

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

*(Troisième édition, Septembre 2016)*

(Version 6 de Dtm-Vic)

**Ludovic Lebart**

**Marie Piron**



# Sommaire

<b>Introduction .....</b>	<b>5</b>
<b>I. Présentation générale de Dtm-Vic.....</b>	<b>8</b>
1. Mise en place des fichiers de données	
2. Techniques d'analyse de données	
3. Visualisation des résultats	
4. La boîte à outils	
5. Format interne des fichiers de données	
<b>II. Données numériques :</b>	
<b>Prise en main de Dtm-Vic à partir de trois exemples.....</b>	<b>22</b>
➤ Analyse en Composantes Principales	
➤ Analyse des Correspondances	
➤ Analyse des Correspondances Multiples	
<b>III. Données textuelles et mixtes :</b>	
<b>Prise en main de Dtm-Vic à partir de trois exemples.....</b>	<b>63</b>
1. Simples textes : Série de poèmes	
2. Analyse Textuelle de questions ouvertes	
3. Analyse directe de réponses libres	
<b>IV. Importation, création et exportation des fichiers .....</b>	<b>101</b>
1. Fichiers numériques et textuels à partir d'Excel (r)	
2. Saisie manuelle de données numériques	
3. Exportation vers Excel.	
<b>V. Recodage, archivage, outils divers .....</b>	<b>115</b>
1. Recodage, archivage	
2. Intervention élémentaire sur une base de données	
3. Prétraitements numériques	
4. Prétraitements textuels	
5. Lemmatisation	
<b>VI. Autres analyses avec Dtm-Vic.....</b>	<b>131</b>
1. Données numériques : Semiométrie	

2. Données numériques : Contiguïté (Iris de Fisher / Anderson)
3. Description de graphes
4. Reconstitution d'images

**VII. Annexe : Notions de statistique multidimensionnelle .....175**

1. Rappels des principes des méthodes exploratoires
2. Les méthodes factorielles, aspects techniques
3. Analyse en composantes principales (ACP)
4. Analyse des correspondances (AC)
5. Analyse des correspondances multiples (ACM)
6. Autres méthodes
7. Classification hiérarchiques, Arbre de longueur minimale
8. Partitions, cartes auto-organisées
9. Classification mixte (ou : hybride)
10. Méthodes de validation

**Références bibliographiques sommaires .....212**

## **Dtm-Vic**

### ***Data and text mining***

### ***Visualization, inference, Classification***

**Logiciel d'analyse exploratoire multidimensionnelle  
de données numériques et textuelles**

Librement téléchargeable sur : [www.dtmvic.com](http://www.dtmvic.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 Dtm-Vic s'intitule: "*Statistique Exploratoire Multidimensionnelle*"<sup>1</sup>.

Les **données textuelles** sont, en particulier, des données à la fois 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*"<sup>2</sup> (téléchargeable à partir de [www.dtmvic.com](http://www.dtmvic.com)).

---

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

<sup>2</sup> *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.

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

On sait que les explorateurs découvrent souvent autre chose que ce qu'ils cherchent. Les utilisateurs de Dtm-Vic ont souvent l'occasion de le vérifier: 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 6 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 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.

Le sixième chapitre présente des applications plus approfondies, mettant notamment en œuvre de nouvelles options des procédures de visualisation. Ce chapitre VI aborde la Sémiométrie, les analyses de contiguïté, les descriptions de graphes, et illustre les capacités de compression d'images par les techniques factorielles.

Enfin, le septième et dernier chapitre contient des rappels d'analyse multidimensionnelle, première étape vers un approfondissement des méthodes.

Toutes ces phases de l'apprentissage supposent que le logiciel et le recueil de données servant d'exemples dans ce manuel aient été copiés ou téléchargés, depuis le site<sup>3</sup>: <http://www.dtmvic.com>

---

<sup>3</sup> 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 exemple) par l'utilisateur. On obtient l'écran d'accueil suivant:



L'affichage de la *galerie des précurseurs* ne dure que quelques secondes... mais peut être réactivé, comme la bibliographie historique (petit bouton « *and their seminal papers* »).

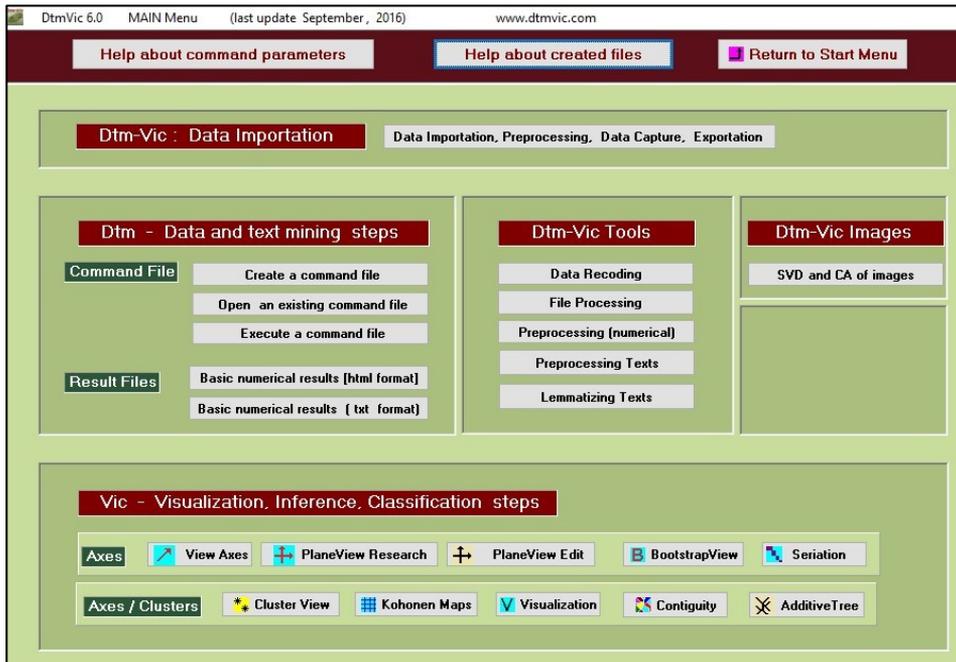
Cet écran préliminaire contient des informations générales (About DtmVic), une description du format des données interne à DtmVic (Data Format), l'accès au Tutorial (voisin de ce manuel, mais sans images, et avec des compléments sur certaines analyses de texte).

Le bouton « Books and User's guides on line » renvoie sur le site [dtmvic.com](http://dtmvic.com).

Le bouton « Recent features » résume les apports récents.

Enfin, le pavé « Statistical tools: Some reminders » donne accès aux notions théoriques de base.

Il faut cliquer sur « Start DtmVic » pour accéder à un écran principal (Main Menu) voisin de l'ancien écran d'accueil de DtmVic .



Après la rubrique d'importation : **Dtm Vic Data Importation** qui comprend les procédures de mise en place des données (importation, saisie, exportation), Dtm-Vic est structuré en deux étapes principales :

I – La première étape **Dtm – Data and Text mining** comprend 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 pré-traitement, de recodage, de stockage des données, et la rubrique **DtmVic Images** consacrée à certaines analyses d'images (optique pédagogique).

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. Toutes les analyses relèvent d'un même enchaînement d'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, cliquer sur les boutons **Help**, dans la barre du haut. Le bouton « Help about created files » s'affiche alors en rouge. Pour supprimer l'affichage de cette rubrique d'aide cliquer à nouveau sur le bouton.

## I.1 Mise en place des fichiers de données :

- Cette mise au format interne DtmVic des données a lieu une fois pour toute. Les analyses qui suivront utiliseront les données dans le format interne.
- Cliquer sur **Data Importation, Preprocessing, Data Capture, Exportation**.

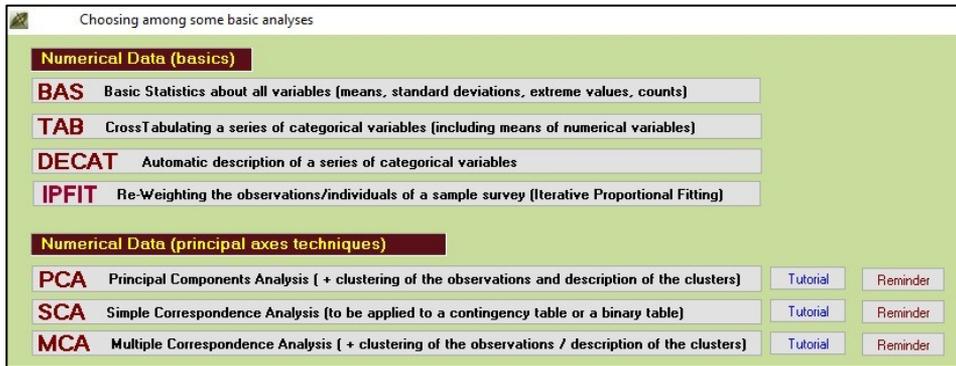
Une fenêtre suggérant différentes procédures apparaît :

<div style="background-color: #2d3748; color: white; padding: 5px; text-align: center; font-weight: bold;">Importation of variables, observations and texts</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px;">Importing Dictionary, Data, and Texts</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px;">Specific preprocessing and importation tools</div>
<p>- Importation de fichiers de données numériques ou textuelles et constitution des fichiers dictionnaire, données et textes dans le format Dtm-Vic. <b>Voir chapitre IV</b></p> <p>- Quelques outils de pré-traitement.</p>
<div style="background-color: #2d3748; color: white; padding: 5px; text-align: center; font-weight: bold;">Building the dictionary of variables and creating the data file</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px; border: 1px dashed gray;">Building the dictionary (manually)</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px;">Creating the data file (manually)</div>
<p>Modules de saisie de données : construction du dictionnaire des variables et création du fichier de données. <b>Voir chapitre IV.</b></p>
<div style="background-color: #2d3748; color: white; padding: 5px; text-align: center; font-weight: bold;">Exporting a DTM file to R or to Excel(r)</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px;">Exporting dtm data (and dictionary) to R or Excel (r)</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px;">Exporting dtm data, dictionary, and texts into a unique XML file</div>
<p>Exportation de fichiers de données en format Excel, R ou XML.. <b>Voir chapitre IV</b></p>
<div style="background-color: #2d3748; color: white; padding: 5px; text-align: center; font-weight: bold;">Dtm_tools: Amending or updating data and dictionary</div> <div style="background-color: #f5f5dc; padding: 5px; text-align: center; margin-top: 5px;">Dtm_tools</div>
<p>Création de nouvelles variables, sélection d'un sous-échantillon ou concaténation de plusieurs fichiers. <b>Voir l'accès direct à la boîte à outils <a href="#">DtmVic Tools</a> et chapitre V</b></p>

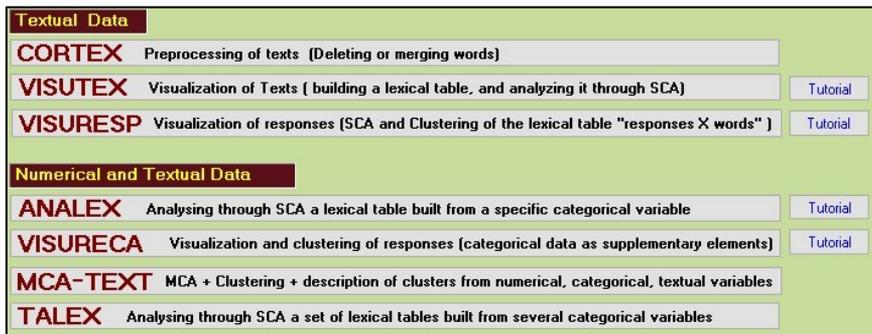
## I.2 Techniques d'analyse des données

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

Une fenêtre affichant différentes techniques d'analyse possibles apparaît. La partie supérieure de cette fenêtre traite des données numériques :



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



Pour un certain nombre de méthodes, on a accès directement au tutoriel extrait du tutoriel général de l'écran d'accueil (Start Menu).

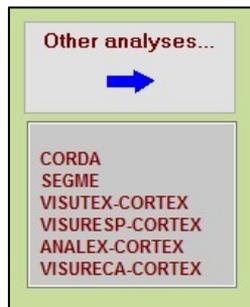
Pour les trois méthodes en axes principaux de base, on a accès aux rappels théoriques (Reminder) qui sont également consultables à partir de l'écran d'accueil.

Un pavé « Other Analyses » réservé aux données textuelles donne accès à quelques étapes plus spécialisées.

Explicitations sommaires des traitements:

<p><b>Numerical Data (basics)</b></p> <p><b>BAS</b> Basic Statistics about numeric</p> <p><b>TAB</b> CrossTabulating a series of c</p> <p><b>DECAT</b> Automatic description of</p> <p><b>IPFIT</b> Re-Weighting the observatic</p>	<p>Analyse descriptive univariée, <b>BAS</b> ;</p> <p>Demande de tableaux croisés <b>TAB</b> des variables continues ou nominales;</p> <p>Description automatique d'une variable par une série de variables nominales <b>DECAT</b>. Redressement de l'échantillon, <b>IPFIT</b> (Iterative Proportional Fitting).</p>
<p><b>Numerical Data (principal a</b></p> <p><b>PCA</b> Principal Components Analys</p> <p><b>SCA</b> Simple Correspondence Anal</p> <p><b>MCA</b> Multiple Correspondence Anc</p>	<p>Analyse statistique exploratoire de données numériques : Enchaînement d'une analyse factorielle (Analyse en Composantes Principales <b>PCA</b>), (Analyse des Correspondances Simples <b>SCA</b>), (Analyse des Correspondances Multiples <b>MCA</b>) et d'une classification (k-means et classification ascendante hiérarchique). Voir chapitre II.</p>
<p><b>Textual Data</b></p> <p><b>CORTEX</b> Preprocessing of texts</p> <p><b>VISUTEX</b> Visualization of Texts</p> <p><b>VISURESP</b> Visualization of resp</p>	<p>Analyse statistique exploratoire d'un corpus de textes: <b>CORTEX</b> supprime ou regroupe des mots (lemmatisation sommaire empirique); <b>VISUTEX</b> réalise une analyse des correspondances simples d'une table lexicale (voir chapitre III); <b>VISURESP</b> réalise une analyse directe de réponses ouvertes.</p>
<p><b>Numerical and Textual Data</b></p> <p><b>ANALEX</b> Analysing through SCA a lexic</p> <p><b>VISURECA</b> Visualization and clustering</p> <p><b>MCA-TEXT</b> MCA + Clustering + descri</p> <p><b>TALEX</b> Analysing through SCA a set of le</p>	<p>Analyse statistique exploratoire de questions ouvertes (voir chapitre III): <b>ANALEX</b> réalise une analyse des correspondances simples d'une table lexicale agrégée; <b>VISURECA</b> réalise une analyse analogue à VISURESP, mais l'illustre avec des variables nominales ; <b>MCA-TEXT</b> : Analyse des correspondances multiples (variables nominales), classification illustrées par les variables lexicales. <b>TALEX</b> : Analyse d'une juxtaposition de tables lexicales.</p>

D'autres techniques d'analyse textuelle sont proposées dans le pavé :



➤ Si l'on clique sur le bouton « **Other Analyses** », 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**.

<p><b>CORDA</b> Concordances of a</p> <hr/> <p><b>SEGME</b> Lists of repeated se</p>	<p><b>CORDA</b> fournit les concordances d'une liste de mots.</p> <p><b>SEGME</b> donne les listes de segments répétés.</p>
<p><b>VISUTEX-CORTEX</b> ▾</p> <hr/> <p><b>VISURESP-CORTEX</b></p>	<p><b>VISUTEX-CORTEX</b> réalise l'analyse VISUTEX précédente, après correction de textes similaire à CORTEX.</p> <p><b>VISURESP-CORTEX</b> réalise l'analyse VISURESP après CORTEX.</p>
<p><b>ANALEX-CORTEX</b> A</p> <hr/> <p><b>VISURECA-CORTEX</b></p>	<p><b>ANALEX-CORTEX</b> réalise simultanément les procédures CORTEX et ANALEX</p> <p><b>VISURECA-CORTEX</b> réalise simultanément les procédures CORTEX et VISURECA</p>

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, avec ces analyses composites, 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'enchaînement 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 »:

**VIC - Visualization, Inference, Classification**

## 1.3 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 :

-  ViewAxes : axes factoriels.  
Classements, pour chaque axe, des coordonnées des individus, des variables actives, supplémentaires, etc. pour une évaluation rapide des résultats.
-  PlaneView Research : plans factoriels.  
Description des plans factoriels pour tous les types d'éléments impliqués dans les analyses (Max : 30 000 points) (sauvegarde en format bmp). (Cadrage sur l'écran ou échelles identiques sur les axes).
-  PlaneView Edit: plans factoriels avec étiquettes mobiles.  
Description éditable des plans factoriels pour tous les types d'éléments impliqués dans les analyses (Max : 900 points) (sauvegarde en format bmp). (Cadrage sur l'écran ou échelles identiques sur les axes).
-  Bootstrap : Bootstrap (BootstrapView).  
Zones de confiance (ellipses ou enveloppes convexes) dans les plans factoriels pour les éléments sélectionnés (Voir le rappel (*Reminder*) dans la fenêtre Bootstrap).
-  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 sur les plans factoriels.  
Représentation des positions des centres de classes (*clusters*) 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 (*Self Organizing Map*).  
Cartes auto-organisées des individus, des variables, et simultanées des individus et des variables à partir des coordonnées factorielles (Grilles carrées 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 contiguïté. Analyse locale, structure de graphe.

L'analyse de Contiguïté 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 la section VI.2, chapitre VI de ce manuel.

-  Additive Trees : Arbres additifs utilisant le logiciel SplitsTree<sup>4</sup>.

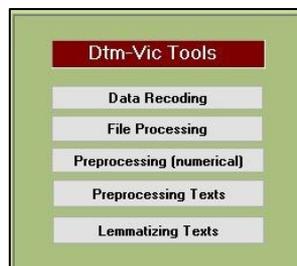
L'analyse arborée (Additive trees) produit un arbre dont les sommets sont les objets à classer, mais qui contient plus d'information que la classification hiérarchique classique (dont l'arbre de longueur minimale fait partie). Les distances entre objets (variables, individus) dans l'espace global sont approximées par les distances « longueurs du plus court chemin sur le graphe ».

L'analyse de contiguïté est abordée dans la section VI.2, chapitre VI de ce manuel.

**Important :** On peut accéder directement à tous les boutons de cette phase de visualisation **VIC** (pour une analyse réalisée antérieurement) à condition d'ouvrir simplement le fichier de commande, à partir du bouton « **Open an existing command file** ». Il n'est alors pas nécessaire de procéder à une nouvelle exécution, puisque tous les fichiers intermédiaires sont sauvegardés.

## 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).



<sup>4</sup> Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

➤ Cliquer sur **Data Recoding**

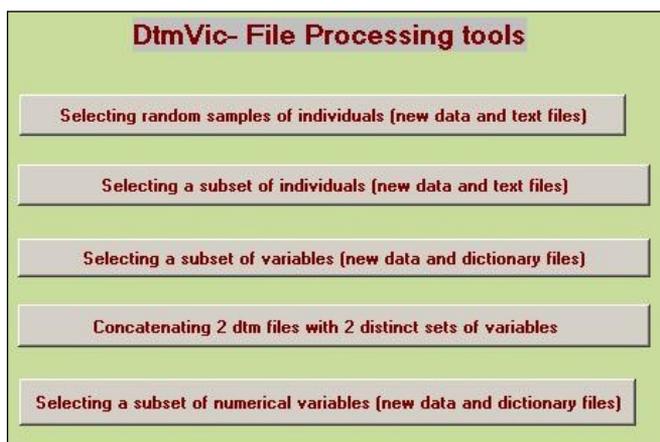
Le premier menu qui apparaît concerne le recodage des données et l'archivage de certains résultats.



*Création ou recodage de variables nominales :*

- i) Regroupement de modalités ;
- ii) Création d'une variable nominale par croisement de 2 variables nominales ;
- iii) Transformation d'une variable continue en variable nominale ;
- iv) Archivage des axes factoriels et des partitions.

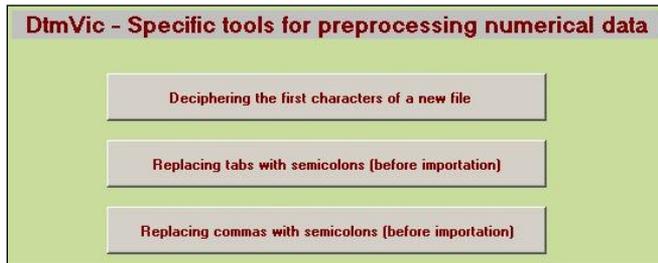
Le second groupe d'actions concerne le bouton : **File Processing**



- *Il propose des modifications de la base de données par :* (Voir Chapitre V)

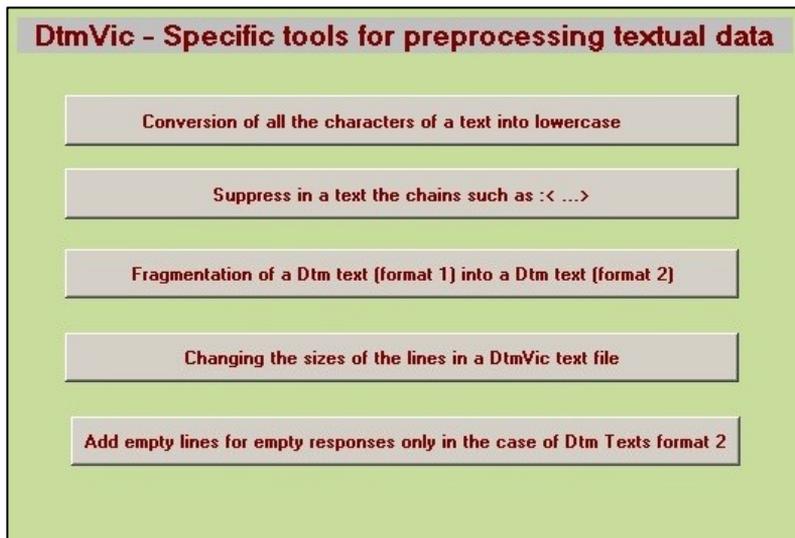
- i) Sélection d'un sous-ensemble aléatoire d'individus (lignes) ;
- ii) Sélection d'un sous-ensemble d'individus (lignes) à partir d'un filtre ;
- ii) Sélection d'un sous-ensemble de variables (colonnes) ;
- iii) Concaténation de deux bases de données (variables différentes).
- iv) Sélection d'un sous-ensemble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

Le menu suivant (bouton : **Preprocessing (numerical)**) 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 menu suivant ouvert par le bouton **Preprocessing Texts** 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) 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.

iv) Changement de longueur des lignes de texte. Au départ, format DtmVic sans limitation pour la longueur des lignes (format 1 ou 2). 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).

v) Cette dernière procédure très limitée et spécialisée permet de faire respecter la contrainte « une ligne vide par réponse ouverte vide » pour des fichiers textuels DtmVic qui utiliseraient deux séparateurs consécutifs.

vi) Lemmatisation d'un fichier de type DtmVic (type 1 ou 2) utilisant le logiciel **TreeTagger**<sup>5</sup>. La procédure permet de faire une analyse morpho-syntaxique, puis de lemmatiser un texte en supprimant certaines catégories grammaticales (prépositions, articles, ...). Valable pour les textes anglais, français, espagnols, italiens.

\*  
\* \*

La rubrique **DtmVic Images**, surtout 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 d'accueil].

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. Les noms des fichiers sont libres, mais l'extension .txt est commode pour une consultation rapide du contenu des fichiers. 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 interne à Dtm-Vic actionné par le bouton "**Open an existing command file**" du menu principal.

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\_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).
- **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).

<sup>5</sup> Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Un cas d'application qui montre toutes les possibilités du logiciel est un recueil de données d'enquête par sondage, comportant des réponses aux questions fermées (< 1200) et des réponses aux questions ouvertes (< 12). Les questions fermées peuvent donner lieu à des variables continues (ou encore quantitatives) ou à des variables nominales (ou qualitatives).

```

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

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 libre 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 – (autres) 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 caractères, une procédure de Dtm-Vic-Tools permet de les reformater.

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

```

---- 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)**

Commentaires du tableau 4 : 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.

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 30 000, limite du nombre d'observations de Dtm-Vic dans la version actuelle). Le texte total d'un individu est alors limité à 100000 caractères.

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

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. La liste des fichiers créés par DtmVic est présentée grâce au bouton « **Help about Created Files** » du menu principal (accessible à partir du menu d'accueil par le bouton « **Start DtmVic** »).

## II. Données numériques :

### Prise en main de Dtm-Vic à partir de trois exemples

---

Les trois exemples 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 *supprap*.
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 est situé dans le répertoire :

**DtmVic-Exemples\_A\_Start/ EX\_A01.PrinCompAnalysis.**

Il vise à décrire un ensemble de variables continues par l'Analyse en Composantes Principales (*sur la méthodologie de l'ACP, voir aussi la section VII.3 de ce manuel*).

### II.1.1. Les données et fichiers Dtm-Vic :

(Exemple : 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](http://www.cesp.org)) en 1992 auprès de 18 000 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 ici 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 activé par le bouton du menu principal : « **Open an existing command file** »).

Ident	Sexe	Caract. socio-éco			Activités							Médias	
		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" (facultatif, peut être remplacé par 4 blancs) 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. Rappelons que les espaces vides (blancs) 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
```

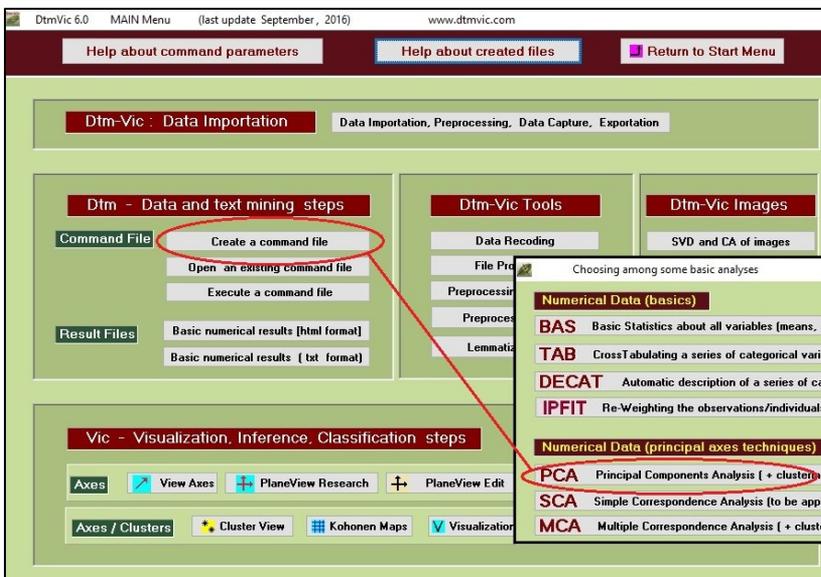
Ce fichier de données comprend 96 lignes et 45 colonnes. Pour une ligne *i*, la première valeur (entre quotes - guillemets simples) correspond à l'identifiant de l'individu *i*, c'est-à-dire ici le groupe *i* de répondants, et les 44 autres valeurs correspondent aux réponses des 44 variables séparées 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

- Cliquer sur le bouton **Create a Command file** de **Command File**

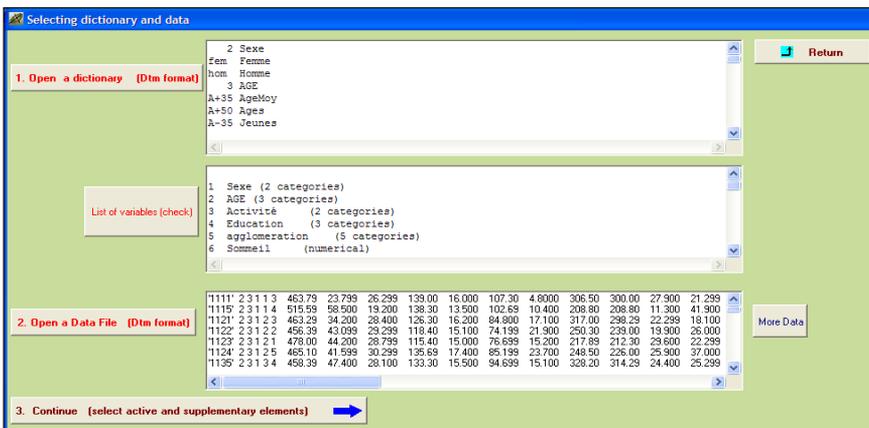


- Sélectionner l'analyse : **PCA** – Principal Components Analysis dans la rubrique **numerical data (principal axes techniques)**

Une fenêtre "Selecting dictionary and data" apparaît.

### Etape 2 : Sélection des fichiers dictionnaire et données

- Cliquer 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.



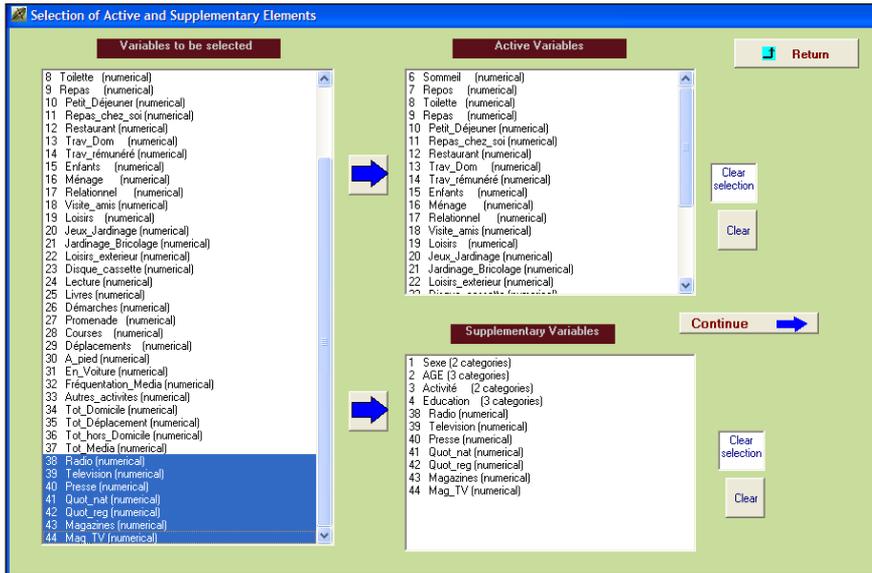
Cliquer sur le bouton : **Open a Data File**. Toujours dans le répertoire **DtmVic\_Examples\_A\_Start \EX\_A01.PrinCompAnalysis**, ouvrir le fichier **PCA\_dat.txt** qui s'affiche dans une troisième fenêtre.

- Cliquer sur : **3. Continue** → . Une fenêtre "Selection of active and supplementary elements" apparaît alors.

### 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 :

- 1 "Variables to be selected" où figure l'ensemble des variables
- 2 "Active Variables" qui reçoit les variables actives sélectionnées
- 3 "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

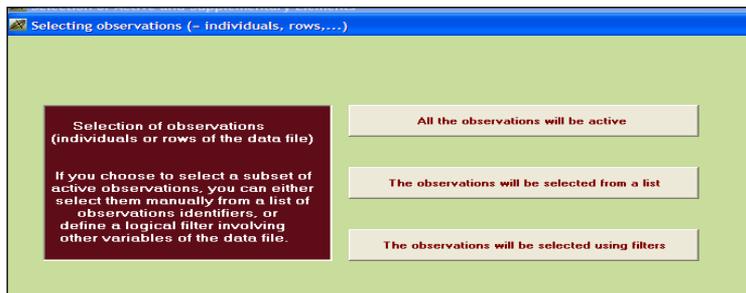
➤ Cliquer sur : **Continue** ➔

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.

Cliquer 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 (doivent !) 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

- Cliquer sur **1-Select some options**

Une fenêtre "Options : Bootstrap and/or Clustering of observations" apparaît.

- Cliquer sur : "yes" pour la procédure "bootstrap" ; indiquer le nombre de réplifications (par défaut 25) puis **Enter**. C'est le bootstrap partiel qui est appliqué par défaut. Si le bootstrap n'est pas adopté, cliquer sur : "no".

**Note technique : Les différents types de *bootstrap* pour variables non-textuelles dans Dtm-Vic (voir aussi section VII.10 : Validation)**

**a\_ *Bootstrap* partiel pour les variables actives**

Avec ce type de *bootstrap*, le plan initial sert d'espace de référence pour accueillir les réplifications, 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 réplifications 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**

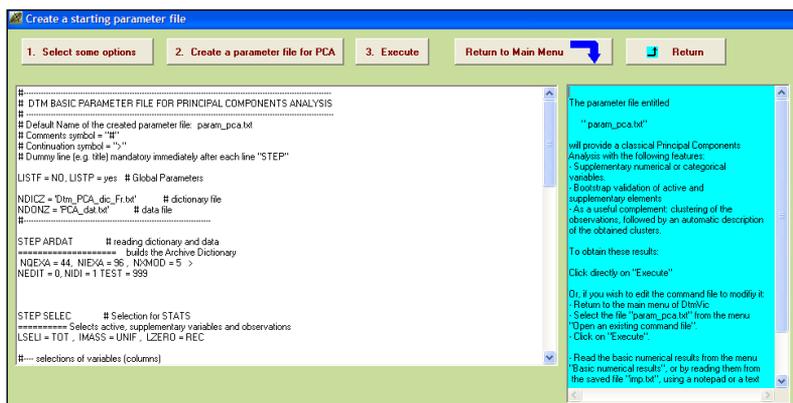
Rappelons que dans ce cas, chaque réplification donne lieu à une analyse en composantes principales spécifique. Il existe trois implémentations du *bootstrap* total dans Dtm-Vic.

- Le *bootstrap* de type 1 (simples corrections du signe des axes pour les analyses des réplifications).
- 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.

- Sélectionner le nombre de classes souhaité (nous suggérons 7 classes) puis cliquer sur **Enter**.

- Cliquer sur **Continue** ➔

La fenêtre : "Create a starting parameter file" réapparaît.



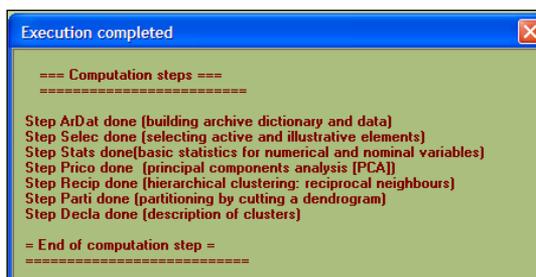
### b. Création du fichier paramètre

- Cliquer 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** (dossier **DtmVic\_Examples\_A\_Start**). Pour le conserver en vue d'analyses ultérieures, il faudra, après avoir quitté Dtm-Vic, le renommer.

### c. Exécution

- Cliquer 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 bouton "Help about parameters" disponible à partir de l'éditeur interne et du menu principal (Main Menu)).

### II.1.3 Fichier de résultats

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

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



- ou cliquer 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\_08.07.16\_14.45.txt" signifie le 8 juillet 2016, 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.

Revenir 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 tous 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.

- Cliquer sur :  **ViewAxes**.

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)							Coordonnées des variables nominales supplémentaires (Suppl categories)				
Active variables   Suppl. Categories   Individuals (observ)							Active variables   Suppl. Categories   Individuals (observ)				
<input type="button" value="View"/> <input type="button" value="Exit"/>							<input type="button" value="View"/> <input type="button" value="Exit"/>				
Identifiant	axis 1	axis 2	axis 3	axis 4	axis 5		Identifiant	axis 1	axis 2	axis 3	axis 4
Repas_chez_s	731	48	-559	-189	116		actifs	-1667	-97	-1024	393
Démarches	708	158	157	486	194		AgeMoy	-495	166	-1434	441
Repas	689	43	-492	-123	188		Agés	1866	-1475	246	377
Courses	683	149	-72	483	150		Femme	1312	1197	-955	-90
Petit_Déjeun	666	-268	-8	347	104		Homme	-1486	-1356	968	103
Television	635	-231	28	-497	17		inactifs	1970	115	1212	-463
Ménage	620	541	-277	-284	-29		Jeunes	-1486	1373	1006	-776
Fréquentatio	570	-439	412	-254	-28		primaire	939	-1185	-1070	-1215
Repos	566	-541	39	-35	-164		secondaire	-119	68	239	-277
Mag_TV	467	126	-52	-74	37		supérieur	-555	802	503	1258
Promenade	432	-28	492	19	155						
Lecture	386	252	446	573	-294						
Toilette	381	196	50	481	82						

**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*.

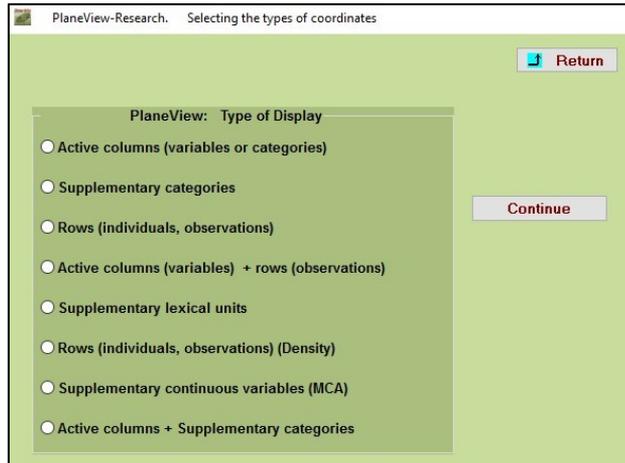
- Cliquer sur l'onglet des éléments à examiner, Active variables par exemple puis sur **View**. Il est possible d'ordonner les coordonnées sur un axe donné, en cliquant sur le libellé "axis x" en haut de l'axe x.
- Cliquer sur : **Exit** pour sortir de cet outil.

## 2- Plans factoriels

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

- **2.1 Cliquer sur :**  **PlaneView Research**

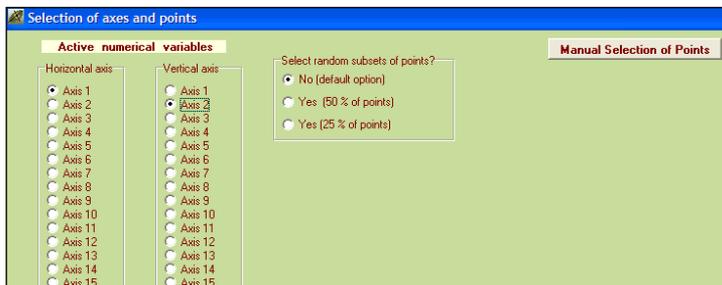
Une fenêtre propose différentes visualisations de plans factoriels.



Dans cet exemple particulier 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.

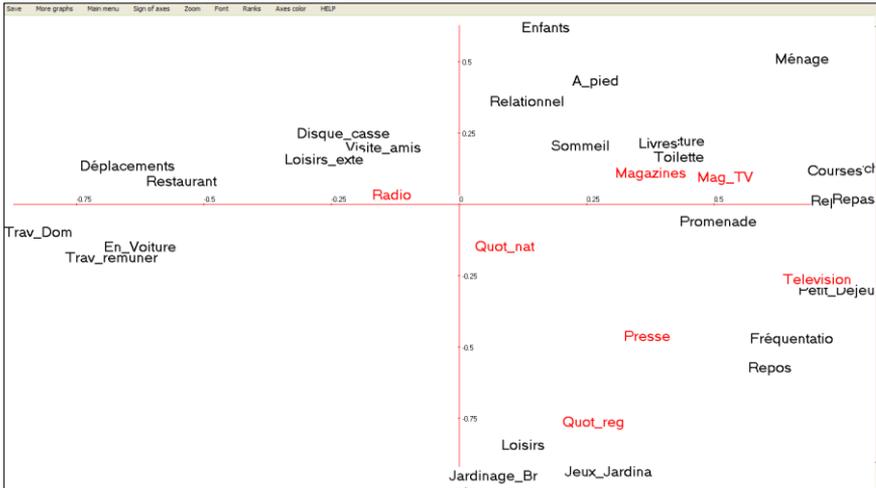
Sélectionner la rubrique "**Actives columns (variables or categories)**".

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



- Choisir les axes 1 et 2 puis cliquer sur : **Display**. Il est possible de ne faire figurer sur les plans que certaines variables. Cliquer alors sur : **Manual Selection of points**. Sélectionner les variables et les transférer dans la seconde fenêtre ( **Select**).

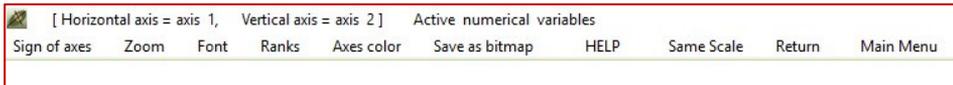
La fenêtre du plan factoriel apparaît.



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).

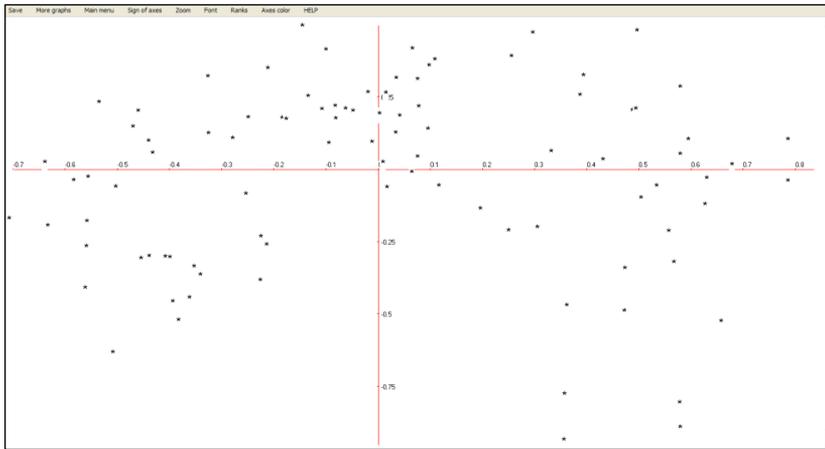
#### Bandeau de PlaneView Research



**Note :** Pour chaque graphique, le bandeau du haut contient des options :

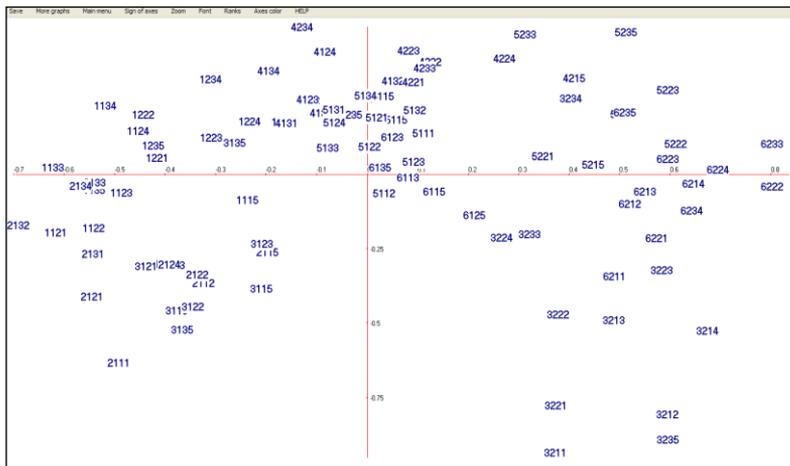
- « Sign of axes » permet d'inverser les axes ; « Zoom » possible (1,5 ; 2) ;
- « Font » offre la possibilité de modifier la police et la couleur des caractères ;
- « Rank », est utile seulement dans le cas des affichages très complexes, (ce qui n'est pas le cas ici) : ce bouton convertit les deux coordonnées 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 distorsion substantielle de l'affichage).
- « Axes color » change la couleur des axes ;
- « Save as bitmap » sauvegarde le graphique en format « .bmp » ;
- « Same scale » abandonne le cadrage sur la taille de l'écran pour donner la même échelle aux deux axes.

- Pour fermer le graphique, cliquer sur la croix en haut à droite puis sur :  **Return**, ou directement sur la rubrique du bandeau "Main menu".
- Retourner ensuite sur :  **PlaneView Research** pour sélectionner une autre représentation.

a) Rubrique "Individus actifs (densité)"

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

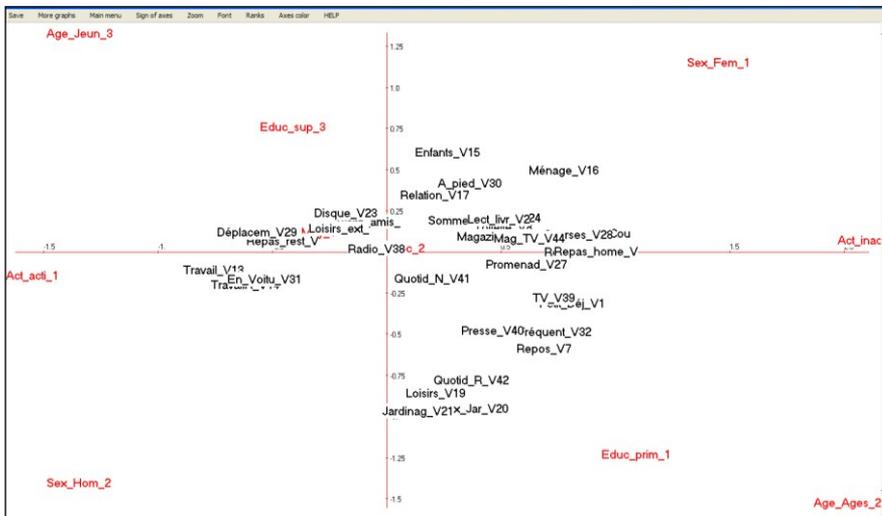
**Remarque :** Les identifiants des individus sont remplacés par un caractère simple [cas de nombreux individus, plusieurs milliers par exemple]. Cet affichage montre la forme du nuage des individus et d'éventuels individus aberrants. Les identifiants d'origine peuvent s'afficher en cliquant sur le bouton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

b) 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).

c) Rubrique "colonnes actives + catégories supplémentaires" :



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

## 2.2 Le bouton : PlaneView Edit :

Ce type de représentation (limité à 900 points) permet de déplacer (raisonnablement...) les étiquettes des points du graphique pour une meilleure lisibilité des étiquettes, ou en vue d'une publication.

- Cliquer sur  PlaneView Edit puis sur **Continue**

Une fenêtre apparaît. Choisir par exemple "actives columns (variables) (with continuous supplementary variables)", cliquer sur **Continue** et sélectionner le plan factoriel.

Il faut cliquer sur « **View** » (bandeau) pour que l'image s'affiche.

Toujours sur ce bandeau, un bouton « **Redraw** » permet de retracer les axes qui auraient pu être effacés par les déplacements d'étiquettes.

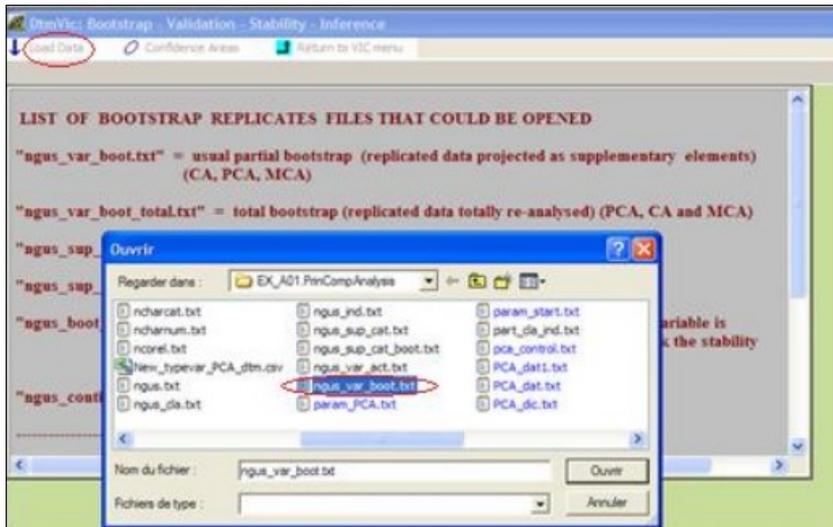


Plan factoriel (1,2) – rubrique "PlaneView Edit" puis bouton:  
*"active columns (variables) with continuous supplementary variables"*

### 3- Validation Bootstrap

Cet outil permet de valider la position des variables sur les plans factoriels.

- Cliquer sur : **B Bootstrap**

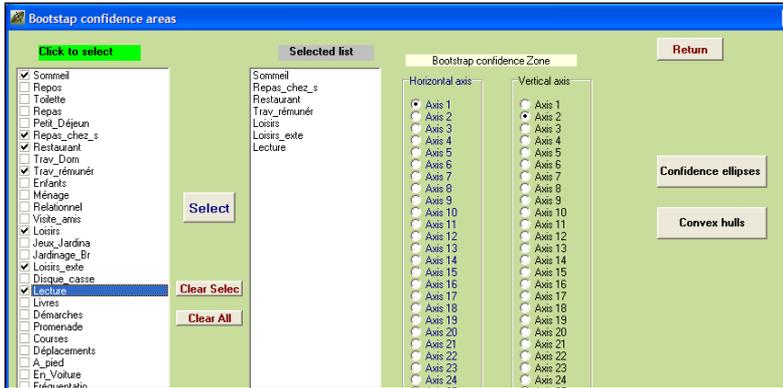


La fenêtre contient un bouton « **Tutorial** » qui affiche un article de synthèse (en Anglais) sur les différentes possibilités de *bootstrap* dans DtmVic.

Une fenêtre "DtmVic – Bootstrap – Validation – Stability – Inférence" apparaît.

- Cliquer sur **Load Data** puis ouvrir dans le répertoire le fichier des réplifications selon le bootstrap choisi. Sélectionner le fichier **ngus\_var\_boot.txt** pour un bootstrap partiel. Répondre **OK** à la fenêtre "Set of principal coordinates loaded" qui s'affiche.
- Puis cliquer sur **Confidence Areas**.

Une fenêtre "Bootstrap confidence areas" s'affiche:

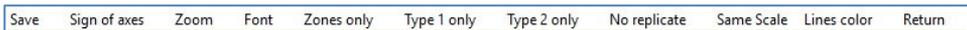


- Sélectionner dans la rubrique "Click to Select" les variables dont on veut visualiser les ellipses. Les transférer avec **Select**, dans la fenêtre "selected list".
- Choisir ensuite le plan factoriel puis cliquer sur **Confidence ellipses** pour obtenir l'affichage graphique des variables actives (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 (voir plus bas).

- Fermer la fenêtre et choisir maintenant le bouton : **Convex Hulls**. Les ellipses sont remplacées par les enveloppes convexes des réplifications bootstrap. Les enveloppes convexes prennent en considération les points périphériques, tandis que les ellipses sont dessinées en utilisant la densité des nuages des réplifications. Les deux informations sont complémentaires.

### Bandeau de la fenêtre Graphique / Bootstrap



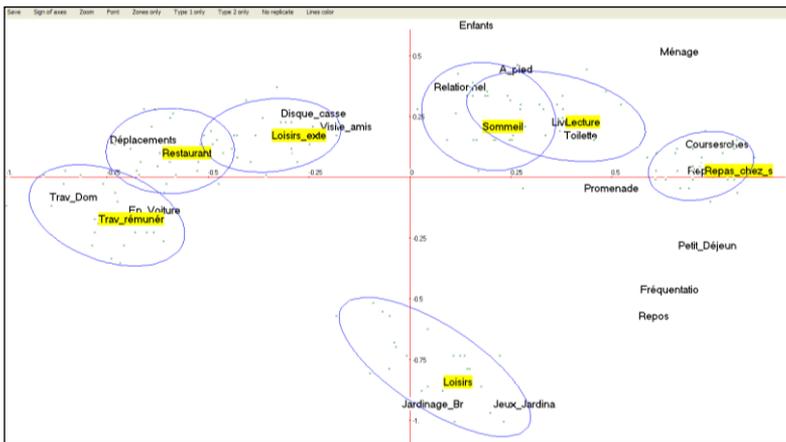
« Save » : Sauvegarder en format **.bmp** ; « Sign of axes » : change le sens des axes.

« Zoom » : possibilités de zoom (1,5 ; 2) ; « Font » : Changement de police ;

« Zones only » : Seules les zones sélectionnées sont représentées ;

Les boutons « Type 1 (2) only » ne sélectionnent qu'un bloc (cas de lignes/colonnes, ou de actives/supplémentaires) ; « No replicate » efface les petites étoiles représentant les réplifications. « Same Scale » impose la même échelle sur les deux axes ; « Lines color » permet de changer la couleur des tracés.

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



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 (comportant des milliers d'individus) non regroupé donnerait lieu à des ellipses de confiance beaucoup plus petites.

#### 4- Classification (ClusterView)

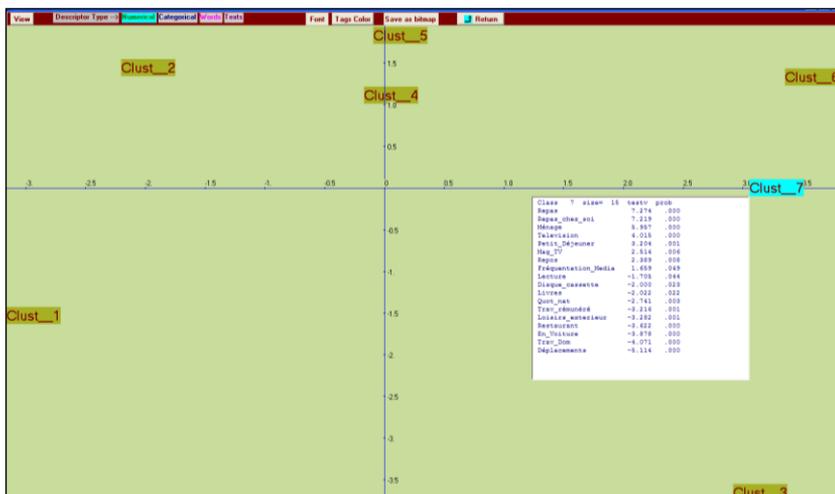
Cette option permet de visualiser les centres des classes, qui sont projetés sur le plan factoriel.

- Cliquer sur  **ClusterView**. Choisir les axes (1 et 2 pour commencer), et **Continue**.

La fenêtre "DTM-Display of clusters" apparaît.

- Cliquer sur **View**. Les centres des 7 classes apparaissent sur le plan factoriel. Cliquer 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 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.



**Commentaire :** En actionnant ce bouton "numerical" (bandeau du haut), puis en opérant un clic droit sur les points représentant les classes, 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 particulier, les autres rubriques du menu principal ne sont pas appropriées.

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

Ce deuxième exemple vise à décrire un petit tableau de contingence par l'analyse des correspondances simple (les données sont dans le répertoire :

**DtmVic-Exemples\_A\_Start/ EX\_A02. SimpleCorAnalysis).**

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

(Exemple : *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
'Employé'	511	593	57	217	172	306
'Ouvrier_qualif'	385	457	42	174	104	220
'Ouvrier_non_qualif'	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

Rappel : Dans ce format interne de Dtm-Vic, les libellés des catégories commencent à la colonne 6, [une police à intervalle fixe telle que le "courrier" peut être employée pour faciliter l'utilisation de ce genre de format]. Attention : Pas d'espaces vides dans les identifiants (individus et variables) !

## 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, cliquer : **Create a command file (Command File)**.

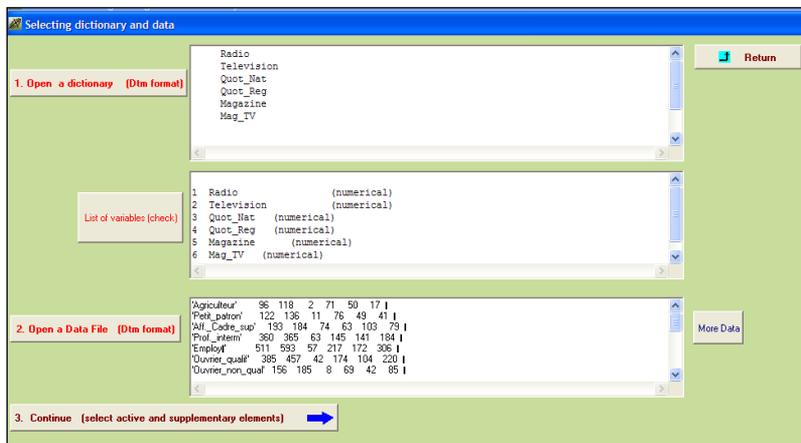
Une fenêtre "Choosing among some basic analyses" apparaît.

- Sélectionner 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

- Cliquer 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.

➤ Cliquer 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.*

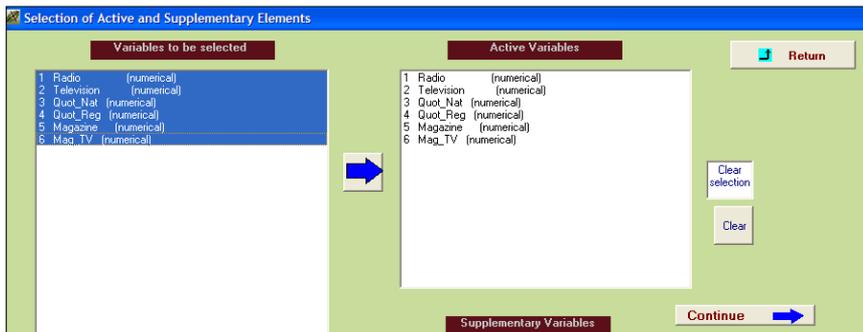
➤ Cliquer sur : **3. Continue** ➔

Une fenêtre "Selection of active and 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.

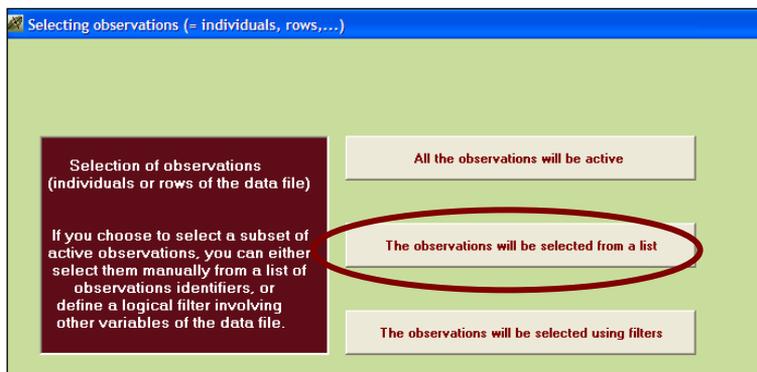
➤ Sélection des variables continues actives : V1 à V6 à transférer dans la fenêtre "Active Variables" (en cliquant sur la flèche bleue).



➤ Cliquer sur : **Continue** ➔ . Une fenêtre "Selecting observations" apparaît.

### Etape 4 : Sélection des observations (individus)

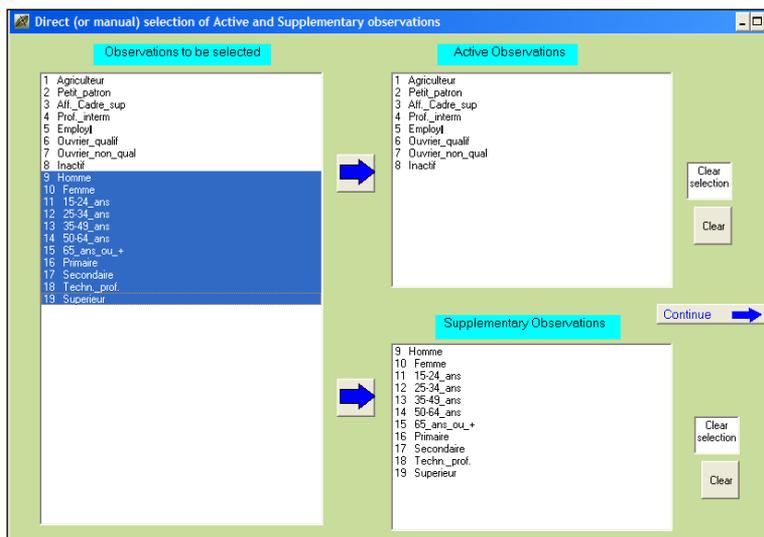
Les lignes ne représentent pas ici des observations ou individus comme pour l'ACP ou l'Analyse des Correspondances Multiples (plus loin) mais des modalités de variables. Aussi de la même manière que l'on considère des variables actives et/ou supplémentaires, on procède à la sélection des modalités actives 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.



- Cliquer sur: **The observations will be selected from a list**

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

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



- Cliquer 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*.

(Sinon, cliquer directement sur : **2-Create a parameter file for SCA**).

- Cliquer 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*.

- Cliquer sur "yes" pour la procédure *bootstrap* ; indiquer le nombre de réplifications (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 cliquer sur : **Enter**. Nous ne voulons pas effectuer de classification.
- Cliquer sur : **Continue** ➔

La fenêtre : "Create a starting parameter file" réapparaît.

- Cliquer 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).

- Cliquer 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).

