Bulletin de méthodologie sociologique. **n. 23, June 1989, pp. 3 - 46**

L'ANALYSE DES DONNEES MULTIDIMENSIONNELLES PAR LE LANGAGE D'INTERROGATION DES DONNEES (LID):

Au delà de l'analyse des correspondances

par:

Jean-Marc Bernard. Brigitte Le Roux & Henry Rouanet

(Groupe Mathématique et Psychologie. CNRS et Université René Descartes, 12 rue Cujas, 75OO5 Paris)

et Marie-Ange Schiltz

(Centre de Mathématiques Sociale, 54 boulevard Raspail, 75006 Paris)

Summary. Multidimensional data analysis with LID data interrogation language. Chapter I describes the three stages which lead to the analysis of a questionnaire by the language LID; correspondence analysis, data structuring, data exploration with LID. Chapter II describes the language LID and the interface program INTERF between EyeLID-1 and ADDAD. Chapter III presents two concrete examples of application: the first is the reanalysis of a British survey; the second is the analysis of a French survey. Chapter IV, in the framework of a ternary frequency table, presents a study of data interaction. Multivariate Data Analysis. Factorial Correspondence Analysis. Data Interrogation. LID.

Résumé. Le Chapitre I décrit les trois étapes qui conduisent i analyser un questionnaire par le langage LID: analyse des correspondances, structuration des données, explorations à l'aide du LID. Le Chapitre II décrit le langage LID et le programme INTERF d'interface entre EyeLID-1 et ADDAD. Le Chapitre III présente deux applications concrètes: la première prolonge la réanalyse d'une enquête britannique, la deuxième prolonge l'analyse d'une enquête française. Le Chapitre IV présente, sur le paradigme du tableau de contingence ternaire, une étude de l'interaction.
Analyse des données multivariées. Analyse factorielle des correspondances. Interrogation des données. LID.

AVANT-PROPOS

Les données dont il sera question dans les chapitres qui suivent sont des données d'observation, typiquement des résultats de questionnaires. L'analyse des questionnaires, en sociologie, est communément effectuée par les méthodes d'analyse multidimensionnelle exploratoire, en particulier en France, l'analyse des correspondances multiples effectuée à partir du tableau "Individus x Questions" mis sous forme disjonctive complète, nous dirons tableau "Individus x Modalités". C'est ce paradigme que, pour fixer les idées, nous prendrons pour guider la discussion qui suit. Pour une comparaison avec d'autres méthodes, en particulier anglo-saxonnes, voir le compte-rendu d'ATP 1988 [2] et le numéro spécial de la *Revue de Statistique Appliquée* [3].

Ce texte s'adresse principalement à ceux qui, pour analyser les questionnaires, utilisent l'analyse des correspondances, dont les bases seront ici supposées connues [4], [7], [24]. Comme on sait, l'analyse des correspondances du tableau "Individus x Modalités" conduit à la détermination des variables principales, appelées souvent, à la suite de Benzécri, "facteurs". Comme on verra, dans le langage d'interrogation de données, un facteur sera une variable explicative catégorisée ; c'est pourquoi pour éviter toute confusion, nous dirons toujours dans ce texte "variable principale", ou "variable factorielle", et non pas "facteur".

L'emploi de "facteur" en analyse des correspondances a été discuté dans la référence [19J.

Les variables principales, prises en nombre suffisant, fournissent un résumé des données et servent de base à l'interprétation. Pour chaque variable principale, on a une valeur (coordonnée) par individu, et également une valeur par modalité: d'où deux nuages: le nuage des individus et le nuage des modalités. En pratique, dans l'analyse des données sociologiques, l'interprétation des axes factoriels se fait surtout à partir du nuage des modalités, en termes d'opposition entre modalités: "Le premier axe oppose les modalités homme et femme ; le deuxième, les plus jeunes aux plus âgés: etc."

Il est notoire que généralement cette interprétation des axes factoriels reste intuitive et dépourvue de précisions quantitatives. En outre, surtout quand le nombre d'individus est élevé, l'interprétation fait rarement appel au nuage des individus.

