "**3**" (Analyse Discriminante Classique) dans le troisième bloc "meth = method".
- Dans ce cas particulier d'analyse discriminante, les paramètres suivants n'ont pas de sens. Dtm-Vic vous demande de les ignorer *(Remettre à 0 les compteurs si nécessaire).*

L'analyse de contiguïté sera effectuée en utilisant le graphique associé à la partition en 3 espèces de fleurs. (Toutes les paires d'individus appartenant à la même espèce sont reliées par une arête; il y a aucune arête entre individus appartenant à des espèces différentes)

➢ Cliquez sur VALIDATE.

⊙ Un résumé des paramètres apparaît.

➢ Dans la barre supérieure de la fenêtre, cliquez sur Execute. Les calculs sont effectués.

La rubrique "Results" de cette barre supérieure contient des détails techniques sur les calculs impliqués dans l'analyse de contiguïté. La matrice associée au graphe avec ses trois blocs diagonaux de "1" et avec la valeur "0" est d'ailleurs visible dans cette présentation des résultats.

➢ Cliquez ensuite sur Contiguïty View.

⊙ La fenêtre "Visualization : loading files, selecting axes" correspondant au bouton V Visualization apparaît.

➢ Dans le menu Load coordinates de la nouvelle fenêtre, ouvrez le fichier ngus_contig.txt.

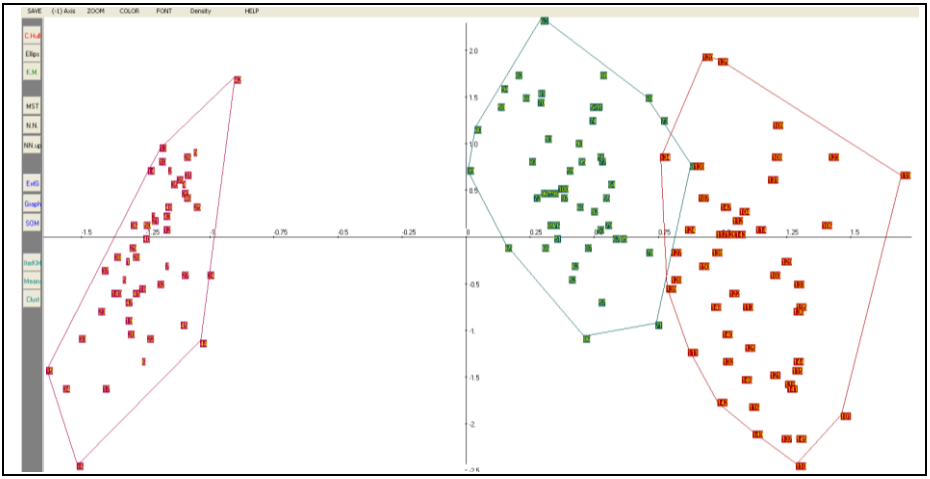
➢ Dans le menu Load or create a partition et dans le sous-menu Load partition File, choisissez le fichier: part_cat.txt (nous allons identifier les trois espèces d'iris)

Nous ne pouvons pas calculer l'arbre de longueur minimale, ni les plus proches voisins à partir du fichier de coordonnées issu de l'analyse de contiguïté: ngus_contig.txt, mais nous pourrions charger des résultats obtenus antérieurement à partir du fichier ngus_ind.txt, issu de l'analyse en composantes principales, résultats qui sont sauvegardés.

➢ Cliquez sur Graphics. Choisissez ensuite les axes 1 et 2 (valeurs par défaut)

- Cliquez sur DISPLAY. Changer les couleurs de l'écran si nécessaire pour obtenir un bon contraste entre les classes, puis verrouiller les couleurs.

- Cliquez sur : C.Hull. Les trois espèces sont encore bien séparées. Mais c'est moins une surprise, puisque l'analyse discriminante linéaire vise précisément à la séparation des classes. Nous sommes ici dans un cas "supervisé". La méthode utilise la connaissance a priori de l'espèce de l'iris pour construire de nouvelles coordonnées (fonctions discriminantes) qui induisent la meilleure séparation des classes.



Comme prévu pour ce jeu de données classique, l'analyse discriminante permet une bonne séparation des classes. Elle utilise la connaissance *a priori* des classes pour les séparer.

VI.3 Description de graphes

Contrairement aux répertoires des exemples précédents, le répertoire **EX_C03.Graphs** contient plusieurs sous-répertoires et plusieurs exemples.
Ces exemples visent à décrire quelques graphes planaires symétriques simples à partir de leurs matrices associées, principalement par analyse des correspondances.

VI.3.1 Vue d'ensemble des dossiers et fichiers

Les fichiers relatifs aux exemples de graphes sont situés dans le dossier : **DtmVic-Examples/DtmVic-Examples_C_NumData/EX_C03.Graphs.**
Ce dossier se compose de trois sous-répertoires :

- **Chessboard** (damier ou échiquier) se rapporte à la description d'un graphe "en forme de damier" (49 sommets correspondant à un damier carré avec 7 lignes et 7 colonnes, la matrice associée est une matrice binaire 49 x 49).
- **Cycle** concerne la description analogue d'un *cycle* (49 sommets).
- **Geography** concerne la description de graphes associés aux cartes géographiques (graphe de régions contiguës du Japon enregistré sous forme textuelle et externe, graphe des départements contigus de France, enregistré également sous forme textuelle et externe).

a. Le dossier Chessboard

La description d'un graphe sous forme de damier peut être obtenue à partir de plusieurs fichiers de données et dictionnaires différents :

a1 - Un fichier de données numériques : Chessboard_numerical
Dans le sous-répertoire **Chessboard**, ouvrir le sous-sous-répertoire **Chessboard_numerical**. Y figurent les fichiers de données, dictionnaire et paramètres (format numérique classique de Dtm-Vic).

- Le fichier de données : Chessboard_7x7_dat.txt contient la matrice d'incidence du graphe, avec 49 lignes et 49 colonnes. Comme toutes

les données classiques dans le format interne de DtmVic, chaque ligne commence par son identifiant. La cellule m (i, j) d'une telle matrice M vaut 1 si i et j sont des sommets reliés par une arête, 0 sinon.

- Les identificateurs de colonnes se trouvent dans le fichier-dictionnaire associé: Chessboard_7x7_dic.txt.
- Ces fichiers seront analysés par l'analyse des correspondances (fichier de commande: Chessboard_CA.Param.txt) puis par l'analyse en composantes principales (fichier de commande: Chessboard_PCA.Param.txt) afin de procéder à une comparaison. La comparaison n'est pas favorable à l'analyse en composantes principales dans ce cas particulier[11].

a.2 - Un fichier de données "externes" : Chessboard_Extern-7x7.txt

Toujours dans le répertoire **Chessboard_numerical**, le fichier: Chessboard_Extern_7x7.txt est un autre codage possible du graphe *Chessboard*, qualifié d'externe car il est différent du format interne général de Dtm-Vic. Il donne, pour chaque sommet (ligne), les numéros des sommets contigus. La première ligne contient le nombre de sommets (49), puis la longueur des identificateurs (4) et le degré maximum du graphe (borne supérieure du nombre d'arêtes adjacentes à un seul sommet) (10). Notez que chaque ligne de nombres se termine avec la valeur conventionnelle 0, indicateur de fin de ligne pour ce format.
Ce format spécifique, très compact, peut conduire directement à une description du graphe dans le sous-menu "contiguïté" de DtmVic.

a.3 - Un fichier de données textuelles : Chessboard_textual_7x7.txt

Le fichier Chessboard_textual_7x7.txt, dans le sous-sous-répertoire **Chessboard_textual**, contient les mêmes informations de base sous une forme tout à fait distincte : le format est celui des réponses à une question ouverte. Chaque sommet du graphe est considéré comme une personne interrogée répondant à la question ouverte fictive : "Quels sont vos