```

Execution completed

=== Computation steps ===
=====
Step ArDat done (building archive dictionary and data)
Step Selec done (selecting active and illustrative elements)
Step Afcor done (correspondence analysis [CA])
Step Defac done (description of principal axes)

= End of computation step =
=====
[Click about here to hide this Memo]

```

*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 la rubrique : **Result Files**

- Cliquer 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 nom du fichier résultat est construit, avec date et heure, selon les mêmes principes que pour l'analyse en composantes principales.

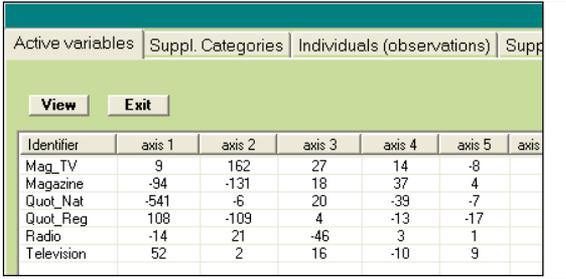
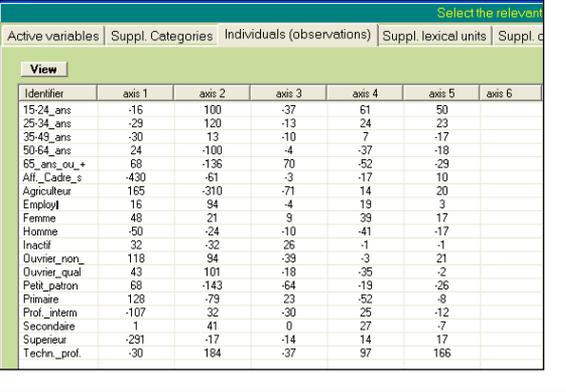
## II.2.4 Visualisation des résultats

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

- Cliquer sur:  **ViewAxes**. 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).

- Cliquer sur: **Active variables** puis sur: **View** pour obtenir les classements des coordonnées des modalités "média". Cliquer ensuite sur: Individuals (observations) puis sur: **View** pour obtenir les classements des coordonnées des modalités actives "statut d'activité" et des modalités supplémentaires.

<p>Coordonnées des modalités de la variable "média" classées Pour chaque axe.</p>	
<p>Coordonnées des modalités de la variable "statut d'activité". (Cette variable est positionnée en ligne et considérée ici comme individus)</p>	
<p><i>L'axe 1 oppose la presse quotidienne nationale aux autres médias et les cadres aux autres catégories</i>  <i>L'axe 2 oppose la presse régionale et magazine à la presse TV, et les agriculteurs et indépendants aux employés et ouvriers</i></p>	

- Cliquer sur : **Exit** pour sortir de cet outil.

## 2- Plans factoriels

- Cliquer sur :  **PlaneView Research.**

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.

- Cliquer sur la rubrique : "**Actives columns + Active rows**" puis Sélectionner les axes principaux désirés (ici les axes 1 et 2). Cliquer ensuite sur : **Display.**

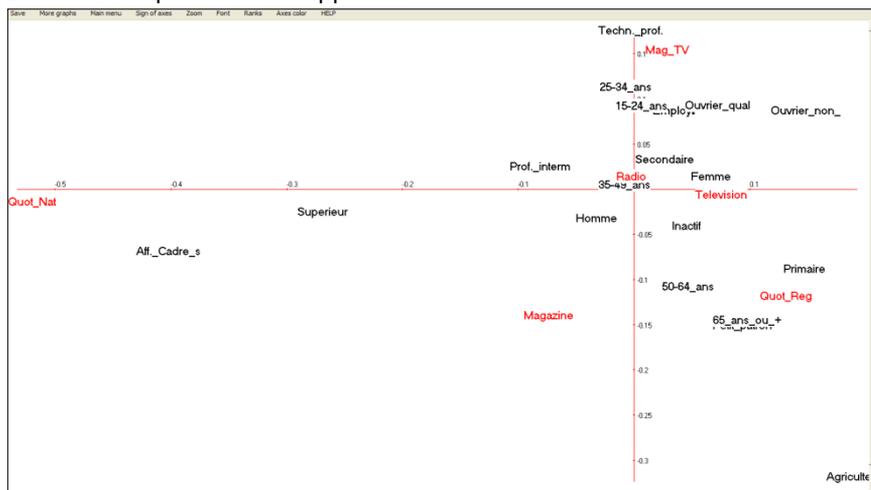
Apparaît une fenêtre pour choisir le plan factoriel suivant la paire d'axes souhaitée.

- Choisir les axes 1 et 2 (choix par défaut) puis cliquer sur : **Display**. Il est possible de ne faire figurer sur les plans que certaines variables. Cliquer 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 :

- « Sign of axes » permet d'inverser les axes ; « Zoom » possible (1,5 ; 2) ;
- « Font » offre la possibilité de modifier la police et la couleur des caractères ;
- « 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 distorsion substantielle de l'affichage).
- « Axes color » change la couleur des axes ;
- « Save as bitmap » sauvegarde le graphique en format « .bmp » ;
- « Same scale » abandonne le cadrage sur la taille de l'écran pour donner la même échelle aux axes.

La fenêtre du plan factoriel apparaît.



**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.

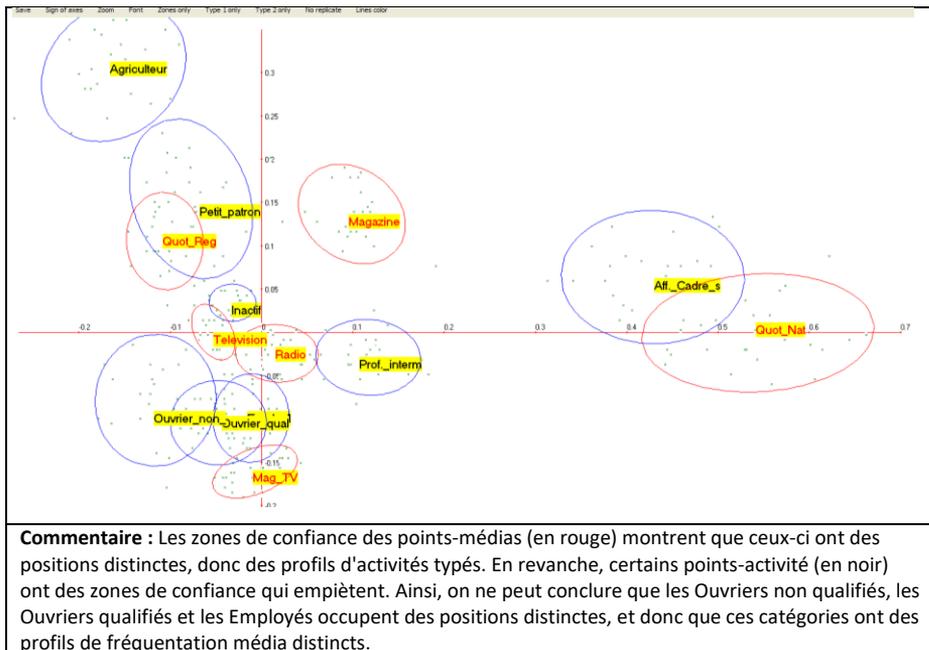
- Retourner ensuite sur : "PlaneView Research" pour sélectionner une autre représentation factorielle. Pour fermer le graphique, cliquer sur : **Return** ou sur la croix en haut à droite, puis sur : **Return** dans la fenêtre de sélection des axes.

### 3- Validation Bootstrap

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

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

- Cliquer ensuite sur : **Load Data** puis ouvrir dans le répertoire le fichier des réplifications 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.
- Sélectionner ( "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 cliquer sur **Confidence ellipses** pour l'affichage graphique des variables actives (fichier **ngus\_var\_boot.txt**).



- Pour fermer le graphique, cliquer sur : **Return**.

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

L'exemple 3 (répertoire : **DtmVic-Exemples\_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 annuelle par sondage effectuée par le CREDOC en 1986 sur "les conditions et aspirations des Français"<sup>6</sup>. Elles traitent des réponses d'un 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-Exemples\_A\_Start**, sont contenus les fichiers dictionnaire et des données en format Dtm-Vic :

#### 1. le fichier dictionnaire : **MCA\_dic.txt (extraits)**

```

      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
AB07 100001-200000
AB08 >200000
AB09 paris.agglo.paris
      2 sexe
AC01 Homme
AC02 Femme
. . . . .

```

<sup>6</sup> 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.

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 (facultatif) 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: MCA\_dat.txt (extraits)

```
'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. 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
'0054' 5. 5. 1. 24. 1. 3. 3. 1. 3. 2. 2. 2. 3. 2. 2. 2. 2. 1. 2. 2. 2..... 4. 7. 4. 7. 5. 7. 4. 3. 3. 3. 1 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 (format libre).

### 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 :

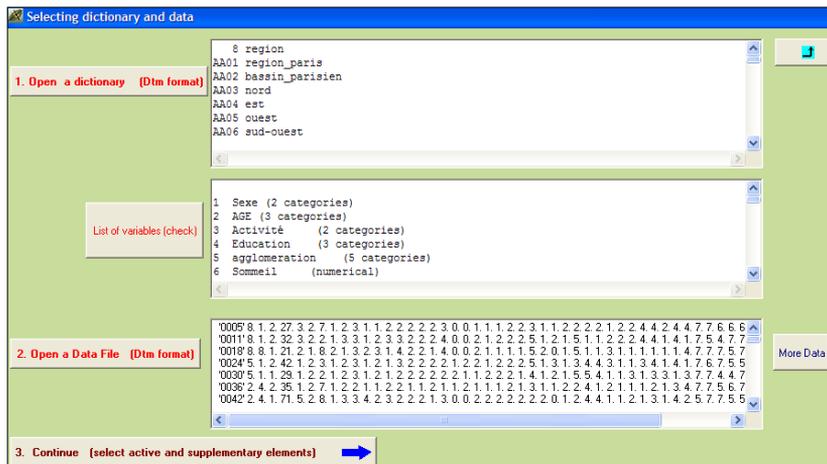
#### **Etape 1 : Sélection de l'analyse**

- Cliquer sur le bouton : **Create a command file** , ligne : **Command File**

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

- Sélectionner 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

➤ Cliquer sur le bouton : **Open a dictionary**. Dans le répertoire :

**DtmVic-Exemples\_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.

➤ Cliquer sur le bouton : **Open a Data File**. Dans le même répertoire, ouvrir le fichier **MCA\_dat.txt** qui s'affiche dans une troisième fenêtre.

➤ Cliquer sur **3. Continue** ➔

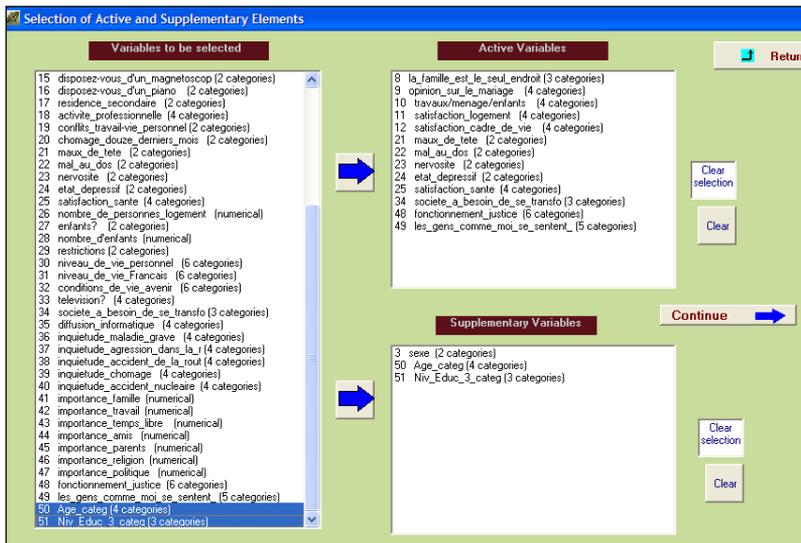
La fenêtre " *Selection of active and supplementary elements* " apparaît.

### 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

➤ Cliquer 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 (*option choisie*)
2. Sélectionner les observations sur une liste
3. Sélectionner les observations par un filtre

➤ Cliquer 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 :

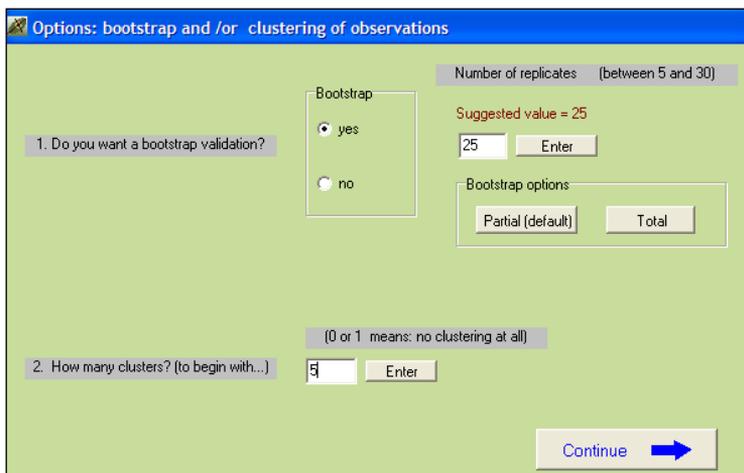
- un *bootstrap* qui permet de valider les positions des variables.
- une classification avec une description automatique des classes.

➤ Cliquer 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 et la section VII.10 de l'annexe.

➤ Cliquer sur : "**yes**" pour la procédure "bootstrap" ; indiquer le nombre de réplifications (par défaut 25) puis : **Enter**. C'est le bootstrap partiel qui est appliqué par défaut. Si le bootstrap n'est pas adopté, cliquer sur "**no**" et passer directement à l'option de classification.



➤ Sélectionner le nombre de classes souhaité (nous suggérons 5 classes) puis cliquer sur : **Enter**.

➤ Cliquer sur **Continue** ➔

La fenêtre "Create a starting parameter file" réapparaît.

- Cliquer 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.
- Cliquer sur **3-Execute**

```

Execution completed

=== Computation steps ===
=====
Step ArDaT done (building archive dictionary and data)
Step SeleC done (selecting active and illustrative elements)
Step Multm done (multiple correspondence analysis [MCA])
Step Recip done (hierarchical clustering: reciprocal neighbours)
Step Parti done (partitioning by cutting a dendrogram)
Step Decla done (description of clusters)

= End of computation step =
=====

[Click about here to hide this Memo]

```

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** (Coupeure 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 un 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**

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

Return	
<b>DtmVic: Main basic numerical results</b>	
Table of content	
<a href="#">Ardat (building archive dictionary and data)</a>	
<a href="#">SeleC (selecting active and illustrative elements)</a>	
<a href="#">Multm (multiple correspondence analysis [MCA])</a>	
<a href="#">Recip (hierarchical clustering: reciprocal neighbours)</a>	
<a href="#">Parti (partitioning by cutting a dendrogram)</a>	
<a href="#">Decla (description of clusters)</a>	

- ou encore : cliquer 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, revenir 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

- Cliquer sur  **ViewAxes**. 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**.
- Cliquer 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".

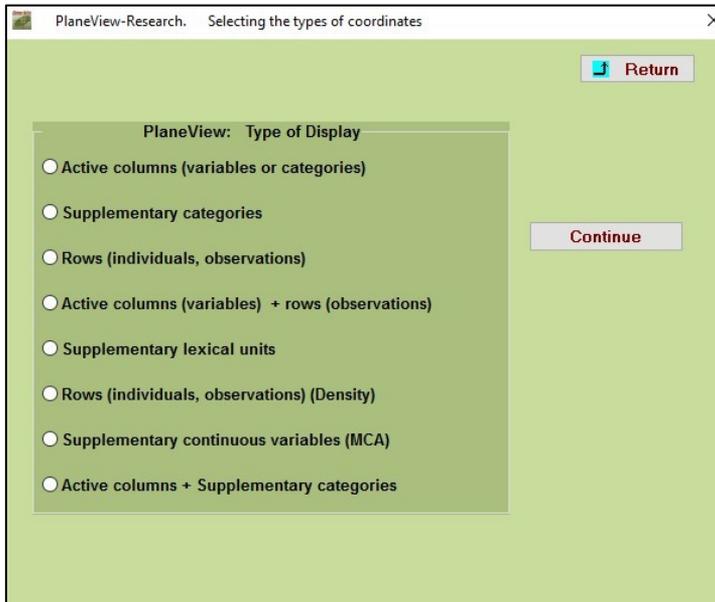
Active variables	Suppl. Categories	Individuals (observ)																																																																																																																																																										
<div style="display: flex; justify-content: space-between;"> <span>View</span> <span>Exit</span> </div> <table border="1"> <thead> <tr> <th>Identifiant</th> <th>axis 1</th> <th>axis 2</th> <th>axis 3</th> <th>ax</th> </tr> </thead> <tbody> <tr><td>satisfaction_sante:p</td><td>-2256</td><td>-300</td><td>-1250</td><td>-</td></tr> <tr><td>satisfaction_sante:p</td><td>-1370</td><td>122</td><td>898</td><td>-</td></tr> <tr><td>etat_depressif_oui</td><td>-1350</td><td>-317</td><td>-569</td><td>-</td></tr> <tr><td>justice:ne_sait_pas</td><td>-1001</td><td>935</td><td>-716</td><td>-</td></tr> <tr><td>mariage:ne_sait_pas</td><td>-906</td><td>1282</td><td>-698</td><td>-</td></tr> <tr><td>la_femme_seule</td><td>-879</td><td>1442</td><td>626</td><td>-</td></tr> <tr><td>transf-soc:ne_sait_p</td><td>-865</td><td>1383</td><td>-307</td><td>-</td></tr> <tr><td>maux_de_tete_oui</td><td>-785</td><td>-145</td><td>-61</td><td>-</td></tr> <tr><td>solitude:assez_d'acco</td><td>-694</td><td>-363</td><td>17</td><td>-</td></tr> <tr><td>solitude:tres_d'acco</td><td>-651</td><td>-848</td><td>995</td><td>-</td></tr> <tr><td>nervosite_oui</td><td>-640</td><td>-160</td><td>-160</td><td>-</td></tr> <tr><td>mal_au_dos_oui</td><td>-570</td><td>150</td><td>54</td><td>-</td></tr> <tr><td>satisf.log.peu</td><td>-358</td><td>-680</td><td>1883</td><td>-</td></tr> <tr><td>justice:refus/repond</td><td>-144</td><td>1110</td><td>-17</td><td>-</td></tr> <tr><td>satisf.log.assez</td><td>-129</td><td>-358</td><td>-19</td><td>-</td></tr> <tr><td>plutot_la_femme</td><td>-119</td><td>280</td><td>54</td><td>-</td></tr> <tr><td>mariage:dissout_si_p</td><td>-83</td><td>50</td><td>-165</td><td>-</td></tr> <tr><td>cdv:assez</td><td>-70</td><td>-123</td><td>231</td><td>-</td></tr> <tr><td>famille:-oui-</td><td>-63</td><td>184</td><td>289</td><td>-</td></tr> </tbody> </table>			Identifiant	axis 1	axis 2	axis 3	ax	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'acco	-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	-	<div style="display: flex; justify-content: space-between;"> <span>View</span> <span>Exit</span> </div> <table border="1"> <thead> <tr> <th>Identifiant</th> <th>axis 1</th> <th>axis 2</th> <th>axis 3</th> <th>ax</th> </tr> </thead> <tbody> <tr><td>Age_super_60</td><td>-333</td><td>374</td><td>363</td><td>-2</td></tr> <tr><td>feminin</td><td>-204</td><td>-54</td><td>-101</td><td>-1</td></tr> <tr><td>Niv_Educ_bas</td><td>-203</td><td>59</td><td>142</td><td>-</td></tr> <tr><td>Age_inf_60</td><td>-85</td><td>104</td><td>87</td><td>-1</td></tr> <tr><td>Niv_Educ_moyen</td><td>14</td><td>-64</td><td>-224</td><td>-1</td></tr> <tr><td>Age_inf_40</td><td>82</td><td>-14</td><td>-264</td><td>1</td></tr> <tr><td>Age_inf_30</td><td>248</td><td>-347</td><td>-133</td><td>2</td></tr> <tr><td>masculin</td><td>261</td><td>70</td><td>129</td><td>-</td></tr> <tr><td>Niv_Educ_haut</td><td>335</td><td>-65</td><td>-115</td><td>-1</td></tr> </tbody> </table>				Identifiant	axis 1	axis 2	axis 3	ax	Age_super_60	-333	374	363	-2	feminin	-204	-54	-101	-1	Niv_Educ_bas	-203	59	142	-	Age_inf_60	-85	104	87	-1	Niv_Educ_moyen	14	-64	-224	-1	Age_inf_40	82	-14	-264	1	Age_inf_30	248	-347	-133	2	masculin	261	70	129	-	Niv_Educ_haut	335	-65	-115	-1
Identifiant	axis 1	axis 2	axis 3	ax																																																																																																																																																								
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'acco	-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	-																																																																																																																																																								
Identifiant	axis 1	axis 2	axis 3	ax																																																																																																																																																								
Age_super_60	-333	374	363	-2																																																																																																																																																								
feminin	-204	-54	-101	-1																																																																																																																																																								
Niv_Educ_bas	-203	59	142	-																																																																																																																																																								
Age_inf_60	-85	104	87	-1																																																																																																																																																								
Niv_Educ_moyen	14	-64	-224	-1																																																																																																																																																								
Age_inf_40	82	-14	-264	1																																																																																																																																																								
Age_inf_30	248	-347	-133	2																																																																																																																																																								
masculin	261	70	129	-																																																																																																																																																								
Niv_Educ_haut	335	-65	-115	-1																																																																																																																																																								
Coordonnées (x 1000) des variables nominales actives			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.

- Cliquer sur :  **PlaneView Research**

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". Le bouton  **PlaneView Edit** reprend certaines des rubriques précédentes et permet de déplacer les points superposés pour rendre plus lisible le graphique (cf. Exemple 1 – PCA).

- Sélectionner : "**colonnes actives + catégories supplémentaires**".

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

- Laisser les axes 1 et 2 (option par défaut) puis cliquer sur : **display**. Il est possible de ne faire figurer sur les plans que certaines variables. Cliquer 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 :



**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.

- Pour revenir au menu principal de Dtm-Vic, cliquer, 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.

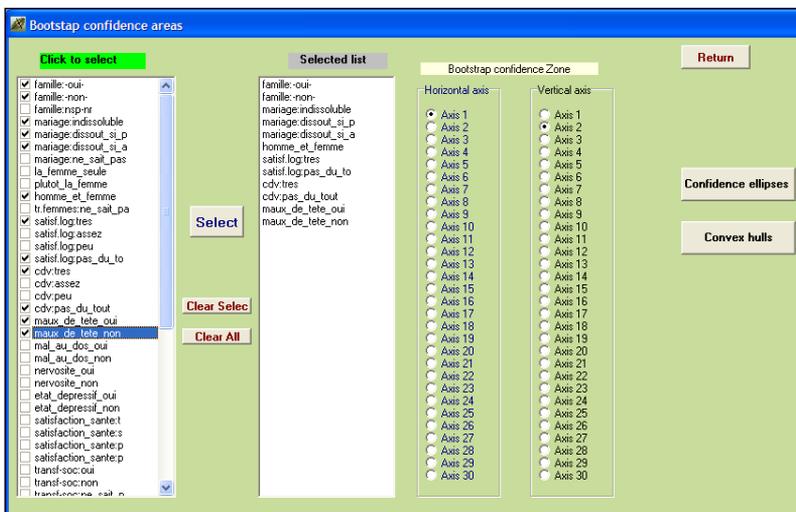
- Cliquer sur **B Bootstrap**

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

- Cliquer sur **Load Data** puis ouvrir dans le répertoire le fichier des réplifications selon la *bootstrap* choisi.
- Sélectionner le fichier **ngus\_var\_boot.txt** pour un bootstrap partiel. Répondre **OK** à la fenêtre "Set of principal coordinates loaded" qui s'affiche.
- Puis cliquer sur **Confidence Ellipse**.

Une fenêtre "Bootstrap confidence areas" s'affiche.

- Sélectionner dans la rubrique "Click to select" les variables dont on veut visualiser les ellipses.
- Les transférer avec **Select**, dans la fenêtre "Selected list".



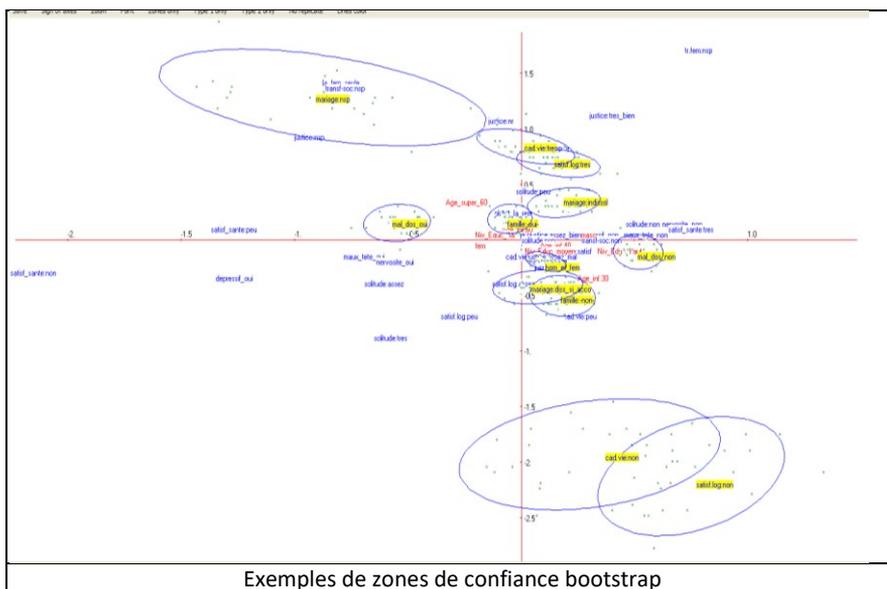
- Choisir ensuite le plan factoriel puis cliquer sur **Confidence ellipses** ou sur **Convex Hulls** pour obtenir l'affichage graphique des variables actives (si le fichier `ngus_var_boot.txt` a été chargé), ou de la catégorie supplémentaire (si le fichier `ngus_sup_cat_boot.txt` a été chargé).

- Les ellipses de confiance prennent en compte la densité du nuage de points-réplifications, 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 réplifications.
- Les enveloppes convexes (*Convex hulls*) enveloppent toutes les réplifications, mais donnent du poids aux points périphériques sans aucune considération de densité<sup>7</sup>.

#### Rappel sur les boutons du bandeau « Bootstrap »

Save Sign of axes Zoom Font Zones only Type 1 only Type 2 only No replicate Same Scale Lines color Return

« Save » : Sauvegarder en format `.bmp` ; « Sign of axes » : change le sens des axes.  
 « Zoom » : possibilités de zoom (1,5 ; 2) ; « Font » : Changement de police ;  
 « Zones only » : Seules les zones sélectionnées sont représentées ;  
 Les boutons « Type 1 (2) only » ne sélectionnent qu'un bloc (cas de lignes/colonnes, ou de actives /supplémentaires) ; « No replicate » efface les petites étoiles représentant les réplifications. « Same Scale » impose la même échelle sur les deux axes ; « Lines color » permet de changer la couleur des tracés.



<sup>7</sup> 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, cliquer, selon la fenêtre, soit sur la croix en haut à droite, soit sur **Return**.

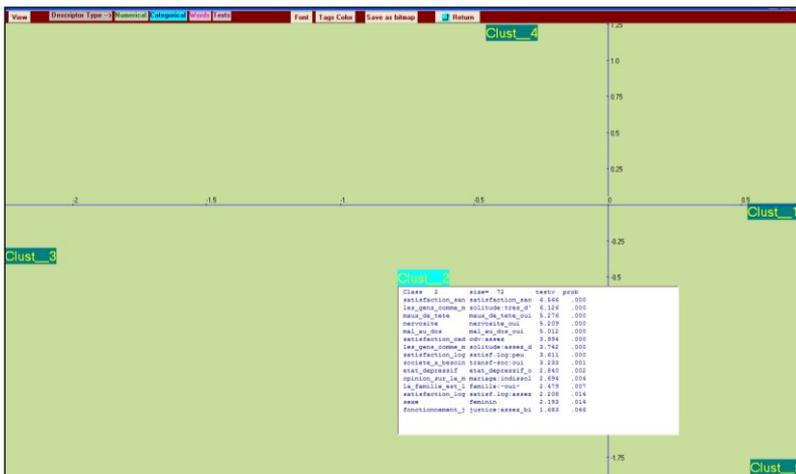
#### 4- Classification

Cette option positionne les classes obtenues sur un plan factoriel, et affiche une description automatique des classes par un clic droit de la souris.

- Cliquer sur  **ClusterView** . Choisir les axes (1 et 2 pour commencer), et : **Continue**.

La fenêtre "DTM-Display of clusters" apparaît.