Le but des méthodes que nous allons présenter est précisément de fonder l'interprétation non seulement sur le nuage des modalités, mais aussi sur celui des Individus, et également de fournir une aide à l'interprétation au moyen de résultats qualitatifs et quantitatifs. Pour parvenir à cet objectif, le **Langage d'interrogation de données** est. selon nous, l'instrument privilégié.

Le langage d'interrogation des données est étroitement lié à l'analyse des comparaisons, méthode développée à partir des années 1970, autour du programme d'analyse de variance VAR3, de Lebeaux, Lépine et Rouanet : voir [13], [20], [21], [22], [11], [14], [9]. L'analyse des comparaisons, en tant que prolongement de l'analyse de variance. s'est développée dans le cadre de l'analyse des données planifiées, c'est-à-dire des données recueillies à l'aide d'un plan (d'expérience ou d'enquête) décrit au moyen de facteurs. Le programme VAR3 d'analyse univariée (ANOVA) met déjà en oeuvre le langage d'interrogation de données: toute demande d'analyse, formulée en termes des facteurs du plan, conduit à calculer la somme des carrés associée.

Après le programme VAR3, l'analyse des comparaisons a été généralisée aux données multidimensionnelles, prolongeant ainsi l'analyse de variance multivariée (MANOVA): voir par exemple le traitement de données multidimensionnelles dans [17]. Il est alors apparu que l'analyse des comparaisons, une fois dépouillée de ses prolongements inférentiels (tests de signification, etc.), pouvait devenir une puissante méthode d'exploration des données d'observation, l'idée de base étant tout simplement qu'une variable initiale (question d'un questionnaire) peut toujours en principe recevoir, dans une problématique descriptive, le statut de facteur (au sens de l'analyse des comparaisons), même en l'absence de planification préalable. Ainsi dans une enquête, à partir de la question "Sexe", on définira le facteur Sexe (à deux modalités homme et femme) ; de même on définira le facteur Age (dont les modalités sont des classes d'âge), etc. Dans ce contexte, la notion de données planifiées fait alors place à celle de données structurée (voir [1], [16] & [17]). Signalons que l'analyse statistique des données structurées est enseignée (cours semestriel) dans le cadre de la maîtrise MST ISASH (2ême année) depuis 1979 (par H. Rouanet et B. Le Roux).

A l'heure actuelle, l'analyse multidimensionnelle des comparaisons peut être effectuée grâce au logiciel interactif EyLID-1 de Baldy et Bernard ([2], [5]). Ce logiciel met en oeuvre je langage d'interrogation de données ("LID"), qui à toute demande d'analyse spécifique associe le tableau des données pertinentes, calcule inerties, covariances etc.. et construit les graphiques pour les paires de variables sélectionnées par l'utilisateur. Les modules graphiques ont été élaborés en vue d'une exploration visuelle commode et approfondie ("Eye").

Le logiciel EveLID-1 a été développé dans le cadre d'une recherche conjointe franco-britannique [2], et il a été appliqué à deux volets de cette recherche conjointe (Dossier Statut au chapitre III et Dossier Suicides au chapitre IV).

L'organisation de cet article sera la suivante:

Au chapitre I, on décrit, sous une forme générale, les trois étapes qui conduisent à analyser un questionnaire par le langage LID: analyse des correspondances, structuration des données, exploration à l'aide du LID.

Au chapitre II, on expose un résumé du langage LID utile pour l'analyse des données d'enquête, puis on décrit brièvement le programme INTERF d'interface entre EyeLID-1 [30] et ADDAD [29].

Au chapitre III, on présente deux applications concrètes des apports du langage LID utilisé à l'issue de l'analyse standard. La première de ces applications, le dossier Statut, prolonge la réanalyse d'une enquête britannique, effectuée par M.A. Schiltz dans le cadre de l'ATP 1988 [27]. La deuxième de ces applications, le dossier Dénonciation, prolonge l'analyse d'une enquête effectuée par Boltanski et al.,[6].

Le chapitre IV présente le dossier Suicides, lui aussi emprunté à l'ATP 1988 (cf. le chapitre de B. Le Roux, H. Rouanet, C. Taylor in [18]). Il illustre, sur le paradigme du tableau de contingence ternaire, comment on peut mener l'étude descriptive d'un concept clé de l'analyse des comparaisons, souvent négligé dans l'analyse des données d'observation, celui d'interaction.