[11] Voir, par exemple: Exploring Textual Data (1998), par L. Lebart, A. Salem, L. Berry, Kluwer Academic Publisher. Cette comparaison avait déjà été faite dans l'article : "Introduction à l'analyse des données", (L. Lebart) *Consommation*, n°4, 1969, p. 65-87, Dunod.

voisins ?". Au lieu d'une matrice binaire M, nous avons affaire ici à un tableau beaucoup plus petit contenant l'adresse (numéro de colonne) des "1" dans la matrice M. Les commandes de Chessboard_Textual.Param.txt conduisent aux mêmes résultats que l'analyse des correspondances de l'alinéa précédent, en utilisant toutefois une séquence d'étapes bien distinctes de Dtm-Vic. C'est un "exemple pédagogique" de pont entre les mesures numériques et textuelles du DtmVic. Attention ! Avec ce type de données, les chiffres ne sont pas considérés comme des nombres au sens mathématique du terme, mais comme de simples séquences de caractères. [Voir ci-dessous l'exemple des cartes du Japon et de France, où les numéros des sommets sont remplacés par les noms des régions et des départements en clair]. Ce dossier contient également le même fichier Chessboard_Extern-7x7.txt que le dossier précédent.

b. Le dossier "Cycle"

Ce sous-répertoire **Cycle** est voisin de celui relatif au graphe *Chessboard*. On y trouve de la même façon que pour le dossier *Chessboard*, un codage numérique et externe. Seule la forme du graphique est différente. Le codage textuel et le fichier de commandes de l'Analyse en composantes principales ont été omis dans ce cas.

c. Le dossier Geography

Les deux sous-répertoires du répertoire **Geography** sont les homologues de l'exemple textuel du dossier **Chessboard**. Les répertoires **Japan_map** et **France_map** illustrent le "codage textuel" dans le cas des graphes décrivant les différentes régions du Japon et des départements de France.

Dans le cas du Japon, par exemple, les deux premières lignes du fichier Japan_map_textual.tex.txt indiquent que les provinces d'*Akita* et d'*Iwate* sont contigües à la province d'*Aomori*, etc. Le fichier de commande correspondant est le fichier Japan_map_textual_Param.txt. Il est similaire au fichier Chessboard_Textual.Param.txt.

Dans le cas de la France, par exemple, les deux premières lignes du fichier France_Text.txt indiquent que le département de l'*Ain* est contigu aux départements *Isère, Jura, Rhône, Hte_Saône, Savoie, Hte_Savoie*. Le fichier France_Param.txt est le fichier de commande correspondant.

Le fichier France_extern.txt représente la carte de France dans le format externe défini dans la section **a.2** ci dessus. Il permettra de tracer le graphe initial dans les plans factoriels.

VI.3.2 Exécution de l'exemple "Chessboard_numerical"

(Répertoire **Chessboard_numerical** dans **EX_C03.Graphs/Chessboard**).

Dans ce dossier, figurent les fichiers de base :

a) Fichier de données: Chessboard_7x7_dat.txt
b) Fichier Dictionnaire: Chessboard_7x7_dic.txt.
c) Fichiers de commandes: Chessboard_CA.Param.txt [Analyse des Correspondances] et Chessboard_PCA.Param.txt [analyse en composantes principales]

Il est possible de réaliser soit une analyse des correspondances classique ou une analyse en composantes principales.

a. Ouverture et Exécution du fichier paramètre de l'AC

Nous commencerons par exécuter l'analyse des correspondances.

- Cliquez sur le bouton : **Open an existing command file** de Command File (menu principal). Puis recherchez le dossier **Chessboard_numerical** dans **DtmVic-examples /DtmVic-Examples_C_NumData**, puis le fichier de commande Chessboard_CA.Param.txt

> Notez encore que ces "fichiers de commande" peuvent être facilement générés en cliquant sur le bouton "Create" du menu principal (Basic Steps). Une fenêtre "Select a basic analysis" apparaît. Cliquez ensuite sur le bouton: SCA - Simple Correspondence Analysis ou sur le bouton : PCA – Principal Components Analysis – les deux situés dans la rubrique "Numerical Data", et suivez les instructions comme indiqué dans le chapitre II.

Après avoir identifié et vérifié les fichiers de données et du dictionnaire, trois étapes vont être effectuées: ARDAT (Archivage des données), SELEC (sélection des éléments actifs et supplémentaires), AFCOR (analyse des correspondances).

- Cliquez sur Return to execute dans le bandeau pour revenir au menu principal.

- Cliquer sur le bouton : Execute de Command File

➢ Cliquer sur le bouton : Basic numerical results de Result Files

Le bouton ouvre le fichier HTML nommé "imp.html" qui contient les principaux résultats des étapes précédentes de calcul de base. Après lecture de ces résultats numériques, retournez au menu principal.

b. Visualisation et lecture des résultats

Nous allons maintenant visualiser directement le graphique dans l'étape VIC : *Visualization, Inference, Classification steps*.

➢ Cliquez sur V Visualization (on n'utilisera pas ici les boutons "AxeView", "PlaneView", etc.)

⊙ Une fenêtre intitulée "DTM-visualization: loading files, selecting axes" apparaît.

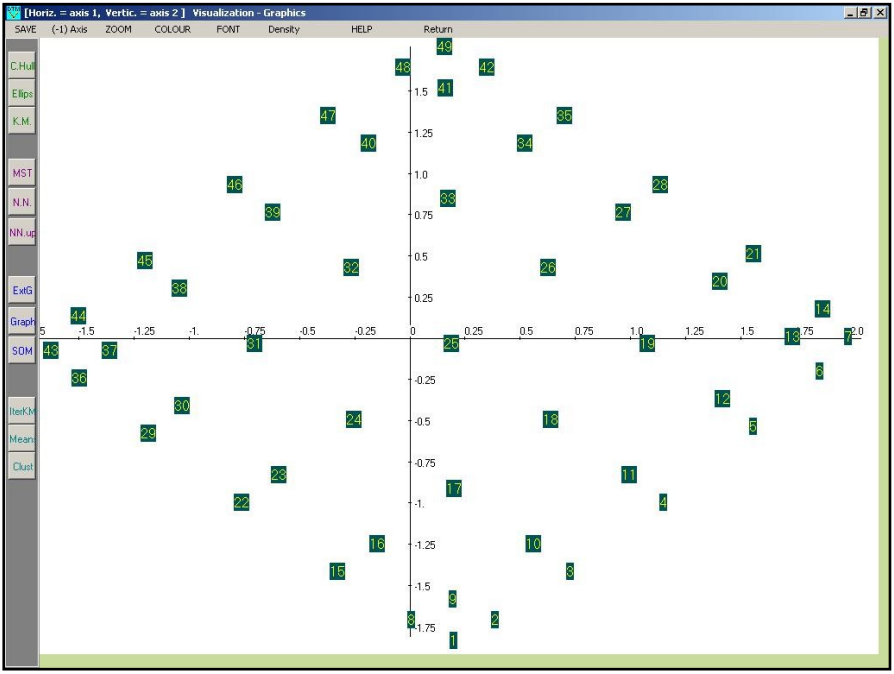


FIgure VI.1 Plan factoriel principal (Analyse des correspondances) pour le graphe "Damier" (après changement de police (bouton "Font") et changement de couleur (bouton "Colour").

➢ Cliquez sur Load coordinates. Dans le sous-menu correspondant,

choisir le fichier: ngus_ind.txt (individus ou observations). Les principales coordonnées des individus (lignes) sont sélectionnées. [En fait, ici, la matrice de données est symétrique, il est équivalent, dans ce cas très particulier, de choisir ngus_var_act. txt].

- Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, sélectionnez No partition.