- Cliquer sur **View**. Les centroides des 5 classes apparaissent sur le plan factoriel.
- Actionner le bouton **Categorical** du bandeau. Puis en cliquant (clic 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 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.**

### 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<sup>8</sup>. 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 ripper 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 beseege thy brow,
and dig deep trenches in thy beauty's field,
thy youth's proud livery, so gazed on now,
will be a tatter'd weed, of small worth held:
then being ask'd where all thy beauty lies,
where all the treasure of thy lusty days,
to say, within thine own deep-sunken eyes,
were an all-eating shame and thriftless praise.
how much more praise deserved thy beauty's use,
if thou couldst answer 'this fair child of mine

**** 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.
```

<sup>8</sup> Pour un ensemble plus important de sonnets et les commentaires attenants, se reporter au site : <http://www.shakespeare-online.com/sonnets/>.

```
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.

====
```

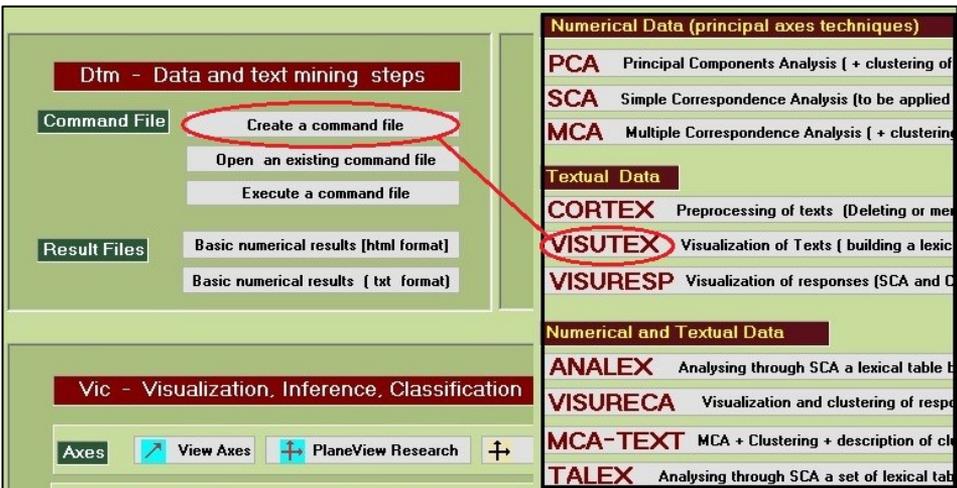
Les textes pouvant avoir des longueurs très différentes, une ligne spécifique sépare un sonnet d'un autre. Elle est caractérisée par des séparateurs "\*\*\*\*" suivis de 4 espaces blancs et du nom du texte. Le symbole "====" indique la fin du fichier. Comme tous les fichiers de données en format Dtm-Vic, 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](http://www.dtmvic.com)].

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

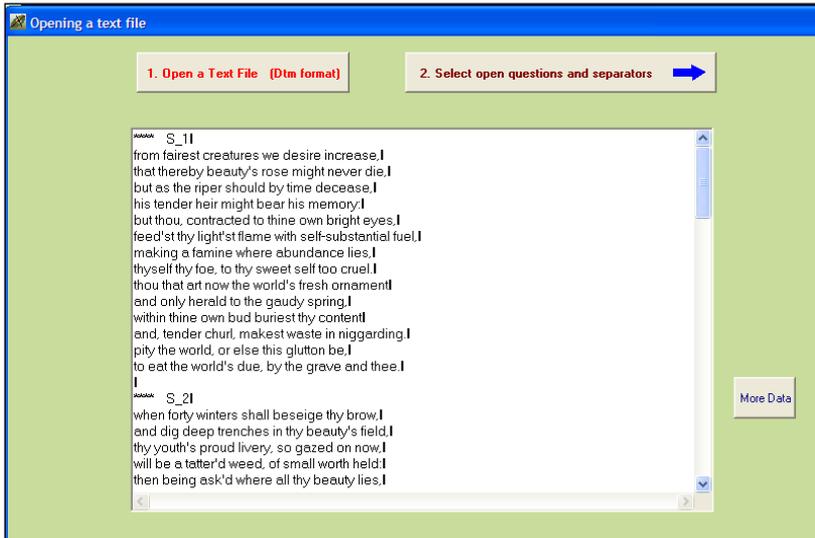
Le fichier de commande, ou fichier paramètre, est créé en 4 étapes.



#### Etape 1 : Sélection de l'analyse

➤ Dans la fenêtre du menu principal, cliquer sur le bouton : **Create a command file** de la rubrique: **Command File**.

- Une fenêtre "Choosing among some basic analyses" apparaît.
  - Choisir l'analyse : **VISUTEX – Visualization of texts** (rubrique : **Textual Data**)
- Une fenêtre : "Opening a text file" apparaît.



### Etape 2 : Sélection du fichier texte

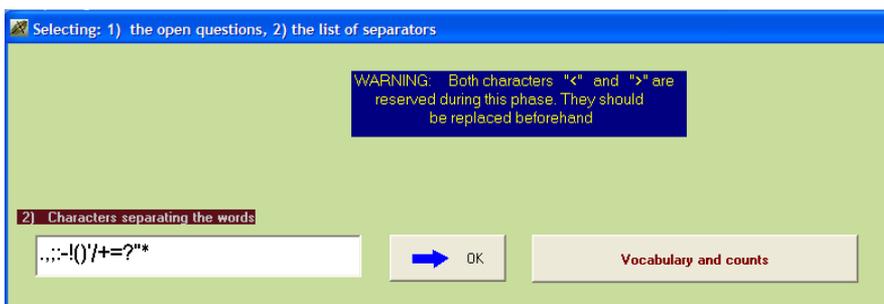
- Cliquer sur le bouton : **1. Open a text File**. Dans le répertoire **EX\_A04.Text-Poems**, ouvrir le fichier **Sonnet\_LowerCase.txt**.

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

- Cliquer ensuite sur : **2. Select Open questions and separators** ➔.

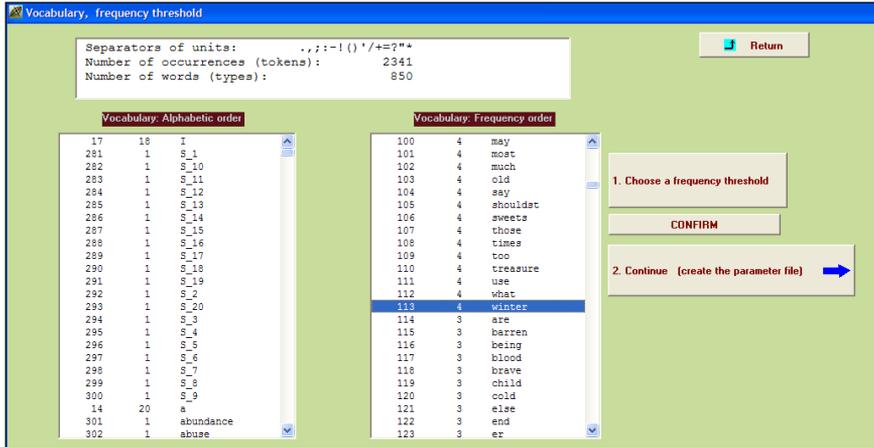
### Etape 3 : Sélection des questions, mots et vocabulaire

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.



- Cliquer directement sur : **vocabulary and counts**

La fenêtre suivante présente le vocabulaire (ordre alphabétique, à gauche, et ordre de fréquence à droite).



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...).

- Sélectionner cette ligne 113 puis cliquer sur **CONFIRM**. La fréquence apparaît. Répondre **OK** à la boîte de message.
- Cliquer 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).



- Cliquer sur **1-Select some options**

Une fenêtre "Options : Bootstrap and/or Clustering of observations" apparaît.

- Cliquer sur **"yes"** pour la procédure "bootstrap" ; indiquer le nombre de répliquions (par défaut 25) puis **Enter**. Si le bootstrap n'est pas adopté, cliquer sur **"no"**.
- Cliquer sur **Continue** →

Options: bootstrap and /or clustering of observations

1. Do you want a bootstrap validation?

Bootstrap

yes

no

Number of replicates (between 5 and 30)

Suggested value = 25

25 Enter

Bootstrap options

Partial (default) Total

Continue →

La fenêtre "Create a parameter file" apparaît de nouveau.

➤ Cliquer sur **2-Create a first parameter file.**

Create a parameter file for the sequence of processing: Vitex

1 - Select some options 2 - Create a first parameter file Execute Return to Main Menu Return

```

#####
# DTM BASIC COMMAND FILE FOR TEXTUAL DATA ANALYSIS
#####
# Default Name of the created command file: param_VISUTEX.txt
# Comments symbol = "#"
# Continuation symbol = "\n"
# Dummy line (e.g. title) mandatory immediately after each line "STEP"
LISTF = NO, LISTP = yes # Global Parameters
#####
NTEXZ = 'Sonnet_LowerCase.txt' # name of text file (free name)
#####
STEP ARTEX
##### Archive - Texts or responses to open ended questions
ITYP=1 LIREP=1 NCDL= 80
STEP SEL0X
##### Selection of open questions (irrelevant here, but necessary...)
#####

```

The command file (or: parameter file) entitled "param\_VISUTEX.txt" will provide a numerical coding of the texts (list of words with their frequencies), together with a correspondence analysis of of lexical table (words x texts), with possible bootstrap confidence areas for points. Characteristic words and lines for each text will be provided. To obtain these results: - Click on "Execute". 0: - Return to the main menu of DtmVic - Select the file " param\_VISUTEX.txt" " from the menu "Open an existing command file". - Click on "Execute". - Read the results from the button "Main basic numerical results". - Use the VIC tools (PlaneView, ClusterView, Bootstrap, etc.) to visualize the results.

Un fichier de commande (*parameter file*) est 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).

➤ Cliquer 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\_VISUTEX.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 an existing command file » (avec aussi l'aide du bouton "**Help about parameters**" disponible dans le menu principal [Main Menu]).

### III.1.3 Fichier de résultats

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

Cliquer 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 cliquer 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_VISUTEEX.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.

**Rappel** : On peut accéder directement à tous les boutons de cette phase de visualisation **VIC** (pour une analyse exécutée antérieurement) à condition d'ouvrir simplement le fichier de commande, à partir du bouton « **Open an existing command file** ». Il n'est alors pas nécessaire de procéder à une nouvelle exécution, puisque tous les fichiers intermédiaires sont sauvegardés.

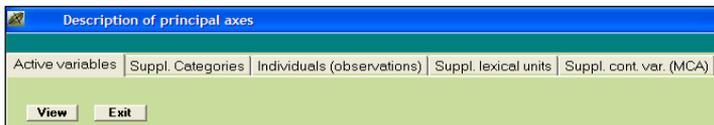


### 1- Axes factoriels

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

➤ Cliquer sur : **ViewAxes** .

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).



➤ Cliquer 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.

➤ Cliquer : **Exit** pour sortir de cet outil.

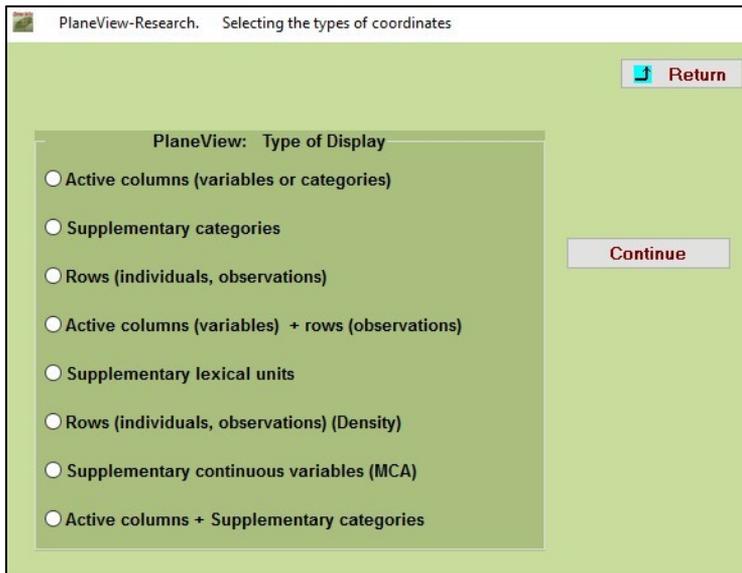
Active variables						Individuals (observations)							
View						View							
Identifiant	axis 1	axis 2	axis 3	axis 4	axis 5	Identifiant	axis 1	axis 2	axis 3	axis 4	axis 5	axis 6	axis 7
S_1	-263	237	4	-214	580	a	316	14	-504	-114	157	321	-75
S_10	-340	-360	273	-9	634	age	83	582	-442	-776	221	-1047	1031
S_11	-321	-158	246	-296	-136	all	-8	483	301	393	-256	64	120
S_12	68	744	331	370	-583	an	-17	172	-910	-75	783	172	-617
S_13	1402	-799	50	-298	-46	and	87	328	90	158	-156	-35	12
S_14	-61	535	442	465	-17	another	-713	-177	414	-470	212	-686	120
S_15	574	337	25	104	-239	art	-601	-370	221	578	736	123	138
S_16	1156	-236	247	-81	119	as	34	418	289	259	-39	-2	4
S_17	583	-98	-172	108	137	be	-648	-774	-222	143	279	239	294
S_18	-64	370	20	540	-59	bear	565	-505	832	-615	402	104	-435
S_19	25	319	354	74	4	beauty	-149	68	-423	90	-266	216	90
S_2	-136	202	-196	381	197	but	250	104	-174	-182	43	187	-250
S_20	-135	-10	-195	50	211	by	-61	293	365	-100	223	270	-430
S_3	-307	34	70	-208	381	can	386	740	495	933	-314	319	340
S_4	-741	-612	-237	-750	-683	change	-114	-203	218	86	810	634	-307
S_5	104	9	-1052	167	-837	d	-72	-35	-486	246	-193	-188	298
						day	691	686	-59	-28	-488	-391	495
						death	-4	-1704	-87	1006	-179	-683	252

### 2- Plans factoriels

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

➤ Cliquer sur **PlaneView Research**

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.



- Sélectionner la rubrique "**Actives columns (variables) + rows (observations)**".

Une fenêtre destinée à sélectionner le plan factoriel suivant la paire d'axes souhaitée apparaît.

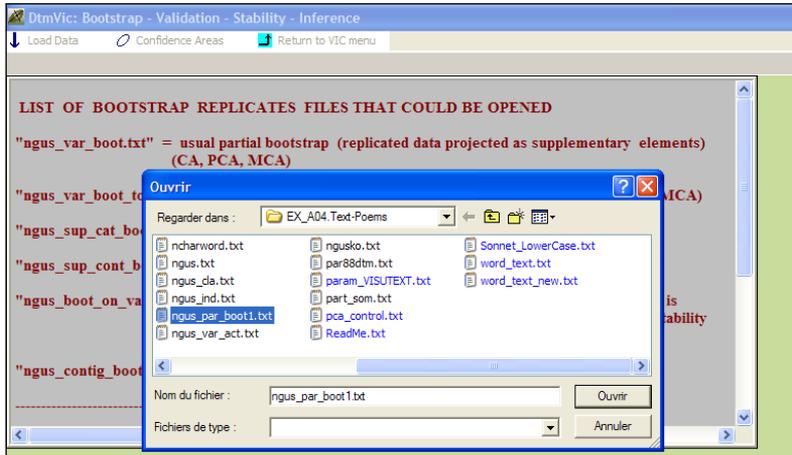
- Choisir les axes 1 et 2 puis cliquer sur : **Display**. Il est possible de ne faire figurer sur les plans que certaines variables. Cliquer alors sur : **Manual Selection of points**. Sélectionner les variables et les transférer dans la seconde fenêtre en cliquant sur : **Select**.

La fenêtre du plan factoriel apparaît.

**Rappel:** Pour chaque graphique, le bandeau du haut contient des options :

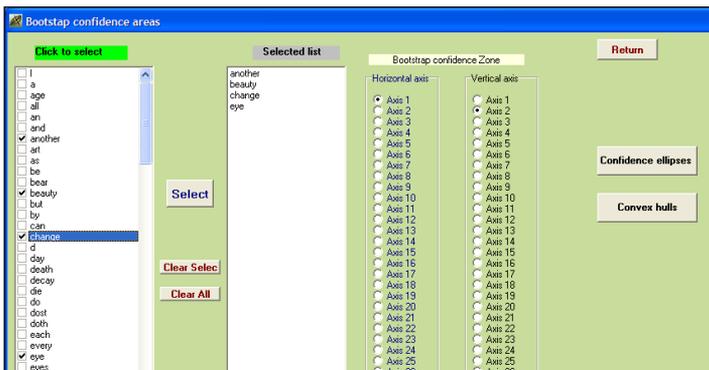
- « Sign of axes » permet d'inverser les axes ; « Zoom » possible (1,5 ; 2) ;
- « Font » offre la possibilité de modifier la police et la couleur des caractères ;
- « Rank », est utile seulement dans le cas des affichages très complexes, (ce qui n'est pas le cas ici) : ce bouton convertit les deux coordonnées 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 distorsion substantielle de l'affichage).
- « Axes color » change la couleur des axes ;
- « Save as bitmap » sauvegarde le graphique en format « .bmp » ;
- « Same scale » abandonne le cadrage sur la taille de l'écran pour donner la même échelle aux deux axes.





- Puis cliquer sur : **Confidence Areas**.

Une fenêtre : "*Bootstrap confidence areas*" s'affiche



- Sélectionner 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 : **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" (ellipse allongée sur la droite) a une position typée sur le premier axe (et neutre sur le second).





- Choisir la carte "map 5x5" (25 cellules) puis **Continue** et répondre OK à la boîte de message : "SOM map completed".

Une nouvelle fenêtre s'affiche.

- Actionner **Draw**. La Carte de Kohonen apparaît.

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.

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.

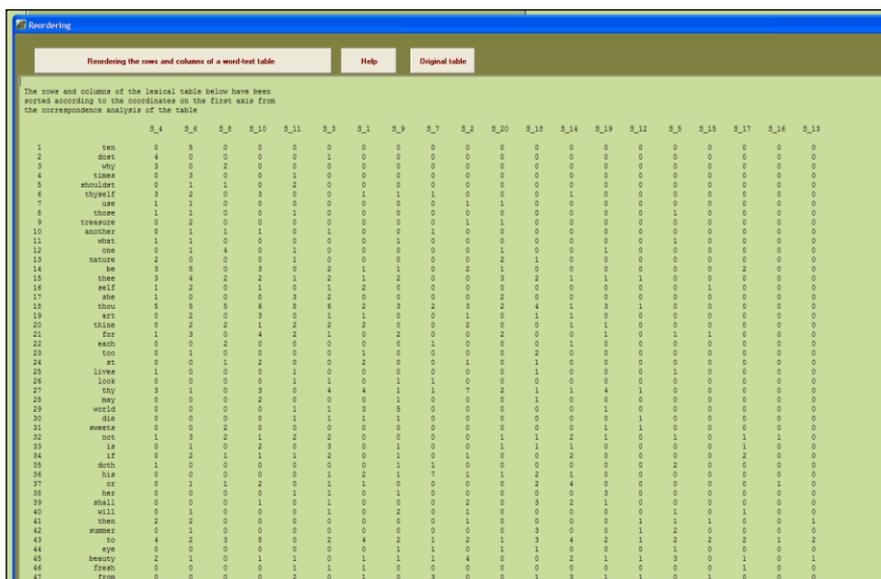
## 5- Sériation

(Voir l'encadré du paragraphe 1.3 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).

- Cliquer sur  **Seriation**.
- La fenêtre "Reordering" apparaît.
- Cliquer sur **Reordering the rows and the columns of a word-text table**.
- Répondre **OK** à la boîte de message: "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	start	4	0	0	0	1	0	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	time	0	3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	shoulder	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	thyeif	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	1	1	0	0	0	0	1	0	0	0	0	0
8	thone	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	1	1	0	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	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	one	0	1	4	0	1	0	0	0	0	0	1	0	0	0	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	se	3	8	0	3	0	2	1	1	0	2	1	0	0	0	0	0	0	2	0	0
15	three	3	4	2	2	1	2	1	2	0	0	3	2	1	1	1	0	0	0	0	0
16	shell	1	2	0	1	0	1	2	0	0	0	0	0	0	0	0	0	0	1	0	0
17	thou	5	5	5	6	8	6	2	3	2	3	2	4	1	3	1	0	0	0	0	0
18	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	0	2	0	0	1	1	0	0	0	0	0
21	far	1	3	0	4	2	1	0	0	2	0	0	2	0	0	0	1	0	0	4	2
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	er	0	1	2	0	0	0	0	0	1	0	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	0	1	0	0	0	1	0	0	0	0	0	0	0
29	world	0	0	0	0	1	1	0	0	0	0	0	0	0	0	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	newce	0	0	2	0	0	0	0	0	0	0	0	0	0	0	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	0	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	1	2	1	0	7	1	1	2	1	0	0	0	0	1	0	0
37	or	0	1	1	2	0	1	1	0	0	0	0	2	4	0	0	0	0	0	0	1
38	her	0	0	0	0	1	1	0	1	0	0	0	0	0	0	3	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	2	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0
41	thou	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	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	so	4	2	3	5	0	2	4	2	1	2	1	3	4	2	1	2	2	1	2	0
44	eye	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	1	0	0	0
45	honesty	2	1	0	1	1	0	1	0	1	4	0	0	2	1	1	3	0	1	0	1
46	fresh	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0
47	from	0	0	0	0	2	0	3	0	3	0	0	1	3	1	1	0	0	1	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.

On remarque que les identificateurs des textes (sonnets) [colonnes du tableau réordonné] sont tronqués aux quatre premiers caractères. Il est donc important que ces 4 premiers caractères puissent suffire à identifier les textes (des améliorations sont prévues).

\*

\* \*

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.

Enfin, le fichier texte aurait pu être lemmatisé plus automatiquement, en utilisant le bouton « **Lemmatizing Texts** » de la rubrique « **Dtmvic Tools** » du menu principal, qui fait appel au logiciel WinTreeTagger. Des boutons d'aide assez détaillée (en Anglais, Espagnol, Français, Italien) sont disponibles lors de l'exécution de cette procédure à laquelle est consacrée la section V.5 du chapitre V.

## 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<sup>9</sup>. 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** 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).

---

<sup>9</sup> 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.

## 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	EDUM MEDIUM
MALE MALE	EDUH HIGH
FEMA FEMALE	3 WILL_PEOPLE_BE_HAPPIER?
12 AGE_CODE	HAP1 Happier
AGE1 18_19	HAP2 LESS_happy
AGE2 20_24	HAP3 About_the_same
AGE3 25_29	4 PEOPLE_PEACE_OF_MIND...
AGE4 30_34	PEA1 INCREASES
AGE5 35_39	PEA2 DECREASES
AGE6 40_44	PEA3 NOT_CHANGES
AGE7 45_49	PEA4 OTHER
AGE8 50_54	3 MORE_OR_LESS_FREEDOM
AGE9 55_59	FRE1 MORE_FREEDOM
AG10 60_65	FRE2 LESS_FREEDOM
AG11 65_70	FRE3 THE_SAME
AG12 71_et_+	3 Age_3_classes
0 AGE	-30 less_than_30
3 EDUCATION	3055 from_30_to_55
EDUL LOW	+ 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<sup>10</sup>.

Des zones de confiance *bootstrap* pourront être dessinées autour des mots et des catégories d'individus.

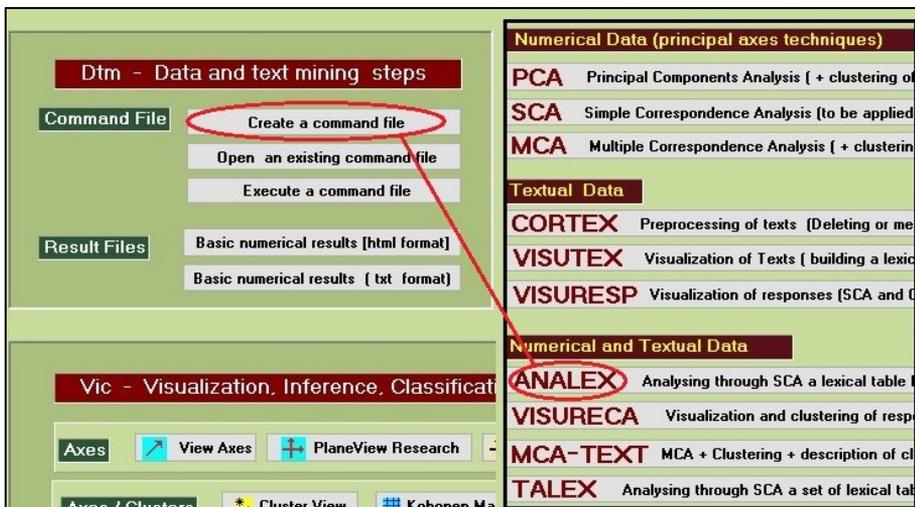
### III.2.2. Mise en œuvre de l'analyse textuelle sur tableau lexical agrégé – ANALEX

Le fichier paramètre est créé en 5 étapes :

#### Etape 1 : Sélection de l'analyse

➤ Dans le *menu principal*, cliquer sur : **Create a command file**

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



➤ Sélectionner 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.

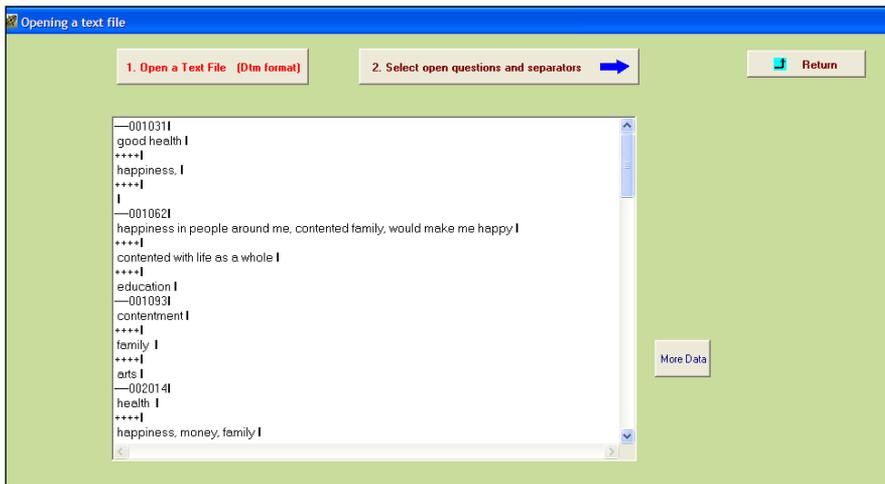
#### Etape 2 : Sélection du fichier texte

➤ Cliquer sur le bouton : **Open a text File**. Dans le répertoire **EX\_A05.Text-Responses**, ouvrir le fichier : **TDA\_tex.txt**.

<sup>10</sup> 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). Voir aussi, à propos d'exemples similaires, « *Statistique Textuelle* » (L. Lebart, A. Salem), téléchargeable à partir de [www.dtmvic.com](http://www.dtmvic.com).

Une boîte 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.

- Cliquer sur : **OK**, le fichier texte en format Dtm-Vic de type 2 s'affiche dans une première fenêtre.

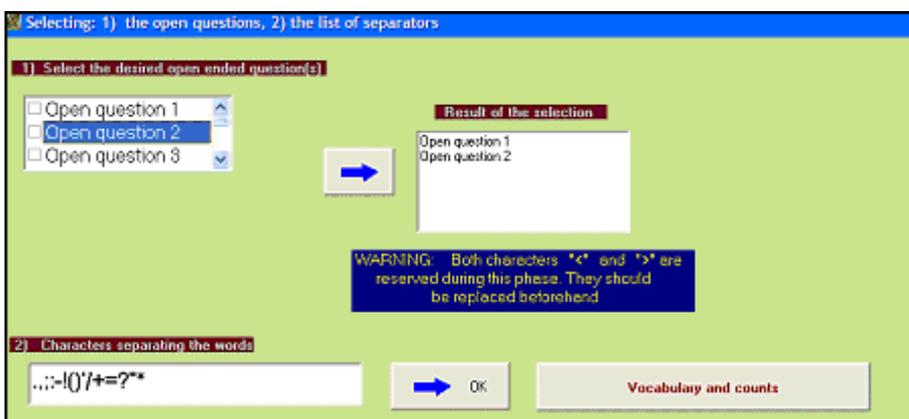


- Cliquer sur : **2. Select Open questions and separators**

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

### Etape 3 : Sélection des questions ouvertes

- Sélectionner les questions ouvertes 1 et 2 et les transférer dans "Result of the selection" (la question 2 étant une relance de la première, on peut en effet agréger les deux réponses). Puis choisir les séparateurs. Ici, nous adoptons ceux proposés par défaut. Cliquer alors sur **Vocabulary and counts**.



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

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

- Sélectionner cette ligne puis : **CONFIRM**. La fréquence apparaît. Répondre **OK**

Vocabulary, frequency threshold

Separators of units: .,:-!()'/\*=?\*  
 Number of occurrences (tokens): 13919  
 Number of words (types): 1365

Return

Vocabulary: Alphabetic order

666	1	1
667	1	100
668	1	14
669	1	12
257	6	2
670	1	3
671	1	30
672	1	6
9	286	I
673	1	If
674	1	Improving
675	1	Independence
676	1	Indoor
472	2	Ireland
473	2	It
8	300	a
296	5	ability
44	55	able
677	1	abled
70	31	about
398	3	above
474	2	abroad
678	1	absence
475	2	abuse

Vocabulary: Frequency order

132	16	long
133	16	make
134	16	own
135	16	worries
136	15	ne
137	15	personal
138	15	relationship
139	15	social
140	14	am
141	14	marriage
142	14	or
143	14	sufficient
144	14	together
145	14	without
146	13	animals
147	13	got
148	13	know
149	13	making
150	13	now
151	13	old
152	13	one
153	13	order
154	13	parents
155	13	religion

1. Choose a frequency threshold

CONFIRM

2. Continue (create the parameter file) →

- Cliquer sur **2. Continue (create the parameter file)**.

Une fenêtre d'ouverture "fichiers dictionnaires et données" apparaît.

Selecting dictionary and data

Return

1. Open a dictionary (Dtm format)

```

2 GENDER
MALE MALE
FEMA FEMALE
12 AGE_CODE
AGE1 18_19
AGE2 20_24
AGE3 25_29
  
```

List of variables (check)

1	GENDER	(2 categories)
2	AGE_CODE	(12 categories)
3	AGE	(numerical)
4	EDUCATION	(3 categories)
5	S1_CHANGE_IN_THE_STANDARD_OF_L	(5 categories)
6	S2_CHANGE_IN_YOUR_STANDARD_OF_L	(5 categories)

2. Open a Data File (Dtm format)

```

_1_ 1 12 80 1 2 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 98 2 2 1 3 1 1 1 2 5 5 5
_6_ 1 12 80 1 2 3 4 2 2 3 3 3 1 3
_7_ 2 7 46 2 4 3 0 0 2 1 2 5 5 5
  
```

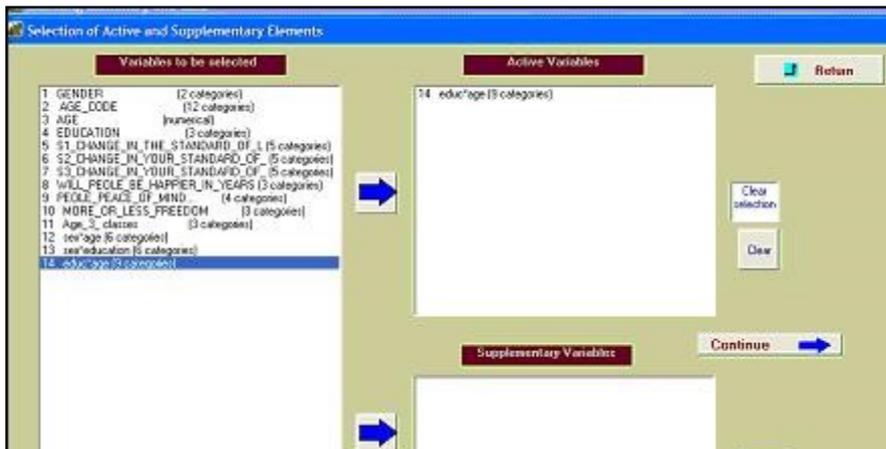
More Data

3. Continue (select active and supplementary elements) →

### Etape 4 : Sélection des fichiers dictionnaire et de données

- Cliquer sur le bouton : **Open a dictionary**. Dans le répertoire **EX\_A05.Text-Responses**, ouvrir le fichier **TDA\_dic.txt**. Il s'affiche dans une fenêtre. Le statut (nominal ou numérique) des variables est indiqué dans une deuxième fenêtre

- Cliquer sur le bouton : **Open a Data File**. Dans le répertoire **EX\_A05.Text-Responses**, ouvrir le fichier **TDA\_dat.txt** (troisième fenêtre).
- Cliquer sur : **3. Continue** ➔



### **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 :

Une fenêtre : " Selection of active et supplementary elements " apparaît.

Elle comprend trois sous-fenêtres :

- "Variables to be selected" où figure l'ensemble des variables
- "Active Variables" qui reçoit les variables actives sélectionnées
- "Supplementary Variables" qui reçoit les variables sup-plémentaires.

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". En effet, dans le cas d'ANALEX, les classes de la procédure ClusterView seront les catégories (qui sont des agrégats d'individus).

- Cliquer sur : **Continue** ➔

Une fenêtre : "Selecting observations" apparaît.

### **Etape 6 : Sélection des observations (individus)**

Trois cas de figure sont possibles :

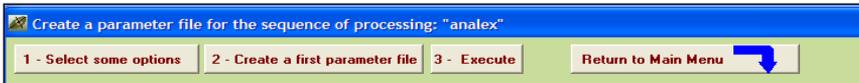
1. Considérer l'ensemble des observations.
2. Sélectionner les observations sur une liste.
3. Sélectionner les observations par un filtre.

Nous considérons ici l'ensemble des observations.

- Cliquer sur: **All the observations will be active**

Une fenêtre : "Create a starting parameter file" apparaît.

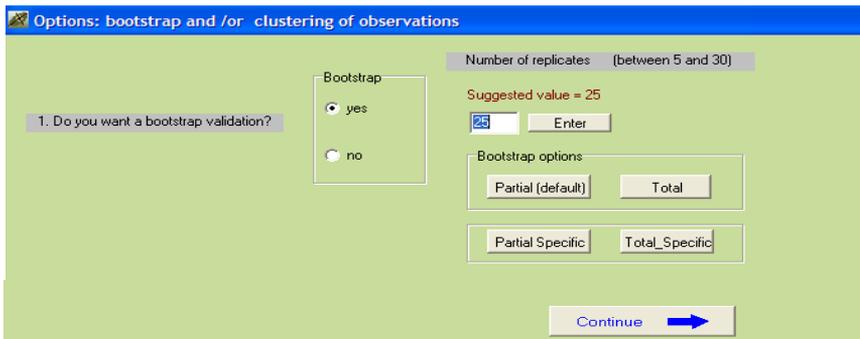
### Etape 7 : Création du fichier paramètre



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 dans les plans factoriels.

- Cliquer sur **1-Select some options**

Une fenêtre: "Options : Bootstrap and/or Clustering of observations" apparaît.



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

- Cliquer sur : **Continue** ➔

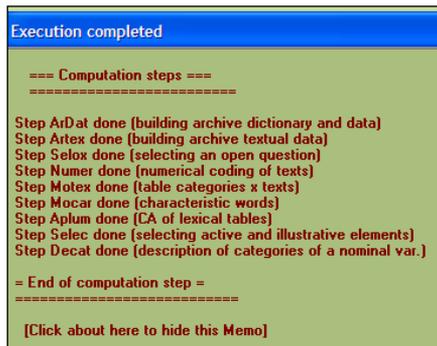
La fenêtre : "Create a starting parameter file" réapparaît.

- Cliquer 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-Exemples\_A\_Start**.

- Cliquer sur **3-Execute**

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