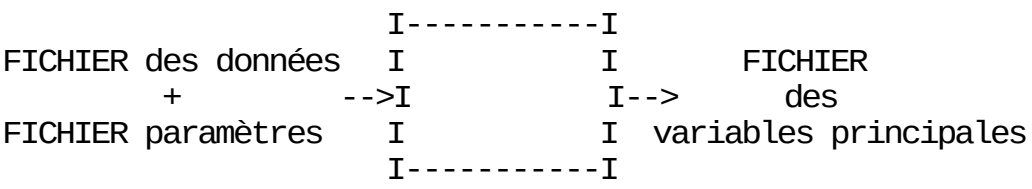
CHAPITRE I: DE L'ANALYSE DES CORRESPONDANCES À L'ANALYSE DES DONNÉES STRUCTURÉES

Nous allons maintenant décrire les trois étapes que nous proposons de mettre en oeuvre pour analyser un questionnaire.

1} une analyse des correspondances;
2) la structuration des données;
3) l'exploration quantitative et qualitative, en particulier graphique, des données préalablement structurées.

PREMIÈRE ÉTAPE: ANALYSE DES CORRESPONDANCES

Ici, nous avons effectué les analyses des correspondances à l'aide du logiciel ADDAD [29]. A partir des données de base, et après recodage éventuel de ces données grâce aux "outils" de l'ADDAD, (pour les dossiers **Statut** et **Dénonciation,** les données sont mises sous forme disjonctive complète), nous disposons, d'une part des variables principales sur les individus (lignes du tableau disjonctif), d'autre part des variables principales sur les modalités (colonnes du tableau disjonctif), ce que résume le schéma ci-après:



Les résultats de l'analyse des correspondances (variables principales) fournissent soit un **résumé descriptif** des données initiales, lorsqu'on retient les premières variables principales (exemples **Statut** et **Dénonciation),** soit un **recodage** sophistiqué mais puissant lorsqu'on retient **toutes** les variables principales (exemple **Suicides).** Rappelons que les variables principales donnent les coordonnées, sur chacun des axes factoriels retenus, des points représentatifs des modalités et aussi des points représentatifs des individus. Dans tous les cas, les variables principales serviront de variables de base pour les analyses ultérieures.

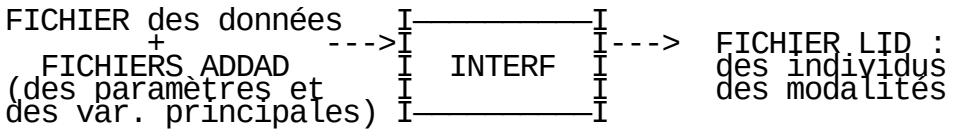
DEUXIÈME ÉTAPE: STRUCTURATION DES DONNÉES

Parmi les variables initiales d'un questionnaire, on peut souhaiter faire jouer à certaines (sexe, âge, etc.) le rôle de variable explicatives, afin de cerner les traits saillants des données.

Ainsi la structuration des données revient fondamentalement à distinguer deux types de variables: les **facteurs** (au sens du langage d'interrogation de données) qui joueront le rôle de variables explicatives (nous emploierons toujours facteur dans ce sens désormais), et les **variables** à proprement parler, qui sont des variables à expliquer. Dans toute la suite, les variables à expliquer seront, bien entendu, les variables principales; les variables explicatives proviendront des variables initiales du questionnaire (questions extraites du fichier des données de base), ces questions pouvant être ou non des "variables actives" de

l'analyse des correspondances. A l'aide du langage d'interrogation de données, on pourra facilement analyser les données préalablement structurées et donner des réponses à des questions spécifiques.

Pour structurer les données, on utilisera le programme INTERF (chapitre II) qui a été conçu pour construire, à partir du fichier des données de base et des fichiers de l'ADDAD (fichiers paramètres et des variables principales), deux fichiers lisibles par le logiciel EyeLID-1. Le programme INTERF est donc un programme d'interface entre ADDAD et EyeLID-1.



Nous décrirons brièvement, dans le cas d'un questionnaire mis sous forme disjonctive complète, les deux fichiers obtenus avec INTERF puis nous présenterons les premiers résultats qu'on peut, alors, facilement obtenir avec LID.