- Cliquez sur Graphics puis choisissez les axes 1 et 2 (par défaut) dans la petite fenêtre "Sélection des axes" et cliquez sur "Continue" puis sur DISPLAY.
  - ⊙ Dans une nouvelle fenêtre intitulée "Vizualisation - Graphics", le plan factoriel principal s'affiche (voir figure VI.1 précédente).

***

Dans la barre d'outils verticale de la fenêtre "Graphics", le bouton ExtG va nous permettre de tracer le graphe initial à partir du codage externe.

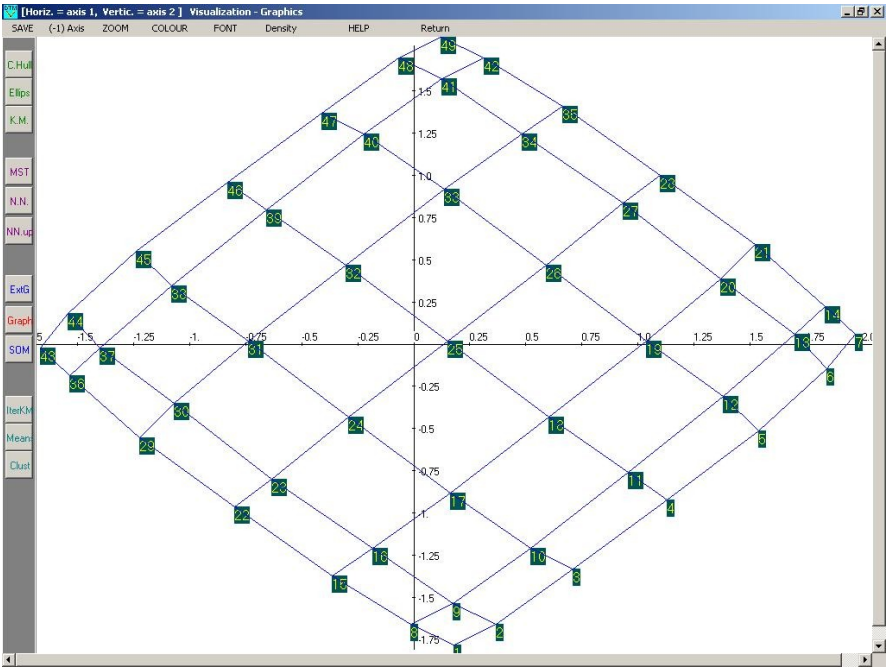


Figure VI.2. Même plan factoriel principal pour le graphe "Damier" avec tracé du graphe initial (après changement de police (bouton "Font") et de couleur (bouton "Colour").

- Pour représenter les arêtes du graphe d'origine, cliquez sur le bouton ExtG (graphe externe) de la barre verticale.
- Ouvrez le fichier Chessboard_Extern_7x7.txt.
- Cliquez sur le bouton Graph.

On obtient alors une représentation du graphe original avec une représentation des arêtes originales (Figure VI.2).
Cette représentation permet aussi d'observer les déformations du graphe planaire dans les espaces engendrés par les paires d'axes de rangs 3 à 12. On observe un effet Guttman multidimensionnel[12].

- Retournez au menu principal en quittant la fenêtre du plan factoriel, puis en cliquant sur Return puis quittez Dtm-Vic.

c. Ouverture et Exécution du fichier paramètre de l'ACP

Reprendre les opérations des sections a et b en ouvrant cette fois-ci le fichier de commande: Chessboard_PCA.Param.txt (PCA : analyse en composantes principales). Répétez toutes les opérations précédentes.
On voit à travers le graphique produit par cet exemple que l'Analyse en Composantes Principales décrit de façon moins fidèle la structure du graphe que l'Analyse des Correspondances (Figure VI.3)**.**

VI.3.3 Exécution de l'exemple "Chessboard_textual"

Cette section concerne l'exécution de l'exemple **Chessboard_textual** du répertoire **DtmVic-Examples_C_NumData/EX_C03.Graphs/Chessboard** et la lecture des résultats.

Nous sommes dans le cadre d'une analyse textuelle similaire à celui de l'exemple qui vise à décrire les réponses à une question ouverte dans une enquête par sondage (Exemple III.2 du chapitre III).

[12] [Voir Benzécri, (1973) «L'analyse des données", Tome II B, chapitre 10, "Sur l'analyse de la correspondance définie par un graphe", pp 244 - 261]

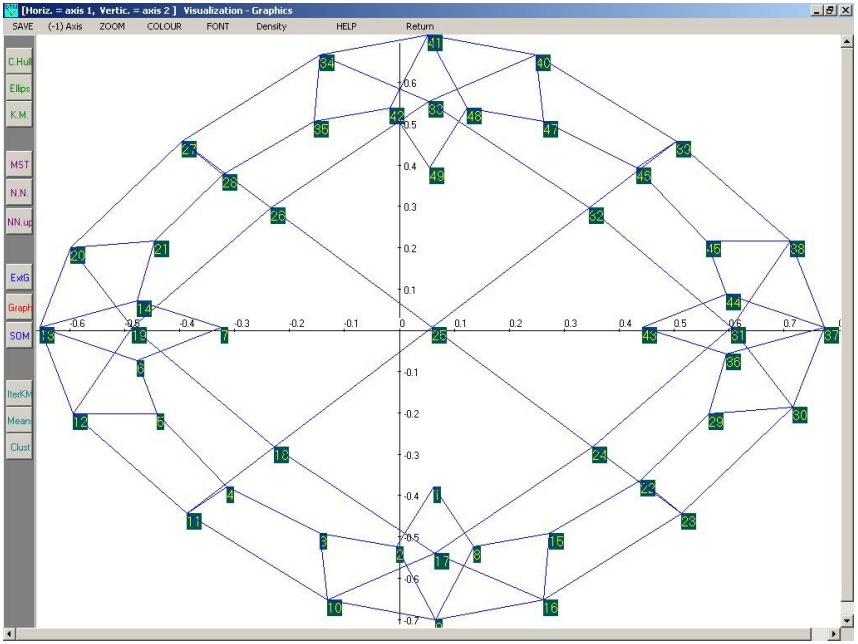


Figure VI.3 Cas de l'analyse en composantes principales. Plan factoriel principal pour le graphe "Damier" avec tracé du graphe initial (après changement de police (bouton "Font") et changement de couleur (bouton "Colour"). Le traitement dissymétrique des lignes et des colonnes et la normalisation opérée par l'ACP ne permettent pas d'obtenir une description satisfaisante de ce type de graphes

On trouve dans ce répertoire le "fichier texte" et le "fichier de commandes". (Dans ce contexte particulier, il n'y a ni fichiers de données ni fichier-dictionnaire : le questionnaire comprend une "pseudo question ouverte", posée à chaque sommet: "Quels sont vos sommets voisins?").

1. Fichier texte: Chessboard_textual_7x7.txt

Le format est le même que celui décrit au paragraphe I.5 (Chapitre 1, §5, tableau 4, dans le cas d'une seule question ouverte). Étant donné que les réponses peuvent avoir des longueurs très différentes, les séparateurs sont utilisés pour distinguer les individus (ou: les personnes interrogées). Les individus (ici: les nœuds) sont séparés par la chaîne de caractères "----" (à partir de la colonne 1) éventuellement suivi d'un identificateur. Attention, les 49 numéros de sommets sont ici considérés comme des mots, ils pourraient être remplacés par 40 noms distincts avec les mêmes

calculs et le même résultat final pour le tracé du graphe.

2. Fichier de commandes: Chessboard_Textual.Param.txt

La phase de calcul de l'analyse est décomposée en "étapes". Chaque étape nécessite quelques paramètres décrits brièvement dans le menu principal de DtmVic (bouton: "Help about parameters").

a. Ouverture et Exécution du fichier decommande

➢ Cliquez sur le bouton : **Open an existing command file** de Command File (menu principal) et ouvrez le fichier paramètre Chessboard_Textual.Par.txt

Quatre étapes sont effectuées:
ARTEX (textes d'archivage), SELOX (sélection de la question ouverte), NUMER (codage numérique du texte), ASPAR (analyse des correspondances du tableau de contingence ["répondants x mots]").