```

Execution completed

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

= End of computation step =
=====
[Click about here to hide this Memo]

```

**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", dans le menu principal et dans le 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.

- Cliquer 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	
<b>DtmVic: Main basic numerical results</b>	
Table of content	
<a href="#">Ardat (building archive dictionary and data)</a>	
<a href="#">Artex (building archive textual data)</a>	
<a href="#">Selox (selecting an open question)</a>	
<a href="#">Numer (numerical coding of texts)</a>	
<a href="#">Motex (table categories x texts)</a>	
<a href="#">Mocar (characteristic words)</a>	
<a href="#">Aplum (CA of lexical tables)</a>	
<a href="#">Selec (selecting active and illustrative elements)</a>	
<a href="#">Decat (description of categories of a nominal var.)</a>	
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 13919 mots dont 1365 mots distincts. Avec un seuil de fréquence de 16 (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.



**Rappel :** On peut accéder directement à tous les boutons de cette phase de visualisation **VIC** (pour une analyse exécutée antérieurement) à condition d'ouvrir simplement le fichier de commande, à partir du bouton « **Open an existing command file** ». Il n'est alors pas nécessaire de procéder à une nouvelle exécution, puisque tous les fichiers intermédiaires sont sauvegardés.

#### 1- Axes factoriels

- Cliquer sur  **ViewAxes**.

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" (catégories) et les "observations" (qui correspondent aux mots).

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

Active variables					Suppl. Categories					Individuals (ok)				
View					Exit					View				
Identifiant	axis 1	axis 2	axis 3	axis 4	Identifiant	axis 1	axis 2	axis 3	axis 4	axi...	axis 6	axis		
+55/high	-86	279	279	462	a	-112	-52	12	93	-57	56	61		
+55/low	305	-111	70	-14	able	-4	-127	87	-114	-27	96	101		
+55/medium	114	217	8	-71	about	160	-564	68	-208	-122	126	-68		
-30/high	-337	-377	219	-35	after	541	-79	-261	100	-75	1	59		
-30/low	-101	-209	-71	783	all	32	254	7	8	35	-61	-76		
-30/medium	-208	-149	-199	-29	and	43	-43	41	9	-29	19	59		
30-55/high	-296	104	268	-148	anything	405	-136	197	-128	226	-232	8		
30-55/low	39	115	-150	-12	are	317	135	26	-115	224	-171	-14		
30-55/medium	-131	177	79	23	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		

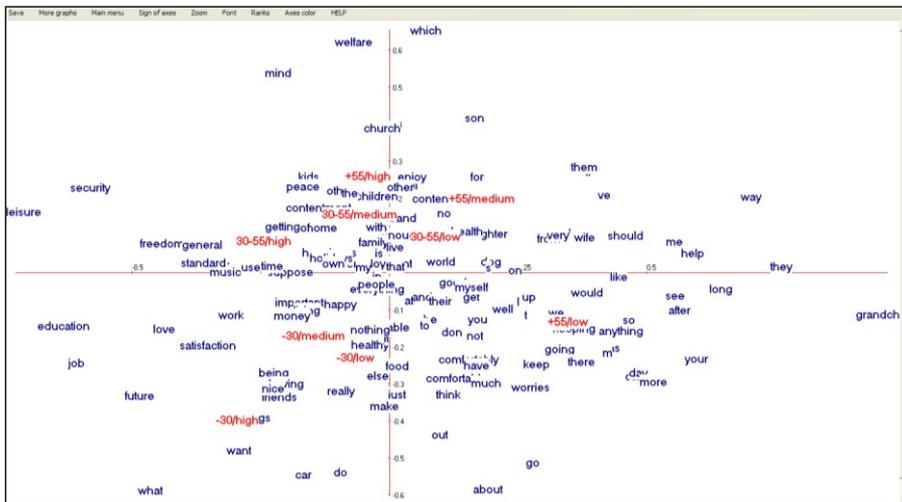
## 2- Plans factoriels

➤ Cliquer sur  **PlaneView Research**.

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.

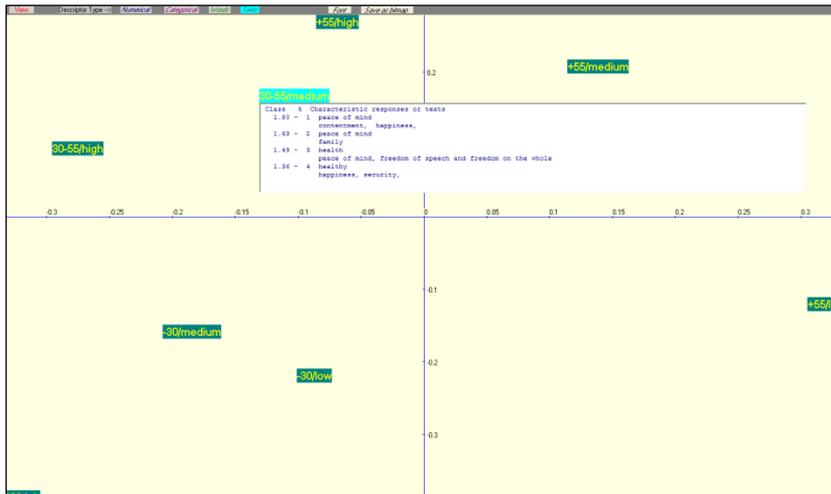
Apparaît alors une fenêtre pour sélectionner la paire d'axes souhaitée. Choisir les axes 1 et 2 puis cliquer sur **Display**. Le plan factoriel apparaît.



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).







**5- Carte auto-organisée : (Kohonen map)**

- Cliquer sur **Kohonen Map.**

Une première fenêtre "Selection of elements" apparaît.

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 any 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	we keeping going are	very them on	people from +55/high
you food church +55/medium	well way ve should our	your there me like grandchildre as anything +55/low	they	would much help day

Carte auto-organisée (de Kohonen). 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.

- Cliquer sur "**Rows + columns**"

Une fenêtre "Kohonen map or SOM map" apparaît.

- Choisir la carte "map 5x5" puis **Continue** et répondre **OK** à la boîte de message : "SOM map completed"

Une nouvelle fenêtre "Kohonen map" s'affiche

- Actionner **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).

## 6- Sériation

(Voir l'encadré du paragraphe 1.3 du chapitre 1)

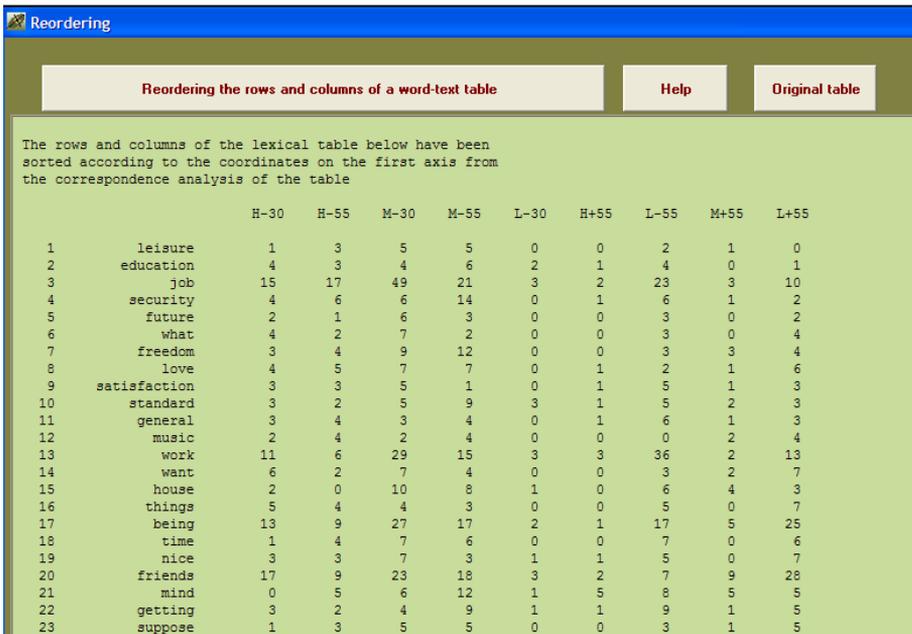
La sériation est appliquée ici à la table lexicale croisant les 9 catégories de répondants et les mots choisis (mots apparaissant au moins 16 fois dans le corpus). Dans cette version de Dtm-Vic, la sériation peut être obtenue seulement après les deux types d'analyse : VISUTEX et ANALEX. Ces deux approches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

- Cliquer sur  **Seriation**.

La fenêtre "Reordering" apparaît.

- Cliquer 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.



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

Tableau réordonné à la fois en ligne et en colonne

On peut lire sur ce tableau réordonné 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 (partie droite 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 aucun 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 (comme dans le cas de la procédure **ANALEX** de la section précédente).

La section III.2 a donné 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"

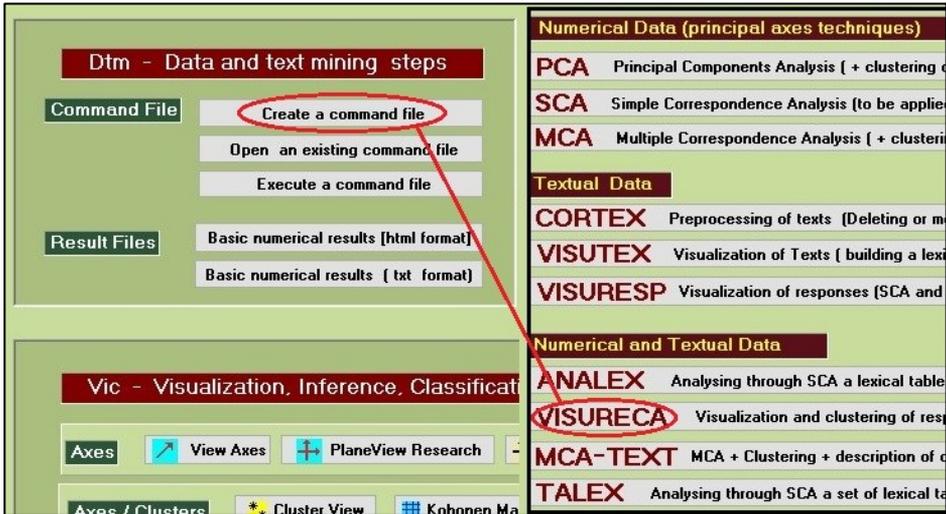
Le fichier paramètre est créé en 5 étapes :

##### **Etape 1 : Sélection de l'analyse**

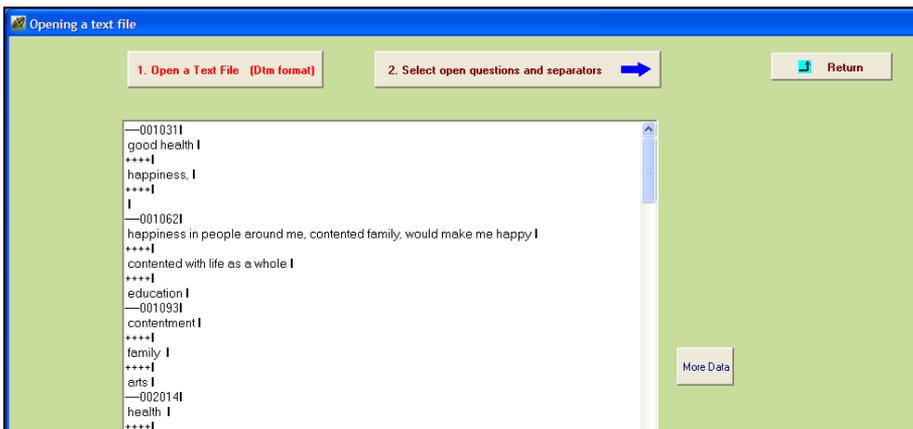
➤ Dans le *menu principal*, cliquer sur : **Create a Command file** de **Command File**.

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

➤ Sélectionner l'analyse **VISURECA** – Visualization and Clustering of responses with categorical data as supplementary elements dans la rubrique **Numerical and Textual Data**.



Une fenêtre : "Opening a text file" apparaît.



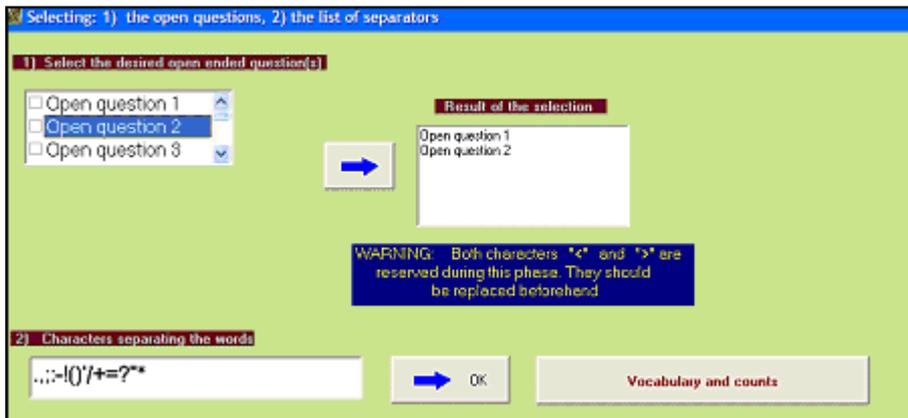
## Etape 2 : Sélection du fichier texte

- Cliquer 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 boîte 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.
- Cliquer sur : **OK**, le fichier s'affiche dans une première fenêtre.

Un deuxième bouton : **2.Select Open questions and separators** apparaît.

- Cliquer sur ce bouton.

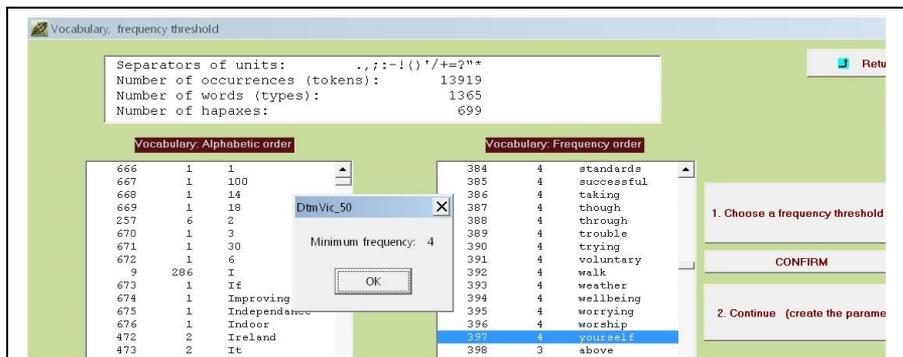
Une nouvelle fenêtre: "Selecting : 1) the open questions, 2) the list of separators" se présente.



### Etape 3 : Sélection des questions ouvertes

- Sélectionner 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. Cliquer alors sur **Vocabulary and counts**.

La fenêtre suivante présente le vocabulaire (alphabétique et 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 avons 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.

- Sélectionner cette ligne puis : **CONFIRM**. La fréquence apparaît. Répondre **OK**.

- Cliquer 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**

- Cliquer 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
- Cliquer 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 est la même que pour l'exemple II.2.
- Cliquer 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 les 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.
- Cliquer sur : **Continue →**

Une fenêtre : "Selecting observations" apparaît.

#### **Etape 6 : Sélection des observations (individus)**

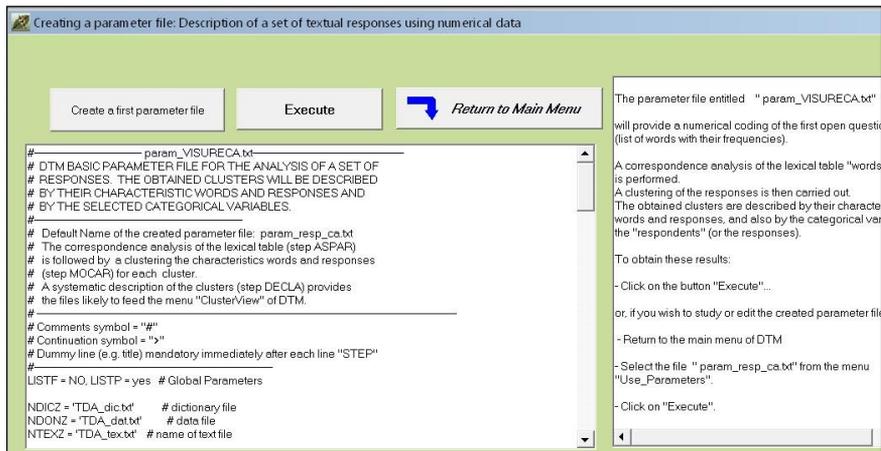
Nous considérons ici l'ensemble des observations.

- Cliquer sur: **All the observations will be active**

Une fenêtre : "Create a starting parameter file" apparaît.

#### **Etape 7 : Création du fichier paramètre**

- Cliquer sur : **2-Create a first parameter file**.
- Un fichier paramètre est 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, la validation *bootstrap* est réalisée par défaut. 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 `param_VISURECA.txt` (ou fichier paramètre) avant l'exécution, paramètres des étapes (STEP) "PARTI" et "DECLA"].

➤ Cliquer sur **Execute**.

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



**Affichage des étapes de calcul après l'exécution**

**Commentaires sur les étapes de calcul :**

**Ardat** (Archivage des données), **Artex** (Archivage des textes), **Selox** (sélection des questions ouvertes), **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

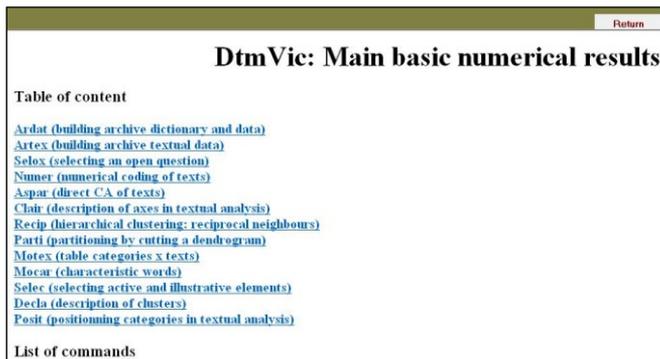
partition des individus), **Decla** (description automatique des classes à partir des variables supplémentaires nominales et continues), enfin **Posit** (positionnement des variables nominales supplémentaires dans les plans factoriels construits, rappelons-le, avec les mots des réponses aux questions ouvertes actives).

**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.

- Cliquer 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.







## 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 Dtm-Vic 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 (qu'il faudra sauvegarder en fichiers « textes ») ou en format ".csv" issu d'Excel pour les données numériques et textuelles, ou encore simplement en format texte (codes ASCII). Le software notepad++ (gratuit) permet de convertir des fichiers en Unicode ou UTF8 en format ANSI.

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 (un petit utilitaire, dans DtmVic-Tools [rubrique « Preprocessing texts »] permet de ramener la longueur des lignes à 200 caractères ou moins).

➤ Cliquer sur le bouton :

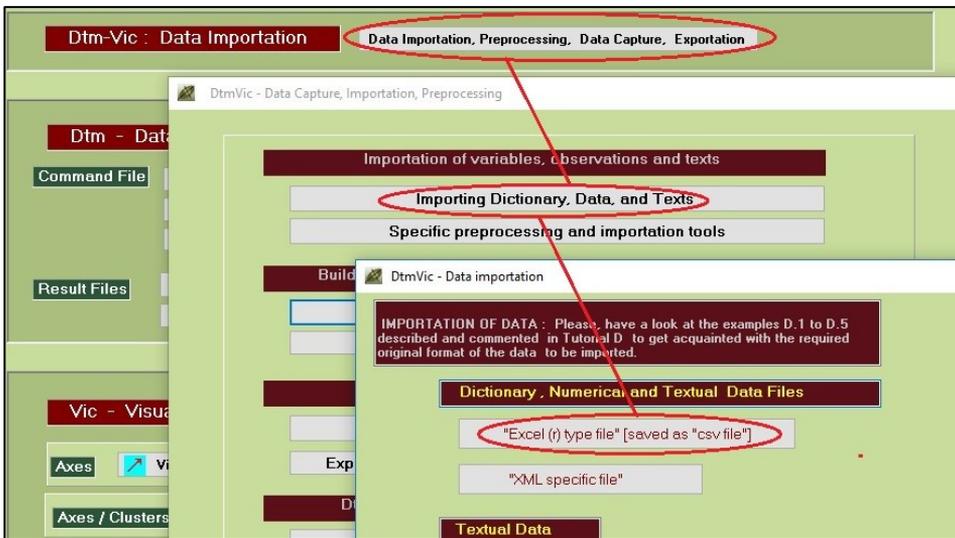
**Data Importation, Preprocessing, Data Capture, Exportation.**

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 dictionary 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.
- La procédure, **Exporting a DTM file to R or to Excel(r)** concerne l'exportation, alors que, dans le menu principal, **Dtm\_tools** permet les recodages et l'archivage des données.



Menu principal, fenêtre « Data Capture, Importation, Preprocessing », fenêtre « Data importation »

## IV.1. Importation de fichiers Excel<sup>®</sup>

### 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), une 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	people_be_happier?	people_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 daught	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, togetherr	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 c	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 colonne et la première ligne contiennent respectivement les identifiants 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.

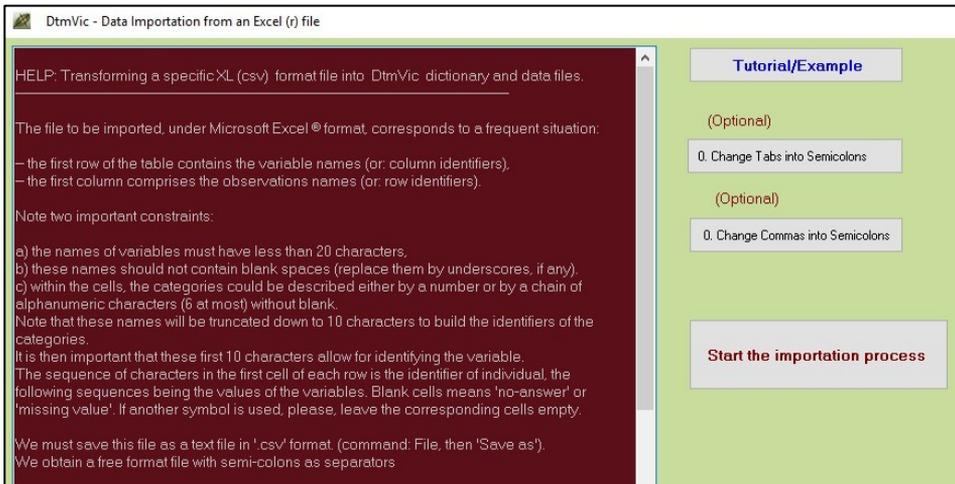
Dans certaines versions d'Excel, notamment les versions anglophones, le séparateur, pour le format ".csv", n'est pas le point virgule, mais la virgule. La procédure d'importation de DtmVic prévoit une possibilité de changement des séparateurs. De fait, tout comme les espaces vides, les points-virgules 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 remplacés par les points décimaux habituels dans les notations anglo-saxonnes et dans les langages de programmation.

## IV.1.2. Procédure d'importation

➤ Sélectionner, dans le menu principal, **Data Importation, Preprocessing, Data Capture, Exportation** puis **Importing Dictionary, Data and Texts** dans **Importation of variables, observations and texts**.

➤ Cliquer 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, cliquer 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].

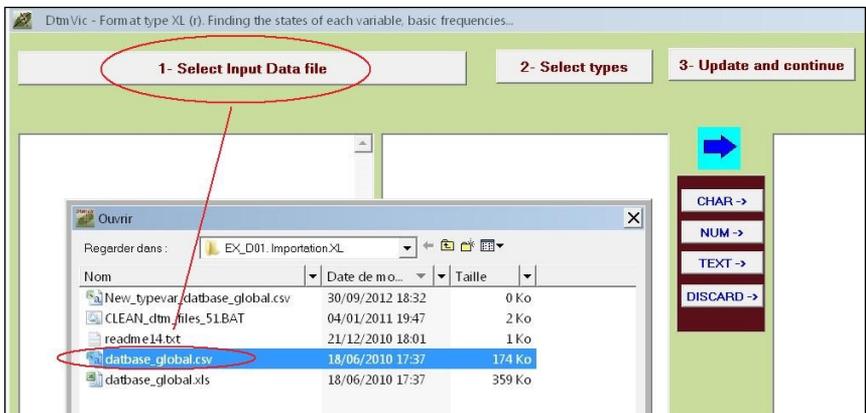
Dans ce cas, Sélectionner le fichier Excel sauvegardé avec des tabulations ou des virgules, et le convertir. Un nouveau nom est donné au fichier créé. Le procédé d'importation continuera d'employer ce nouveau fichier.

Dans tous les cas :

- Cliquer sur le bouton **Start the importation process**.

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

- Cliquer sur **1.Select Input Data file** et ouvrir 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.



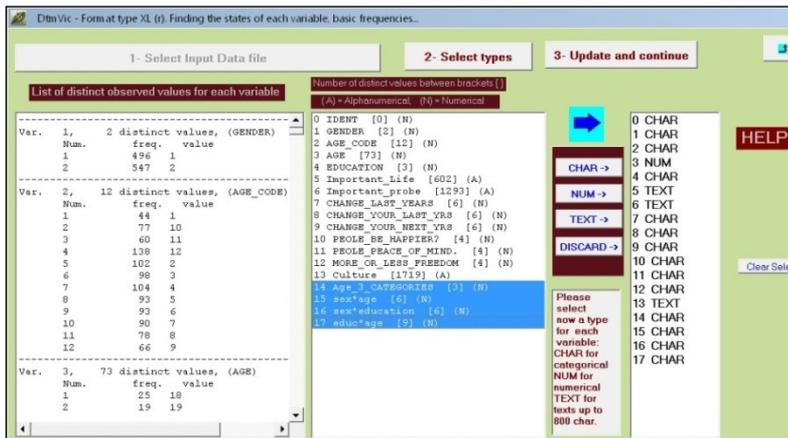
Le descriptif des variables s'affiche dans la fenêtre de gauche. Dans la fenêtre centrale, nous pouvons lire entre crochets le nombre de valeurs distinctes observées dans le fichier et entre parenthèses une lettre A ou N (cf. figure suivante).

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** : Sélectionner une ou plusieurs variables dans la liste de la fenêtre centrale puis spécifier 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)](cette variable peut avoir été codée sous forme numérique, comme un numéro de département ou une classe d'âge par exemple).
  - **NUM ->** pour variable numérique (ou continue, ici la variable 3-Age)
  - **TEXT ->** pour les variables textuelles (des textes) : les réponses aux questions ouvertes (variables 5, 6, 13).
  - **DISCARD ->** pour abandonner des variables.
- Une fois l'attribution du statut accompli, cliquer sur **3.Updating and continue** puis

répondre **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 Dtm-Vic, 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.

- Cliquer 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.

DtmVic - Format type XL (r). Finding the states of each categorical variable, basic frequencies...

1) Values and Counts (as a global check of the whole file content)      2) Create dictionary and data

total number of variables 17

```

0, IDENT, Char, 30, 1
1, GENDER, Char, 6, 1
2, AGE_CODE, Char, 6, 1
3, AGE, Num, 6, 1
4, EDUCATION, Char, 6, 1
5, Important_Life, Text, 8000, 1
6, Important_probe, Text, 8000, 1
7, CHANGE_LAST_YEARS, Char, 6, 1
8, CHANGE_YOUR_LAST_YRS, Char, 6, 1
9, CHANGE_YOUR_NEXT_YRS, Char, 6, 1
10, PEOPLE_BE_HAPPIER?, Char, 6, 1
11, PEOPLE_PEACE_OF_MIND., Char, 6, 1
12, MORE_OR_LESS_FREEDOM, Char, 6, 1
13, Culture, Text, 8000, 1
14, Age_3_CATEGORIES, Char, 6, 1
15, sex*age, Char, 6, 1
16, sex*education, Char, 6, 1
17, educ*age, Char, 6, 1

```

Var. 1, 2 distinct values, (GENDER)

Num.	freq.	value
1	496	1
2	547	2

Var. 2, 12 distinct values, (AGE\_CODE)

Num.	freq.	value
1	44	1
2	77	10
3	60	11
4	138	12
5	102	2
6	98	3
7	104	4
8	93	5
9	93	6
10	90	7
11	78	8
12	66	9

Var. 3, numerical, (AGE)

mean	sd	min	max
45.868	18.383	18.000	90.0

- Cliquer sur **Create dictionary and data**.

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

DtmVic - Creating a dictionary and a data file

Name for the new dictionary

Name for the new data file

Name for the new text file

Create new dictionary

Create data and text files

Create a DTM Parameter file

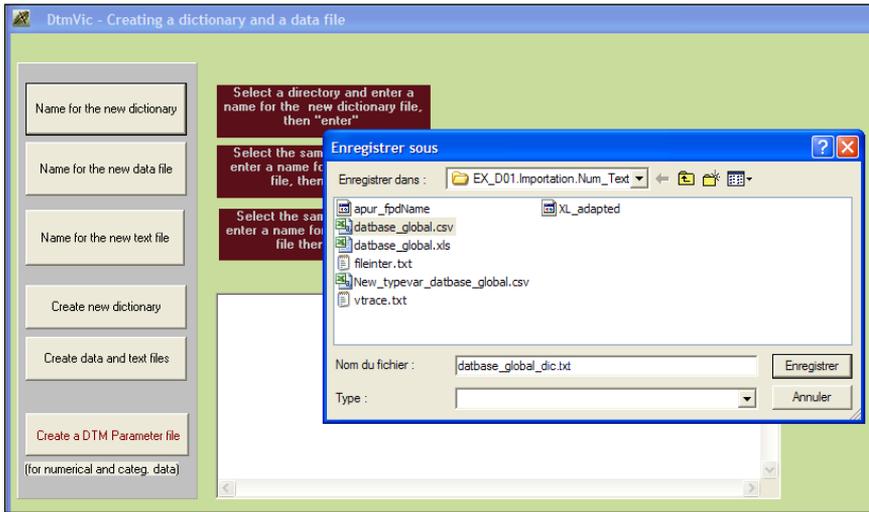
(for numerical and categ. data)

Select a directory and enter a name for the new dictionary file, then "enter"

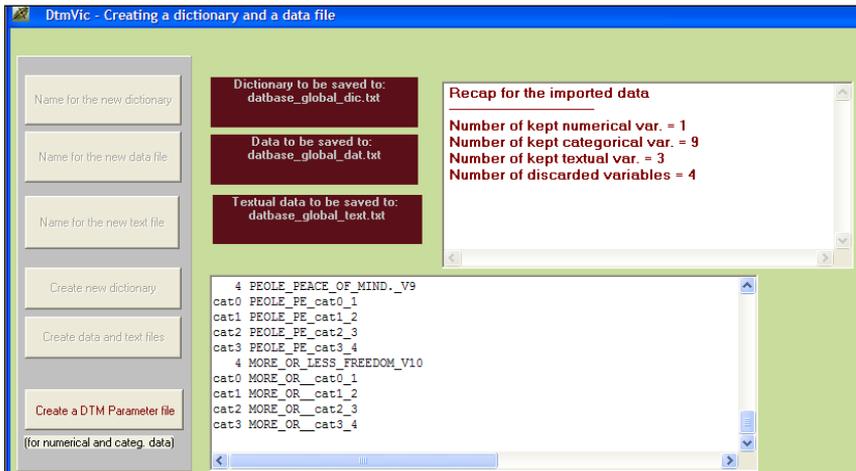
Select the same directory and enter a name for the new data file, then "enter"

Select the same directory and enter a name for the new textual file then "enter"

- Cliquer sur **Name for the new dictionary**. Entrer le nom du fichier dictionnaire **Database\_global\_dic.txt** (par exemple) et enregistrer.



- Cliquer ensuite sur **Name for the new data file**. Entrer le nom du fichier de données **Dabase\_global\_dat.txt** (par exemple) et enregistrer.
- Cliquer sur **Name for the new text file**. Entrer le nom du fichier dictionnaire **Dabase\_global\_text.txt** (par exemple) et enregistrer. S'il n'y a pas de données textuelles, passer à l'étape suivante.



- Cliquer 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 créé. Une boîte de message donne le nombre d'individus. Répondre **OK**. En cas de présence de questions ouvertes, cliquer sur **Create new text file**.

Un récapitulatif des données importées apparaît dans une nouvelle fenêtre.

L'importation est maintenant terminée. La suite est facultative.

- Cliquer sur le bouton **Create a DTM Parameter file**, pour obtenir des statistiques élémentaires uniquement sur les variables numériques et nominales.

Une fenêtre "create a first parameter file" apparaît sur l'écran.

- Cliquer 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.

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*. Il est automatiquement sauvegardé sous le nom de **param\_start.txt** dans le dossier de travail.

- Cliquer enfin sur **Execute**.

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

```
#
# DTM BASIC PARAMETER FILE : param_start.txt
#
# Comments symbol = "#"
# Continuation symbol = ">"
# Dummy line (e.g. title) mandatory immediately after each line "STEP"

LISTF = NO, LISTP = yes # Global Parameters

NDICZ = 'database_global_dic.txt' # dictionary file
NDONZ = 'database_global_dat.txt' # data file

STEP ARDAT # reading dictionary and data
===== builds the Archive Dictionary
NQEXA = 10, NIEXA = 1043, NXMOD = 12 >
NEDIT = 0, NIDI = 1 TEST = 999

STEP SELEC # Selection for STATS
```

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**

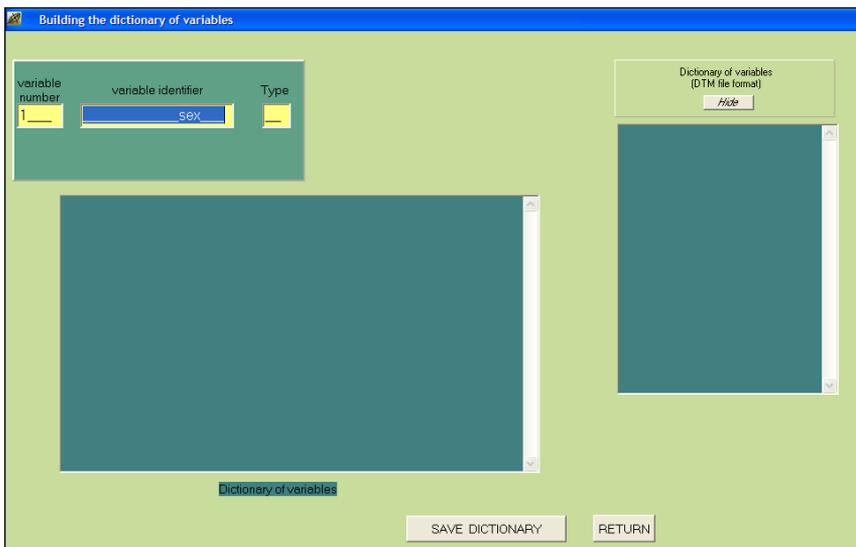
- Cliquer sur **Basic numerical results** (par exemple) pour ouvrir le fichier en format html puis sur **Return** pour en sortir et revenir au menu principal.

## 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

- Sélectionner, 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 1<sup>ère</sup> fenêtre jaune affiche : "1", le numéro de la 1<sup>ère</sup> variable à saisir. Dans la deuxième fenêtre, taper le nom de la variable puis dans la 3<sup>ème</sup> fenêtre donner le "Type" de la variable c'est-à-dire le nombre de modalités si la variable est nominale ou taper "0" si la variable est continue.
- Un bouton **ENTER** s'affiche à l'issue de la saisie du type de la variable. Si celle-ci est continue, continuer 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, cliquer sur **ENTER** (ou appuyer sur la touche "entrée"). Continuer 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. Étant 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.

Cliquer sur **ENTER** (ou presser la touche "Entrée") après chaque saisie.

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

On peut le nommer : `Database_dic.txt`. Cliquer ensuite sur **RETURN**.

## IV.2.2. Le fichier des données

Une fois le fichier dictionnaire créé :

- Sélectionner, **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.

- Cliquer sur **LOAD DICTIONARY** et ouvrir 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 1<sup>ère</sup> variable s'affiche dans la fenêtre.
- Sélectionner la modalité correspondant à l'individu avec le menu déroulant puis cliquer sur **Enter** (ou appuyer sur "Entrée" sur le clavier).

La 2<sup>ème</sup> variable s'affiche. Il s'agit de la saisir de la même façon. Une fois les variables capturées pour l'individu 1, 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 :

- sauvegarder le fichier en cliquant sur **SAVE DATA** et enregistrer le nom du fichier de données : **Database\_dat.txt** (par exemple) 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 (simples statistiques de base pour les données saisies). Les fichiers saisis (dictionnaire, données) sont prêts pour les analyses.

### IV.3. Exportation de fichiers de données numériques en format "Excel<sup>®</sup>" (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 (csv) ou vers un format très 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). En fait, la procédure R : « read.csv() » rend caduque cette option.

- Cliquer sur **Exportation dtm data** dans **Exporting a DTM file to R or to Excel**. Une fenêtre apparaît.



- Cliquer sur **Open a dictionary**. Ouvrir alors le fichier « MCA \_dic.txt », à titre d'exemple, dans « EX\_A03.MultCorAnalysis ».

Une première fenêtre affiche le libellé des variables et des modalités.

- Cliquer ensuite sur **Open a Data file** et ouvrir le fichier "MCA\_dat.txt" dans "EX\_A03.MultCorAnalysis". Puis cliquer sur **List of variables**.

Il est possible d'exporter soit en format Excel<sup>®</sup> 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 boîte de message: "New data file created".

Le nouveau fichier attendu : **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

- Cliquer sur le bouton **Data Recoding** dans le pavé **DtmVic-Tools** du menu principal. Le menu qui apparaît concerne le recodage des données et l'archivage de certains résultats.



#### ***Création ou recodage de variables nominales :***

- Regroupement de modalités ;
- Création d'une variable nominale par croisement de deux variables nominales ;
- Transformation d'une variable continue en variable nominale ;
- 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

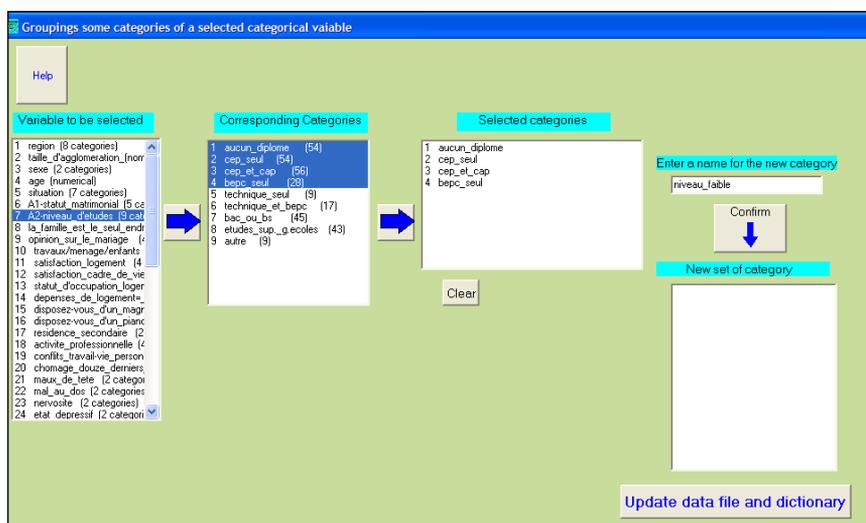
Lors du dépouillement de données d'enquête et à l'occasion de tris à plat effectués sur les variables nominales, on doit parfois regrouper certaines modalités d'une variable nominale pour satisfaire, dans la mesure du possible, certaines règles de recodage : éviter des modalités à faible effectif, équilibrer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

- Cliquer 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.



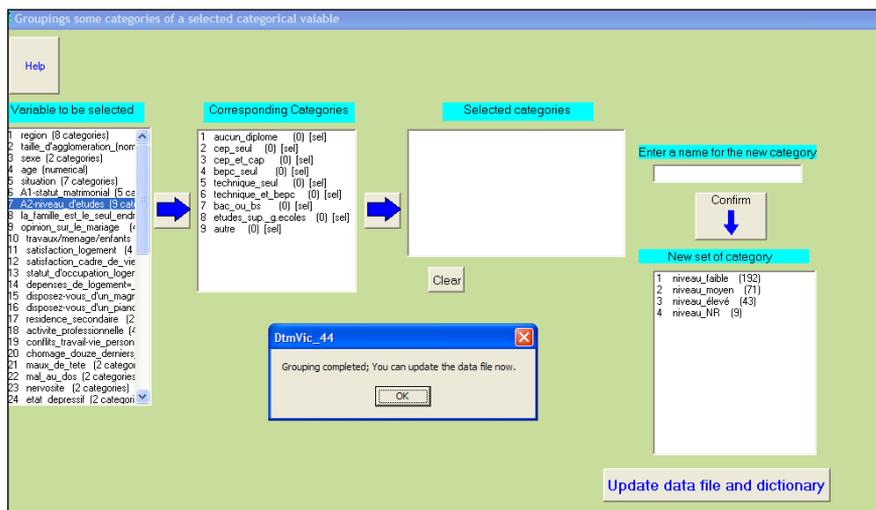
- Sélectionner la variable à recoder. Ici nous choisissons, dans la 1<sup>ère</sup> fenêtre, la variable "7-niveau d'étude" en 9 catégories. Les catégories (modalités) de cette variable s'affichent dans une 2<sup>ème</sup> fenêtre. Sélectionner l'ensemble des modalités à regrouper qui apparaissent dans une 3<sup>ème</sup> fenêtre. Entrer le nom de la nouvelle modalité dans la 4<sup>ème</sup> fenêtre puis confirmer. La nouvelle modalité apparaît dans la 5<sup>ème</sup> fenêtre.
- Recommencer 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 8<sup>ème</sup> modalité en "niveau\_élevé" et la 9<sup>ème</sup> en "niveau\_NR" (Non-réponse).

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

- Une fois les regroupements terminés, répondre : OK puis cliquer 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).

- Cliquer sur : **Return** . L'opération de regroupement des modalités est terminée.

## V.1.2. Croisement de deux variables nominales

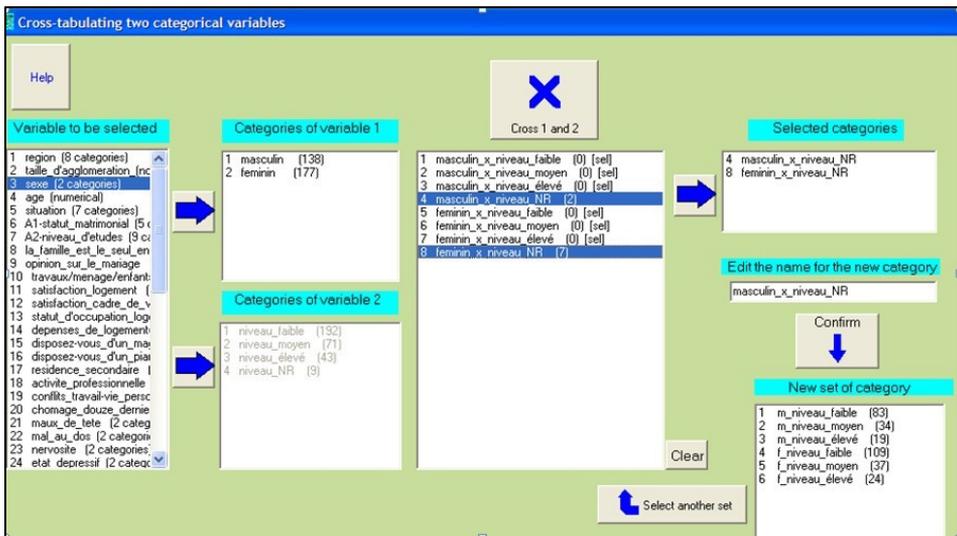
On souhaite dans ce cas augmenter les possibilités d'analyse et d'interprétation en créant une nouvelle variable nominale à partir du croisement de deux variables nominales (Exemple : sexe X âge).

- Cliquer 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.



Sélectionner les modalités à regrouper ou à valider qui apparaissent dans une 3<sup>ème</sup> fenêtre. En effet, un regroupement peut être nécessaire car certaines modalités issues du croisement peuvent correspondre à des effectifs trop faibles. Entrer l'étiquette de la nouvelle modalité dans la 4<sup>ème</sup> fenêtre puis confirmer. La nouvelle modalité apparaît dans la 5<sup>ème</sup> fenêtre.

Recommencer 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 cliquer 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.

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

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

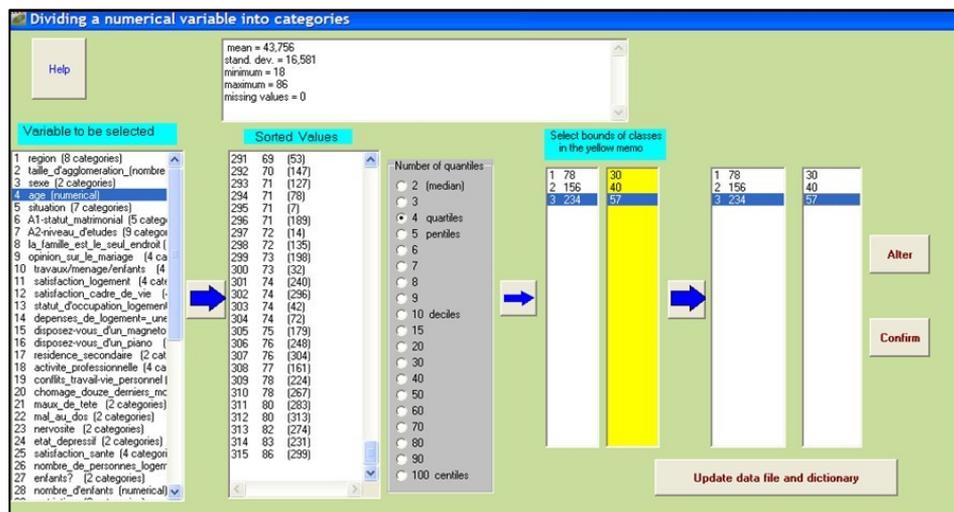
Cette procédure permet de transformer une variable continue en une variable nominale, en regroupant les valeurs numériques en classes. Ce regroupement en k classes se fait à partir d'un découpage préalable en n quantiles (n classes d'effectifs égaux), n étant beaucoup plus grand que k. Ce découpage est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

Cliquer 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.



- Sélectionner la variable continue (V4\_age) et transférer la dans la 2<sup>ème</sup> 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).

- Transférer en cliquant sur . Confirmer et répondre **OK** lors de l'affichage du nombre de modalités.
- Une fois les regroupements terminés, répondre **OK** puis cliquer 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. Cliquer 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 pas 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 en section V.2.1.

- Cliquer 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 (axe factoriel)

- 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**

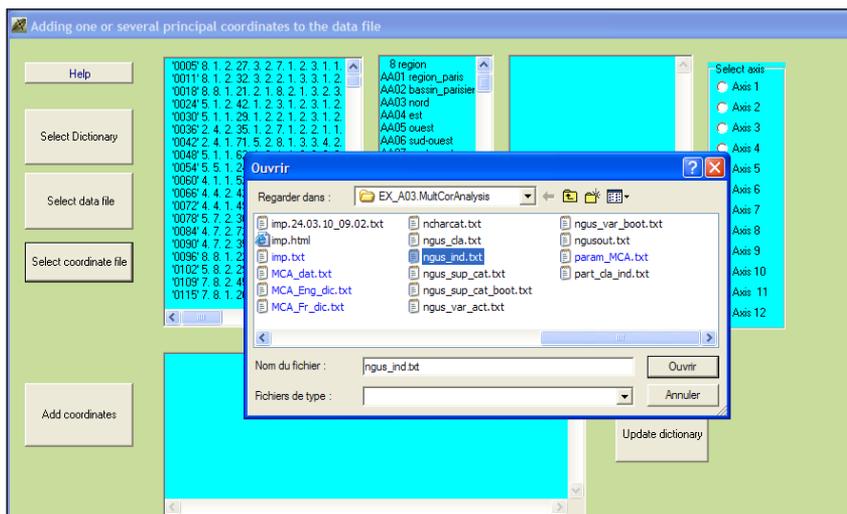
- ouvrir le fichier **ngus\_ind.txt**, puis Sélectionner l'axe à archiver.

Les coordonnées factorielles apparaissent dans la 3<sup>ème</sup> fenêtre.

- Cliquer sur **Add coordinates**.

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

- Cliquer sur **Update dictionary** et répondre **OK** dans la boîte 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

- Cliquer sur **Select partition file**

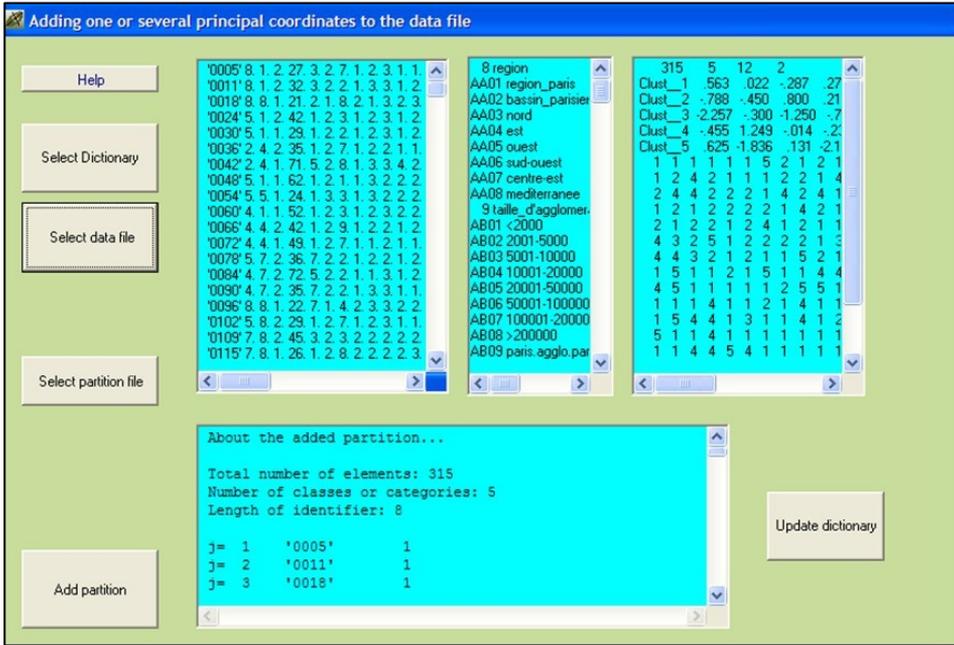
Une fenêtre du dossier : **EX\_A03.MultCorAnalysis** s'affiche où figure le fichier : **part\_cla\_ind.txt** du stockage automatique de la partition créée lors de la procédure : **MCA – Multiple Correspondances Analysis** et dont le nombre de classes a été spécifié lors du paramétrage de l'analyse.

- Ouvrir, 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).
- Cliquer 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.



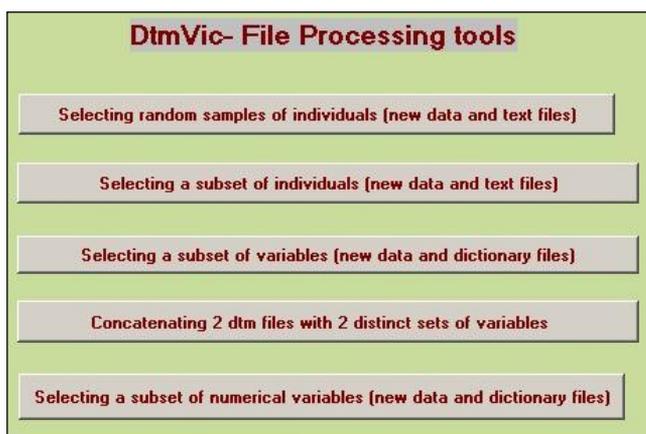
- Cliquer sur : **Update dictionary** et répondre : **OK** dans la fenêtre : "Dictionary update" qui s'affiche.

Les nouveaux fichiers du 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 le bouton **File Processing** dans le pavé **DtmVic-Tools** du menu principal.



- 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 détaillée ici. Elles comportent des rubriques « HELP » qui doivent faciliter la tâche des utilisateurs.

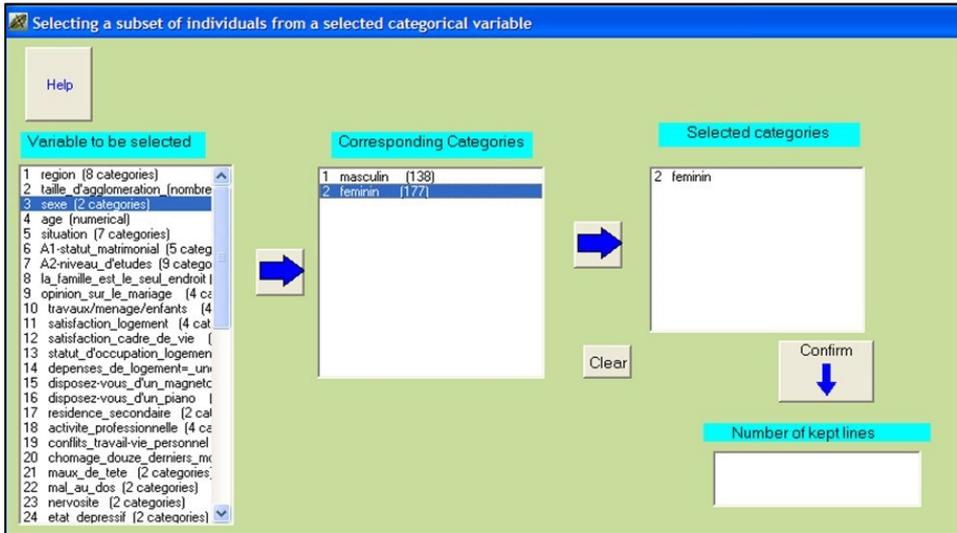
La section i) permet de diviser par 2 ou 4 la taille de l'échantillon de départ (formé de la réunion des 2 ou 4 groupes). Ceci permet de tester des analyses de façon plus économique, mais aussi de valider des structures observées.

La section v) est très particulière et répond à la situation pratique suivante : Si les données comportent un grand ensemble homogène 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 par le câble 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.

- Cliquer 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 fenêtre interne apparaît.
- Sélectionner la variable nominale dans la 1<sup>ère</sup> fenêtre (par exemple 3-Sexe), la transférer dans la 2<sup>ème</sup> fenêtre.
- Sélectionner la modalité de filtrage (par exemple "féminin").
- Cliquer 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.
- Cliquer sur **Update data file and text file**.

Un fichier dont le nom par défaut est : **dtm\_data\_Subset.txt** est créé 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

- Cliquer 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 1<sup>ère</sup> fenêtre l'ensemble des variables à conserver dans la nouvelle base, les transférer dans la 2<sup>ème</sup> 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.**

- Cliquer 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.
- Cliquer sur **Merge Sorted Files**.

Une série de fenêtres s'affichent successivement. Les deux premières précisent l'intégration des deux fichiers de données

Au message : « In file, 0 individuals have no counterparts » : répondre : **OK**.

Une troisième fenêtre donne le nombre d'individus du fichier créé : Répondre **OK**.

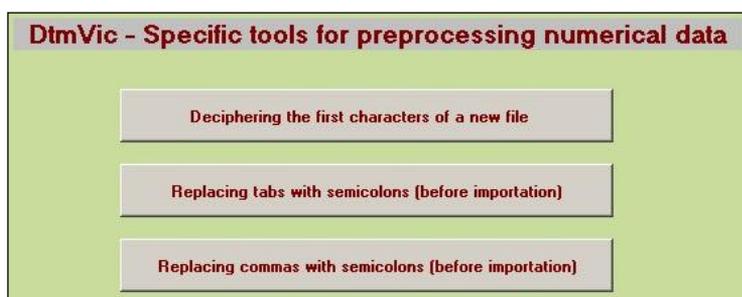
Enfin, une quatrième fenêtre indique que la procédure "merge" des deux fichiers de données est effectuée : répondre **OK**. Les identifiants des deux fichiers apparaissent dans la fenêtre du bas.

- Cliquer sur **Merge dictionaries**.

Une fenêtre indique que la procédure "merge" des dictionnaires est effectuée : répondre : **OK**, et cliquer sur : **Exit**. Deux fichiers `dtm_dico_new` et `dtm_data_new` sont alors créés. Ils sont prêts à être utilisés.

## V.3. Pré-traitement numérique

Le bouton **Preprocessing (numerical)** du pavé : **DtmVic-Tools** du menu principal 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.).

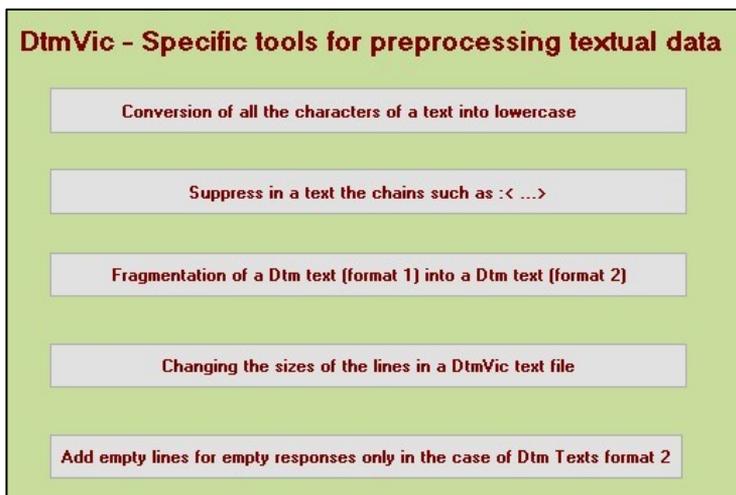
Le premier bouton "**Deciphering the characters of a new file**" nous donne le code ASCII correspondant aux 6000 premiers caractères d'un fichier, opération aussi utile (parfois) qu'élémentaire.

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.4. Pré-traitement textuel

- Le bouton **Preprocessing texts** du pavé **DtmVic-Tools** du menu principal 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 Lexico par exemple). Toutefois, ce type de transformation doit intervenir après que le texte ait été segmenté à partir de certaines balises.

### *iii) 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*).

#### **iv) 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).

#### **v) Addition de lignes vides quand nécessaire**

Enfin le dernier bouton « **Add empty lines for empty responses only in the case of dDtm texts of format 2** » 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 Lemmatisation d'un fichier texte de type 2 (type : fichier d'enquête).

## V.5. Lemmatisation

### **Lemmatisation avec WinTreeTagger d'un fichier de type Dtmic (type 1 ou 2) et ré-importation du fichier lemmatisé dans DtmVic .**

Le bouton : **Lemmatizing texts** 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. La phase d'analyse morpho-syntaxique fait appel au logiciel WinTreeTagger intégré dans l'interface de Dtm-Vic.

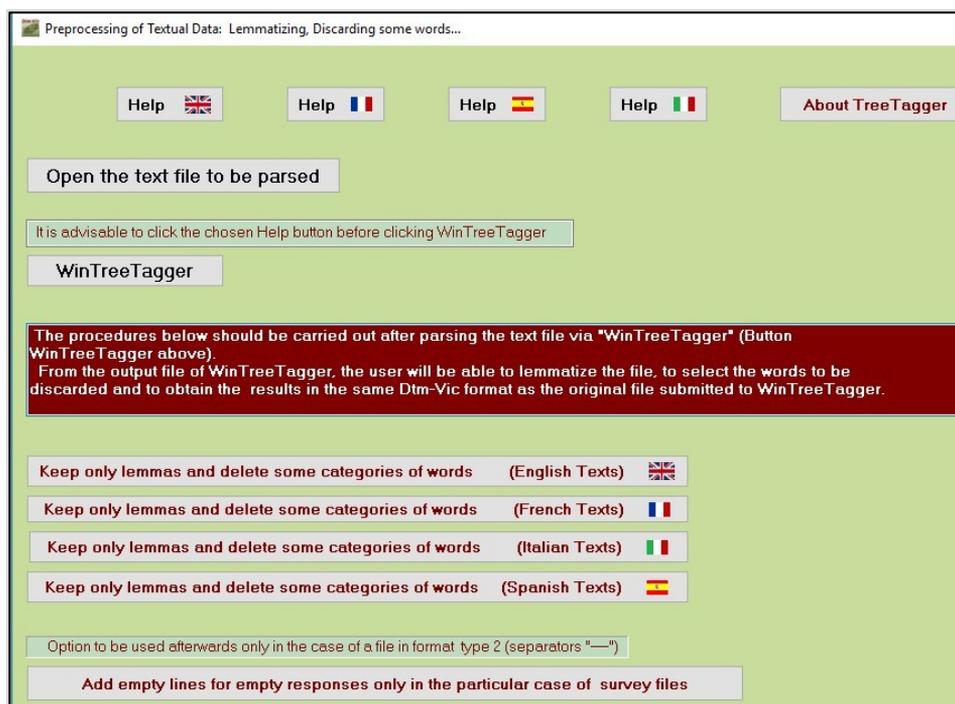
TreeTagger : Auteur: Helmut Schmid, IMS, University of Stuttgart, TreeTagger est un analyseur morpho-syntaxique indépendant des langues dans son principe. Les informations et le téléchargement se font à partir du site web:

<http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/>

On notera que TreeTagger n'a pas d'interface graphique. Comme suggéré par Helmut Schmid, on utilise l'interface Windows plus conviviale WinTreeTagger réalisée par Ciarn O'Duibhin.:

<http://www.smo.uhi.ac.uk/~oduibhin/oideasra/interfaces/winttinterface.htm>

Cliquant sur le bouton : **Lemmatizing texts** , une fenêtre apparaît :



On trouvera ci-dessous le contenu du bouton « **Help** » français.

Le fichier alimentant WinTreetagger *doit impérativement être un fichier texte au format Dtm-Vic*:

- Soit de type 1 (séparateurs: \*\*\*\*)
- Soit de type 2 (séparateurs: ---- avec aussi, dans le cas de plusieurs questions ouvertes: ++++)

Il faut ouvrir ce fichier (bouton "**Open**"). Il faudra l'ouvrir à nouveau lors de l'exécution de WinTreeTagger.

Cliquer sur le bouton « **WinTreeTagger** ». L'interface du logiciel apparaît.

Il faut ouvrir le fichier texte à traiter, donner le nom du fichier de sortie, et cocher l'option suivante "**Replace unknown words with the original tokens**".

Il importe en effet de conserver ces mots inconnus (pour TreeTagger). Sinon, WinTreetagger les remplace par le mot unique : "Unknown".

Le nom du fichier de sortie est, on l'a vu, à spécifier dans WinTreeTagger.

### Précaution très importante:

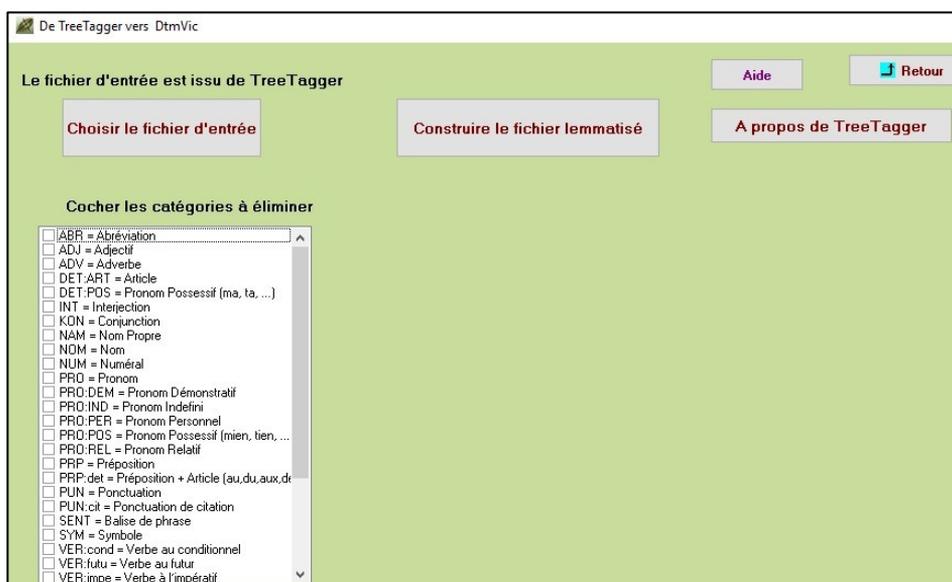
Pour que WinTreetagger ne détruise pas les éléments permettant de reconstituer la structure Dtm-Vic du fichier initial, il faut que les identificateurs de textes ne soient pas de simples nombres (qui seraient alors remplacés par un symbole unique). Ils doivent commencer par une lettre et être séparés par au moins un espace des séparateurs "----".

Il ne faut pas éliminer les noms ou les abréviations dont la suppression pourrait détruire des identificateurs.

Noter 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 réimporte en format Dtm-Vic.

Après exécution, fermer la fenêtre de WinTreeTagger, et cliquer sur le bouton :  
 « **Keep only lemmas and delete some categories of words** » (français)

La fenêtre suivante : *De TreeTagger vers DtmVic* apparaît (dans le cas français).



Le fichier à ouvrir avec le bouton « **Choisir le fichier d'entrée** » est le fichier de sortie de WinTreeTagger.

Il faut cocher les catégories à éliminer (articles, prépositions...) mais il faut garder à l'esprit que la lemmatisation et l'élimination des mots outils détruisent le contexte. Certains mots n'ont pas la même connotation au singulier et au pluriel, précédés ou non d'un article défini, au passé ou au futur pour les verbes.

Cliquer ensuite sur « **Construire le fichier lemmatisé** ».

Le fichier final du texte lemmatisé a pour nom : « Lemme\_*[nom du fichier d'entrée]* ».

# 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 Visualization, Inference, Classification 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 est 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**, analyse un ensemble de variables numériques ("données sémiométriques") par analyse en composantes principales. Les principaux axes sont complétés par une classification, avec description automatique des classes.

La procédure "Visualization" propose différents outils de visualisation (enveloppes convexes ou ellipses de densité pour les classes, tracé de l'arbre de longueur minimale (*Minimum Spanning Tree*) et 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 "Visualization".

### 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<sup>11</sup>. 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 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-Exemples\_C\_NumData/EX\_C01. PCA\_Semio**.

#### 1. le fichier de données : **PCA\_semio.dat.txt**

---

<sup>11</sup> Pour de plus amples informations, se référer à l'ouvrage : "La sémiométrie" par L. Lebart, M. Piron, J-F Steiner; Editeur: Dunod, Paris, 2003. Ce livre peut être téléchargé à partir du site: [www.dtmvic.com](http://www.dtmvic.com) (rubrique "Publications").

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 (au lieu des 210 mots du questionnaire sémiométrique). 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 la lecture de 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 a command file** du menu principal (étapes de base), comme indiqué au chapitre 2 de ce manuel. Une fenêtre "*Select a basic analysis*" s'affiche. Cliquer ensuite sur : **Principal Components analysis** situé dans la rubrique "**Numerical Data**", et suivre 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

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

Ensuite, rechercher le dossier **DtmVic-Examples\_C\_NumData** dans **DtmVic-examples**. Dans ce répertoire (ou dossier), ouvrir le répertoire **EX\_C01.PCA\_Semio**.

Ouvrir le fichier de paramètres: **EX\_C01\_Param.txt**. Le fichier paramètre (esquissé ci-dessous) s'affiche dans la fenêtre de l'éditeur de texte :