Le fichier des individus contient les variables principales sur I (lignes du tableau disjonctif) et les facteurs choisis par l'utilisateur pour décrire les individus, avec un facteur (créé automatiquement) indexant les Individus et appelé I. Il est usuel de fonder l'interprétation sur les individus dont les contributions relatives sont supérieures à un certain seuil; on pourra alors introduire un facteur technique "contribution" qui sera particulièrement utile pour les représentations graphiques.

Le fichier des modalités contient les variables principales sur les modalités (colonnes du tableau disjonctif) et deux facteurs formels notés Q et R qui serviront à coder l'ensemble des modalités: Q indexant les questions et R les modalités à l'intérieur des questions. Ainsi, à partir des deux variables initiales "Sexe" (à 2 modalités) et "Age" (à 4 modalités), on notera: qlrl la modalité "homme", qlr2 la modalité "femme"; q2rl, q2r2, q2r3 et q2r4 les 4 modalités de l'Âge.

TROISIÈME ÉTAPE: EXPLORATION DES DONNÉES STRUCTURÉES

Nous procéderons maintenant à un survol du langage d'interrogation de données tel qu'il est mis en oeuvre dans le logiciel EyeLID-1 en vue de le raccorder aux résultats familiers d'une analyse des correspondances.

Rappelons que EyeLID-1 est un logiciel interactif d'analyse des données multidimensionnelles structurées: "Eye" pour exploration visuelle des données et "LID" pour Langage d'Interrogation de Données. Le langage renvoie à la structuration des données, laquelle peut soit provenir d'un plan d'expérience soit, comme ici, être introduite sur des données de simple observation en vue de leur exploration.

Une question posée sur les données se traduit par une demande d'analyse qui comprend:

- un mot-clé indiquant la procédure statistique à mettre en jeu (exemples VAR pour variance, TAB pour tableau des données, etc.);
- une formule du langage d'interrogation de données, qui indique les données spécifiques sur lesquelles va être appliquée la procédure statistique (exemple: dans Statut la formule I<gl> désigne les individus du groupe gl);
- la liste des variables numériques (variables à expliquer) su lesquelles portera la procédure statistique.

Toute **demande d'analyse** se traduira par des résultats qualitatifs ou quantitatifs.

Du point de vue **qualitatif,** les résultats seront sous forme de graphiques; un accent particulier ayant été mis, dans EyeLID-1. sur les possibilités d'exploration graphique: sélection d'options graphiques, d'attributs graphiques, choix de couleurs, liaison de certains points entre eux, etc., ce qui permet de faire ressortir les traits pertinents des données.

Du point de vue **quantitatif,** on obtiendra des résultats numériques tels que inertie d'un nuage, d'un sous-nuage, etc. Par exemple, à partir d'un facteur "Groupe" (G), on pourra décomposer l'inertie selon les axes, l'intérêt d'une telle décomposition étant de choisir les plans des axes qui discriminent au mieux les groupes. Si à la comparaison G, on associe la comparaison intra-G, notée I(G), on obtient la **double décomposition des inerties** (voir dossiers **Suicides** et Statut, et [17]).

En résumé, le logiciel EyeLID-1 permet d'utiliser de façon conjointe et complémentaire **quantification** et **visualisation.**

PREMIERS RÉSULTATS OBTENUS AVEC EYELID-1

On peut, non seulement, retrouver les principaux résultats de l'analyse des correspondances mais aussi aller immédiatement plus loin; c'est ce que nous commenterons maintenant pour chacun des deux fichiers "Individus" et "Modalités" produit par le programme INTERF.

Dans ce qui suit, VI, V2, V3, etc. désignent les variables principales successives, et V l'ensemble des variables principales retenues dans l'analyse.