Notez que ce fichier de commande peut également être généré en cliquant sur le bouton "Create" de la rubrique "Command file" du menu principal ("Basic Steps"). Une fenêtre "Select a Basic Analysis" apparaît. Cliquez ensuite sur le bouton : VISURESP, situé dans la rubrique "Textual Data", et suivez les instructions comme indiqué dans les chapitres II et III.
Notez également que dans ce cas de données simples (une seule "question ouverte"), il est possible de considérer chaque réponse comme un texte. Dans un tel cas, le séparateur "----" doit être remplacé par le séparateur "****", comme dans l'exemple III.1 du chapitre III. Au lieu de l'analyse "VISURESP" (Visualization of responses), il est alors nécessaire d'effectuer l'analyse "VISUTEX" (Visualization of texts).

➢ Cliquez sur Return to execute dans le bandeau pour revenir au menu principal.

➢ Cliquer sur le bouton : Execute de Command File

Cette phase exécute les étapes de calcul présentes dans le fichier de commande : Numérisation du "texte" et analyse des correspondances du tableau lexical.

➢ Cliquer sur le bouton : Basic numerical results de Result Files

Le bouton ouvre le fichier HTML nommé "imp.html" qui contient les principaux résultats des étapes précédentes de calcul de base.

L'étape NUMER, nous apprend, par exemple, que nous avons 49 "réponses", avec un nombre total de mots (occurrences = ici: arêtes du graphe) de 217, impliquant 49 mots distincts (ici: les sommets voisins sur le damier). Notez que chaque sommet a aussi été considéré comme son propre voisin.

Après lecture de ces résultats numériques, retour au menu principal.

b. Visualisation et lecture des résultats

Nous allons maintenant visualiser les résultats avec les outils de l'étape VIC *: Visualization, Inference, Classification steps*.

Pour tracer le graphe : Cliquez sur Visualization

Toutes les étapes de la section précédente peuvent être réalisées de la même façon. Les graphiques obtenus sont identiques à ceux de la section VI.3.2.b. Il n'y a pas lieu de les reproduire.

VI.3.4 Exécution directe de l'exemple "Chessboard_Extern"

Il n'y a ni fichier de commandes, ni fichier de dictionnaire pour ce type d'analyse utilisant directement le format "Externe". Pour ce type de codage du graphe ("codage externe de graphe"), il est prévu une entrée directe dans le menu "Contiguity".

- Cliquez sur Contiguity dans l'étape VIC *: Visualization, Inference, Classification steps*
- Cliquez sur Parameter/Edit. Choisissez l'élément "Create"

  Nous allons établir les paramètres nécessaires à une description graphique:

  - Dans le premier bloc intitulé "*ncoord = Input coordinate file*", cochez "**0**" (*File ngus_ind: coordinates of individuals/observations*). Aucun fichier de coordonnées (simple description d'un graphe externe).

  - Dans le deuxième bloc intitulé "*npart = partition file*" cochez "**0**" (*no partition*)

  - Dans le troisième bloc intitulé "*meth = method*", cochez "4"(graphe de contiguïté externe).

- Cliquez sur VALIDATE.
- Dans la barre supérieure de la fenêtre, cliquez sur Execute
  - Une nouvelle fenêtre apparaît, et vous êtes invités à choisir le fichier du graphe externe Chessboard_Extern_7x7.txt du

répertoire **EX_C04.Graphs/ Chessboard/ Chessboard-Extern**.

⊙ Une autre fenêtre "Reading an external graph" apparait.

➢ Cliquez sur CONTINUE

⊙ Une série de fenêtres apparaissent indiquant les détails techniques des calculs impliqués dans l'analyse des correspondances de la matrice M associée au graphe (Ces résultats sont enregistrés dans le fichier imp_contig.txt, sauvegardé dans le répertoire de travail).

➢ Cliquez sur V Visualization

⊙ La fenêtre intitulée "DTM-visualization..." apparaît.

➢ Cliquez sur Load coordinates. Dans le sous-menu correspondant, choisir le fichier: anagraf.txt , qui contient les coordonnées factorielles pour les analyses directes de graphes.

➢ Cliquez ensuite sur Load or create a partition. Dans le sous-menu correspondant, sélectionnez No partition. Puis procédez comme pour l'exemple *Chessboard*.

➢ 3- Cliquez sur Graphics puis choisissez les axes 1 et 2 (par défaut) dans la fenêtre "Sélection des axes" et cliquez sur Continue puis sur DISPLAY.

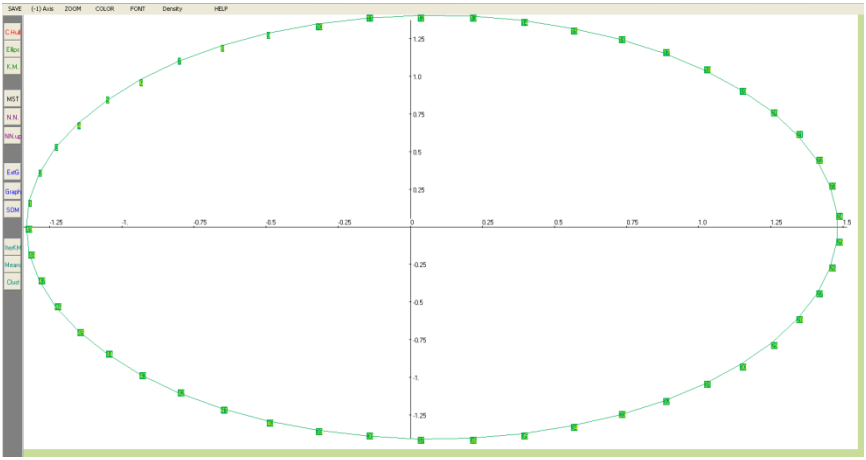
⊙ Dans une nouvelle fenêtre intitulée "Vizualisation - Graphics", le plan factoriel principal s'affiche

Une fois de plus, toutes les étapes de la section précédente pourront être réalisées. Les graphiques obtenus sont encore identiques à ceux de la section VI.3.2.b. Ils ne sont donc pas reproduits.

VI.3.5 Exécution des exemples "Cycle"

Cette section est en tout point identique à la section VI.3.2 (exécution de l'exemple "Chessboard_Numerical") et VI.3.4. Le graphique a la forme d'un cycle, avec le même nombre de sommets.
Les fichiers homologues des fichiers Chessboard_7x7_dat.txt Chessboard_7x7_dic.txt, Chessboard_Extern_7x7.txt et Chessboard_ CA_Param.txt sont maintenant respectivement Cycle_49_dat.txt,

Cycle_49_dic.txt, Cycle_Extern_49.txt et Cycle_CA_Param.txt. Ils peuvent être trouvés dans le répertoire **Cycle**.



Plan factoriel principal pour le graphe "Cycle" avec tracé du graphe initial (après changement de police (bouton "Font") et changement de couleur (bouton "Colour").