```
#----- Extraits du fichier de commande -----
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 du fichier paramètre sont cohérents avec ceux du répertoire. Neuf "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 sont 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).

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 proposé avec l'exemple.

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

## b. Exécution du fichier de commande (fichier paramètre)

- Cliquer sur : **Execute** de **Command File**

Les étapes de calcul de base présentes dans le fichier de commande sont exécutées : archivage de données et du dictionnaire, choix des éléments actifs et supplémentaires, statistiques élémentaires, analyse en composantes principales de la table sélectionnée, répliquations "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 :

- Cliquer 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 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  **ViewAxes** et  **PlaneView**, la validation des positions des points sur les graphiques par *Bootstrap*, avec :  **BootstrapView**, la classification avec :  **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)

- Cliquer sur le bouton  **Visualization**

Une fenêtre intitulée "DTM-visualization: loading files, selecting axes" apparaît.

- Cliquer 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.

- Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, Sélectionner 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*).

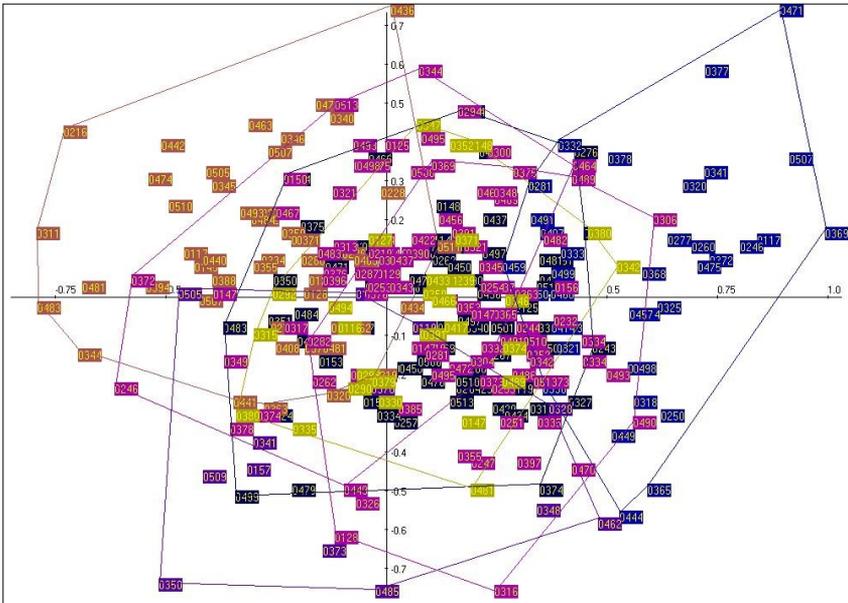
- Cliquer sur **Graphics** puis, dans la fenêtre "Sélection des axes", choisir les axes 2 et 3 (qui constituent le premier "plan sémio-métrique", car l'axe 1 est un "axe de notation" – voir l'ouvrage « La Sémiométrie » précité).

- Cliquer 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. Presser 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 (ces étapes se situent à la fin du fichier de commande. Noter que l'étape EXCAT n'est pas prévu dans les générations automatiques par menu des fichiers de commandes, et s'obtient directement à partir d'une édition du fichier de commande).

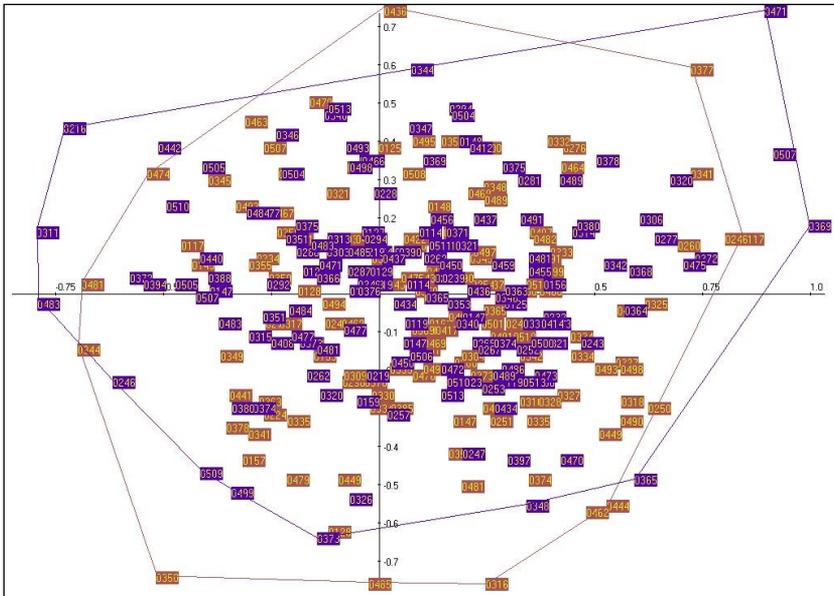
- Cliquer à nouveau sur  **Visualization**
- Dans la fenêtre intitulée "DTM-visualization: Loading files, Selecting axes", cliquer sur **Load coordinates**

Dans le sous-menu correspondant, choisir à nouveau le fichier: "**ngus\_ind.txt**". Les coordonnées des individus (lignes) sont sélectionnées.

- Cliquer ensuite sur **Load or create a partition**.

Dans le sous-menu correspondant, choisir le fichier "**part\_cat.txt**". La partition induite par les catégories de la variable 76 (sexe) est chargée.

- Cliquer sur **Graphics** puis choisir encore les axes 2 et cliquer sur **Continue** puis sur **DISPLAY**. Le Plan factoriel (2, 3) s'affiche.
- Cliquer 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)

- Cliquer sur **Visualization**

Une fenêtre intitulée "DTM-visualization: loading files, selecting axes" apparaît.

- Cliquer sur **Load coordinates**. Dans le sous-menu correspondant, choisir 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.

- Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, Sélectionner la partition obtenue précédemment à l'étape de calcul. Choisir alors **No partition**.
- 1 - Cliquer sur **Min. Span. Tree** (*Minimum Spanning Tree*). Choisir 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- Cliquer sur **N.N** (recherche de plus proches voisins [*Nearest Neighbours*] limité à 20 NN). Répondre **OK** à la recherche des plus proches voisins.
- 3- Cliquer sur **Graphics** puis choisir 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 cliquer sur **Continue** puis sur **DISPLAY**.

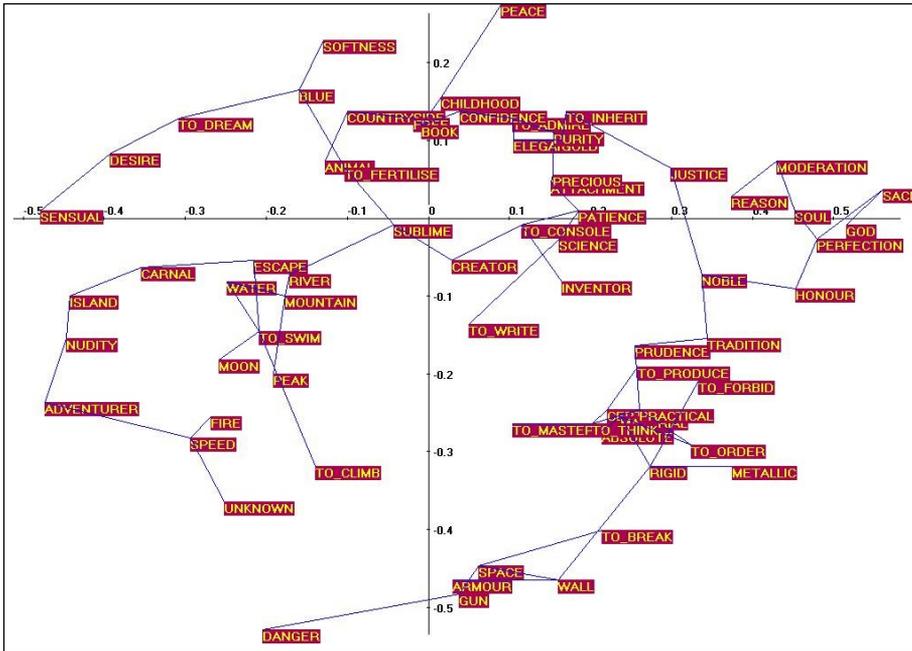
Le Plan factoriel (2, 3) s'affiche.

Dans le bandeau de gauche de la fenêtre "Graphics" figurent quatre familles de boutons :

- 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 ( $\leq 20$ ).
- 

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)

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.



#### 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" des variables (ou des individus).

- Cliquer sur **V Visualization**

La fenêtre intitulée "DTM-visualization: Loading files, Selecting axes" apparaît.

- Cliquer sur **Load coordinates**. Dans le sous-menu correspondant, choisir le fichier: **ngus\_var\_act.txt** pour une classification des variables actives ; Pour un regroupement d'individus, Sélectionner le fichier: **ngus\_ind.txt**.
- Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, Sélectionner l'option "**Create a new k-means partition**".
- Ensuite sélectionner (figure ci-après) le nombre de classes désirées, le nombre de coordonnées principales pour les calculs de distances, le nombre maximum d'itérations (généralement < 12 ) et cocher "**yes**" pour visualiser les itérations.
- 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**.



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).

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. Noter 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 (cf. section VII.8 de l'annexe statistique VII).

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).

\*  
\* \*

## 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  $M$  ( $m_{ij}=1$  si les sommets  $i$  et  $i'$  sont reliés par une arête,  $m_{ij}=0$  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 (G, codé par M) sur les données X.

Dans cet exemple, nous allons traiter la situation dans laquelle la structure du graphe G et la matrice M et 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 contiguïté. *Publications de l'ISUP*. XVIII, 81-112.

Lebart, L. (2000): Contiguity Analysis and Classification, In: W. Gaul, O. Opitz and M. Schader (Eds):*Data Analysis*. Springer,Berlin, 233--244.

Lebart L. (2006): Assessing Self Organizing Maps via Contiguity Analysis. *Neural Networks*, **19**, 847-854.

## VI.2.2 Les données "Iris" de Fisher / Anderson :

Pour les données numériques en format texte de Dtm-Vic, chercher le répertoire **DtmVic\_Examples**. Dans ce répertoire, ouvrir le dossier : **DtmVic\_Examples\_C\_NumData**. Puis ouvrir 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 mesures sont les longueurs des différents constituants des fleurs: Longueur et largeur des sépales, longueur et 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 – *courier*, par exemple - représente 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 a command file**, rubrique **Command File** du menu principal. 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

- Cliquer sur le bouton : **Open an existing command file** de la rubrique **Command File** (menu principal).
- Rechercher dans **DtmVic\_Examples**, **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).

Noter que le bouton: **Help about parameters** est également accessible à partir de cet éditeur de texte pour expliciter (en Anglais) les paramètres de chaque étape.

Dans ce fichier de commandes, on peut lire, après avoir identifié les deux fichiers (données NDONZ et dictionnaire NDICZ), 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 SELEC)

## b. Exécution du fichier de commande (fichier paramètre)

Revenir au menu principal et exécuter les étapes de calcul de base.

- Cliquer 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. 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  Visualization. Nous allons visualiser les différentes espèces de fleurs (variable n° 5) dans le plan engendré par les premiers axes principaux de l'ACP.

### a. Visualisation à partir d'une partition induite par une variable nominale (espèce d'iris)

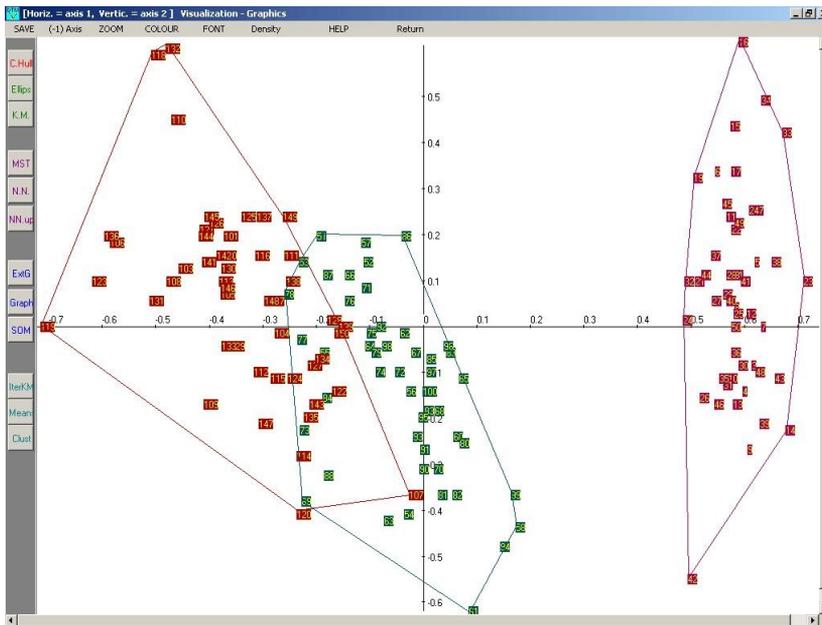
- Cliquer sur  **Visualization**

Une fenêtre intitulée "DTM-visualization ..." apparaît.

- Cliquer 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.
- Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, choisir 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.
- Cliquer sur **Graphics** puis choisir les axes 1 et 2 (par défaut) dans la petite fenêtre "Sélection des axes" et cliquer sur **Continue** puis sur **DISPLAY**.

Dans la nouvelle fenêtre intitulée "Visualization - Graphics" sont affichés les individus dans le plan des 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)



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.

- Le bouton **Density**, par souci de lisibilité, 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).

- Presser le bouton **C.Hull** (*Convex Hull* = enveloppe convexe) qui trace l'enveloppe convexe de chaque classe. Le tracé apparaît ci-dessous.

À cette étape, nous avons obtenu un affichage des 150 individus, avec les enveloppes convexes correspondant aux trois espèces.

C'est l'affichage classique (pour les statisticiens) 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.

➤ Cliquer sur **V Visualization**

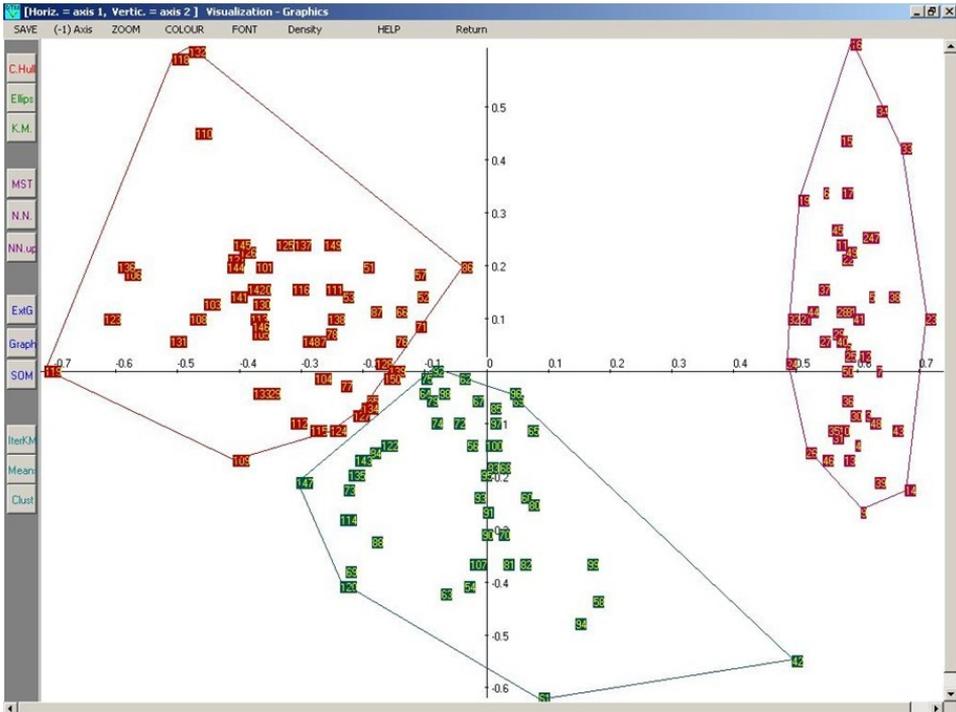
La fenêtre intitulée "*DTM-visualization...*" apparaît.

➤ Cliquer sur **Load coordinates**. Dans le sous-menu correspondant, choisir le fichier: **ngus\_ind.txt**. Les principales coordonnées des individus (lignes) sont sélectionnées.

➤ Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, choisir alors **Load partition File** et ouvrir le fichier **part\_cla\_ind.txt** (partition en 3 classes issue des phases RECIP et PARTI).

Après le chargement de cette partition, les trois dernières opérations précédentes (cf VI.2.5.a.1. à VI.2.5.a.3), c'est-à-dire les opérations "**Minimum Spanning Tree**", "**N.N**" et "**Graphics**", peuvent être effectuées à nouveau. Il est intéressant de visualiser les individus dans le plan engendré par les axes 1 et 2, avec les ellipses de densité des trois classes, ou encore, comme ci-dessus, les enveloppes convexes de ces classes.

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.



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.

Retour vers : [VIC : Visualization, Inference, Classification steps](#)

## VI.2.5. Analyse de contiguïté

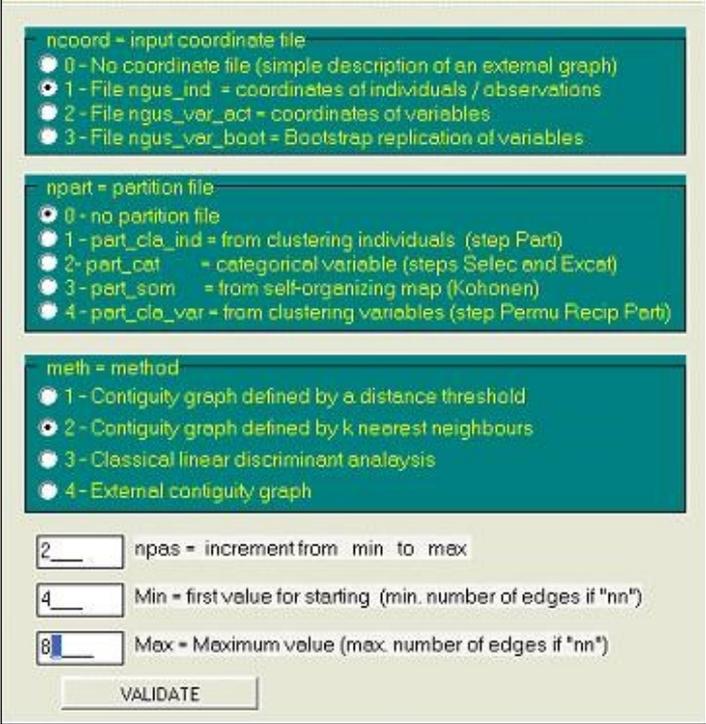
Deux analyses de contiguïté vont être exécutées. La première, non supervisée, utilise le graphe des plus proches voisins. C'est l'analyse de contiguïté intrinsèque. La seconde, supervisée, utilise le graphe formé de trois cliques disjointes correspondant aux trois espèces d'iris (tous les couples d'individus 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 toujours pas prise en compte. Il s'agit donc d'une approche non-supervisée.

- Cliquer sur le bouton :  **Contiguity**.
- Cliquer sur **Parameter/Edit**. Choisir l'élément **Create**

La fenêtre suivante apparaît.



The dialog box contains the following options:

- ncoord = input coordinate file**
  - 0 - No coordinate file (simple description of an external graph)
  - 1 - File `ngus_ind` = coordinates of individuals / observations
  - 2 - File `ngus_var_act` = coordinates of variables
  - 3 - File `ngus_var_boot` = Bootstrap replication of variables
- npart = partition file**
  - 0 - no partition file
  - 1 - `part_clo_ind` = from clustering individuals (step Part)
  - 2 - `part_cat` = categorical variable (steps Selec and Excat)
  - 3 - `part_som` = from self-organizing map (Kohonen)
  - 4 - `part_clo_var` = from clustering variables (step Permu Recip Part)
- meth = method**
  - 1 - Contiguity graph defined by a distance threshold
  - 2 - Contiguity graph defined by *k* nearest neighbours
  - 3 - Classical linear discriminant analysis
  - 4 - External contiguity graph

Below the sections are three input fields:

- npas = increment from min to max
- Min = first value for starting (min. number of edges if "nn")
- Max = Maximum value (max. number of edges if "nn")

A **VALIDATE** button is located at the bottom.

Nous allons établir les paramètres nécessaires à une analyse de contiguïté:

- Dans le premier bloc intitulé "**ncoord = Input coordinate file**", cocher "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**" cocher "0" (*no partition*)
- Dans le troisième bloc intitulé "**meth = method**" cocher "2" (*Contiguity graph defined by k nearest neighbours*).
- Ensuite, nous aurons à entrer les valeurs numériques suivantes :
- npas = 2 (incrément 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 ).

- Cliquer sur **VALIDATE**.

➤ Dans la barre supérieure de la fenêtre, cliquer sur **Execute**. Les calculs sont effectués.

La rubrique **Results** permet de consulter les détails techniques des calculs impliqués dans l'analyse de contiguïté.

➤ Cliquer ensuite sur **Contiguïty View**.

La fenêtre "Visualization : loading files, selecting axes" qui correspondait au bouton  **Visualization** apparaît.

➤ Dans le menu **Load coordinates** de la nouvelle fenêtre, ouvrir 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**.

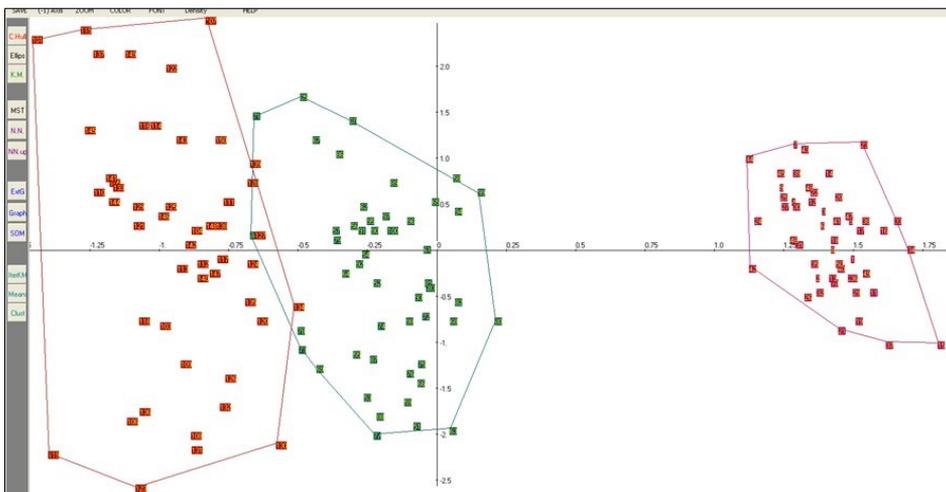
➤ Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu **Load partition File**, Sélectionner le fichier: **part\_cat.txt**. (Avec ce fichier, nous allons identifier les espèces). Nous ne pouvons pas calculer l'arbre de longueur minimale ("minimum Spanning Tree"), ni les plus proches voisins à partir du fichier : **ngus\_contig.txt**.

➤ Cliquer sur **Graphics**. Choisir ensuite les axes 1 et 2 (qui sont d'ailleurs les valeurs par défaut)

➤ Choisir (cocher) 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).

➤ Cliquer sur **DISPLAY**. Changer les couleurs, si nécessaire.

➤ Cliquer 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 (section b ci-dessous).

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 connues *a priori*, et utilisées par la méthode.

L'analyse de contiguïté réussit à séparer assez correctement les trois variétés d'Iris. Les excellents résultats sont dus 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 cliquer sur  **Contiguity**.
- Cliquer sur **Parameter/Edit**. Choisir l'élément **Create**.
- Cocher :
  - "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)

- Cliquer sur **VALIDATE**.

➤ Dans la barre supérieure de la fenêtre, cliquer sur **Exécute**.

➤ 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.

➤ Cliquer ensuite sur **Contiguïty View**.

La fenêtre "**Visualization : loading files, selecting axes**" correspondant en fait au bouton :  **Visualization** apparaît.

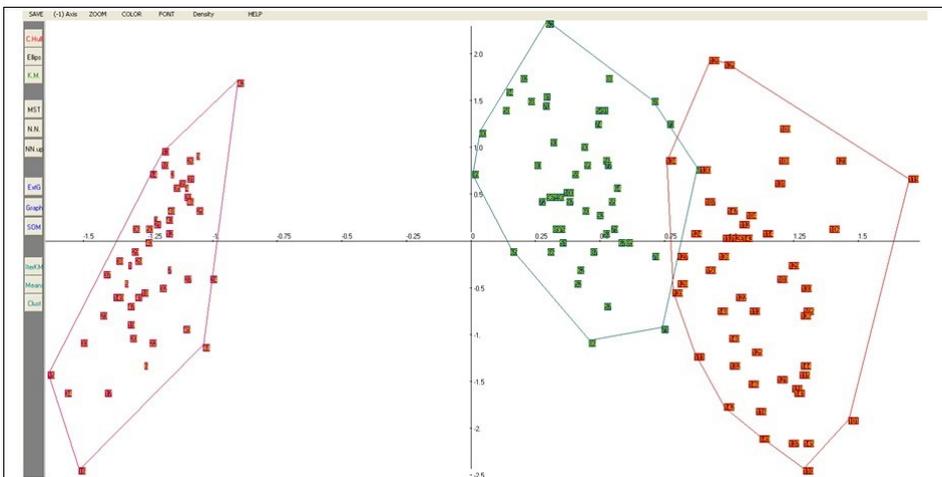
➤ Dans le menu **Load coordinates** de la nouvelle fenêtre, ouvrir le fichier `ngus_contig.txt`.

➤ Dans le menu **Load or create a partition** et dans le sous-menu **Load partition File**, choisir 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'ACP, résultats qui sont sauvegardés.

➤ Cliquer sur **Graphics**. Choisir ensuite les axes 1 et 2 (valeurs par défaut ).

➤ Cliquer sur **DISPLAY**. Changer les couleurs de l'écran si nécessaire pour obtenir un bon contraste entre les classes, puis verrouiller les couleurs.



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.

➤ Cliquer 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.

## 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 :

**a) 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).

**b) Cycle** concerne la description analogue d'un *cycle* (49 sommets).

**c) 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-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 (le fichier de commande s'appelle maintenant : **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<sup>12</sup>.

### 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). Noter 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-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 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.

---

<sup>12</sup> Voir, par exemple: Exploring Textual Data (1998), par L. Lebart, A. Salem, L. Berry, Kluwer Academic Publisher. Cette même comparaison avait déjà été faite dans l'article : "Introduction à l'analyse des données", (L. Lebart) *Consommation*, n°4, 1969, p. 65-87, Dunod.

### 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 contiguës 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 :

- Fichier de données: **Chessboard\_7x7\_dat.txt**
- Fichier Dictionnaire: **Chessboard\_7x7\_dic.txt**.
- 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.

- Cliquer sur le bouton : **Open an existing command file** de **Command File** (menu principal). Puis rechercher le dossier **Chessboard\_numerical** dans **DtmVic-examples /DtmVic-Examples\_C\_NumData**, puis le fichier de commande **Chessboard\_CA.Param.txt**

Noter encore que ces "fichiers de commande" peuvent être facilement générés en cliquant sur le bouton "Create a command file" du menu principal. Une fenêtre "Select a basic analysis" apparaît. Cliquer 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 suivre 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).

- Cliquer 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, retourner au menu principal.

## b. Visualisation et lecture des résultats

Nous allons maintenant visualiser directement le graphique dans l'étape :

### **VIC : Visualization, Inference, Classification steps.**

- Cliquer sur **Visualization** (on n'utilisera pas ici les boutons "ViewAxes", "PlaneView", etc.)

Une fenêtre intitulée "DTM-visualization: loading files, selecting axes" apparaît.

- Cliquer 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**].
- Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, Sélectionner **No partition**.
- Cliquer sur **Graphics** puis choisir les axes 1 et 2 (par défaut) dans la petite fenêtre "Sélection des axes" et cliquer sur **Continue** puis sur **DISPLAY**.

Dans une nouvelle fenêtre intitulée "Vizualisation - Graphics", le plan factoriel principal s'affiche (voir figure VI.1).

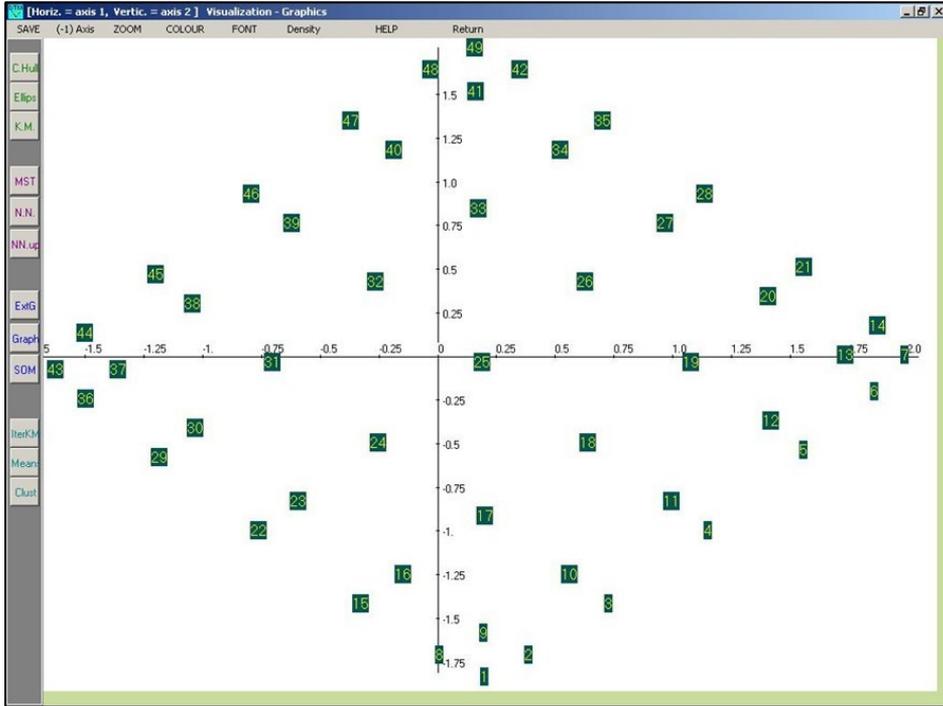


Figure VI.1 Plan factoriel principal (Analyse des correspondances) pour le graphe "Damier" (après changement de police (bouton "Font") et changement de couleur (bouton "Colour").

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.

- Pour représenter les arêtes du graphe d'origine, cliquer sur le bouton **ExtG** (graphe externe) de la barre verticale.
- Ouvrir le fichier Chessboard\_Extern\_7x7.txt.
- Cliquer 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<sup>13</sup>.

- Retourner au menu principal en quittant la fenêtre du plan factoriel, puis en cliquant sur **Return** puis quitter Dtm-Vic.

<sup>13</sup> [Voir Benzécri, (1973) «L'analyse des données», Tome II B, chapitre 10, "Sur l'analyse de la correspondance définie par un graphe", p 244 - 261]

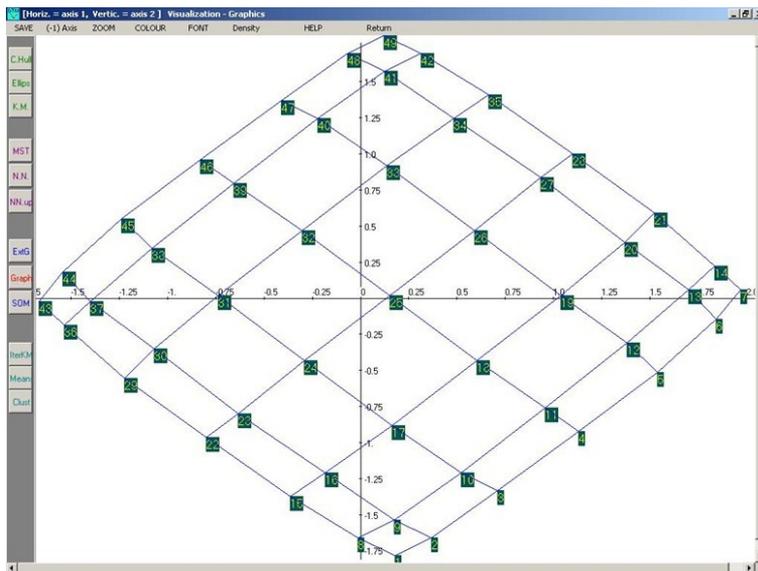


Figure VI.2. Même plan factoriel principal pour le graphe "Damier" avec tracé du graphe initial (après changement de police (bouton "Font") et de couleur (bouton "Colour")).