VI.3.6 Exécution de l'exemple "France_map"

(Dossier : **Geography**)

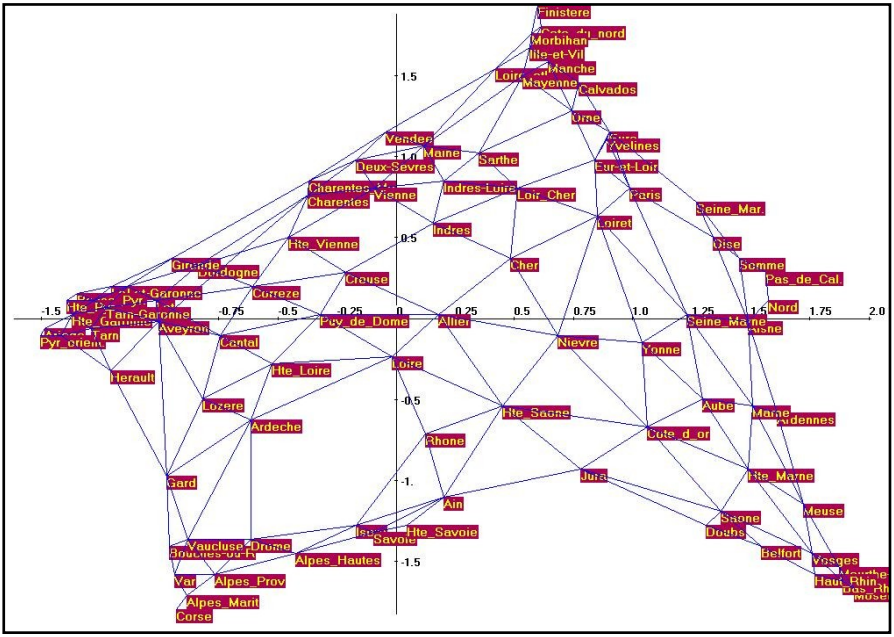
Cette section est identique à la section VI.3.3 (Exécution de l'exemple *Chessboard_Textual*). Le graphique est maintenant une schématisation d'une carte de France, présentée comme une suite de réponses à la question ouverte : "Quelles sont vos départements voisins ?", les "personnes interrogées" étant les départements français.

```
****    Ain
     Ain  Isere  Jura
     Rhone  Hte_Saone  Savoie  Hte_Savoie
****    Aisne
     Aisne  Ardennes  Marne
     Nord  Oise  Seine_Marne  Somme
****    Allier
     Allier  Cher  Creuse
     Loire  Nievre  Puy_de_Dome  Hte_Saone
```

Extrait du fichier de données textuelles : France_Text.txt (trois premiers départements)

L'homologue du dossier **Chessboard_Textual** est : **France_map**, tandis que les homologues des trois fichiers Chessboard_textual_7x7.txt,

Chessboard_Extern_7x7.txt et Chessboard_textual_Param.txt sont respectivement les trois fichiers : France_Text.txt, France_extern.txt et France_Param.txt.



Plan factoriel principal pour le graphe "France" avec tracé du graphe initial (après changement de police (bouton "Font") et changement de couleur (bouton "Colour"). Le signe des axes (arbitraire) peut être changé, pour retrouver l'orientation initiale.

VI.3.7 Exécution de l'exemple "Japan_map"

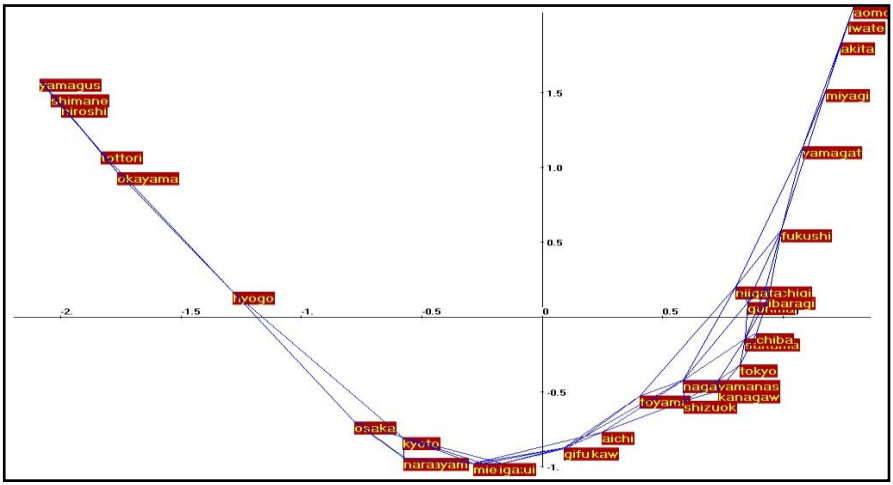
(Dossier : **Geography**)

Cette section est identique à la précédente, ainsi qu'à la section VI.3.3 (Exécution de l'exemple "Chessboard_Textual"). Le graphique est maintenant une esquisse d'une carte du Japon, codée comme les réponses à la question ouverte "Quelles sont vos régions voisines", les "répondants (fictifs)" étant les mêmes régions du Japon. Le dossier **Japan_map** contient les trois fichiers homologues des précédents (texte, externe et paramètre) : Japan_map_Textual.tex.txt, Japan_map_Extern.txt et Japan_map_Textual.Param.txt.

```
---- aomori
akita iwate
---- akita
aomori iwate yamagata miyagi
---- iwate
aomori akita miyagi
---- yamagata
akita miyagi niigata fukushima
```

Extrait du fichier de données textuelles : Japan_map_Textual.tex.txt (trois premières régions). Ici, les régions sont considérées comme des individus (séparateur ----) alors que les départements ont été considérés comme des textes (séparateur ****). Les deux codages sont possibles dans cette configuration simple.

La même séquence d'opération conduit au graphique suivant, dont la forme parabolique est en partie imputable à la forme de l'archipel, mais aussi à un effet Guttman marqué, déjà évoqué en section VI.3.2.b, à propos des axes 3 et suivants, et accentué ici par une différence d'échelle entre les axes . Cet effet Guttman dès le second axe apparaît évidemment pour les graphes en forme de chaînes ou de tresses (premier axe dominant, les axes suivants étant des fonctions polynomiales du premier).



Plan factoriel principal pour le graphe "Japon" avec tracé du graphe initial (après changement de police (bouton "Font") et changement de couleur (bouton "Colour"). Le signe des axes est arbitraire. Il peut aussi être changé, pour retrouver l'orientation géographique initiale.

VI.4. Reconstitution d'images

(parenthèse méthodologique)

Les exemples cette section VI.4 sont principalement des exemples pédagogiques qui servent à illustrer les propriétés de compression des analyses en axes principaux (en gardant un nombre limité d'axes principaux provenant d'une décomposition aux valeurs singulières ou d'une analyse des correspondances) dans le domaine de l'analyse d'images (domaine peu familier pour certains utilisateurs actuels de Dtm-Vic). Une comparaison est faite avec les séries de Fourier discrètes (en gardant un nombre limité de termes de l'expansion) qui, elles, prennent en compte les positions relatives des pixels.

VI.4.1 Format des fichiers image

Ce type de traitement ne fait pas usage des données en format-texte interne Dtm-Vic, car il traite d'images numérisées. Un simple tableau rectangulaire de nombres entiers suffit: il n'est pas nécessaire d'avoir des identificateurs de lignes ou colonnes (dictionnaire).

En fait, trois formats particuliers seront utilisés : tableaux rectangulaires de niveaux de gris (format texte simple : "txt"), format "pgm" (acronyme de "Portable Gray Map" ou "Portable Grey Map" en Anglais britannique) et pour les images couleur, format "ppm" (acronyme de "Portable Pixel Map").

On trouvera les fichiers d'exemple dans le dossier EX_C05.Images du dossier **DtmVic_Examples_C_NumData**. Dans ce répertoire, ouvrez le répertoire (dossier) de l'exemple C.5: EX_C05. Images. Quatre sous-répertoires correspondent aux quatre exemples:

- "1_Cheetah_txt",
- "2_Baalbeck_pgm",
- "3_Cardinal_ppm_color",
- "4_Extra_pgm_ppm" .

Tous les fichiers contenus dans ces sous-répertoires peuvent être

examinés avec un éditeur de texte (tel que "Notepad", inclus dans Windows, "UltraEdit", ou un logiciel libre tel que "Notepad + +" ou "TotalEdit", etc.). Pour les images en niveaux de gris, deux formats d'entrée sont disponibles:

1 - **Le format de texte simple**. [Voir l'exemple 1, c'est-à-dire l'image cheetah.txt[13] du dossier **1_cheetah.txt**]. Le tableau de données contient des entiers positifs inférieurs ou égaux à 255 qui sont les valeurs du niveau de gris pour chaque pixel (pas d'identificateur). Un tel format qui ne contient pas explicitement la taille de l'image est le plus simple. En raison de sa rusticité, il n'est ni utilisé ni fourni par les logiciels de traitement d'images usuels.