### 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). Répéter toutes les opérations précédentes.

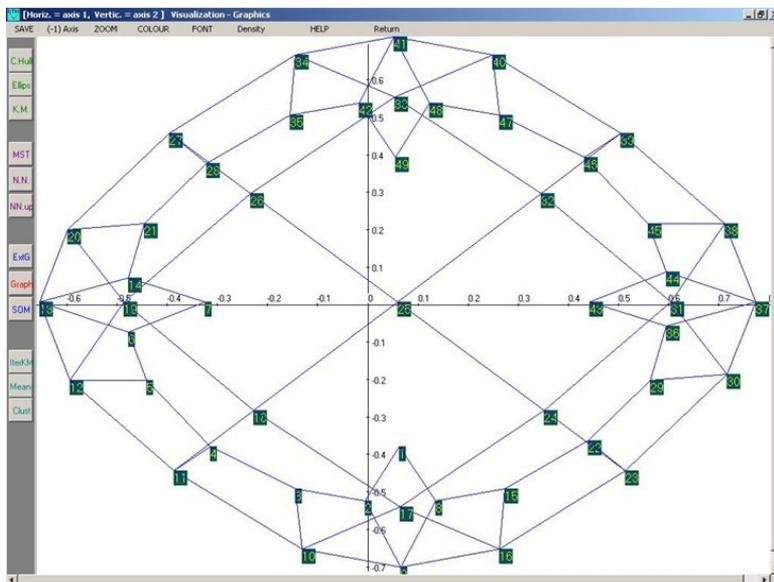


Figure VI.3 Cas de l'analyse en composantes principales. Plan factoriel principal pour le graphe.

La figure VI.3 représente le "Damier" avec aussi un tracé du graphe initial (après changement de police (bouton "**Font**") et changement de couleur (bouton "**Colour**"). 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). Le traitement dissymétrique des lignes et des colonnes opéré par l'ACP ne permet pas d'obtenir une description satisfaisante de ce type de graphes.

### 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).

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 de commande

- Cliquer sur le bouton : **Open an existing command file** de **Command File** (menu principal) et ouvrir 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"]).

Noter que ce fichier de commande peut également être généré en cliquant sur le bouton "Create a command file" de la rubrique **Command file** du menu principal. Une fenêtre "Select a Basic Analysis" apparaît. Cliquer ensuite sur le bouton : **VISURESP**, situé dans la rubrique **Textual Data** et suivre les instructions comme indiqué dans les chapitres II et III.

Noter 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).

- Cliquer 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). Noter 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 : Cliquer 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"), il est prévu une entrée directe dans le menu "Contiguïté".

- Cliquer sur  **Contiguity** dans l'étape **VIC : Visualization, Inference, Classification**
- Cliquer sur **Parameter/Edit**. Choisir l'élément "Create"

Nous allons établir les paramètres nécessaires à une description graphique:

- Dans le premier bloc intitulé "*ncoord = Input coordinate file*", cocher "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*" cocher "0" (*no partition*)

- Dans le troisième bloc intitulé "*meth = method*", cocher "4" (*graphe de contiguïté externe*).

➤ Cliquer sur **VALIDATE**.

➤ Dans la barre supérieure de la fenêtre, cliquer 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*" apparaît.

➤ Cliquer sur **CONTINUE**

Une série de fenêtres apparaît 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).

➤ Cliquer sur  **Visualization**

La fenêtre intitulée "*DTM-visualization...*" apparaît.

➤ Cliquer sur **Load coordinates**. Dans le sous-menu correspondant, choisir le fichier: **anagraf.txt**, qui contient les coordonnées factorielles pour les analyses directes de graphes.

➤ Cliquer ensuite sur **Load or create a partition**. Dans le sous-menu correspondant, Sélectionner **No partition**. Puis procéder comme pour l'exemple *Chessboard*.

➤ Cliquer sur **Graphics** puis choisir les axes 1 et 2 (par défaut) dans la fenêtre "Sélection des axes" et cliquer 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.





## 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.

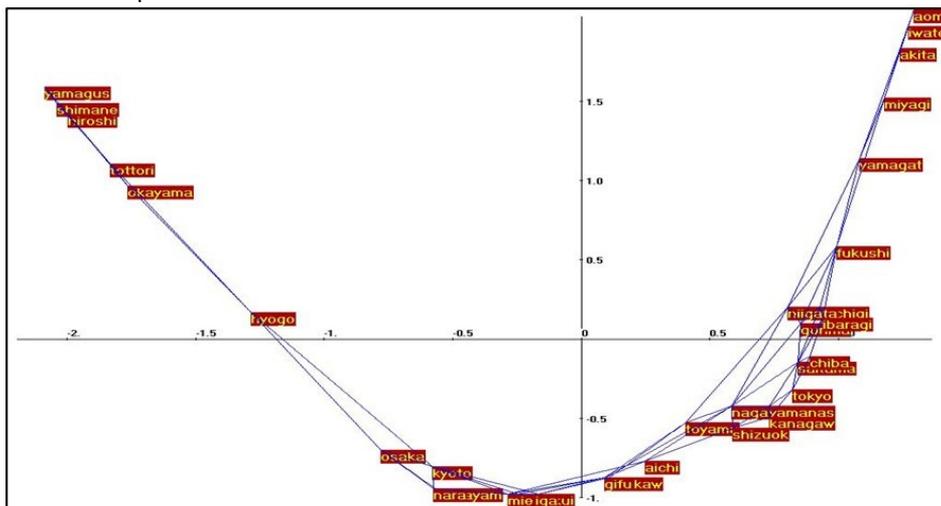


Figure VI.6. Plan factoriel principal pour le graphe "Japon" avec tracé du graphe initial (après changement de police (bouton "Font") et changement de couleur (bouton "Colour"). Le signe des axes est arbitraire. Il peut aussi être changé, pour retrouver l'orientation géographique initiale.

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). Un tel effet est en effet décrit par Guttman (1941) dans un article séminal, très antérieur à l'apparition des ordinateurs, article qui contient un véritable formulaire de l'analyse des correspondances multiples, sans toutefois entrevoir toutes les possibilités exploratoires de la méthode.

## VI.4. Reconstitution d'images

*(Méthodologie - pédagogie)*

Les exemples cette section VI.4 sont principalement des exemples pédagogiques qui servent à illustrer les propriétés de compression des analyses en axes principaux dans le domaine de l'analyse d'images (domaine peu familier pour certains utilisateurs actuels de Dtm-Vic). Cette compression se réalise en gardant un nombre limité d'axes principaux provenant d'une décomposition aux valeurs singulières ou d'une analyse des correspondances. 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, ouvrir 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", « PilotEdit », 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`<sup>14</sup> 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). Ce 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).

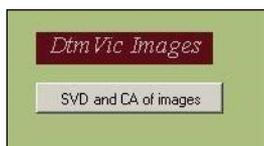
Le format pgm est un format simple et transparent, en niveaux de gris. La première ligne contient l'identificateur de format: P2. Les deuxième et troisième lignes contiennent trois entiers: nombre de colonnes, nombre de lignes, et la valeur maximale (255). Ensuite, le tableau est affiché par ligne. Chaque 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<sup>15</sup>.

**3 - le format ppm.** Pour les (petites) images couleur, le format d'entrée est le format texte ppm ("portable pixel map"). Consulter 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 essentiellement pédagogique, limitation à 1000 pour le nombre de pixels en ligne ou en colonne].

## VI.4.2 Analyse pour la compression d'images

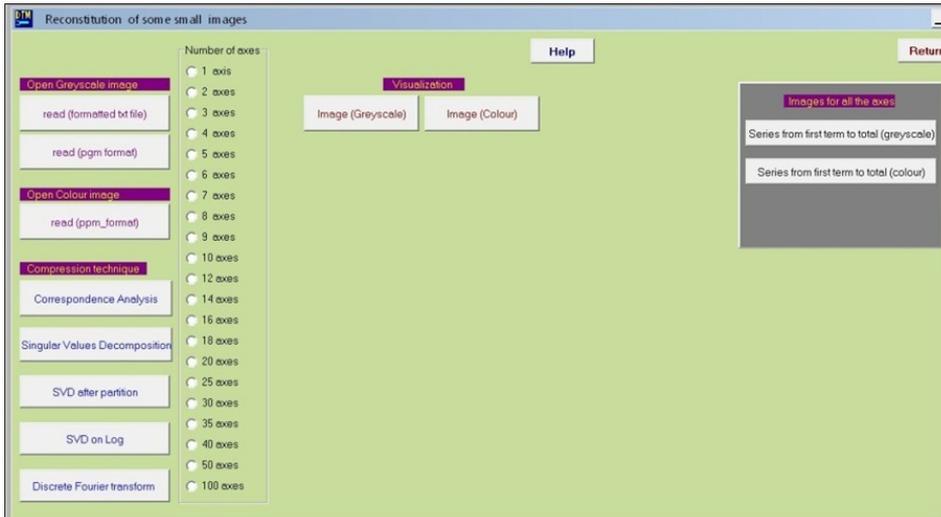
- Cliquer sur le bouton : `SVD and CA of images`, dans la rubrique "DtmVic Images" du menu principal.



Une fenêtre apparaît, dont la partie supérieure figure ci-dessous.

<sup>14</sup> Cette image est adaptée du livre " *La compression de données*", Mark Nelson, M & T Publishing Inc, 1992.

<sup>15</sup> Pour plus d'informations sur un tel format, veuillez consulter (par exemple): <http://netpbm.sourceforge.net/doc/pgm.html>.



### 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).

Plus bas, dans la même colonne, cinq boutons (bleus) correspondant aux cinq méthodes de compressions choisies :

- 1) *Correspondence Analysis* : Analyse des correspondances simple du tableau de niveaux de gris considéré comme une table de contingence.
- 2) *Singular Values Decomposition* : (ou SVD : Décomposition aux valeurs singulières).
- 3) *SVD after partition* : Analyse après partition préalable de l'image. 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  $p = 2 \times 2, 3 \times 3, 4 \times 4, 5 \times 5$ , etc.).
- 4) *SVD on Log* : Analyse logarithmique. Cette variante consiste à faire une transformation logarithmique préalable, puis à procéder à une SVD du tableau doublement centré en ligne et en colonne (cf §VII.6).
- 5) *Discrete Fourier Transform* : Séries de Fourier discrètes. Développements en séries de Fourier simples des profils de niveaux de gris des lignes (ou des colonnes) du tableau décrivant l'image.

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) :

- Cliquer, selon l'extension du fichier image, sur un des boutons **Read**. (txt format, ou : pgm format, ou : ppm\_format). Répondre **OK** aux boîtes 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 boîte de message « *End of computation* ».
- Sélectionner le nombre d'axes. Répondre **OK** au mémo : *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 créés (image originale, et images reconstituées à partir d'un nombre variable d'axes principaux) sont 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 *bitmap*, mais surtout de les sauvegarder en format JPEG, plus économique en espace.

- Cliquer sur **Exit**.

### VI.4.3 Exécution d'un premier exemple

(format de texte simple : Exemple : *Tête de guépard* : **1\_Cheetah\_txt**)

- Cliquer sur le bouton : **SVD and CA of images**, dans la rubrique **DtmVic- Images** du menu principal.

La fenêtre "*Reconstitution of some small images*", décrite plus haut, apparaît.

a. Cliquer sur le premier bouton **Read (formatted txt file)** dans la rubrique **Open Greyscale image**.

- Dans le répertoire **EX\_CO4\_Image**, ouvrir le sous-répertoire **1\_Cheetah\_txt**. Dans ce répertoire, ouvrir le fichier **Cheetah.txt**. Une boîte de message rappelle les dimensions du fichier image.

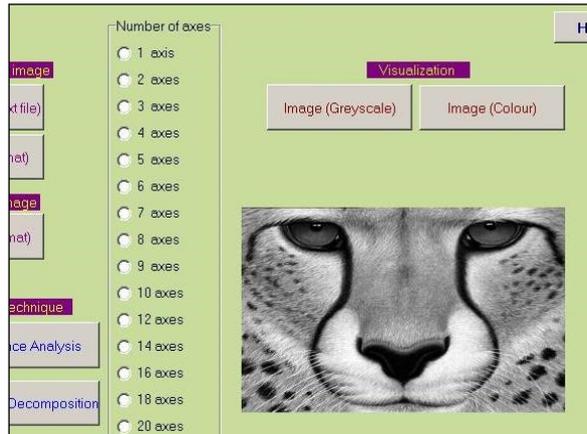


Figure VI.7. Portion de fenêtre présentant l'image originale **Cheetah.txt** avant le choix du nombre d'axes.

- b. Pour visualiser l'image d'origine, dans la rubrique **Visualization**, cliquer sur : **Image (Greyscale)**. L'image apparaît alors au centre de la fenêtre, comme indiqué ci-dessus.

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 de rechercher une compression optimale, mais de comparer deux approches hiérarchiques simples (bases de vecteurs propres *versus* bases de fonctions trigonométriques).

c. Le cas des méthodes factorielles

Dans la partie inférieure gauche de la fenêtre, dans la rubrique : **Compression technique**, cliquer sur le bouton: **Correspondence Analysis** (pour commencer). L'analyse s'effectue.

- c1. Pour obtenir un aperçu de la reconstitution des données, de 1 à 100 axes, cliquer 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. (cf. Figure VI.8).

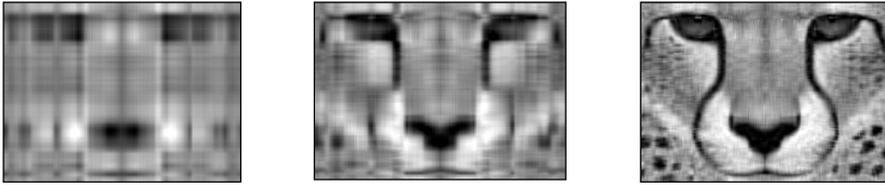


Figure VI.8. Cas de l'analyse des correspondances : Images reconstituées successivement avec un axe principal, quatre axes et 16 axes. Dans le cas d'un seul axe, la formule de reconstitution contient deux termes : le terme correspondant à l'hypothèse d'indépendance (« axe 0 ») et le premier axe.

**c3.** A la place de l'analyse des correspondances, on peut choisir la méthode de "Singular Value Decomposition" (Décomposition aux Valeurs Singulières), et refaire les opérations **c1.** et **c2.** (cf. Figure VI.9)

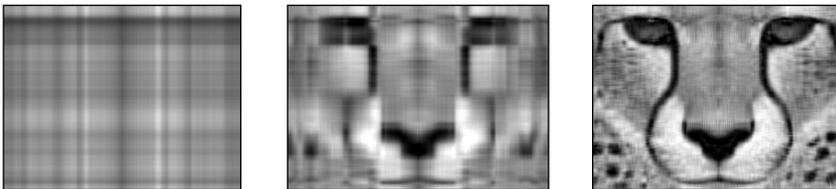


Figure VI.9. Cas de la décomposition aux valeurs singulières: Images reconstituées successivement avec un axe principal, quatre axes et 16 axes. Dans ce cas, pour un axe, la formule de reconstitution ne contient qu'un seul terme, d'où un "retard" par rapport à l'analyse des correspondances, retard qui 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**, cliquer sur le bouton: **Discrete Fourier Transform**. Une fenêtre s'affiche.



Portion de la fenêtre de commande des compressions par séries de Fourier discrètes.

d1. Ensuite, sélectionner le mode de calcul de la série de Fourier, en ligne ou en colonne ("Row-wise" ou "Columnwise"). Sélectionner "Row-wise", par exemple.

d2. Puis, comme précédemment, pour obtenir un aperçu de la reconstitution des données lorsque le nombre de termes augmente, cliquer 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. Pour un nombre de termes particulier (parmi les termes de la sélection suggérée), sélectionner le nombre requis dans la liste verticale correspondante, et visualiser chaque image avec l'analogie du bouton utilisé en **b**.



Figure VI.10. 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

- Cliquer 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 "Baalbek"

- a. Cliquer sur le premier bouton **Read (pgm format)** dans la rubrique **Open Greyscale image**.

Dans le répertoire **EX\_CO4\_Image**, ouvrir le sous-répertoire **2\_Baalbek\_pgm**. Dans **2\_Baalbek\_pgm**, ouvrir le fichier **Baalbek.pgm**. Une boîte de message rappelle les dimensions du fichier image.

b. Pour visualiser l'image d'origine, dans la rubrique **Visualization**, cliquer sur:

**Image (Greyscale)**.

c. Puis, dans la partie inférieure gauche de la fenêtre, dans la rubrique : **Compression technique**, cliquer 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 Baalbek) peut contaminer la reconstitution dans le cas des axes principaux (cf. Figure VI.11).

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...

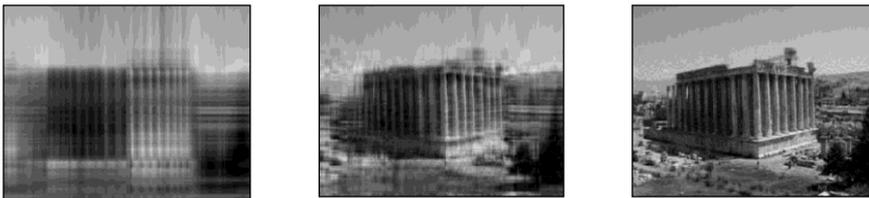


Figure VI.11. Temple de Baalbek. Cas de l'analyse des correspondances : Images reconstituées successivement avec deux axes principaux, neuf axes et 50 axes. Les traits structuraux captés par les premiers axes se répercutent sur les axes suivants, et il faut atteindre près de 50 axes pour obtenir un ciel conforme à celui de l'image initiale.

#### VI.4.4.2 Exemple "Cardinal"

Pour ouvrir le fichier couleur du Cardinal de l'île Maurice, cliquer sur le troisième bouton **Read (ppm format)** dans la rubrique **Open colour image**.

Dans le répertoire **EX\_CO4\_Image**, ouvrir le sous-répertoire **3\_Cardinal\_ppm\_color**, puis ouvrir 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).



Figure VI.12. 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"

Ce dernier exemple contient les deux formats d'image pgm et ppm.

Dans le répertoire **EX\_CO4\_Image**, ouvrir le sous-répertoire **4\_Extra\_pgm\_ppm**, puis ouvrir le fichier **broom.pgm**. Une boîte de message rappelle les dimensions du fichier image.



Figure VI.13. 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.



Figure VI.14. Enfant balayant une cour. Cas de l'analyse des correspondances : Images couleur (ppm) reconstituées successivement avec deux 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 manche du balai que tient l'enfant n'apparaît pas avant le 20<sup>ème</sup> axe : les traits structuraux diagonaux sont défavorisés par la formule de reconstitution des données...

## VII. Annexe statistique

### ➤ QUELQUES NOTIONS DE STATISTIQUE MULTIDIMENSIONNELLE

Les méthodes d'analyse statistique exploratoire utilisées par le logiciel Dtm-Vic visent à mettre en forme de vastes ensembles de données, à en dégager des structures et aussi à valider ces structures. Elles relèvent de la statistique exploratoire multidimensionnelle, de l'analyse des données, ou encore du *Data Mining*, ces trois désignations étant actuellement à peu près équivalentes. On utilise parfois à leur propos l'expression *statistique structurale* pour marquer l'importance accordée à la phase de validation des structures. Ces méthodes généralisent la statistique descriptive classique et utilisent des outils mathématiques assez intuitifs, mais plus complexes que les moyennes, variances et coefficients de corrélations empiriques de la statistique descriptive.

Sont présentés dans cette annexe les principes des techniques utilisées en insistant sur l'analyse en composantes principales, la technique d'analyse factorielle de base la plus répandue. Certains développements de l'ouvrage noté [SEM 2006]<sup>16</sup> seront repris ; ils seront complétés par des travaux sur les méthodes de validation, et en particulier sur les techniques dites de *bootstrap*, sur les cartes de Kohonen, ou sur des techniques d'analyse moins utilisées comme l'analyse logarithmique.

Les rappels de statistique multidimensionnelle de ce chapitre sont adaptés de l'annexe 1 de l'ouvrage « La sémiométrie », [Ludovic Lebart, Marie Piron, Jean-François Steiner. Dunod, 2003] et de l'ouvrage : « Statistique Exploratoire Multidimensionnelle » [Ludovic Lebart, Marie Piron, Alain Morineau, Dunod, 2006]. L'ouvrage « La sémiométrie » comme sa version anglaise, sont librement téléchargeables sur le site : [www.dtmvic.com](http://www.dtmvic.com), rubrique « Publications ».

### VII.1 Rappel des principes des méthodes exploratoires multidimensionnelles

Les méthodes exploratoires multidimensionnelles recouvrent un grand nombre de techniques qui ont pour objectif de décrire et synthétiser l'information contenue dans de vastes tableaux de données.

Au départ, les données se présentent sous forme de grands tableaux rectangulaires, notés **X**. Les lignes ( $i=1, \dots, n$ ) du tableau représentent les  $n$  individus, les sujets enquêtés par exemple, et les colonnes ( $j=1, \dots, m$ ) les  $m$  variables qui peuvent être des mesures, des caractéristiques ou encore des notes relevées sur les individus.

<sup>16</sup> *Statistique Exploratoire Multidimensionnelle*, [Visualisation et inférence en fouille de données], 4<sup>ème</sup> ed. L.Lebart, M. Piron, A. Morineau. Dunod, 2006.

### > VII.1.1 REPRESENTATION GEOMETRIQUE ET NUAGES DE POINTS

Afin de comprendre le principe des méthodes de statistique exploratoire multidimensionnelle, il est utile de représenter de façon géométrique l'ensemble des  $n$  individus ( $n$  lignes) et l'ensemble des  $m$  variables ( $m$  colonnes) comme deux *nuages de points*, chacun des deux ensembles étant décrit par l'autre. On définit alors, pour les deux nuages, des distances entre les points-lignes et entre points-colonnes qui traduisent les associations statistiques entre les individus (lignes) et entre les variables (colonnes).

**Tableau A.1 :**  
**Exemple de tableau X de notes (de 1 à 7)**  
**attribuées à :  $m = 7$  mots, par  $n = 12$  répondants**

<i>mots Répondants</i>	arbre	cadeau	danger	morale	orage	politesse	sensuel
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

Dans le cas de la sémiométrie<sup>17</sup>, une variable (un mot) est un point dont les coordonnées sont les notes données par les  $n$  individus (répondants) : le nuage des  $m$  mots se situe dans un espace à  $n$  dimensions. De même, un individu est un point dont les coordonnées sont les notes attribuées aux  $m$  mots ; le nuage des  $n$  individus se trouve dans un espace à  $m$  dimensions.

Les figures A.1 et A.2 illustrent, à partir du tableau A.1 contenant les notes attribuées à 7 mots par 12 répondants, la représentation de ces deux nuages de points intrinsèquement liés.

Le nuage des points-mots est construit dans l'espace des individus, ici à partir seulement de deux individus, R04 et R08, car deux dimensions rendent possible un graphique dans un plan (*cf.* figure A.1).

<sup>17</sup> Cf. l'ouvrage téléchargeable précité (« La sémiométrie »), et le jeu de données de l'exemple de la section VI.1 (« Données numériques : Sémiométrie ») du chapitre VI de ce manuel.

	arbre	cadeau	danger	morale	orage	politesse	sensuel
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

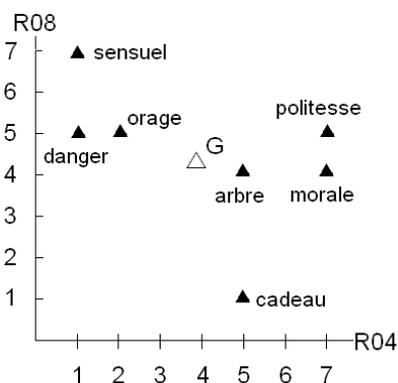


Figure A.1 : Représentation du nuage des mots dans l'espace des deux répondants « R04 » et « R08 »

De la même façon, le nuage des 12 répondants est construit dans l'espace des variables, ici à partir de deux mots, *Morale* et *Sensuel*, c'est-à-dire dans un espace de deux dimensions (cf. figure A.2).

Pour chacun des nuages est représenté le *point moyen* appelé aussi *centre de gravité*. Il s'agit de G pour le centre de gravité des notes attribuées par les répondants (cf. figure A.1) et de G' pour celui des répondants ayant notés les deux mots retenus.

	arbre	cadeau	danger	morale	orage	politesse	sensuel
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

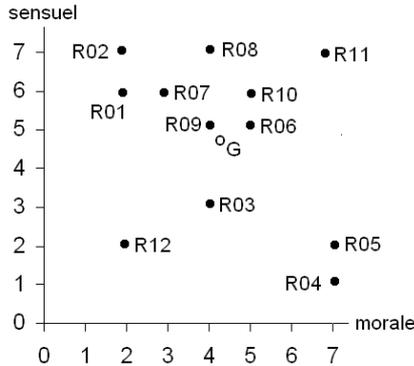


Figure A.2 : Représentation du nuage des répondants dans l'espace engendré par les deux mots « Sensuel » et « Morale »

➤ **VII.1.2 PRINCIPE ET METHODES D'ANALYSE**

S'il est toujours possible de calculer des distances entre les lignes et des distances entre les colonnes d'un tableau **X**, il n'est pas possible de les visualiser de façon immédiate (les représentations géométriques associées impliquant en général des espaces à plus de deux ou trois dimensions) : il est nécessaire de procéder à des transformations et des approximations pour en obtenir une représentation plane.

Les tableaux de distances associés à ces représentations géométriques (simples dans leur principe, mais complexes en raison du grand nombre de dimensions des espaces concernés) peuvent être décrits par les deux grandes familles de méthodes que sont les méthodes factorielles et les méthodes de classification. La première famille se propose de rechercher les directions principales selon lesquelles les points s'écartent le plus du point moyen. La seconde famille va rechercher des groupes ou classes d'individus qui soient les plus homogènes (figure A.3).

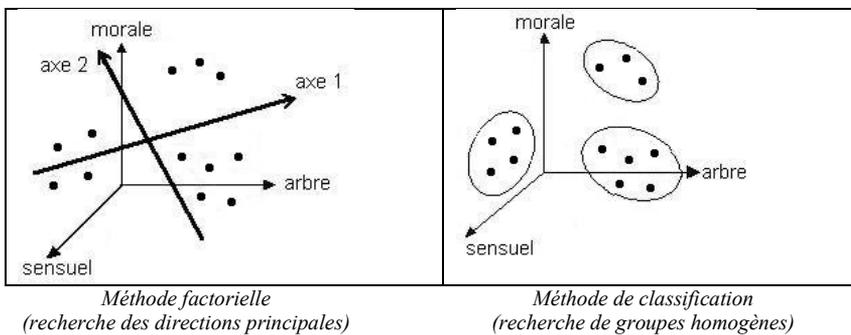


Figure A.3 : Deux grandes familles de méthodes

Ces méthodes impliquent souvent de la même manière les individus (lignes) et les variables (colonnes). La confrontation des espaces d'individus et de variables enrichit les interprétations.

## ➤ VII.2 LES METHODES FACTORIELLES : ASPECTS TECHNIQUES

Les méthodes factorielles<sup>18</sup> permettent de gérer simultanément des quantités importantes de données et leur système de corrélations et, par une technique réalisant une sorte de *compression*, d'en dégager la structure interne, notamment sous forme de graphique-plans.

### ➤ VII.2.1 Recherche des sous-espaces factoriels

L'objectif est de rechercher des sous-espaces de dimensions réduites (entre trois et dix, par exemple) qui ajustent au mieux le nuage de points-individus et celui des points-variables, de façon à ce que les proximités mesurées dans ces sous-espaces reflètent autant que possible les proximités réelles. On obtient ainsi un espace de représentation, l'espace factoriel, défini par les axes principaux d'inertie et l'on représente les points du nuage dans ce système d'axes (*cf.* figure A.4). Ces axes réalisent les meilleurs ajustements de l'ensemble des points selon le critère classique des moindres carrés, qui consiste à rendre minimale la somme des carrés des écarts entre les points et les axes.

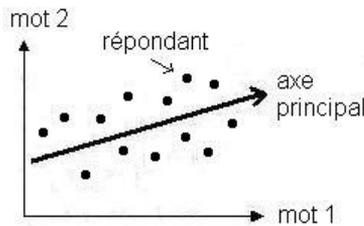


Figure A.4 : Ajustement du nuage des points-individus dans l'espace des mots

Le premier de ces axes correspond à la droite d'allongement maximum du nuage, le second axe maximise le même critère en étant assujéti à être orthogonal au premier, et ainsi de suite pour les axes suivants qui sont tous orthogonaux entre eux. Cette orthogonalité traduit l'indépendance (en fait, la non-corrélation) des axes.

$X$  désigne le tableau de données ayant subi des transformations préliminaires (variables centrées réduites, par exemple),  $X'$  son transposé.

Soit  $u_1$  le vecteur unitaire qui caractérise le premier axe.  $u_1$  est alors le vecteur propre de la matrice  $X'X$  correspondant à la plus grande valeur propre  $\lambda_1$  [*cf.* SEM 2006].

Plus généralement, le sous-espace à  $q$  dimensions qui ajuste au mieux (au sens des moindres carrés) le nuage est engendré par les  $q$  premiers vecteurs propres de la matrice  $X'X$  correspondant aux  $q$  plus grandes valeurs propres.

La procédure d'ajustement est exactement la même pour les deux nuages. On démontre

<sup>18</sup> Elles comprennent dans la littérature statistique française des trente dernières années toutes les techniques de représentation utilisant des « axes principaux »: analyse en composantes principales, analyse des correspondances simples et multiples, analyse factorielle dite classique (en anglais : *factor analysis*) ou analyse en facteurs communs et spécifiques.

alors qu'il existe des relations simples liant les axes calculés dans les deux espaces, celui des individus et celui des variables (*relations dites de transition*). Cette relation s'exprime de la façon suivante :

$$\mathbf{u}_q = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}' \mathbf{v}_q$$

où  $\mathbf{u}_q$ ,  $\mathbf{v}_q$  sont respectivement les  $q$ -èmes vecteurs propres de  $\mathbf{X}'\mathbf{X}$  et  $\mathbf{X}\mathbf{X}'$  et  $\lambda_q$  la valeur propre associée.

Le vecteur des coordonnées des points sur chacun des axes, appelé *facteur*, est une combinaison linéaire des variables initiales. On dénote par  $\boldsymbol{\psi}_\alpha$  et  $\boldsymbol{\varphi}_\alpha$  les facteurs correspondant à l'axe  $\alpha$  respectivement dans l'espace noté  $\mathbb{R}^m$  (espace dont les  $n$  points ont pour coordonnées sont les  $m$  mots) et dans l'espace noté  $\mathbb{R}^n$  (espace dont les  $m$  points ont pour coordonnées sont les  $n$  individus).

Les deux nuages de points, celui des mots et celui des répondants, sont intrinsèquement liés et révèlent exactement les mêmes structures : dans un cas, les facteurs décriront les corrélations entre les mots, dans l'autre les associations entre les répondants.

Les plans factoriels de visualisation utilisés tout au long de cet ouvrage correspondent chacun à un couple de facteurs.

Le plan le plus utilisé est le plan  $(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2)$ .

Les éléments (mots ou individus) qui participent au calcul des axes sont les *éléments actifs*. On introduit aussi dans l'analyse des *éléments supplémentaires* (ou *illustratifs*) qui ne participent pas à la formation des axes, mais qui sont projetés *a posteriori* dans les plans factoriels et peuvent aider à leur interprétation (cf. section VII.10.3).

## ➤ VII.2.2 Techniques de base, méthodes dérivées

La nature des informations, leur codage dans le tableau de données, les spécificités du domaine d'application vont introduire des variantes au sein des méthodes factorielles. Celles qui sont utilisées ici ne sont en fait que des dérivées de deux techniques fondamentales, l'analyse en composantes principales et l'analyse factorielle des correspondances.

L'analyse en composantes principales s'applique à un tableau de mesures numériques. Elle est utilisée, dans le cadre de l'exemple II.1 du chapitre II de ce manuel, pour analyser des durées en minutes (enquête budget-temps), et dans le cadre de la sémiométrie (section VI.1), pour traiter un tableau de notes.

La plupart des exemples d'analyse de données textuelles présentés au chapitre III de ce manuel reposent sur l'analyse factorielle des correspondances appliquée aux tableaux de contingence lexicaux.

## VII.3 L'ANALYSE EN COMPOSANTES PRINCIPALES: ASPECTS TECHNIQUES

L'Analyse en Composantes Principales (Hotelling, 1933) s'applique à des variables à valeurs numériques (des mensurations, des taux, des notes, de durées, etc.) représentées sous forme d'un tableau rectangulaire de mesures  $\mathbf{R}$  de terme général  $r_{ij}$  dont les colonnes sont les variables et les lignes représentent les individus sur lesquels ces variables sont mesurées. En sémiométrie par exemple, les variables sont donc les mots; les lignes les répondants et les valeurs numériques, les notes.

### ➤ VII.3.1 INTERPRETATIONS GEOMETRIQUES

Les représentations géométriques entre les lignes d'une part et entre les colonnes d'autre part du tableau de données permettent de visualiser les proximités respectivement entre les individus et entre les variables (cf. figures A.1 et A.2 ci-dessus). Dans  $\mathbb{R}^m$ , deux points-individus sont très voisins si, dans l'ensemble, leurs  $m$  coordonnées sont très proches. Les deux répondants concernés sont alors caractérisés par des valeurs presque égales pour chaque variable. La distance utilisée est la distance euclidienne usuelle.

Dans  $\mathbb{R}^n$ , si les valeurs prises par deux variables particulières sont très voisines pour tous les répondants, ces variables seront représentées par deux points très proches dans cet espace. Cela peut vouloir dire que ces variables mesurent une même chose ou encore qu'elles sont liées par une relation particulière.

Mais les unités de mesure des variables peuvent être très différentes et rendre alors nécessaire des transformations du tableau de données.

### ➤ VII.3.2 PROBLEME D'ECHELLES DE MESURE ET TRANSFORMATION DES DONNEES

On veut que la distance entre deux individus soit indépendante des unités des variables pour que chaque variable joue un rôle identique. Pour cela, on attribue à chaque variable  $j$  la même dispersion en divisant chacune de ses valeurs par leur écart-type  $s_j$  avec :

$$s_j^2 = \frac{1}{n} \sum_{i=1}^n (r_{ij} - \bar{r}_j)^2.$$

Par ailleurs on s'intéresse à la manière dont les individus s'écartent de la moyenne. On place alors le point moyen au centre de gravité du nuage des individus. Les coordonnées du point moyen sont les valeurs moyennes des variables notées :

$$\bar{r}_j = \frac{1}{n} \sum_{i=1}^n r_{ij}$$

Prendre ce point comme origine revient à centrer les variables, c'est-à-dire à soustraire à chaque variable  $j$  sa moyenne  $\bar{r}_j$ .

On corrige ainsi les échelles en transformant le tableau de données  $\mathbf{R}$  en un nouveau tableau  $\mathbf{X}$  de la façon suivante :

$$x_{ij} = \frac{r_{ij} - \bar{r}_j}{s_j \sqrt{n}}$$

Les variables ainsi réduites et centrées ont toutes une variance,  $s^2(x_j)$ , égale à 1 et une moyenne,  $\bar{x}_j$ , nulle et deviennent comparables. D'autres transformations préalables sont possibles.

### ➤ VII.3.3 ANALYSE DU NUAGE DES N REpondANTS

La transformation des données nous conduit à effectuer une translation de l'origine au centre de gravité de ce nuage et à changer (dans le cas de l'analyse dite normée) les échelles sur les différents axes.

Pour réaliser l'analyse du nuage des points-répondants dans  $\mathbb{R}^m$ , la matrice  $\mathbf{X}'\mathbf{X}$  à diagonaliser dans cet espace, est la matrice des corrélations (dont la figure A.4 fournit un exemple) qui a pour terme général :

$$c_{jj'} = \sum_{i=1}^n x_{ij} x_{ij'} = \frac{1}{n} \sum_{i=1}^n \frac{(r_{ij} - \bar{r}_j)(r_{ij'} - \bar{r}_{j'})}{s_j s_{j'}}$$

$c_{jj'}$  est le coefficient de corrélation entre les variables  $j$  et  $j'$ .

Les coordonnées des  $n$  points-individus sur l'axe factoriel  $\mathbf{u}_\alpha$  sont les  $n$  composantes du vecteur  $\boldsymbol{\psi}_\alpha = \mathbf{X}\mathbf{u}_\alpha$ .

Tableau de notes (1 à 7) données à 7 mots par 12 répondants

	arbre	cadeau	danger	morale	orage	politesse	sensuel
R01	7	4	2	2	3	1	6
R02	6	3	1	2	4	1	7
R03	4	5	3	4	3	4	3
R04	5	5	1	7	2	7	1
R05	4	5	2	7	1	6	2
R06	5	7	1	5	2	6	5
R07	4	2	1	3	5	3	6
R08	4	1	5	4	5	4	7
R09	6	6	2	4	7	5	5
R10	6	6	3	5	3	6	6
R11	7	7	6	7	7	6	7
R12	2	2	1	2	1	3	2

La figure A.4-a illustre la représentation du nuage des répondants pour le tableau de 12 répondants ayant noté 7 mots (tableau déjà présenté en section A.1) dans le plan principal (2, 3)<sup>19</sup>. Les répondants R01 et R02 ont donné, de la même façon, des notes très contrastées et ont donné des notes élevées à *Arbre* et *Sensuel* et des notes faibles à

<sup>19</sup> Dans le cas très particulier de la sémiométrie, le plan (2, 3) est considéré comme le plan sémiométrique principal compte tenu du caractère particulier du premier axe (axe dit *de taille*, cf. « La sémiométrie », chapitre 5).

*Morale* et *Politesse* ; ils sont par conséquent proches dans le plan et se différencient des répondants R05 et R04 qui se sont exprimés de façon inverse sur les mots. Le répondant R08 se distingue en ayant très bien noté *Danger* sans pour autant bien noter les autres mots, alors que R11 a bien noté tous les mots.

**Matrice des corrélations**

!	arbr	cade	dang	mora	orag	poli	sens
-----	-----	-----	-----	-----	-----	-----	-----
arbr !	1.00						
cade !	.55	1.00					
dang !	.29	.14	1.00				
mora !	.16	.62	.36	1.00			
orag !	.51	.09	.54	-.01	1.00		
poli !	.00	.63	.23	.91	-.05	1.00	
sens !	.56	-.08	.45	-.30	.68	-.37	1.00

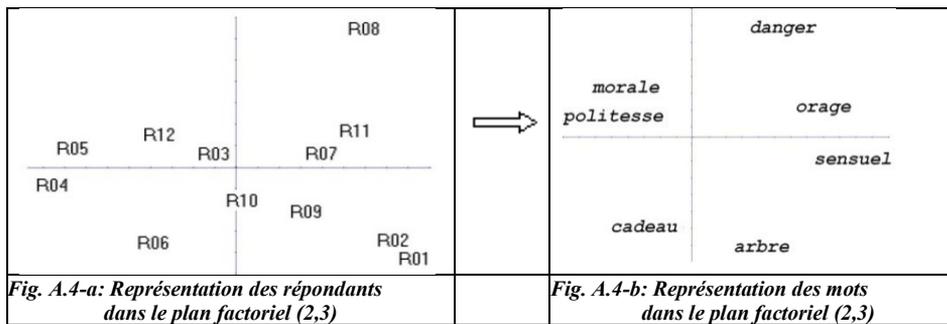


Figure A.4 : Analyse en composantes principales sur le tableau des notes de 7 mots par 12 répondants

➤ **VII.3.4 ANALYSE DU NUAGE DES VARIABLES**

Les coordonnées factorielles  $\varphi_{\alpha j}$  des points-variables sur l'axe  $\alpha$  sont les composantes de  $\mathbf{u}_\alpha \sqrt{\lambda_\alpha}$  et l'on a :  $\varphi_{\alpha j} = cor(j, \psi_\alpha)$

La coordonnée  $\varphi_{\alpha j}$  d'un point-variable  $j$  sur un axe  $\alpha$  n'est autre que le *coefficient de corrélation* de cette variable avec le facteur  $\psi_\alpha$  (combinaison linéaire des variables initiales) considéré lui-même comme une variable artificielle dont les coordonnées sont constituées par les  $n$  projections des individus sur cet axe.

Les axes factoriels étant orthogonaux deux à deux, on obtient ainsi une série de variables artificielles non corrélées entre elles, appelées *composantes principales*, qui synthétisent les corrélations de l'ensemble des variables initiales.

Sur la figure A.4-b, comme sur la matrice de corrélations correspondante, *Politesse* et *Morale* sont très corrélés et dans une moindre mesure *Orage* et *Sensuel*. On retrouve

bien les comportements des répondants où R01 et R02 vont dans la direction des « bons noteurs » d'*Arbre* et de *Sensuel* et des « mauvais noteurs » de *Morale* et *Politesse* à l'inverse des répondants R04 et R05.

Les variables fortement corrélées avec un axe vont contribuer à la définition de cet axe<sup>20</sup>. Cette corrélation se lit directement sur le graphique puisqu'il s'agit de la coordonnée du point-variable  $j$  sur l'axe  $\alpha$ . On s'intéresse surtout aux variables présentant les plus fortes coordonnées et l'on interprétera les composantes principales en fonction des regroupements de certaines de ces variables et de l'opposition avec les autres<sup>21</sup>.

On notera alors que tous les points-variables sont sur une sphère de rayon 1 centrée à l'origine des axes<sup>22</sup>. Les plans d'ajustement couperont la sphère suivant de grands cercles (de rayon 1), les *cercles de corrélations*, à l'intérieur desquels sont positionnés les points-variables. Dans ce manuel, les cercles ne sont pas tracés dans les plans factoriels représentant les mots pour une meilleure lisibilité des libellés (le cadrage des plans factoriels peut en effet entraîner une forte réduction d'échelle).

Pour des développements plus techniques, on se reportera à l'ouvrage SEM-2006 ou au bouton « PCA » (Principal Component Analysis) du pavé : « Statistical tools, some reminders » du menu d'accueil de Dtm-Vic.

## ➤ VII.4 L'ANALYSE DES CORRESPONDANCES

L'analyse des correspondances<sup>23</sup> s'applique en premier lieu à une table de contingence  $\mathbf{K}$ , appelé aussi tableau croisé, à  $n$  lignes et  $m$  colonnes, qui ventile une population selon deux variables qualitatives à  $n$  et  $m$  modalités. Les lignes et les colonnes jouent donc des rôles similaires.

Dans la section II.2 de ce manuel, l'analyse est appliquée à un tableau croisant 8 statuts d'activité en ligne avec 6 types de médias en colonne.

Dans la section III.1, elle est appliquée au tableau lexical croisant en ligne les 114 mots les plus fréquents dans les 20 premiers sonnets de Shakespeare avec, en colonne, ces 20 sonnets.

Dans la section III.2, l'analyse des correspondances porte sur la table de contingence lexicale croisant les 136 mots apparaissant plus de 16 fois (dans les réponses de 1043

<sup>20</sup> L'exemple n'est bien évidemment pas suffisamment représentatif pour que le plan puisse être interprété. Il a juste vocation à rapprocher le tableau de données des résultats.

<sup>21</sup> L'analyse en composantes principales ne traduit que des liaisons linéaires entre les variables. Un coefficient de corrélation faible entre deux variables signifie donc que celles-ci sont indépendantes linéairement, alors qu'il peut exister une relation non linéaire.

<sup>22</sup> L'analyse du nuage des points-variables dans  $\mathbb{R}^n$  ne se fait pas par rapport au centre de gravité du nuage, contrairement à celui des points-individus mais par rapport à l'origine. La distance d'une variable  $j$  à l'origine  $O$  s'exprime par :  $d^2(O, j) = \sum_{i=1}^n x_{ij}^2 = 1$ .

<sup>23</sup> Présentée et étudiée de façon systématique comme une technique souple d'analyse exploratoire de données multidimensionnelles par J.-P. Benzécri (1973), l'analyse des correspondances s'est trouvée depuis d'autres précurseurs, en particulier C. Hayashi (1956), et a donné lieu à des travaux dispersés et indépendants les uns des autres.

individus à une question ouverte) avec 9 catégories de répondants (âge – scolarité).

### ➤ VII.4.1 Notations

Soit  $k = \sum_{i,j} k_{ij}$  la somme de tous les éléments  $k_{ij}$  de la table de contingence  $\mathbf{K}$ .

On note  $f_{ij} = k_{ij} / k$  les fréquences relatives avec  $\sum_i \sum_j f_{ij} = 1$ .

On note :  $f_{i.} = \sum_j f_{ij}$ ,  $f_{.j} = \sum_i f_{ij}$ , les fréquences marginales relatives.

La table de contingence  $\mathbf{K}$  est transformé en un tableau de profils-lignes  $f_{ij} / f_{i.}$  et un tableau de profils-colonnes  $f_{ij} / f_{.j}$ .

Le point  $i$  de  $\mathbb{R}^m$  a pour coordonnées :  $f_{ij} / f_{i.}$  pour tout  $j \leq m$ .

De même, le point  $j$  de  $\mathbb{R}^n$  a pour coordonnées :  $f_{ij} / f_{.j}$  pour tout  $i \leq n$ .

Notons une différence importante entre l'analyse des correspondances et l'analyse en composantes principales : les transformations opérées sur le tableau dans les deux espaces sont identiques (car les ensembles mis en correspondance jouent des rôles analogues).

### ➤ VII.4.2 Distance du Chi-deux et équivalence distributionnelle

Les distances entre deux points-lignes  $i$  et  $i'$  d'une part et entre deux points-colonnes  $j$  et  $j'$  d'autre part sont données par les équations suivantes :

$$d^2(i, i') = \sum_{j=1}^{j=m} \frac{1}{f_{.j}} \left[ \frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right]^2 \quad d^2(j, j') = \sum_{i=1}^{i=n} \frac{1}{f_{i.}} \left[ \frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right]^2$$

La distance du  $\chi^2$  offre l'avantage de vérifier le principe d'équivalence distributionnelle. Ce principe assure la robustesse des résultats de l'analyse des correspondances vis-à-vis de l'arbitraire du découpage en modalités des variables nominales.

Il s'exprime de la façon suivante : si deux lignes (resp. colonnes) du tableau de contingence ont même profil (sont proportionnelles) alors leur agrégation n'affecte pas la distance entre les colonnes (resp. lignes). On obtient alors un nouveau point-ligne (resp. point-colonne) de profil identique et affecté de la somme des fréquences des deux points-lignes (resp. points-colonnes).

Cette propriété est importante car elle garantit une certaine invariance des résultats vis-à-vis de la nomenclature choisie pour la construction des modalités d'une variable qualitative.

➤ **VII.4.3 Formulaire et propriétés**

Les deux nuages de points (dans l'espace des colonnes et dans l'espace des lignes) sont construits de manière analogue. Nous récapitulons ici les éléments de base de l'analyse qui vont permettre la construction des facteurs.

Les éléments de base de l'analyse : récapitulation

Nuage de n points-lignes dans l'espace $R^m$	← Eléments → de base	Nuage de m points-colonnes dans l'espace $R^n$
m coordonnées (point-ligne i) $\frac{f_{ij}}{f_i}$ , pour $j=1, 2, \dots, m$ .	Analyse du tableau	n coordonnées (point-colonne j) $\frac{f_{ij}}{f_j}$ , pour $i=1, 2, \dots, n$ .
$d^2(i, i') = \sum_{j=1}^m \frac{1}{f_j} \left[ \frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_i} \right]^2$	avec la distance du $\chi^2$	$d^2(j, j') = \sum_{i=1}^n \frac{1}{f_i} \left[ \frac{f_{ij}}{f_j} - \frac{f_{i'j}}{f_j} \right]^2$
n masses des n points i : $f_i$ .	et les masses	m masses des m points j : $f_j$ .

➤ **Remarques**

Il existe une différence fondamentale avec l'analyse en composantes principales : les transformations faites sur les données brutes dans les deux espaces sont identiques (car les ensembles mis en correspondance jouent des rôles analogues).

**Les coordonnées factorielles sont centrées :**

$$\sum_{i=1}^n f_i \psi_{\alpha i} = \sum_{j=1}^m f_j \varphi_{\alpha j} = 0$$

**et de variance égale à  $\lambda_\alpha$  :**

$$\sum_{i=1}^n f_i \psi_{\alpha i}^2 = \sum_{j=1}^m f_j \varphi_{\alpha j}^2 = \lambda_\alpha$$

➤ **Relations de transition (ou quasi-barycentriques)**

Notons les relations fondamentales existant entre les coordonnées des points-lignes et des points-colonnes sur l'axe  $\alpha$  les relations quasi-barycentriques :

$$\begin{cases} \psi_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{j=1}^m \frac{f_{ij}}{f_i} \varphi_{\alpha j} \\ \varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_{i=1}^n \frac{f_{ij}}{f_j} \psi_{\alpha i} \end{cases}$$

Ainsi, au coefficient de dilatation  $\frac{1}{\sqrt{\lambda_\alpha}}$  près, les projections des points représentatifs d'un nuage sont, sur un axe, les *barycentres* des projections des points représentatifs de l'autre nuage.

La matrice de terme général  $\begin{pmatrix} f_{ij} \\ f_i \end{pmatrix}$  permettant de calculer les coordonnées d'un point  $i$  à partir de tous les points  $j$  n'est autre que le tableau des profils-lignes.

➤ **Représentation simultanée des lignes et colonnes**

Les relations quasi-barycentriques justifient la représentation simultanée des lignes et des colonnes.

Si les méthodes factorielles sont fondées sur le calcul des distances entre points-lignes d'une part, et entre points-colonnes d'autre part. La distance entre un point-ligne et un point-colonne n'a pas de sens puisque ces points sont dans des espaces différents.

L'analyse des correspondances offre cependant la possibilité de positionner et d'interpréter **un point** d'un ensemble relatif à un espace **par rapport à l'ensemble des autres points** définis dans l'autre espace.

➤ **Formule de reconstitution des données**

$$f_{ij} = f_i \cdot f_j \sum_{\alpha=1}^{\alpha=m} \sqrt{\lambda_\alpha} \varphi_{\alpha j} \psi_{\alpha i}$$

qui s'écrit aussi, en faisant intervenir la première valeur propre qui vaut 1, et les facteurs correspondants

$$f_{ij} = f_i \cdot f_j \left( 1 + \sum_{\alpha=2}^{\alpha=m} \sqrt{\lambda_\alpha} \varphi_{\alpha j} \psi_{\alpha i} \right)$$

A titre d'exemple, c'est cette formule qui est utilisée (dans le cas de l'analyse des correspondances) pour reconstituer les images (section VI.4 de ce manuel dévolue à la reconstitution d'images) Pour des développements plus étendus, on se reportera à l'ouvrage SEM-2006 ou au bouton « CA » (Correspondence Analysis) de la barre verticale « Statistical tools, some reminders » du menu d'accueil de Dtm-Vic.

## VII.5 L'ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)

L'analyse des correspondances peut se généraliser de plusieurs façons au cas où plus de deux ensembles sont mis en correspondance. Une des généralisations la plus simple et la plus utilisée est l'*analyse des correspondances multiples* qui permet de décrire de vastes tableaux binaires, dont les fichiers d'enquêtes socio-économiques constituent un exemple typique : les lignes de ces tableaux sont en général des individus ou observations (limités à 30000 dans Dtm-Vic) ; les colonnes sont des modalités de

variables nominales, le plus souvent des modalités de réponses à des questions (limités à 1200 dans Dtm-Vic).

L'analyse des correspondances multiples est une analyse des correspondances simple appliquée non plus à une table de contingence, mais à un *tableau disjonctif complet*. Les propriétés d'un tel tableau sont intéressantes, les procédures de calculs et les règles d'interprétation des représentations obtenues sont simples et spécifiques. Les principes de l'ACM remontent à Guttman (1941) et Burt (1950).

L'extension du domaine d'application de l'analyse des correspondances se fonde sur l'équivalence suivante : si pour  $n$  individus, on dispose des valeurs (réponses) prises par deux variables nominales ayant respectivement  $p_1$  et  $p_2$  modalités, il est alors équivalent (à des normalisations près) de soumettre à l'analyse des correspondances le tableau de contingence  $(p_1, p_2)$  croisant les deux variables ou d'analyser le tableau binaire à  $n$  lignes et  $(p_1 + p_2)$  colonnes décrivant les réponses.

L'analyse de ce dernier tableau se généralise immédiatement au cas de plus deux variables nominales.

### VII.5.1 Tableau disjonctif complet, tableau de Burt

On désigne par  $p$  le nombre total des modalités de  $s$  questions (la question  $q$  ayant  $p_q$  modalités). On a :

$$p = \sum_{q=1}^s p_q$$

On construit, à partir du tableau de données  $\mathbf{R}$  à  $n$  lignes et  $s$  colonnes donnant les numéros des modalités choisies par  $n$  individus, le tableau  $\mathbf{Z}$  à  $n$  lignes et  $p$  colonnes décrivant les  $s$  réponses des  $n$  individus par un codage binaire.

Le tableau  $\mathbf{Z}$  est la juxtaposition de  $s$  sous-tableaux :

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_q, \dots, \mathbf{Z}_s]$$

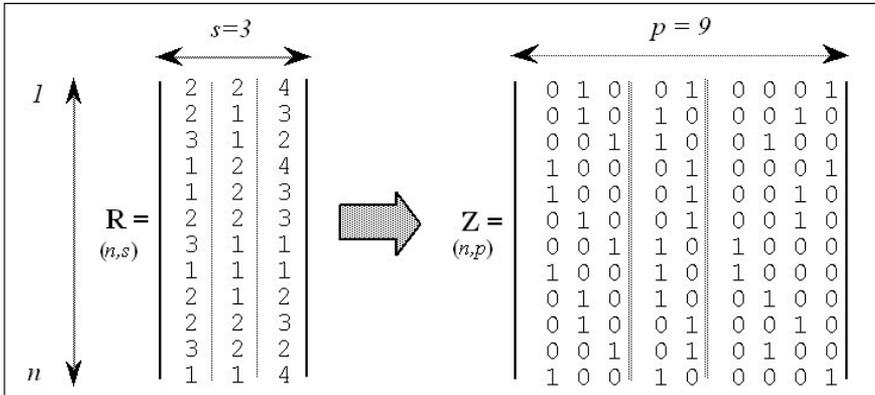


Figure A5. Construction du tableau disjonctif complet  $\mathbf{Z}$  ( $n$  individus,  $s$  questions,  $p$  modalités en tout)

### ➤ **Tableau de contingence de Burt**

L'ensemble des  $pq$  modalités de réponse à une question permet de partitionner l'échantillon en  $pq$  classes. La donnée de deux questions mises sous forme disjonctive complète permet de réaliser deux partitions de l'ensemble des individus et l'on obtient un tableau de contingence. L'analyse du tableau croisant les deux partitions peut être généralisée au cas de  $s$  partitions,  $s$  étant un entier supérieur à 2.

On construit, à partir du tableau disjonctif complet  $\mathbf{Z}$ , le tableau symétrique  $\mathbf{B}$  d'ordre  $(p.p)$  qui rassemble les croisements deux à deux de toutes les variables :

$$\mathbf{B} = \mathbf{Z}'\mathbf{Z}$$

$\mathbf{B}$  est appelé *tableau de contingence de Burt* associé au tableau disjonctif complet  $\mathbf{Z}$ .

Le terme général de  $\mathbf{B}$  s'écrit :

$$b_{jj'} = \sum_{i=1}^n z_{ij}z_{ij'}$$

$\mathbf{B}$  est une juxtaposition de tableaux de contingence.

Les marges sont, pour tout  $j \leq p$ :

$$b_j = \sum_j^p b_{jj'} = sz_{.j}$$

et l'effectif total  $b$  vaut :

$$b = s^2 n$$

Le tableau  $\mathbf{B}$  est formé de  $s^2$  blocs où l'on distingue :

- le bloc  $\mathbf{Z}'_q\mathbf{Z}_{q'}$  indicé par  $(q, q')$ , d'ordre  $(pq, pq')$  qui n'est autre que la table de contingence croisant les réponses aux questions  $q$  et  $q'$ .
- le  $q^{\text{ième}}$  bloc carré  $\mathbf{Z}'_q\mathbf{Z}_q$  obtenu par le croisement d'une variable avec elle-même. C'est une matrice d'ordre  $(pq, pq)$ , diagonale puisque deux modalités d'une même question ne peuvent être choisies simultanément.

Les termes diagonaux sont les effectifs des modalités de la question  $q$ .

### ➤ **VII.5.2 Principes de l' ACM**

L'analyse des correspondances multiples est l'analyse des correspondances d'un tableau disjonctif complet.

Ses principes sont donc ceux de l'analyse des correspondances à savoir :

- mêmes transformations du tableau de données en profils-lignes et en profils-colonnes;
- même critère d'ajustement avec pondération des points par leurs profils marginaux;
- même distance, celle du  $\chi^2$ .

L'analyse des correspondances multiples présente cependant des propriétés particulières dues à la nature même du tableau disjonctif complet.

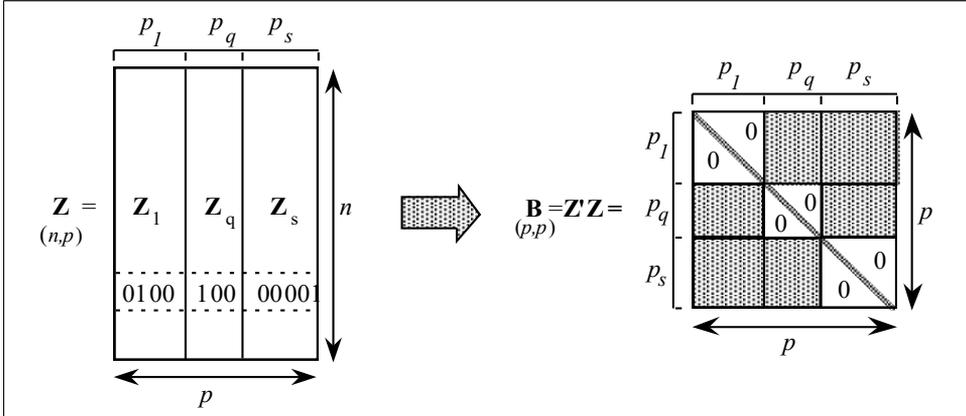


Figure A.6. Construction du tableau de Burt  $B$  à partir du tableau disjonctif complet  $Z$

➤ **Règles d'interprétation**

Dire qu'il existe des affinités entre réponses, c'est dire aussi qu'il existe des individus qui ont choisi simultanément toutes ou presque toutes ces réponses.

L'analyse des correspondances multiples met alors en évidence des types d'individus ayant des profils semblables quant aux attributs choisis pour les décrire. Compte tenu des distances entre les éléments du tableau disjonctif complet et des relations barycentriques particulières, on exprime :

- *la proximité entre individus en termes de ressemblances :*  
Deux individus se ressemblent s'ils ont choisi globalement les mêmes modalités.
- *la proximité entre modalités de variables différentes en termes d'association :*  
Ces modalités correspondent aux points moyens des individus qui les ont choisies et sont proches parce qu'elles concernent globalement les mêmes individus ou des individus semblables.
- *la proximité entre deux modalités d'une même variable en termes de ressemblance :*  
par construction, les modalités d'une même variable s'excluent. Si elles sont proches, cette proximité s'interprète en termes de ressemblance entre les groupes d'individus qui les ont choisies (vis-à-vis d'autres variables actives de l'analyse).

Les règles d'interprétation des résultats (coordonnées, contributions, cosinus carrés) concernant les éléments actifs d'une analyse des correspondances multiples sont sensiblement les mêmes que celles d'une analyse des correspondances simple. On calcule la contribution et la qualité de représentation de chaque modalité et de chaque individu, si ceux-ci ne sont pas anonymes pour l'analyse. En revanche, les règles d'interprétation des valeurs propres et des taux d'inertie sont différentes.

**VII.5.3 Cas de 2 questions**

Dans le cas de deux questions  $q_1$  et  $q_2$ , le tableau disjonctif complet s'écrit :

$$Z = [Z_1, Z_2]$$

et nous ramène directement à l'analyse du tableau de contingence. Il est alors équivalent, au point de vue de la description des associations entre modalités, d'effectuer :

- [1] l'analyse des correspondances du tableau  $\mathbf{Z}$  d'ordre  $(n,p)$ ;
- [2] l'analyse des correspondances du tableau  $\mathbf{B}$  d'ordre  $(p,p)$ ;
- [3] l'analyse des correspondances du tableau  $\mathbf{K} = \mathbf{Z}'_1\mathbf{Z}_2$  d'ordre  $(p_1, p_2)$ .

L'équivalence entre l'analyse des correspondances du tableau disjonctif complet  $\mathbf{Z}$  et celle du tableau des correspondances multiples  $\mathbf{B}$  a été donnée dans le cas général de plusieurs questions.

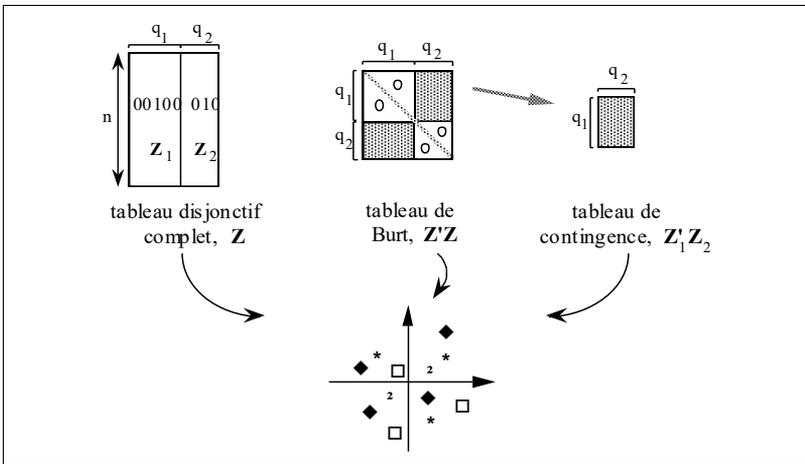


Figure A.7. Equivalence des 3 analyses des correspondances dans le cas de 2 questions

## VII.6 Autres méthodes

On présente ici deux méthodes qui utilisent une réduction par axes principaux : l'analyse logarithmique qui fournit des résultats très proches de l'analyse des correspondances (proposée dans Dtm-Vic comme une des méthodes de compression d'images), et l'analyse factorielle classique (ou : analyse en facteur comuns et spécifiques), pour son rôle historique et son cadre conceptuel.

### VII.6.1 L'analyse logarithmique

L'analyse logarithmique, proposée par J.-B. Kazmierczak (1985), réalise la propriété de l'équivalence distributionnelle de l'analyse des correspondances sur des tableaux qui ne sont pas obligatoirement des tables de contingence. J.-B. Kazmierczak reprend et généralise le principe de Yule qui stipule que l'on ne change pas la distance entre deux lignes ni la distance entre deux colonnes d'un tableau en remplaçant les lignes et les colonnes de ce tableau par d'autres lignes et colonnes qui leur sont proportionnelles (il

s'agit en fait d'une généralisation du principe d'équivalence distributionnelle).

L'analyse logarithmique consiste à prendre les logarithmes des données (après addition éventuelle d'une constante en cas de données négatives), puis, après les avoir centrées à la fois en ligne et en colonne, à les soumettre à une analyse en composantes principales non normée, qui coïncide ici avec une *décomposition aux valeurs singulières* [SEM-2006].

Ainsi, si  $\mathbf{R}$  est un tableau de données  $(n, m)$  et si  $\mathbf{A}$  et  $\mathbf{B}$  sont deux matrices diagonales respectivement de dimensions  $(n, n)$  et  $(p, p)$  à éléments diagonaux positifs, la matrice  $\mathbf{ARB}$  donne lieu à la même analyse logarithmique que la matrice  $\mathbf{R}$ . Une méthode voisine, mais non identique (*Spectral mapping*), a été proposée par Greenacre et Lewi (2009).

### ➤ VII.6.2 L'analyse en facteurs communs et spécifiques

L'analyse factorielle en facteurs communs et spécifiques (*factor analysis*) est probablement le modèle linéaire de variables latentes le plus ancien<sup>24</sup>. Ces modèles ont été essentiellement développés principalement par les psychologues et psychométriciens. Les développements auxquels ils donnent lieu sont complexes et diversifiés. On pourra consulter sur ce point les ouvrages classiques de Harman (1967), Mulaik (1972)<sup>25</sup>.

Mentionnons également les travaux d'Anderson et Rubin (1956) et de Lawley et Maxwell (1963) qui ont placé l'analyse factorielle en facteurs communs et spécifiques dans un cadre inférentiel classique.

#### ➤ Le modèle de l'analyse factorielle

Ce modèle se propose de reconstituer, à partir d'un petit nombre  $q$  de facteurs, les corrélations existant entre  $m$  variables observées. On pose un modèle *a priori* :

$$\mathbf{x}_i = \mathbf{\Gamma} \mathbf{f}_i + \mathbf{e}_i$$

$(m,1) \quad (m,q) \quad (q,1) \quad (p,1)$

Dans cette écriture  $\mathbf{x}_i$  représente le  $i$ -ème vecteur observé des  $m$  variables ;  $\mathbf{\Gamma}$  est un tableau  $(m, q)$  de coefficients inconnus (avec  $q < m$ ) ;  $\mathbf{f}_i$  est la  $i$ -ème valeur du vecteur aléatoire et non observable de  $q$  *facteurs communs* ; et  $\mathbf{e}_i$  la  $i$ -ème valeur du vecteur non observable de résidus, lesquels représentent l'effet combiné de *facteurs spécifiques* et d'une perturbation aléatoire.

On désigne par  $\mathbf{X}$  le tableau  $(n,p)$  dont la  $i$ -ème ligne représente l'observation  $i$ . De même  $\mathbf{F}$  désigne le tableau  $(n,q)$  non observable dont la  $i$ -ème ligne est  $\mathbf{f}_i'$  et  $\mathbf{E}$  le tableau  $(n,p)$  non observable dont la  $i$ -ème ligne est  $\mathbf{e}_i'$ . Le modèle liant l'ensemble des observations aux facteurs hypothétiques s'écrit :

$$\mathbf{X} = \mathbf{F} \mathbf{\Gamma}' + \mathbf{E}$$

$(n,m) \quad (n,q) \quad (q,m) \quad (n,m)$

<sup>24</sup> A l'origine des principes de la méthode se trouvent Spearman (1904) (analyse monofactorielle), puis Garnett (1919) et Thurstone (1947) (analyse multifactorielle).

<sup>25</sup> En économétrie, on distingue habituellement les modèles fonctionnels, ou à effet fixes (comme la régression multiple et le modèle linéaire dans son ensemble), et les modèles structurels ou à effet aléatoire (modèles de variables latentes).

Dans cette écriture, seul $X$ est observable, et le modèle est par conséquent indéterminé.
--

L'identification de ce modèle et l'estimation des paramètres posent des problèmes complexes. Une cascade d'hypothèses *a priori* supplémentaires permet cette identification.

## ➤ VII.7 CLASSIFICATION HIERARCHIQUE, ARBRE DE LONGUEUR MINIMALE

Les techniques de classification automatique<sup>26</sup> sont destinées à produire des groupements d'objets ou d'individus décrits par un certain nombre de variables ou de caractères. Les circonstances d'utilisation sont sensiblement les mêmes que celles des méthodes d'analyse factorielle descriptive présentées aux sections précédentes. Dans la plupart des enchaînements proposés dans le menu « *Create a command file* » de Dm-Vic, la classification est un complément systématique des analyses en axes principaux.

Il existe plusieurs familles d'algorithmes de classification : les *algorithmes hiérarchiques* qui fournissent une hiérarchie de partitions des objets et les algorithmes conduisant directement à des *partitions* comme les méthodes d'agrégation autour de centres mobiles (section VII.8 ci-après). Les *modèles mixtes* (systématiquement mis en œuvre dans Dtm-Vic) combinent les deux approches (section VII.9 ci-après).

### ➤ VII.7.1. L'algorithme de base de la classification hiérarchique (CAH)

Les principes communs aux diverses techniques de classification ascendante hiérarchique sont simples. Il s'agit de créer, à chaque étape de l'algorithme, une partition obtenue en agrégeant deux à deux les éléments les plus proches.

L'algorithme de base de la CAH produit une hiérarchie en partant de la partition dans laquelle chaque élément à classer constitue une classe, pour aboutir à la partition formée d'une seule classe réunissant tous les éléments.

Pour  $n$  éléments à classer, il est composé de  $n$  étapes. A la première étape, il y a donc  $n$  éléments à classer. On construit la matrice de distances entre les  $n$  éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément.

On construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants. On se trouve dans les mêmes conditions qu'à l'étape  $l$ , mais avec seulement  $(n-l)$  éléments à classer.

On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

---

<sup>26</sup> La classification est une branche de l'analyse des données qui constitue une étape fondamentale dans beaucoup de disciplines scientifiques. Elle a donné lieu à des publications nombreuses et diversifiées dont : Sokal et Sneath (1963) et Benzécri (1973).

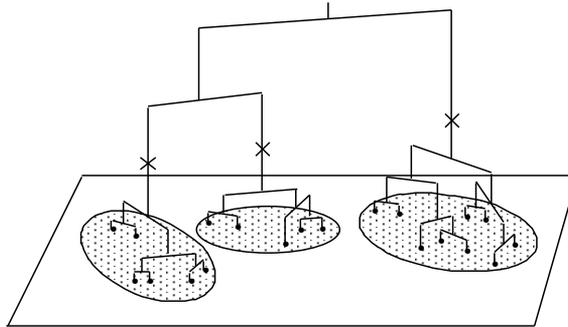


Figure A.8: Dendrogramme ou arbre hiérarchique

L'algorithme ne fournit pas une partition en  $q$  classes d'un ensemble de  $n$  objets mais une *hiérarchie de partitions*, se présentant sous la forme d'*arbres* appelés également *dendrogrammes* et contenant  $n - 1$  partitions (cf. figure A.8). L'intérêt de ces arbres est qu'ils peuvent donner une idée du nombre de classes existant effectivement dans la population. Chaque coupure d'un dendrogramme fournit une partition.

#### ➤ A\_ DISTANCES ENTRE ELEMENTS ET ENTRE GROUPEES

On suppose au départ que l'ensemble des individus à classer est muni d'une *distance*<sup>27</sup>. Ceci ne suppose donc pas que les distances soient toutes calculées au départ : il faut pouvoir les calculer ou les recalculer à partir des coordonnées des points-individus, celles-ci devant être accessibles rapidement.

Dans Dtm-Vic (Etape RECIP) les distances sont calculées *à la volée* à partir des coordonnées factorielles.

Une fois constitué un groupe d'individus, il convient de se demander ensuite sur quelle base on peut calculer une distance entre un individu et un groupe et par la suite une distance entre deux groupes.

Ceci revient à définir une stratégie de regroupements des éléments, c'est-à-dire se fixer des *règles de calcul des distances entre groupements* disjoints d'individus, appelées *critères d'agrégation*.

Cette distance entre groupements pourra en général se calculer directement à partir des distances des différents éléments impliqués dans le regroupement.

Par exemple, si  $x$ ,  $y$ ,  $z$  sont trois objets, et si les objets  $x$  et  $y$  sont regroupés en un seul élément noté  $h$ , on peut définir la distance de ce groupement à  $z$  par la plus petite distance des divers éléments de  $h$  à  $z$  :

$$d(h,z) = \text{Min} \{d(x,z), d(y,z)\}$$

Cette distance s'appelle le *saut minimal (single linkage)* (Sneath, 1957 ; Johnson, 1967) et constitue un critère d'agrégation.

<sup>27</sup> Il s'agira parfois simplement d'une mesure de dissimilarité. Dans ce cas, l'inégalité triangulaire  $d(x,y) \leq d(x,z) + d(y,z)$  n'est pas exigée.

Une autre règle simple et fréquemment employée est celle de la *distance moyenne* ; pour deux objets  $x$  et  $y$  regroupés en  $h$  :

$$d(h, z) = \frac{\{d(x, z) + d(y, z)\}}{2}$$

Plus généralement, si  $x$  et  $y$  désignent des sous-ensembles disjoints de l'ensemble des objets, ayant respectivement  $n_x$  et  $n_y$  éléments,  $h$  est alors un sous-ensemble formé de  $n_x + n_y$  éléments et on définit :

$$d(h, z) = \frac{\{n_x d(x, z) + n_y d(y, z)\}}{n_x + n_y}$$

### ➤ **B\_ ALGORITHME DE CLASSIFICATION**

L'algorithme fondamental de classification ascendante hiérarchique se déroule de la façon suivante :

- ▶ Étape 1 : il y a  $n$  éléments à classer (qui sont les  $n$  individus);
- ▶ Étape 2 : on construit la matrice de distances entre les  $n$  éléments et l'on cherche les deux plus proches, que l'on agrège en un nouvel élément. On obtient une première partition à  $n-1$  classes;
- ▶ Étape 3 : on construit une nouvelle matrice des distances qui résultent de l'agrégation, en calculant les distances entre le nouvel élément et les éléments restants (les autres distances sont inchangées). On se trouve dans les mêmes conditions qu'à l'étape 1, mais avec seulement  $(n-1)$  éléments à classer et en ayant choisi un critère d'agrégation. On cherche de nouveau les deux éléments les plus proches, que l'on agrège. On obtient une deuxième partition avec  $n-2$  classes et qui englobe la première;
- ▶ Étape  $m$  : on calcule les nouvelles distances, et l'on réitère le processus jusqu'à n'avoir plus qu'un seul élément regroupant tous les objets et qui constitue la dernière partition.

## ➤ **VII.7.2 ARBRE DE LONGUEUR MINIMALE : DEFINITION ET ALGORITHMES**

L'ensemble des  $n$  objets à classer peut être considéré comme un ensemble de points d'un espace. Cette représentation est classique si les objets sont décrits par une série de  $p$  variables : on a  $n$  points dans l'espace  $\mathbb{R}^p$  et donc une distance pour chaque paire de points. On représente ainsi l'ensemble des objets et des valeurs de l'indice par un *graphe complet valué*<sup>28</sup>. Mais si le nombre d'objets dépasse quelques unités, ce type de représentation devient inextricable. On cherchera alors à extraire de ce graphe un *graphe partiel* (ayant les mêmes sommets, mais moins d'arêtes) plus aisé à représenter, et permettant néanmoins de bien résumer les valeurs des indices de distance.

<sup>28</sup> Les objets à classer sont alors les nœuds du graphe (non orienté); les lignes continues joignant les paires de points sont les arêtes; et les indices, les valuations de ces arêtes.

Parmi tous les graphes partiels, ceux qui ont une structure d'*arbre*<sup>29</sup> sont particulièrement intéressants, car ils peuvent faire l'objet d'une représentation plane.

Un arbre est un *graphe connexe* (il existe un chemin reliant tout couple de sommets) *sans cycle* (un cycle est un chemin partant et aboutissant au même point sans emprunter deux fois la même arête). On peut définir de façon équivalente un arbre à  $n$  sommets soit comme un graphe sans cycle ayant  $n - 1$  arêtes, soit comme un graphe connexe ayant  $n - 1$  arêtes<sup>30</sup>.

La *longueur* d'un arbre sera la somme des "longueurs" (valeurs de l'indice) de ses arêtes. Parmi tous les graphes partiels qui sont des arbres, l'*arbre de longueur minimale* a retenu depuis longtemps l'attention des statisticiens en raison de ses bonnes qualités descriptives, qui ne sont pas étrangères à sa parenté avec les classifications hiérarchiques. Si l'on désire par exemple déceler rapidement sans ordinateur les traits de structure que peut cacher une matrice de corrélations relative à une trentaine de variables, c'est probablement la plus aisée des procédures à mettre en œuvre.

➤ **ARBRE DE LONGUEUR MINIMALE : ALGORITHME DE KRUSKAL (1956)**

On range les  $n(n - 1)/2$  arêtes dans l'ordre des valeurs croissantes de l'indice. On part des deux premières arêtes, puis on sélectionne successivement toutes les arêtes qui ne font pas de cycle avec les arêtes déjà choisies. On interrompt la procédure dès que l'on a  $n - 1$  arêtes. De cette façon, on est sûr d'avoir obtenu un arbre (graphe sans cycle ayant  $n - 1$  arêtes).

➤ **ARBRE DE LONGUEUR MINIMALE : ALGORITHME DE PRIM (1957)**

On part d'un objet quelconque (sommets du graphe). L'étape  $1$  consiste à chercher l'objet  $v_1$  le plus proche, c'est-à-dire l'arête la plus courte. L'étape  $k$  consiste à adjoindre au recueil d'arêtes déjà constitué  $V_{k-1}$  la plus courte arête  $v_k$  qui touche un des sommets de  $V_{k-1}$ . Il y a  $n - 1$  étapes. Cet algorithme est plus rapide que le précédent. L'arbre obtenu est de longueur minimale car  $V_k$  est à tout moment un arbre de longueur minimale sur les  $k$  sommets concernés. C'est l'algorithme utilisé dans Dtm-Vic.

➤ **VII.8 PARTITIONS, CARTES AUTO-ORGANISEES**

Il s'agit pour l'essentiel des techniques *d'agrégation autour de centres mobiles*, et des *cartes auto-organisées (Self Organising Maps)* appelées encore *cartes de Kohonen*. Ces méthodes sont particulièrement intéressantes dans le cas des grands tableaux car elles sont peu coûteuses en temps calcul et peu gourmandes en espace mémoire.

<sup>29</sup> On ne confondra pas un tel arbre, entendu au sens de la théorie des graphes, et dont les sommets sont les objets à classer, avec l'arbre des parties d'un ensemble (dendrogramme) produit par les techniques de classification hiérarchique, dont les sommets sont des parties (à l'exception des éléments terminaux qui sont les objets à classer eux-mêmes).

<sup>30</sup> On trouvera la démonstration de ces propriétés dans les manuels classiques tels que ceux (historiques) de Berge (1963, 1973).

## VII.8.1 Méthodes de partitionnement



### AGREGATION AUTOUR DE CENTRES MOBILES (OU METHODE K-MEANS)

Bien qu'elle ne fasse appel qu'à un formalisme limité et que son efficacité soit dans une large mesure attestée par les seuls résultats expérimentaux, la méthode d'*agrégation autour de centres mobiles* est probablement la technique de partitionnement la mieux adaptée actuellement aux vastes recueils de données ainsi que la plus utilisée pour ce type d'application. Produisant des partitions des ensembles étudiés, elle est utile aussi bien comme technique de description et d'analyse que comme technique de réduction, généralement en association avec des analyses factorielles et d'autres méthodes de classification.

L'algorithme peut être imputé principalement à Forgy (1965), bien que de nombreux travaux (parfois antérieurs : Thorndike, 1953), le plus souvent postérieurs (MacQueen, 1967; Ball and Hall, 1967) aient été menés parallèlement et indépendamment pour introduire des variantes ou des généralisations. Cette méthode peut être considérée comme un cas particulier de techniques connues sous le nom de *nuées dynamiques* étudiées dans un cadre formel par Diday (1971).

Elle est particulièrement intéressante pour les gros fichiers numériques car les données sont traitées en *lecture directe* : le tableau des données, conservé sur une mémoire auxiliaire (disque) est lu plusieurs fois de façon séquentielle, sans jamais encombrer de zones importantes dans la mémoire vive de l'ordinateur. La lecture directe permet également d'utiliser au mieux les particularités du codage des données, ce qui réduit le temps de calcul dans le cas des codages disjonctifs.



### BASES THEORIQUES DE L'ALGORITHME

Soit un ensemble  $I$  de  $n$  individus à partitionner, caractérisés par  $p$  caractères ou variables.

On suppose que l'espace  $\mathbb{R}^p$  supportant les  $n$  points-individus est muni d'une distance appropriée notée  $d$  (souvent distance euclidienne usuelle ou distance du  $\chi^2$ ). On désire constituer au maximum  $q$  classes.

Les étapes de l'algorithme sont illustrées par l'exemple VI.1 du chapitre VI (section « d » du paragraphe VI.1.3 intitulée : « **Calcul direct d'une partition dans le menu "Visualisation"** »).

• **Étape 0** : On détermine  $q$  centres provisoires de classes (par exemple, par tirage pseudo-aléatoire sans remise de  $q$  individus dans la population à classifier). Les  $q$  centres :

$$\{C_1^0, \dots, C_k^0, \dots, C_q^0\}$$

induisent une première partition  $P^0$  de l'ensemble des individus  $I$  en  $q$  classes :

$$\{I_1^0, \dots, I_k^0, \dots, I_q^0\}$$

Ainsi l'individu  $i$  appartient à la classe  $I_k^0$  s'il est plus proche de  $C_k^0$  que de tous les autres centres<sup>31</sup>.

- **Étape 1** : On détermine  $q$  nouveaux centres de classes :

$$\{C_1^1, \dots, C_k^1, \dots, C_q^1\}$$

en prenant les centres de gravité des classes qui viennent d'être obtenues :

$$\{I_1^0, \dots, I_k^0, \dots, I_q^0\}$$

Ces nouveaux centres induisent une nouvelle partition  $P^1$  de  $I$  construite selon la même règle que pour  $P^0$ . La partition  $P^1$  est formée des classes notées :

$$\{I_1^1, \dots, I_k^1, \dots, I_q^1\}$$

- **Étape  $m$**  : On détermine  $q$  nouveaux centres de classes :

$$\{C_1^m, \dots, C_k^m, \dots, C_q^m\}$$

en prenant les centres de gravité des classes qui ont été obtenues lors de l'étape précédente,

$$\{I_1^{m-1}, \dots, I_k^{m-1}, \dots, I_q^{m-1}\}$$

Ces nouveaux centres induisent une nouvelle partition  $P^m$  de l'ensemble  $I$  formée des classes :

$$\{I_1^m, \dots, I_k^m, \dots, I_q^m\}$$

Le processus se stabilise nécessairement et l'algorithme s'arrête soit lorsque deux itérations successives conduisent à la même partition, soit lorsqu'un critère convenablement choisi (par exemple, la mesure de la variance intra-classes) cesse de décroître de façon sensible, soit encore parce qu'un nombre maximal d'itérations a été fixé *a priori*. Généralement, la partition obtenue finalement dépend du choix initial des centres.

Précisons que la méthode dite des *k-means* (*k-moyennes*) introduite par MacQueen (1967) commence effectivement par un tirage pseudo-aléatoire de centres ponctuels. Cependant la règle de calcul des nouveaux centres n'est pas exactement la même que celle qui vient d'être exposée. On n'attend pas d'avoir procédé à la réaffectation de tous les individus pour modifier la position des centres : chaque réaffectation d'individus entraîne une petite modification de la position du centre correspondant<sup>32</sup>.

En une seule itération, cette procédure peut ainsi donner une partition de bonne qualité.

<sup>31</sup> Les classes sont alors délimitées dans l'espace par les cloisons polyédrales convexes formées par les plans médiateurs des segments joignant tous les couples de centres.

<sup>32</sup> On parle parfois d'algorithme en ligne (*on line*) pour ce type de modification en cours de lecture, alors que la méthode exposée plus haut procède par paquet (*batch*).

Mais celle-ci dépendra de l'ordre des individus sur le fichier.

## ➤ VII.8.2 LES CARTES AUTO-ORGANISEES DE KOHONEN

L'objectif des cartes auto-organisées de Kohonen est de classer un ensemble d'observations de façon à conserver la topologie initiale de l'espace dans lesquelles ces observations sont décrites.

### ➤ VII.8.2.1 Le principe

Les cartes de Kohonen<sup>33</sup> cherchent à représenter dans un espace à deux (parfois trois) dimensions les lignes ou les colonnes d'un tableau en respectant la notion de voisinage dans l'espace des éléments à classer. Tout comme dans le cas de l'analyse en composantes principales, il est utile d'imaginer au départ l'ensemble des données comme un nuage de points dans un espace de grande dimension.

Le principe est de considérer une carte comme une grille rectangulaire (parfois hexagonale) aux mailles déformables, laquelle, une fois dépliée épouse au mieux les formes du nuage de points. Les nœuds de la grille sont les *neurones* de la carte. Chaque point du nuage est projeté sur le nœud dont il est le plus proche. De fait, chaque point, décrit initialement dans un espace multidimensionnel est représenté à la fin par deux coordonnées donnant la position du *neurone* sur la carte : l'espace est réduit. L'ensemble des points affectés à un même *neurone* sont proches dans l'espace initial. Ils décrivent et regroupent des individus semblables.

On définit *a priori* une notion de voisinage entre classes et les observations voisines dans l'espace des variables de dimension  $q$  appartiennent après classement à la même classe ou à des classes voisines. Ces voisinages peuvent être choisis de diverses manières mais en général on les suppose directement contigus sur la grille rectangulaire (ce qui représente alors 8 voisins pour un *neurone*).

### ➤ VII.8.2.2 L'algorithme

L'algorithme d'apprentissage pour classer  $m$  points est itératif<sup>34</sup>. L'initialisation consiste à associer à chaque classe  $k$  un centre provisoire  $C_k$  à  $q$  composantes choisi de manière aléatoire dans l'espace à  $q$  dimensions contenant les  $m$  mots à classer. A chaque étape on choisit un mot  $i$  au hasard que l'on compare à tous les centres provisoires et l'on affecte le mot au centre  $C_{k_0}$  le plus proche au sens d'une distance donnée *a priori*. On rapproche alors du mot  $i$  le centre  $C_{k_0}$  et les centres voisins sur la carte ce qui s'exprime à l'étape  $t$  par :

$$C_k(t+1) = C_k(t) + \varepsilon(i(t+1) - C_k(t))$$

où  $i(t+1)$  est le mot présenté à l'étape  $t+1$ ,  $\varepsilon$  un paramètre d'adaptation positif et

<sup>33</sup> Introduites en 1981 par Teuvo Kohonen, elles font partie des méthodes dites *neurales* (cf. Kohonen, 1989). Elles donnent lieu à plusieurs applications relevant par exemple de l'analyse de textes, les diagnostics médicaux et industriels, les contrôles de processus, la robotique.

<sup>34</sup> On se réfère dans la présentation de l'algorithme au cours de P.Letremy et M.Cottrell (SAMOS-MATISSE, Université Paris I). Voir aussi Thiria et al. (1997).

inférieur à 1. Cette expression n'intervient que pour le centre  $C_{k_0}$  et ses voisins.

Cet algorithme est analogue à celui des centres mobiles, mais dans ce dernier cas, il n'existe pas de notion de voisinage entre classes et on ne modifie à chaque étape que la position du centre  $C_{k_0}$ . L'auto-organisation de la carte de Kohonen est la conséquence de la notion de voisinage. Comme l'algorithme des centres mobiles, cet algorithme est très adapté aux applications où les données sont importantes et où il n'est pas utile de les stocker.

## ➤ VII.9 CLASSIFICATION MIXTE (OU HYBRIDE)

Les algorithmes de classification sont plus ou moins bien adaptés à la gestion d'un nombre important d'objets à classer. Les méthodes de partitionnement (agrégation autour des centres mobiles ou cartes auto-organisées) offrent des avantages incontestables puisqu'elles permettent d'obtenir une partition sur un ensemble volumineux de données à un faible coût, mais elles présentent l'inconvénient de fixer a priori le nombre de classes et de produire des partitions dépendant des premiers centres choisis. Au contraire, la classification hiérarchique est une famille d'algorithmes que l'on peut qualifier de "déterministes" (i.e. qui donnent toujours les mêmes résultats à partir des mêmes données). Par contre si ces algorithmes donnent des indications sur le nombre de classes à retenir ils sont mal adaptés aux vastes recueils de données. Aussi on procède souvent à une classification mixte qui cumule les avantages des deux types de classification.

### ➤ VII.9.1 STRATEGIE DE CLASSIFICATION MIXTE

La classification autour des centres mobiles peut en fait être utilisée comme auxiliaire d'autres méthodes de classification. En fournissant des partitions de vastes ensembles de données, elle permet de réduire la dimension de l'ensemble des éléments à classer en opérant des regroupements préalables.

De ce fait, un algorithme de classification qui paraît actuellement bien adapté au partitionnement d'un ensemble comprenant des milliers ou des dizaines de milliers d'individus est un *algorithme mixte*. L'idée repose sur la combinaison des deux techniques de classification présentées précédemment. Cette idée a été mise en œuvre spontanément par de nombreux praticiens ; elle se trouve, par exemple, sous le nom de *hybrid clustering* dans Wong (1982).

#### ➤ A) LES ETAPES DE L'ALGORITHME

L'algorithme de *classification mixte* procède en trois phases : l'ensemble des éléments à classer subit un partitionnement initial (centres mobiles) de façon à obtenir quelques dizaines, voire quelques centaines de groupes homogènes ; on procède ensuite à une agrégation hiérarchique de ces groupes, dont le dendrogramme suggérera éventuellement le nombre de classes finales à retenir ; et enfin, on optimise (encore par la technique des centres mobiles appliquée à partir des centres de classe déjà trouvés) la ou les partitions correspondant aux coupures choisies de l'arbre. La figure 6.3 - 1

schématise les différentes étapes de l'algorithme de classification mixte.

➤ **1 - Partitionnement initial**

Cette première étape vise à obtenir, rapidement et à un faible coût, une partition des  $n$  objets en  $k$  classes homogènes, où  $k$  est largement plus élevé que le nombre  $s$  de classes désiré dans la population, et largement plus petit que  $n$ . Nous utilisons, pour ce partitionnement initial en quelques dizaines de classes, un algorithme de partitionnement. Ce sera, par exemple, l'algorithme de l'agrégation autour des centres mobiles.