2 - **le format pgm**. ("Portable grey map") (voir l'exemple 2, avec l'image Baalbeck.pgm du dossier **2_Baalbeck_pgm**, en utilisant un éditeur de texte ou un bloc-notes).
Le format pgm est un format simple et transparent en niveaux de gris. La première ligne contient l'identificateur de format: P2. Les deuxième et troisième lignes contiennent trois entiers: nombre de colonnes, nombre de lignes, et la valeur maximale (255). Ensuite, le tableau est affiché par ligne. Chaque pixel de la table est représenté comme un nombre décimal décrivant le niveau de gris (<255). Chaque pixel de la table a au moins un espace blanc avant et après. Aucune ligne ne dépasse 72 caractères[14].

3 - **le format ppm**. Pour les (petites) images couleur, le format d'entrée est le format texte ppm ("portable pixel map"). Consultez l'exemple 3 Cardinal.ppm, via un éditeur de texte ou un bloc-notes (dossier **3_Cardinal_ppm**). Ce format est assez voisin de pgm, mais avec trois entiers (3 niveaux de RGB : Red, Green, Blue) sur une même ligne par pixel. Ce format est également celui de l'exemple 4.

Les fichiers pgm et ppm peuvent être obtenus par une exportation à partir du logiciel libre "Open Office" (préciser pgm, format texte), en utilisant un fichier JPEG en entrée. [Attention, pour ce module, limitation à 1000 pour le nombre de pixels en ligne ou en colonne].

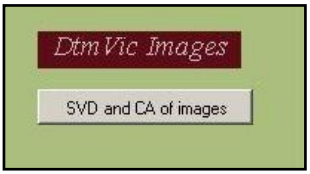
---

[13] Image adaptée du livre " *La compression de données*", Mark Nelson, M & T Publishing Inc, 1992.

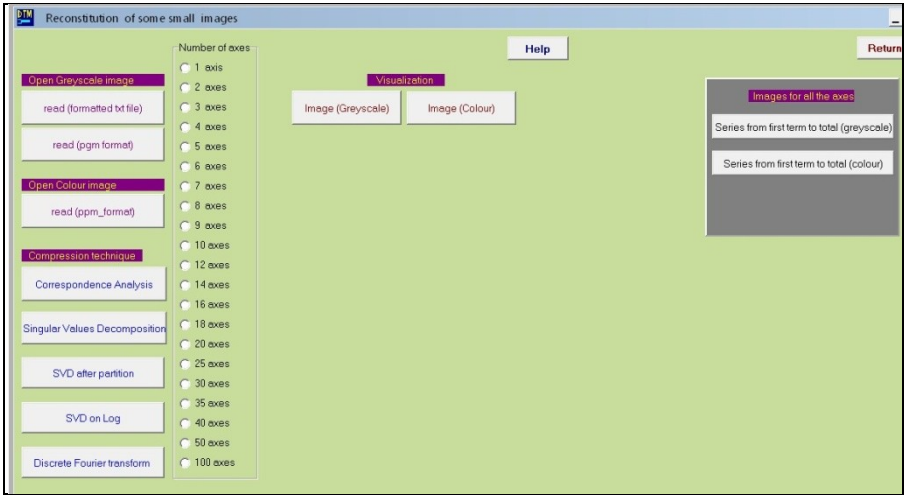
[14] Pour plus d'informations sur un tel format, veuillez consulter (par exemple): http://netpbm.sourceforge.net/doc/pgm.html.

VI.4.2 Analyse pour la compression d'images

- Cliquez sur le bouton : SVD and CA of images, dans la rubrique "DtmVic Images" du menu principal.



  - ⊙ Une fenêtre apparaît, dont la partie supérieure est représentée ci-dessous.



Description de la fenêtre "*Reconstitution of some small images*"

Sur la gauche figurent en colonne trois boutons (rouge foncé) correspondant aux trois formats de fichiers images décrits au paragraphe précédent (format simple de niveaux de gris, format pgm de niveaux de gris, format ppm couleur). Puis, plus bas, cinq boutons (bleus) correspondant aux cinq méthodes de compressions choisies (Analyse des correspondances, SVD - Décomposition aux valeurs singulières, Analyse

après partition préalable de l'image[15], analyse logarithmique[16], Séries de Fourier discrètes). Pour les quatre premières méthodes, le nombre d'axes retenus (de 1 à 100) est à cocher dans la seconde colonne. Si le nombre d'axes retenu est 8, par exemple, ce sont les 8 premiers termes de la formule de reconstitution des données qui sont utilisés pour reconstituer l'image. Les deux boutons centraux déclenchent un affichage des images (gris ou couleur). Les deux boutons du panel gris sur la droite déclenchent un balayage automatique pour tous les axes proposés. Toutes les figures intermédiaires sont sauvegardées en format *Windows bitmap* (.bmp).

Avant d'examiner les exemples, schématisons la suite des opérations à faire dans le cas des analyses en axes principaux (méthodes factorielles) :

- Cliquez, selon l'extension du fichier image, sur un des boutons Read. (txt format, ou : pgm format, ou : ppm_format). Répondre OK aux boites de message number of columns et number of rows qui s'affichent.
- Sélectionner une des méthodes, par exemple l'analyse des correspondances Correspondence Analysis ou la décomposition aux valeurs singulières Singular Values Decomposition. Répondre OK lorsque s'affiche la boite de message End of computation.
- Sélectionner le nombre d'axes. Répondre OK dans la fenêtre number of axes.
- Cliquer sur un des boutons Image selon l'image choisie (noir et blanc ou couleur). En fait, le bouton "Help" permet d'obtenir les informations nécessaires (en Anglais). Les fichiers images crées (image originale, et images reconstituées à partir d'un nombre variable d'axes principaux) sont automatiquement sauvegardés en format ".bmp".

Le logiciel "Paint", du volet "Accessoire" des programmes sous Windows, (ou le logiciel gratuit "IrfanView" par exemple) permet de visualiser ces images et de les sauvegarder en format JPEG, plus économique en espace.

[15] Cette variante consiste à centrer préalablement les niveaux de gris à l'intérieur de p zones rectangulaires avant SVD, puis à ajouter les p moyennes après SVD. (on peut choisir p = 2 x 2, 3 x 3, 4 x 4, 5 x 5, etc.)

[16] Cette variante consiste à faire une transformation logarithmique préalable, puis à procéder à une SVD tu tableau doublement centré en ligne et en colonne.

➢ Cliquer sur Exit .

VI.4.3 Exécution d'un premier exemple

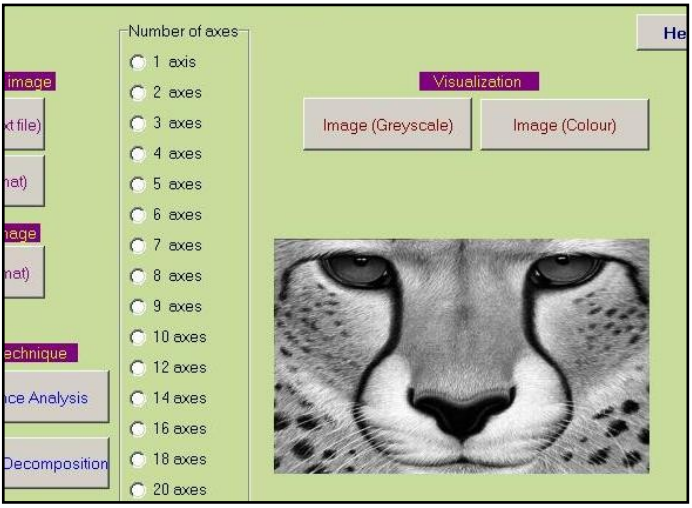
(**format de texte simple : Exemple :** *Tête de guépard :* **1_Cheetah_txt**)

➢ Cliquez sur le bouton : SVD and CA of images , dans la rubrique *DtmVic- Images* du menu principal.

➢ La fenêtre "Reconstitution of some small images" apparaît (cf. ci-dessus).

a. Cliquez sur le premier bouton Read (formatted txt file) dans la rubrique Open Greyscale image.
Dans le répertoire **EX_CO4_Image**, ouvrez le sous-répertoire **1_Cheetah_txt**. Dans ce répertoire, ouvrez le fichier Cheetah.txt. Une boite de message rappelle les dimensions du fichier image.

b. Si vous désirez visualiser l'image d'origine, dans la rubrique Visualization, cliquez sur: Image (Greyscale). L'image apparaît alors au centre de la fenêtre, comme indiqué ci-dessous.



Portion de fenêtre présentant l'image originale Cheetah.txt avant le choix du nombre d'axes.

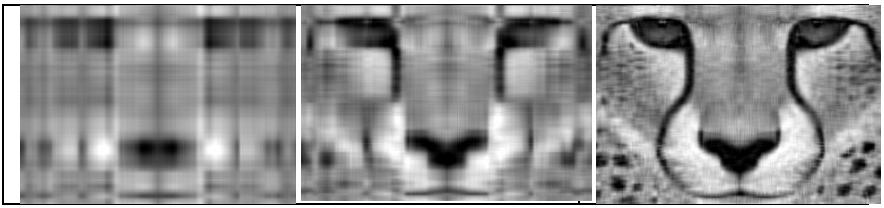
La rubrique "c" ci-après est consacrée aux méthodes factorielles de compression (axes principaux), puis la rubrique "d" qui suivra examinera à titre de comparaison la compression obtenue en ne retenant que les premiers termes des séries de Fourier entières. Il ne s'agit pas ici d'optimiser la compression, mais de comparer deux approches hiérarchiques simples (bases de vecteurs *versus* bases de fonctions trigonométriques).

c. Le cas des méthodes factorielles

Dans la partie inférieure gauche de la fenêtre, dans la rubrique : Compression technique , cliquez sur le bouton: Correspondence Analysis (pour commencer). L'analyse s'effectue.

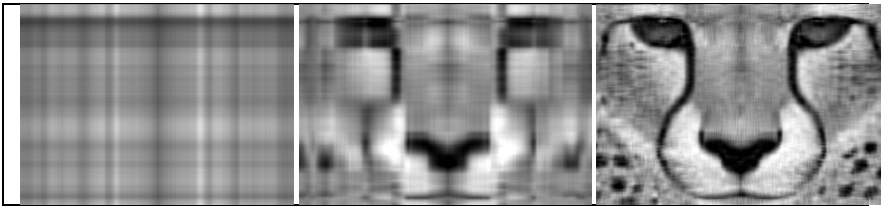
c1. Si vous souhaitez obtenir un aperçu de la reconstitution des données, de 1 à 100 axes, cliquez directement sur le bouton: Series from first term to total (greyscale) , dans le panel : Images for all the axes. On peut alors observer la reconstitution progressive de l'image.

c2. Si vous vous intéressez à un nombre d'axes particulier, sélectionnez le nombre requis dans la liste verticale correspondante, et visualisez chaque image avec le bouton utilisé en **b.**



Cas de l'analyse des correspondances : Images reconstituées successivement avec un axe principal , quatre axes et 16 axes. Dans ce cas, pour un seul axe, la formule de reconstitution contient deux termes : le terme correspondant à l'hypothèse d'indépendance (0 axe) et le premier axe.

c3. A la place de l'analyse des correspondances, vous pouvez choisir la méthode de "Singular Value Decomposition" (Décomposition aux Valeurs Singulières), et refaire les opérations **c1.** et **c2.**



Cas de la décomposition aux valeurs singulières: Images reconstituées successivement avec un axe principal , quatre axes et 16 axes. Dans ce cas, pour un axe, la formule de reconstitution ne contient qu'un seul terme, d'où un "retard" par rapport à l'analyse des correspondances, retard qui s'estompe au fil de l'accumulation des axes.

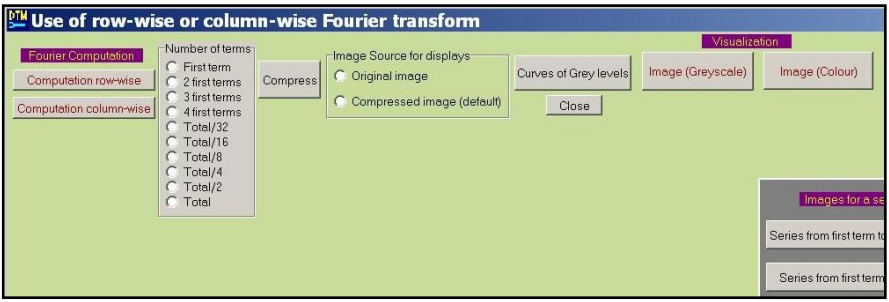
Note : Toutes les images créées sont systématiquement enregistrées au format bitmap (extension: ". bmp") dans le répertoire du fichier de l'image analysée.

*

* *

d. Le cas des séries de Fourier discrètes :

Dans la partie inférieure gauche de la fenêtre, dans la rubrique : Compression technique , cliquez sur le bouton: Discrete Fourier Transform. Une nouvelle fenêtre s'affiche.



Portion de la fenêtre de commande des compressions par séries de Fourier discrètes.

d1. Vous devez ensuite sélectionner le mode de calcul de la série de Fourier, en ligne ou en colonne ("Row-wise" ou "columnwise"). Sélectionnez "Row-wise", par exemple.

d2. Puis, comme précédemment, si vous souhaitez obtenir un aperçu de la reconstitution des données lorsque le nombre de termes augmente, cliquez directement sur le bouton: Series from first term

to total (greyscale) , dans le panel : Images for a series of terms. On peut alors observer la reconstitution progressive de l'image.

d3. Si vous vous intéressez à un nombre de termes particulier (parmi les termes de la sélection suggérée), sélectionnez le nombre requis dans la liste verticale correspondante, et visualisez chaque image avec l'analogue du bouton utilisé en **b.**

Cas des séries de Fourier discrètes (option : ligne par ligne): Images reconstituées successivement avec deux termes , 9 termes et 19 termes. L'analyse colonne par colonne donne des résultats différents, mais avec un pouvoir de compression équivalent dans le cas de cette image.

d4. La comparaison de la reconstitution obtenue (en fonction du nombre de termes conservés dans la décomposition de Fourier) avec la reconstitution précédente (à l'aide de CA ou de SVD) est intéressante.

Note 1: Un affichage graphique des niveaux de gris pour chaque ligne peut être obtenu à partir du bouton "Curves of grey levels" (appuyer plusieurs fois pour balayer toute l'image).

Note 2: Toutes les images créées sont enregistrées au format bitmap (extension: ".bmp") dans le répertoire du fichier de l'image analysée.

Note 3: La compression par SVD ou CA ne dépend pas de l'ordre des lignes et des colonnes de la table (contrairement à la compression de Fourier). Néanmoins, cette compression par axes principaux que l'on peut qualifier de "compression structurelle" (parce qu'elle ignore les positions relatives des éléments) donne des résultats satisfaisants.

VI.4.4 Exécution des autres exemples

- Cliquez sur le bouton : SVD and CA of images, dans la rubrique *DtmVic-Images* du menu principal de Dtm-Vic.
- La fenêtre "Reconstitution of some small images" apparaît (cf. ci-dessus).