➤ **2 - Agrégation hiérarchique des classes obtenues**

La seconde étape consiste à effectuer une classification ascendante hiérarchique où les éléments terminaux de l'arbre sont les  $k$  classes de la partition initiale. Quelques uns de ces groupements peuvent être proches les uns des autres. Ils correspondent à un groupe "réel" qui aurait été coupé artificiellement par l'étape précédente. D'autre part, la procédure crée, en général, plusieurs petits groupes ne contenant parfois qu'un seul élément. Le but de l'étape d'agrégation hiérarchique est de reconstituer les classes qui ont été fragmentées et d'agréger des éléments apparemment dispersés autour de leurs centres d'origine. L'arbre correspondant est construit selon le critère de Ward qui tient compte des masses au moment des choix des éléments à agréger.

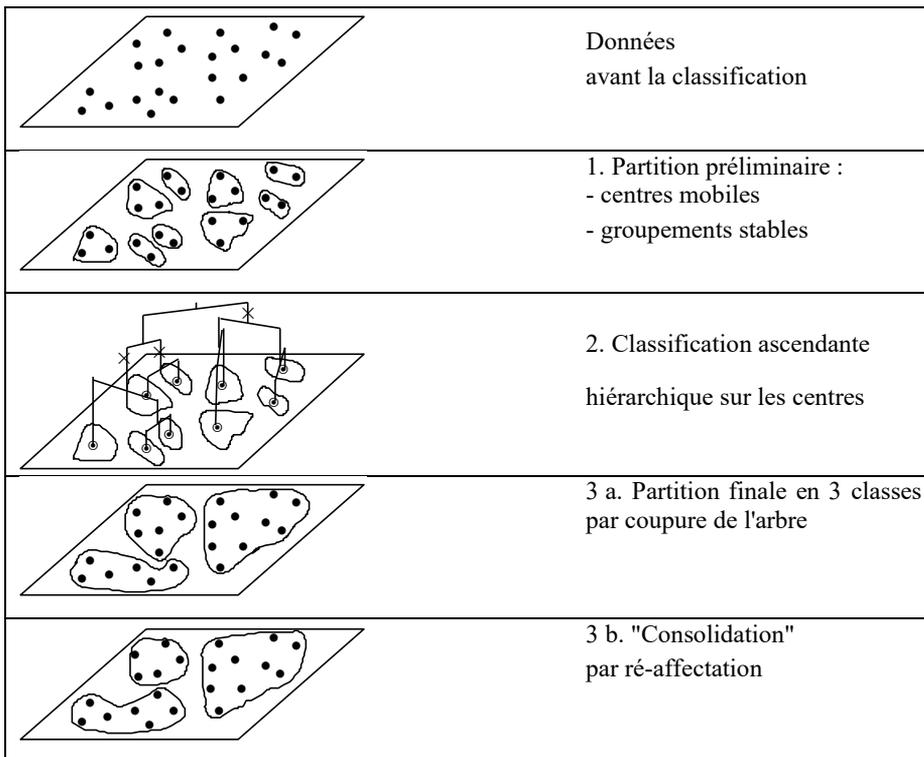


Figure A.10: Schématisation de la classification mixte

➤ **3 - Partitions finales**

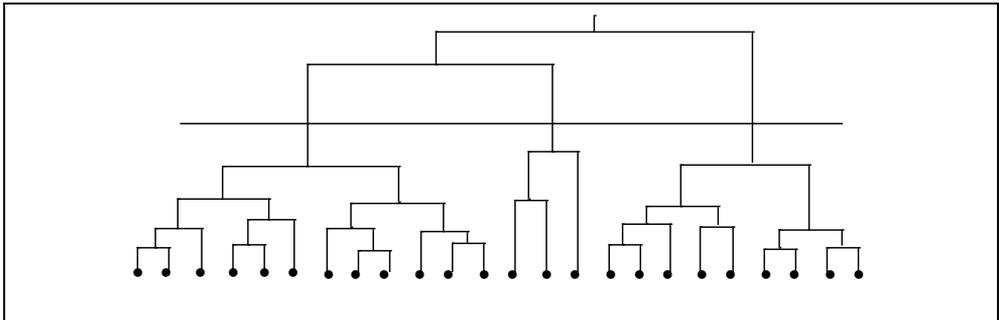
La partition finale de la population est définie par coupure de l'arbre de la classification ascendante hiérarchique. L'homogénéité des classes obtenues peut être optimisée par réaffectations.

➤ **B) CHOIX DU NOMBRE DE CLASSES PAR COUPURE DE L'ARBRE**

Le choix du niveau de la coupure, et ainsi du nombre de classes de la partition, peut être facilité par une inspection visuelle de l'arbre (cf. figures A.11 et A.12) : la coupure doit être faite après les agrégations correspondant à des valeurs peu élevées de l'indice, qui regroupent les éléments les plus proches les uns des autres, et avant les agrégations correspondant à des valeurs élevées de l'indice, qui dissocient les groupes bien distincts dans la population.

En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité, car les individus regroupés auparavant étaient proches, et ceux regroupés après la coupure sont nécessairement éloignés, ce qui est la définition d'une bonne partition.

En pratique, la situation n'est pas aussi clairement définie que le montre la figure A.11. L'utilisateur pourra choisir entre deux ou trois niveaux de coupure possibles et donc entre deux ou trois partitions finales.



**Figure A.11 : Coupure visuelle de l'arbre**

La coupure de l'arbre peut être facilitée par l'examen de l'histogramme des indices croissants de niveau et l'on coupera au niveau pour lequel cet histogramme marque un palier important. Toute barre de cet histogramme indique la valeur de l'indice d'une agrégation c'est-à-dire la perte d'inertie obtenue en passant d'une partition en  $s$  classes à la partition en  $s - 1$  classes.

La situation idéale est montrée par la figure A.12 (a) où l'on observe un palier évident entre le 4<sup>ème</sup> et le 5<sup>ème</sup> indice suggérant ainsi une bonne partition en cinq classes. La figure A.12 (b) est typique de la situation où il est difficile de décider d'un nombre "réel" de groupes dans la population. Mais une telle partition, en  $s$  classes par exemple,

n'est pas la meilleure possible, car l'algorithme de classification hiérarchique n'a pas la propriété de donner à chaque étape une partition optimale. C'est pourquoi une procédure de « consolidation » est nécessaire.

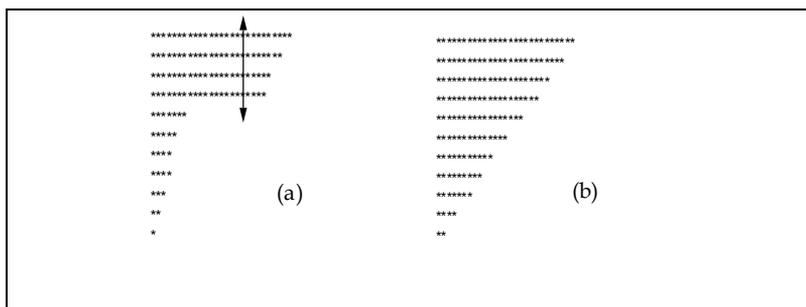


Figure A.12 : Histogrammes des indices de niveau

➤ **C) PROCEDURE DE CONSOLIDATION**

Pour améliorer la partition obtenue, on utilise de nouveau une procédure d'agrégation autour des centres mobiles dont on sait qu'elle ne peut qu'augmenter l'inertie entre les classes à chaque itération. Cette procédure de consolidation a pour effet d'optimiser, par réaffectation, la partition obtenue par coupure de l'arbre hiérarchique. Malgré la relative complexité de la procédure, on ne peut toujours pas être assuré d'avoir trouvé la "meilleure partition en *k* classes" mais on s'en approche vraisemblablement dans beaucoup de situations courantes.

➤ **VII.9.2 DESCRIPTION STATISTIQUE DES CLASSES**

La description automatique des classes constitue en pratique une indispensable étape de toute procédure de classification.

Les aides à l'interprétation des classes sont généralement fondées sur des comparaisons de moyennes ou de pourcentages à l'intérieur des classes avec les moyennes ou les pourcentages obtenus sur l'ensemble des éléments à classer. Pour sélectionner les variables continues ou les modalités des variables nominales les plus caractéristiques de chaque classe, on mesure l'écart entre les valeurs relatives à la classe et les valeurs globales. Ces statistiques peuvent être converties en un critère appelé *valeur-test*<sup>35</sup> permettant d'opérer un tri sur les variables, et de désigner ainsi les variables les plus caractéristiques (cf. Morineau, 1984).

Parmi les variables figurent également celles qui n'ont pas contribué à la construction des classes mais qui peuvent participer à leur description sur le même principe que les variables supplémentaires dans une analyse factorielle.

Ces variables permettent *a posteriori* d'identifier et de caractériser les regroupements établis à partir des variables actives.

<sup>35</sup> Voir la section VII.10 ci-dessous.

## ➤ VII.10 OUTILS DE VALIDATION

Les notions de *valeur-test* et de *variable supplémentaire* jouent un rôle important en analyse descriptive de données. Les *valeurs-test* (section VII.10.1) sont un outil d'inférence statistique élémentaire, mais polyvalent et très utile, surtout si l'utilisateur est averti des problèmes de *comparaisons multiples* qui ne manquent pas d'intervenir (section VII.10.2).

La technique des variables supplémentaires (section VII.10.3) est un outil fondamental de valorisation des méthodes factorielles, qui permet une validation *externe* des résultats, à la fois épreuve de cohérence et enrichissement des interprétations.

Les deux autres outils de validation utilisés dans cet ouvrage sont les *intervalles de confiance d'Anderson* et les procédures de rééchantillonnage *bootstrap*.

Les procédures de rééchantillonnage *bootstrap* (section VII.10.5) sont utilisées dans pratiquement tous les exemples présentés dans ce manuel.

### ➤ VII.10.1 QU'EST-CE QU'UNE VALEUR-TEST ?

La valeur-test est un critère qui permet d'apprécier rapidement si une modalité d'une variable nominale (*i.e.* : une catégorie de répondants) a une position *significative* sur un axe. Pour cela, on teste l'hypothèse selon laquelle un groupe d'individus, correspondant à une modalité donnée d'une variable nominale supplémentaire (comme la modalité *profession libérale, cadre supérieur* pour la variable nominale *catégorie socio-professionnelle*, par exemple), peut être considéré comme tiré au hasard, sans remise, dans la population.

Dans le cas d'un véritable tirage au hasard, le centre de gravité du sous-nuage représentant le groupe (*i.e.* : la modalité) s'éloigne peu du centre de gravité du nuage global correspondant à tout l'échantillon.

On convertit alors la coordonnée de cette modalité sur l'axe en une *valeur-test* qui est, sous cette hypothèse, la réalisation d'une variable normale centrée réduite. Autrement dit, dans l'hypothèse selon laquelle une modalité a une composition *aléatoire*, la valeur-test correspondante a 95% de chances d'être comprise dans l'intervalle  $[-1.96, +1.96]$ .

Supposons qu'une modalité  $j$  concerne  $n_j$  individus. Si ces  $n_j$  individus sont tirés au hasard (c'est ce qu'on appelle l'hypothèse nulle  $H_0$ ) parmi les  $n$  individus analysés (tirage supposé sans remise), la moyenne de  $n_j$  coordonnées tirées au hasard dans l'ensemble fini des  $n$  valeurs  $\psi_{\alpha i}$  (coordonnée du répondant  $i$  sur l'axe  $\alpha$ ) est une variable aléatoire  $X_{\alpha j}$  :  $X_{\alpha j} = \frac{1}{n_j} \sum_{i \in I(j)} \psi_{\alpha i}$  avec pour espérance  $E(X_{\alpha j}) = 0$  et pour variance<sup>36</sup>

<sup>36</sup> Il s'agit de la formule classique donnant la variance d'une moyenne lors d'un tirage sans remise de  $n_j$  objets parmi  $n$ , en fonction de la variance totale  $\lambda_{\alpha}$ , qui est aussi, dans le cas des coordonnées factorielles, la valeur propre correspondant à l'axe  $\alpha$ .

$$\text{Var}_{H_0}(X_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{\lambda_{\alpha}}{n_j}$$

Dans la formule donnant  $X_{\alpha j}$ ,  $I(j)$  est le sous-ensemble des répondants caractérisés par la modalité  $j$  de la variable nominale.

La coordonnée  $\varphi_{\alpha j}$  de la modalité  $j$  est proportionnelle à la variable aléatoire  $X_{\alpha j}$  et s'écrit

$$\text{ainsi : } \varphi_{\alpha j} = \frac{1}{\sqrt{\lambda_{\alpha}}} X_{\alpha j}$$

$$\text{On a donc } E(\varphi_{\alpha j}) = 0 \text{ et } \text{Var}_{H_0}(\varphi_{\alpha j}) = \frac{n - n_j}{n - 1} \frac{1}{n_j}$$

La quantité  $t_{\alpha j}$  :  $t_{\alpha j} = \sqrt{n_j \frac{n-1}{n-n_j}} \varphi_{\alpha j}$  mesure en *nombre d'écart-types* la distance entre la modalité  $j$ , c'est-à-dire le quasi-barycentre des  $n_j$  individus, et l'origine, sur l'axe factoriel  $\alpha$ . On appelle cette quantité « *valeur-test* ». D'après le théorème de la limite centrale (*central limit theorem*), sa distribution tend vers une loi de Laplace-Gauss centrée réduite.

On considère alors comme occupant une *position significative* les modalités dont les valeurs-test sont supérieures à 2 (pour 1.96) en valeur absolue, ce qui correspond approximativement au seuil usuel de probabilité de 5%. Souvent les valeurs-test sont largement supérieures à ce seuil. On les utilise alors pour trier les modalités, des plus significatives au moins significatives. La valeur-test systématise la notion de *t-value* souvent utilisée dans la littérature statistique.

On doit noter que les valeurs-test n'ont de sens que pour les modalités supplémentaires (cf. section suivante), ou des modalités actives ayant des contributions absolues faibles, c'est-à-dire se comportant en fait comme des modalités supplémentaires<sup>37</sup>. Lorsque l'on dispose d'un nombre important de modalités supplémentaires, les valeurs-test permettent de repérer rapidement les modalités utiles à l'interprétation d'un axe ou d'un plan factoriel.

## ➤ VII.10.2 PROBLEMES DE COMPARAISONS MULTIPLES

Le calcul simultané de plusieurs valeurs-test ou de plusieurs seuils de probabilités se heurte à l'écueil des *comparaisons multiples*, bien connu des statisticiens ; cf. O'Neill et Wetherill (1971), Saville (1990), Westfall et Young (1993), Westfall *et al.* (1999), Hsu (1996).

Supposons que l'on projette 100 modalités *supplémentaires* (cf. section suivante VII.10.3) qui soient vraiment tirées au hasard. Les valeurs-test attachées à ces modalités sont alors toutes des réalisations de variables aléatoires normales centrées réduites indépendantes.

Dans ces conditions, *en moyenne*, sur 100 valeurs-test calculées, cinq seront en dehors

<sup>37</sup> Les coordonnées sur un axe des individus correspondant à une modalité active ne peuvent être considérées comme tirées au hasard, puisque la modalité a contribué à construire l'axe.

de l'intervalle  $[-1.96, +1.96]$  et seront, en apparence seulement, significatives. Le seuil de 5% n'a de sens en fait que pour un seul test, et non pour des tests multiples.

On résout en pratique cette difficulté en choisissant un seuil plus sévère<sup>38</sup>. Le seuil le plus sévère et pessimiste que l'on puisse imaginer est le « seuil de *Bonferroni* » (on divise le seuil initial par le nombre de tests : dans le cas de 210 tests :  $0.05 / 210 = 2.4 \cdot 10^{-4}$ ). La valeur-test unilatérale correspondante est de 3.49. Cette valeur nous fournit un garde-fou prudent à l'excès<sup>39</sup>.

*Une solution pragmatique (cas multidimensionnel) : le bootstrap.*

La technique de validation par *bootstrap* dont il sera question plus loin dans cette annexe apporte une contribution intéressante au difficile problème des comparaisons multiples, car les répliqués d'échantillons permettent de prendre en compte toutes les variables simultanément, et donc de prendre en compte l'interdépendance des variables. Il s'agit d'un test global, et non plus de tests séparés pour chaque variable. Une illustration en est donnée, par exemple, par la figure des sections III.1.4 et III.2.4 du chapitre III qui représentent les zones de confiance simultanées des mots, dont certains apparaissent comme significativement distincts. Dans ce cas, les tests ne sont pas réalisés isolément ni en série, mais simultanément.

### ➤ VII.10.3 UTILITE DES ELEMENTS SUPPLEMENTAIRES

L'analyse factorielle permet de trouver des sous-espaces de représentation des proximités entre points-individus ou entre points-variables. Elle s'appuie pour cela sur des éléments (individus ou variables) dits *actifs*.

Il est possible d'introduire en supplémentaire d'autres points (ou éléments) que l'on ne souhaite pas faire intervenir dans la composition et définition des axes mais dont on veut connaître les positions dans les espaces factoriels<sup>40</sup>. On projette alors ces points après la construction des axes factoriels dans ce nouveau repère. Cette projection se fait de façon très simple en utilisant les formules dites *de transition*, que ce soit en analyse en composantes principales ou en analyse des correspondances.

C'est le cas lorsque l'on souhaite caractériser les axes sémiométriques<sup>41</sup> par les critères socio-démographiques (variables nominales) de la population enquêtée (cf. section VI.1). Ces critères définissent en fait des groupes d'individus et sont considérés comme des modalités de variables nominales. Ce sont les centres de gravité de ces groupes qui sont positionnés dans l'espace des variables. La valeur-test permet d'en apprécier la significativité sur l'axe.

<sup>38</sup> Les valeurs-test permettent surtout de *classer* les modalités supplémentaires par ordre d'intérêt décroissant, ce qui constitue une aide précieuse à l'interprétation des facteurs.

<sup>39</sup> Cf., par exemple, Hochberg (1988), Perneger (1998).

<sup>40</sup> On peut citer trois raisons qui peuvent susciter la mise en supplémentaire d'un point : 1) enrichir l'interprétation des axes par des variables (de nature ou de thématique différente de celle des éléments actifs) n'ayant pas participé à leur construction ; 2) adopter une optique de prévision en projetant les variables supplémentaires dans l'espace des individus. Celles-ci seront « expliquées » par les variables actives ; 3) faire ressortir l'essentiel d'une structure masquée par l'existence d'un point actif, de faible masse, mais très excentré qui pourrait déformer le nuage.

<sup>41</sup> Ces axes, rappelons-le, sont définis par les variables actives que sont les mots.

### ➤ VII.10.4 INTERVALLES DE CONFIANCE D'ANDERSON

Anderson (1963) a calculé les lois limites des valeurs propres d'une analyse en composantes principales sans nécessairement supposer que les valeurs théoriques correspondantes sont distinctes.

L'ampleur de l'intervalle donne une indication sur la stabilité de la valeur propre vis-à-vis des fluctuations dues à l'échantillonnage supposé laplacien (normal). L'empiétement des intervalles de deux valeurs propres consécutives suggérera donc l'égalité de ces valeurs propres. Les axes correspondants sont alors définis à une rotation près. Ainsi l'utilisateur pourra éviter d'interpréter un axe instable.

Si les valeurs propres théoriques  $\lambda_\alpha$  de la matrice des covariances théorique  $\Sigma$  sont distinctes, les valeurs propres  $\hat{\lambda}_\alpha$  de la matrice des covariances empirique  $S$  suivent asymptotiquement des lois normales d'espérance  $\lambda_\alpha$  et de variance  $2\lambda_\alpha^2/(n-l)$  où  $n$  est la taille de l'échantillon. On en déduit les intervalles de confiance approchés au seuil 95% :

$$\lambda_\alpha \in \left[ \hat{\lambda}_\alpha \left( 1 - 1.96\sqrt{2/(n-l)} \right) ; \hat{\lambda}_\alpha \left( 1 + 1.96\sqrt{2/(n-l)} \right) \right]$$

Les intervalles de confiance d'Anderson concernent en fait aussi bien les valeurs propres des matrices des covariances que des matrices de corrélations. Les simulations entreprises montrent que les intervalles de confiance obtenus sont en général « prudent » : le pourcentage de couverture de la vraie valeur est le plus souvent supérieur au seuil de confiance annoncé.

Dans tous les cas, la nature asymptotique des résultats et l'hypothèse sous-jacente de normalité<sup>42</sup> font considérer les résultats comme purement indicatifs.

### ➤ VII.10.5 LES TECHNIQUES DE *BOOTSTRAP*

Face aux résultats d'une analyse factorielle, certaines questions sur la validité des axes obtenus se posent naturellement : Existe-t-il des critères pour tester la stabilité d'une structure et la valider ? Quelle est la part de l'échantillonnage des individus mais aussi, notion plus complexe, celle du choix ou de la sélection des variables ?

Pour tenter de répondre partiellement à ces questions, on peut recourir aux méthodes empiriques de validation. Elles consistent à perturber le tableau initial par des ajouts ou retraits d'éléments du tableau, individus ou variables (poids, codage, etc.). L'hypothèse est la suivante : si les perturbations effectuées sur les échantillons n'affectent pas les configurations observées dans les sous-espaces, celles-ci sont supposées stables et la structure mise en évidence est alors « significative ».

Les méthodes de rééchantillonnage se proposent de systématiser cette démarche<sup>43</sup>. Celle

<sup>42</sup> Muirhead (1982) a montré que l'hypothèse d'existence des quatre premiers moments pour la loi théorique de l'échantillon suffisait pour valider ces intervalles.

<sup>43</sup> Ce sont des méthodes de calculs intensifs qui reposent sur des techniques de simulations d'échantillons à partir d'un seul échantillon. Rendues possibles par la puissance de calcul des ordinateurs, ces techniques se

du *bootstrap*, non paramétrique dans sa forme classique, est bien adaptée au problème de la validité des structures observées dans un plan factoriel ; elle calcule, à partir de simulations, des zones de confiance pour les positions des points-lignes et des points-colonnes.

➤ **Principe du bootstrap**

La technique du *bootstrap*, introduite par Efron (1979), consiste à simuler  $s$  échantillons de même taille  $n$  que l'échantillon initial. Le nombre de simulations  $s$  varie selon les situations, dans le cas multidimensionnel qui nous intéresse, une valeur relativement faible ( $10 < s < 30$ ) apparaît suffisante. Ces échantillons sont obtenus par tirage au hasard *avec remise* parmi les  $n$  individus observés au départ, ceux-ci ayant tous la même probabilité  $1/n$  d'être choisis. Certains individus apparaîtront plusieurs fois et auront de ce fait un poids élevé (2, 3,...) alors que d'autres seront absents (poids nul).

Cette méthode est employée pour analyser la variabilité de paramètres statistiques simples en produisant des intervalles de confiance de ces paramètres. Elle peut aussi être appliquée à de nombreux problèmes pour lesquels on ne peut pas estimer analytiquement la variabilité d'un paramètre. Ceci est le cas pour les caractéristiques des méthodes multidimensionnelles où les hypothèses de multinormalité sont rarement vérifiées. L'analyse en composantes principales est un domaine d'application qui a donné à un grand nombre de travaux utilisant les méthodes de rééchantillonnage de *bootstrap*.

Prenons l'exemple de l'estimation du coefficient de corrélation  $r$  entre deux variables ou entre une variable et un facteur. Le principe consiste à calculer le coefficient de corrélation pour chaque échantillon répliqué (pour lequel on effectue un tirage avec remise des *couples* d'observations). On établit alors la distribution des fréquences du coefficient de corrélation (représentée par l'histogramme des  $s$  valeurs du coefficient  $r$  correspondant aux  $s$  réplifications). Puis on calcule à partir de l'histogramme la probabilité pour que le coefficient de corrélation d'un échantillon soit compris dans différentes fourchettes de valeurs définissant ainsi les intervalles de confiance. On obtient une estimation de la précision de la valeur de  $r$  obtenue sur l'échantillon de base sans faire l'hypothèse d'une distribution normale des données. Les bornes de l'intervalle de confiance peuvent être estimées directement par les quantiles de la distribution simulée.

Pour estimer les coordonnées factorielles issus d'une analyse en composantes principales, le principe est le même que pour le coefficient de corrélation ; on effectue sur chaque échantillon simulé, une analyse en composantes principales puis on établit une distribution de fréquences pour chacune des composantes<sup>44</sup>.

La méthode de *bootstrap* donne dans la plupart des cas une bonne image de la précision statistique de l'estimation sur un échantillon. Les recherches théoriques menées par Efron, en particulier, montrent que, pour de nombreux paramètres statistiques, l'intervalle de confiance correspondant à la distribution simulée par *bootstrap* et celui

---

substituent dans certains cas aux procédures plus classiques qui reposent sur des hypothèses contraignantes. Elles sont les seules procédures possibles lorsque la complexité analytique du problème ne permet pas d'inférence classique.

<sup>44</sup> On trouvera des compléments sur l'intérêt et les limites de cette méthode dans les travaux de Diaconis et Efron (1983) et de Young (1994).

correspondant à la distribution réelle sont généralement de même amplitude.

### ➤ *Mise en œuvre et calcul des zones de confiance*

Il existe plusieurs procédures pour tester, par la méthode de bootstrap, la stabilité des coordonnées factorielles. Gifi (1981), Meulman (1982), Greenacre (1984) ont réalisé des premiers travaux dans le contexte de l'analyse des correspondances simples ou multiples. Dans le cas de l'analyse en composantes principales, Diaconis et Efron (1983), Holmes (1989), Stauffer et al. (1985), Daudin et al. (1988) ont posé le problème du choix du nombre d'axes pertinent et ont proposé des intervalles de confiance pour les points du sous-espace défini par les principaux axes. Les paramètres correspondant sont calculés à partir des échantillons répliqués et supposent des contraintes qui dépendent de ces échantillons.

Pour pallier ces difficultés, il faut se référer à un espace factoriel commun. Plusieurs variantes sont possibles.

On présentera brièvement deux techniques : le *bootstrap total* et le *bootstrap partiel*.

Pour des développements plus étendus, on se reportera à l'ouvrage SEM-2006 ou aux boutons « Validation » et « Bootstrap + » de la barre verticale « Statistical tools, some reminders » du menu d'accueil de Dtm-Vic.

Le *bootstrap total* consiste à réaliser autant d'analyses en composantes principales qu'il y a de répliqués, moyennant une série de transformations afin de retrouver des axes homologues au cours des diagonalisations successives des  $s$  matrices de corrélation répliquées  $C_k$  ( $C_k$  correspond à la  $k$ -ème répliquée). Ces transformations sont des changements de signe des axes, rotations ou permutations d'axes. Cette méthode, proposée par Milan et Whittaker (1995) est en défaut s'il existe des valeurs propres très voisines.

Dans le *bootstrap partiel*, proposé par Greenacre (1984) dans le cas de l'analyse des correspondances, il n'est pas nécessaire de calculer les valeurs et vecteurs propres pour l'ensemble des simulations : les axes principaux calculés sur les données originales non perturbées, jouent un rôle privilégié (la matrice des corrélations initiale  $C$  est en effet l'espérance mathématique des matrices perturbées  $C_k$ ).

Le *bootstrap partiel* se fonde sur la projection en tant qu'*éléments supplémentaires* des points répliqués sur les sous-espaces de référence fournis par les axes principaux de la matrice de corrélation  $C = X'X$ , provenant de l'échantillon initial, donnés par :

$$\mathbf{u}_q = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}' \mathbf{v}_q$$

où  $\mathbf{u}_q$ ,  $\mathbf{v}_q$  sont respectivement les  $q$ -èmes vecteurs propres de  $X'X$  et  $XX'$  et  $\lambda_q$  la valeur propre associée.

La projection<sup>45</sup> de la  $k$ -ème répliquée des  $m$  variables (mots) est donnée par le vecteur

<sup>45</sup> La projection des répliqués Bootstrap, dans le contexte de l'analyse en composantes principales, consiste à utiliser le fait que la coordonnée d'une variable sur un axe factoriel n'est autre que son coefficient de corrélation avec la variable « coordonnées des individus sur l'axe ». On calcule donc les répliqués de ce coefficient, ce qui revient à repondérer, pour chaque répliquée, les individus avec les *ponds Bootstrap* qui caractérisent un tirage sans remise. On obtient, comme sous-produit, des répliqués de la variance sur l'axe,

$\mathbf{u}_q(k)$  de  $\mathbb{R}$  tel que :

$$\mathbf{u}_q(k) = \frac{1}{\sqrt{\lambda_q}} \mathbf{X}' \mathbf{D}_k \mathbf{v}_q$$

et  $\mathbf{D}_k$  désigne la matrice diagonale  $(n, n)$  des *poids bootstrap* associée à la  $k$ -ème réplique<sup>46</sup>.

Dans le cas du bootstrap partiel, les analyses des matrices  $\mathbf{C}_k$  ne sont en aucun cas nécessaires puisque les vecteurs propres sont obtenus à partir de l'analyse en composantes principales de la matrice  $\mathbf{C}$ .

La variabilité bootstrap s'observe donc mieux sur le repère fixe initial, qui est d'ailleurs le moins mauvais repère, étant le seul à utiliser des données originales non perturbées. Cette technique, éprouvée empiriquement, répond parfaitement aux préoccupations des utilisateurs dans le cas de l'analyse en composantes principales.

### ➤ *Bootstrap sur l'ensemble des variables (cas de l'ACP)*

Classiquement les répliques sont obtenues par des tirages avec remise dans l'ensemble des  $n$  individus. Dans certains cas assez exceptionnels, on se propose de tester la stabilité des structures vis-à-vis de l'ensemble des variables. On peut alors répliquer cet ensemble par la méthode du *bootstrap total*.

Nous supposons ainsi implicitement que l'ensemble des variables (par exemple l'ensemble des mots du questionnaire dans le cas de la sémiométrie évoqué en section VI.1) constitue un échantillon de  $m$  variables extrait aléatoirement d'un ensemble potentiel de variables (ensemble des mots dans le cas de a sémiométrie).

Nous cherchons à perturber cet échantillon de mots selon les mêmes principes que le *bootstrap* opéré sur les individus.

Pour cela, on appelle  $\mathbf{B}_k$  la matrice diagonale  $(m, m)$  dont les éléments diagonaux sont les poids des mots de la  $k$ -ème réplique Bootstrap  $(1, 0, 2, 0, \dots)$ . La matrice  $\mathbf{X}$  d'ordre  $(n, n)$  initiale étant supposée centrée, la matrice à diagonaliser est la matrice  $\mathbf{T}_k$  qui vaut :  $\mathbf{T}_k = \mathbf{X} \mathbf{B}_k \mathbf{X}' = \mathbf{X} \mathbf{B}_k^{1/2} \mathbf{B}_k^{1/2} \mathbf{X}'$

On obtient donc :

$$\mathbf{X} \mathbf{B}_k \mathbf{X}' \mathbf{v}_q(k) = \lambda_q \mathbf{v}_q(k)$$

en multipliant chaque terme par  $\mathbf{B}_k^{1/2} \mathbf{X}'$  on a :

$$\mathbf{B}_k^{1/2} \mathbf{X}' \mathbf{X} \mathbf{B}_k^{1/2} \mathbf{B}_k^{1/2} \mathbf{X}' \mathbf{v}_q(k) = \lambda_q \mathbf{B}_k^{1/2} \mathbf{X}' \mathbf{v}_q(k)$$

et en posant  $\mathbf{u}_q(k) = \mathbf{B}_k^{1/2} \mathbf{X}' \mathbf{v}_q(k)$  alors :

$$\mathbf{B}_k^{1/2} \mathbf{X}' \mathbf{X} \mathbf{B}_k^{1/2} \mathbf{u}(k) = \lambda_q \mathbf{u}(k)$$

$\mathbf{T}_k = \mathbf{X} \mathbf{B}_k^{1/2} \mathbf{B}_k^{1/2} \mathbf{X}'$  a les mêmes valeurs propres non nulles que la matrice  $\mathbf{T}_k^* = \mathbf{B}_k^{1/2} \mathbf{X}' \mathbf{X} \mathbf{B}_k^{1/2}$ . On diagonalisera la matrice  $\mathbf{T}_k^*$  de dimension  $(m, m)$

En pratique, on remplace les *poids bootstrap* nuls par des poids infinitésimaux, de façon à ce que les variables absentes d'une réplique apparaissent quand même avec le statut de variable supplémentaire.

qui sont évidemment distinctes de ce que seraient des répliques des valeurs propres.

<sup>46</sup> Cf. Chateau et Lebart (1996).

---

Cette dernière épreuve de validation est évidemment très sévère. On montre en effet que le tirage sans remise suscite approximativement, en moyenne, l'abandon d'un tiers des éléments (ici, des variables !) à chaque réplique.

## Références bibliographiques sommaires

(Documents cités ou conseillés)

- Alvarez, R., Bécue M., Valencia O. (2004) Etude de la stabilité des valeurs propres de l'AFC d'un tableau lexical au moyen de procédures de rééchantillonnage. In: « *Le poids des mots* », Purnelle, G., Fairon, C., Dister, A., editors, PUL, Louvain, 42-51.
- Anderberg M.R. (1973) *Cluster Analysis for Applications*. Academic Press, New York.
- Anderson T. W. (1963) Asymptotic theory for principal component analysis, *Ann. Math. Statist.*, 34, p 22-148.
- Anderson T. W., Rubin H. (1956) Statistical inference in factor analysis, *Proc. of the 3rd Berkeley Symp. on Math. Statist.*, 5, p 111-150.
- Balbi S. (1994) *L'Analisi Multidimensionale dei dati negli anni'90*. Dipartimento di Matematica e Statistica. (Univ. Federico II), Rocco Curto Editore, Napoli.
- Ball G.H., Hall D.J. (1967) A clustering technique for summarizing multivariate data. *Behavioral Sciences*, 12, p 153-155 .
- Becue M. (1991) *Analisis de Datos Textuales*. CISIA, Saint-Mandé.
- Benzécri J.-P., Jambu M. (1976) Agrégation suivant le saut minimum et arbre de longueur minimum. *Les Cahiers de l'Analyse des Données*, 1, p 441-452.
- Benzécri J-P. (1973) *L'Analyse des Données*, Tome 1: *La Taxinomie*, Tome 2: *L'Analyse des Correspondances*, Dunod, Paris (2de. éd. 1976).
- Blasius J., Greenacre M., (1998) *Visualization of Categorical Data*. Academic Press, San Diego.
- Bouroche J.-M., Saporta G. (1980) *L'analyse des données*. coll."Que sais-je", n°1854, PUF, Paris .
- Bry X. (1995) *Analyses Factorielles Simples*. Economica, Paris.
- Burt C. (1950) The factorial analysis of qualitative data. *British J. of Statist. psychol.* 3, 3, p 166-185.
- Cazes P. (1982) Note sur les éléments supplémentaires en analyse des correspondances. *Les Cahiers de l'Analyse des Données*, 1, p 9-23; 2, p 133-154.
- Celeux G., Nakache J.-P. (eds) (1994) *Analyse discriminante sur variables qualitatives*. Polytechnica, Paris.
- Chateau F., Lebart L. (1996) Assessing sample variability in the visualization techniques related to principal component analysis: bootstrap and alternative simulation methods, in : *COMPSTAT96*, A. Prats (ed), Physica Verlag, Heidelberg, p 205-210.
- Cottrell M., Ibbou S., Letrémy P., Rousset P. (2003) Cartes auto organisées pour l'analyse exploratoire de données et la visualisation, *Journal de la Soc. Française de Stat. vol. 144*, 4, p 67 - 106.
- Diaconis P., Efron B. (1983) Computer intensive methods in statistics, *Scientific American*, 248, p 116-130.
- Diday E. , Lemaire J.L., Pouget J., Testu F. (1982) *Eléments d'Analyse des Données*. Dunod, 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.

- Florek K. (1951) Sur la liaison et la division des points d'un ensemble fini. *Colloq. Math.*, 2, p 282-285.
- Forgy E. W. (1965) Cluster analysis of multivariate data : efficiency versus interpretability of classifications. *Biometric Society Meetings*, Riverside, California (Abstract in : *Biometrics* 21, 3, p 768).
- Garrett J.-C. (1919) General ability, cleverness and purpose, *British J. of Psych.*, 9, p 345-366.
- Gifi A. (1990) *Nonlinear Multivariate Analysis*. Wiley, Chichester.
- Gordon A.D. (1987) A review of hierarchical classification, *J.R.Statist.Soc.*, A, 150, Part2, p 119-137.
- Govaert G. (2003) *Analyse des données*, Hermès – Lavoisier, Paris.
- Gower J. C. (1968) Adding a point to vector diagram in multivariate analysis. *Biometrika*, 55, p 582-585.
- Gower J. C., Ross G. (1969) Minimum spanning trees and single linkage cluster analysis. *Appl. Statistics*, 18, p 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.
- Greenacre M., Lewi P. (2009). Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements, *Journal of Classification*, Springer, vol. 26(1), p 29-54.
- Grelet Y. (1993) Préparation des tableaux pour l'analyse des données : le codage des variables. In : *Traitement statistique des enquêtes*, Grangé D., Lebart L. (eds), Dunod, Paris.
- Guttman L. (1941) The quantification of a class of attributes: a theory and method of a scale construction. In : *The prediction of personal adjustment* (Horst P., ed.) p 251 -264, SSCR New York.
- Habert B., Nazarenko A., Salem A. (1997) *Les linguistiques de Corpus*. Armand Colin, Paris.
- Harman H.H. (1967) *Modern Factor Analysis*, Chicago University Press, Chicago.
- Hartigan J. A. (1975) *Clustering Algorithms*. J. Wiley, New York.
- Hayashi C., Suzuki T., Sasaki M. (1992) *Data Analysis for Social Comparative research: International Perspective*, North-Holland, Amsterdam
- Hochberg, Y. (1988) A sharper Bonferroni procedure for multiple tests of significance, *Biometrika*, 75, p 800-803.
- Holmes S. (1989) Using the bootstrap and the RV coefficient in the multivariate context, in: *Data Analysis, Learning Symbolic and Numeric Knowledge*, E. Diday (ed.), Nova Science, New York, p 119-132.
- Hotelling H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24, p 417-441, p 498-520.
- Hsu, J. C. (1996) *Multiple Comparisons: Theory and Methods*, Chapman & Hall, London.
- Jambu M. , Lebeaux M-O. (1978) *Classification Automatique pour l'Analyse des Données*. Tome 1: *Méthodes et Algorithmes*, Tome 2: *Logiciels*. Dunod, Paris.
- Johnson S. C. (1967) Hierarchical clustering schemes. *Psychometrika*, 32, p 241-254.
- Jolliffe I. (1986) *Principal Component Analysis*. Springer-Verlag, New York.
- Kaufman L., Rousseeuw P. J. (1990) *Finding Groups in Data*. J. Wiley, New York.

- Kazmierczak J.-B. (1985) Analyse logarithmique : deux exemples d'application. *Revue de Statist. Appl.*, 33, (1), p 13-24.
- Kohonen T. (1989) *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.
- Kruskal J. B. (1956) On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc.*, 7, p 48-50.
- Lambert T. (1986) *Réalisation d'un Logiciel d'Analyse de Données*. (Thèse) Université de Paris-Sud, Dép. Statistique, Orsay.
- Lawley D. N., Maxwell A. E. (1963) *Factor Analysis as a Statistical Method*, Methuen, London.
- Le Roux B., Rouanet H. (2004) *Geometric Data Analysis*. Kluwer Ac. Publ., Dordrecht.
- Le Roux B., Rouanet M. (2009) *Multiple Correspondence Analysis*. Vol. 163, Sage Publication Inc.
- Lebart L., Morineau A. (1982) *SPAD Système Portable pour l'Analyse des Données*. CESIA, 82 rue de Sèvres, 75007 Paris.
- Lebart L., Morineau A. Bécue M. (1989) *SPAD.T Système Portable pour l'Analyse des Données Textuelles*, Manuel de Référence. CISIA, Paris.
- Lebart L., Morineau A. Pleuvret P., Brian E., Aluja T. (1983) *SPAD Système Portable pour l'Analyse des Données*, Tome II. CESIA
- Lebart L., Morineau A., Lambert T., Pleuvret P. (1991) *SPAD.N version 2 Système Portable pour l'Analyse des Données*. CISIA, Saint-Mandé.
- 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., Warwick K.W. (1984) *Multivariate Descriptive Statistical Analysis, Correspondence Analysis and Related Techniques for Large Matrices*. Wiley, New York.
- Lebart L., Piron M., Morineau A., (2006) *Statistique Exploratoire Multidimensionnelle, Visualisation et Inférence en Fouille de Données*. Dunod, Paris. (4<sup>ème</sup> édition, refondue). [à consulter pour une bibliographie plus complète]
- Lebart L., Piron M., Steiner J.-F. (2003) *La Sémiométrie*, Dunod, Paris.
- Lebart L., Salem A. (1994) *Statistique Textuelle*. Dunod, Paris.
- Lebart L., Salem A., Berry L. (1998) *Exploring Textual Data*. Kluwer, Boston.
- Lerman I. C. (1981) *Classification et analyse ordinale des données*. Dunod, Paris.
- MacQueen J. B. (1967) Some methods for classification and analysis of multivariate observations. *Proc. Symp. Math. Statist. and Probability (5th)*, Berkeley, 1, p 281-297, Univ. of Calif. Press, Berkeley.
- 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, p 163-172.
- Marchand P. (1998) *L'Analyse de Discours Assisté par Ordinateur*. Armand Colin, Paris.
- McQuitty L.L. (1966) Single and multiple classification by reciprocal pairs and rank order type. *Educational Psychology Measurements*. 26, p 253-265.
- Meulman J. (1982) *Homogeneity Analysis of Incomplete Data*, DSWO Press, Leiden.
- Milan L., Whittaker J. (1995) Application of the parametric bootstrap to models that incorporate a singular value decomposition, *Appl. Statist.* 44, 1, p 31-49.
- Morineau A. (1984) Note sur la caractérisation statistique d'une classe et les valeurs-tests, *Bull. Techn. du Centre de Statist. et d'Infor. Appl.*, 2, p 20-27.
- Morineau A., Lebart L. (1986) Specific clustering algorithms for large data sets and implementation in SPAD Software. In : *Classification as a tool of research*, Gaul W., Schader M., Eds, North Holland, Amsterdam, p 321-330

- Mulaik S. A. (1972) *The Foundation of Factor Analysis*, McGraw Hill, New York.
- Murtagh F. (2005) *Correspondence Analysis and Data Coding with R*. Chapman and Hall, Boca Raton, USA.
- Nakache J. P., Confais J. (2005) *Approche pragmatique de la classification*. Editions Technip, Paris.
- Ohsumi N. (1988) Role of computer graphics in interpretation of clustering results. In : *Recent Developments in Clustering and Data Analysis*, Diday E. et al. (eds), Academic Press, Boston.
- O'Neill, R., and G. B. Wetherill. (1971) The present state of multiple comparison methods (with discussion), *Journal of the Royal Statistical Society, Series B*, 33, p 218-250.
- Perneger T.,V. (1998) What is wrong with Bonferroni adjustments, *British Medical Journal*, 136, p 1236-1238
- Prim R.C. (1957) Shortest connection matrix network and some generalizations. *Bell System Techn. J.*, 36, p 1389-1401.
- Rao C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya serie A*, 26, p 329-357.
- Rouanet H., Le Roux B. (1993) *Analyse des données Multidimensionnelles*. Dunod, Paris.
- Roux M. (1985) *Algorithmes de Classification*. Masson, Paris.
- Salem A. (1987) *Pratique des segments répétés, Essai de statistique textuelle*, Klincksieck, Paris
- Saporta G. (1990 - 2010) *Probabilités, Analyse des Données et Statistique*. Technip, Paris.
- Saville, D. J. (1990) Multiple comparison procedures: The practical solution, *American Statistician*, 44, p 174-180.
- Sokal R. R., Sneath P. H. A. (1963) *Principles of Numerical Taxonomy*, Freeman and co., San-Francisco.
- Spearman C. (1904) General intelligence, objectively determined and measured, *Amer. Journal of Psychology*, 15, p 201-293.
- Tenenhaus M. (2007) *Statistique*. Dunod, Paris.
- Thiria S., Lechevallier Y., Gascuel O., Canu S. (1997) *Statistique et méthodes neuronales*, Dunod, Paris.
- Thorndike R.L. (1953) Who belongs in the family. *Psychometrika*, 18, p 267-276.
- Thurstone L. L. (1947) *Multiple Factor Analysis*. The Univ. of Chicago Press, Chicago.
- Tuffery S. (2006) *Data Mining et Statistique Décisionnelle*. Technip, Paris
- Volle M. (1980) *Analyse des Données*, Economica, Paris.
- Westfall, P. H., Young S. S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, J. Wiley, New York.
- Wong M.A. (1982) A hybrid clustering method for identifying high density clusters. *J. of Amer. Statist. Assoc.*, 77, p 841-847.
- Young G. A. (1994) Bootstrap: more than a stab in the dark, *Statistical Science*, 9, p 382-418.