VI.4.4.1 Exemple "Baalbeck"

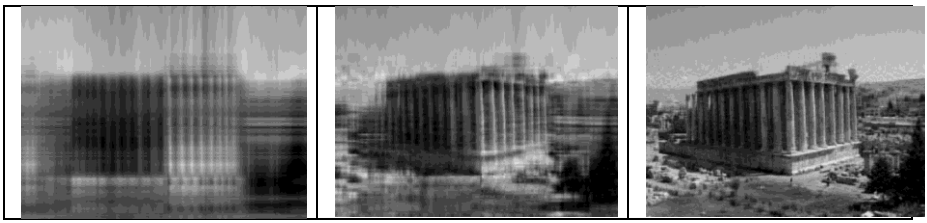
a. Cliquez sur le premier bouton Read (pgm format) dans la rubrique Open Greyscale image.
Dans le répertoire **EX_CO4_Image**, ouvrez le sous-répertoire **2_Baalbeck_pgm**. Dans **2_Baalbeck_pgm**, ouvrez le fichier Baalbeck.pgm. Une boîte de message rappelle les dimensions du fichier image.

b. Si vous désirez visualiser l'image d'origine, dans la rubrique Visualization, cliquez sur: Image (Greyscale).

c. Puis, dans la partie inférieure gauche de la fenêtre, dans la rubrique : Compression technique , cliquez sur le bouton: Correspondence Analysis (pour commencer). L'analyse s'effectue.

Ensuite, refaire toutes les opérations de c.1 à c.3, puis de d.1 à d.4.
Cet exemple est intéressant car il met en évidence le fait qu'une forte structure géométrique de l'image (ici: les colonnes du temple de Baalbeck) peut contaminer la reconstitution dans le cas des axes principaux.

Ce n'est pas le cas de la reconstitution de Fourier ligne par ligne : en reconstituant une ligne de la partie supérieure de l'image (le ciel), on ignore qu'il y a des colonnes plus bas dans l'image. En revanche c'est le cas pour la reconstitution de Fourier colonne par colonne...



Temple de Baalbeck. Cas de l'analyse des correspondances : Images reconstituées

successivement avec deux axes principaux, neuf axes et 50 axes. Les traits structuraux captés par les premiers axes se répercutent sur les axes suivants, et il faut atteindre près de 50 axes pour obtenir un ciel conforme à celui de l'image initiale.

VI.4.4.2 Exemple "Cardinal"

Pour ouvrir le fichier couleur du Cardinal de l'île Maurice, cliquez sur le troisième bouton Read (ppm format) dans la rubrique Open colour image. Dans le répertoire **EX_CO4_Image**, ouvrez le sous-répertoire **3_Cardinal_ppm_color**, puis ouvrez le fichier Cardinal.ppm. Une boîte de message rappelle les dimensions du fichier image.

Note: Rappelons que dans le format ppm, les trois couleurs de base (Rouge, Vert, Bleu) correspondant à chaque pixel ont des emplacements consécutifs sur la même ligne (dont la longueur est donc trois fois le nombre de pixels de la ligne). La compression par SVD ou CA ne dépend pas de l'ordre des colonnes, ce qui signifie que nous n'utilisons même pas le fait que les trois couleurs sont relatives à un même pixel!
Néanmoins, la "compression structurelle" fonctionne. Dans ce cas, la série de Fourier ligne par ligne n'est évidemment pas adaptée (la couleur n'apparaît qu'avec les derniers termes des séries).

Cardinal de l'île Maurice. Cas de l'analyse des correspondances : Images reconstituées successivement avec deux axes principaux, 10 axes et 100 axes.

VI.4.4.3 Exemple "Extra_pgm_ppm"

Cet dernier exemple contient les deux formats d'image pgm et ppm. Dans le répertoire **EX_CO4_Image**, ouvrez le sous-répertoire **4_Extra_pgm_ppm**, puis ouvrez le fichier broom.pgm. Une boîte de message rappelle les dimensions du fichier image.

Enfant balayant une cour. Cas de l'analyse des correspondances : Images en niveaux de gris (pgm) reconstituées successivement avec 2 axes principaux, 10 axes et 100 axes.

Que ce soit en noir ou en couleur, en actionnant le défilement automatique permis par les boutons Series from first term to total , on constate que l'image du balai n'apparaît pas avant le 20ème axe : les traits structuraux diagonaux sont défavorisés par la formule de reconstitution des données...

Enfant balayant une cour. Cas de l'analyse des correspondances : Images couleur (ppm) reconstituées successivement avec deux axes principaux, 10 axes et 100 axes.

Références bibliographiques sommaires

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.

Bouroche J.-M., Saporta G. (1980) *L'analyse des Données.* Coll. Que Sais-je ?, PUF, Paris.

Bry X. (1995) *Analyses Factorielles Simples*. Economica, Paris.

Efron B. (1979) Bootstraps methods : another look at the Jackknife, *Ann. Statist*., 7, p 1-26.

Escofier B., Pagès J. (1988) *Analyses factorielle simple et multiple*. Dunod, Paris.

Gifi A. (1990) *Nonlinear Multivariate Analysis*. Wiley, Chichester.

Gower J.C., Ross G. (1969) Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics,* 18, 54-64.

Gower J.C., Hand D.J. (1996) *Biplots.* Chapman and Hall, London.

Greenacre M. (1984) *Theory and Application of Correspondence Analysis*. Academic Press, London.

Greenacre M., Blasius J. (editors) (2006*) Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC, London.

Habert B., Nazarenko A., Salem A. (1997) *Les linguistiques de Corpus*. Armand Colin, Paris.

Hayashi C., Suzuki T., Sasaki M. (1992) *Data Analysis for Social Comparative research: International Perspective*, North-Holland, Amsterdam

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.

Le Roux B., Rouanet M. (2009) *Multiple Correspondence Analysis*. Vol. 163, Sage Publication Inc.

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., Morineau A. (1982) *SPAD Système Portable pour l'Analyse des Données*. CESIA, 82 rue de Sèvres,75007 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. Bécue M. (1989) *SPAD.T Système Portable pour l'Analyse des Données Textuelles*, Manuel de Référence. CISIA, Paris.

Lebart L., Morineau A., Warwick K.W. (1984) *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York.

Lebart L., Salem A. (1994*) Statistique Textuelle*. Dunod, Paris.

Lebart L., Piron M., Morineau A., (2006) *Statistique Exploratoire Multidimensionnelle, Visualisation et Inférence en Fouille de Données.* Dunod, Paris. (4ème édition, refondue). [à consulter pour une bibliographie plus complète]

Lebart L., Salem A., Berry L. (1998) *Exploring Textual Data*. Kluwer, Boston.

Lebart L., Piron M., Steiner J.-F. (2003) *La Sémiométrie*, Dunod, Paris.

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.

Murtagh F. (2005) . *Correspondence Analysis and Data Coding with R*. Chapman and Hall, Boca Raton, USA.

Roux M. (1985) *Algorithmes de Classification*. Masson, Paris.

Salem A. (1987) *Pratique des segments répétés, Essai de satistique textuelle*, Klincksieck, Paris

Saporta G. (1990 - 2010) *Probabilités, Analyse des Données et Statistique*. Technip, Paris.

Tenenhaus M. (2007) *Statistique*. Dunod, Paris.

Tuffery S. (2006) *Data Mining et Statistique Décisionnelle*. Technip, Paris

Volle M. (1980) *Analyse des Données*, Economica, Paris.

ISBN 978-2-953777-0-8

LUDOVIC LEBART

Télécom-ParisTech

MARIE PIRON

Institut de Recherche pour le Développement

Dtm-Vic

Data and text mining

Visualization, inference, Classification

Logiciel d'analyse exploratoire multidimensionnelle de données numériques et textuelles

Librement téléchargeable sur : www.dtm-vic.com

ISBN *978-2-9537772-0-8*