Fichier des individus

A partir du fichier des individus, on obtient:

- les valeurs propres par la demande d'analyse: VAR I -> V (VAR pour variance)
- le tableau des coordonnées factorielles des individus sur les axes par la demande d'analyse: TAB I -> V (TAB pour tableau). La demande TAB I-> VI,V2 fournit les coordonnées factorielles des points représentatifs des individus.
- le graphique des individus dans le plan des deux premiers axes factoriels par: GRA I -> V1.V2 (GRA pour graphique). La demande GRA I-> V2.V3 produirait le graphique des individus dans le plan des axes 2 et 3 etc.
- la contribution **relative** d'un individu (par exemple i1) à **l'inertie** du premier axe en calculant le rapport des deux nombres obtenus par les demandes:

```
RSS il->V1  et  RSS I->V1 (RSS pour "Raw Sum of Squares").
```

II arrive souvent, en analyse des correspondances, qu'on mette en éléments supplémentaires des lignes qui correspondent à des regroupements d'individus. Ces regroupements peuvent être effectués aussi bien sur un facteur de classification établi lors du recueil des données (exemples: catégorie socioprofessionnelle, sexe, etc.), qu'à partir des résultats d'une classification hiérarchique. Si nous prenons en compte ce facteur "Groupe" (noté G), dans la structuration des données, les coordonnées factorielles des points moyens de chacun des groupes sont obtenues par la demande:

TAB G -> V (voir exemples Statut et Dénonciation) et leur
représentation graphique dans le plan 1-2 par:

GRA G -> V1,V2.

Le logiciel EyeLID-1 permet aussi d'étudier des contributions relatives non directement accessibles à partir des résultats factoriels, par exemple celle d'un sous-groupe

d'individus. Pour les individus du premier groupe g1 (voir Statut), la contribution relative du sous-groupe I<g1> à l'inertie du premier axe factoriel s'obtient comme le rapport du nombre RSS I<g1> −> V1 par RSS I−> V1.

Du point de vue graphique, après avoir choisi le plan des axes qui discriminent au mieux les groupes, on peut facilement révéler les proximités ou les oppositions des groupes à l'aide des attributs graphiques (voir **Statut** ou **Suicides**).

Fichier des modalités

Comme précédemment, on retrouve les valeurs propres et les coordonnées factorielles des modalités, et on établit le graphique factoriel, dans le plan des axes 1-2, à partir des demandes d'analyse suivantes:

VAR Q&R−>V TAB Q&R−>V GRA Q&R−> V1,V2.

La contribution relative de la première modalité (première question) à l'inertie du premier axe sera le rapport des deux nombres résultant des deux demandes:

RSS q1&r1 −> V1 et RSS Q&R −> V1

De plus, on obtiendra facilement la **contribution** relative (à l'inertie du premier axe) **de l'ensemble des modalités** de réponse à la première question (q1). C'est le rapport des deux nombres obtenus par:

RSS q1&R−>V1 et RSS Q&R−> V1

A partir de ces contributions, on pourra dire par exemple: "la contribution du facteur 'Sexe' au 1er axe est de XX%".

On pourra aussi calculer la contribution d'une modalité ou d'une question aux deux premiers axes en faisant le rapport de la somme des deux nombres RSS q1&R −> V1 et RSS q1&R −>V2 par la somme de RSS Q&R −>V1 et de RSS Q&R −> V2. A partir de ces contributions, on pourra dire par exemple: "la contribution du facteur 'Sexe' aux deux premiers axes est de XX%". Ce résultat est important puisqu'il permet de repérer les questions ayant une faible contribution, questions que l'on pourra éventuellement éliminer de l'analyse ou du graphique (voir Statut).

Remarque

Du fait que les facteurs Q et R sont des facteurs formels, certaines demandes d'analyse conduiront à des résultats sans intérêt ou même triviaux.

La demande TAB Q −> V, donnera, pour l'analyse d'un questionnaire mis sous forme disjonctive complète, un tableau remplis de zéros, puisque, pour chaque question, les modalités de réponse ont le même point moyen.

La demande TAB R −> V est dénuée de sens, alors que la demande TAB q3&R −> V fournit les coordonnées factorielles des modalités de la seule question q3.

CHAPITRE II: LE LANGAGE D'INTERROGATION DES DONNÉES ET LE LOGICIEL EYELID-1

1. PRÉSENTATION GÉNÉRALE

Le langage LID implémenté dans le logiciel EyeLID-1 permet l'interrogation d'une base de données statistique, que nous appellerons protocole de base.

Une question spécifique que l'on se pose sur les données est traduite par une demande d'analyse de LID. Cette demande d'analyse déclenche, en premier lieu, la recherche du protocole dérivé du protocole de base, c'est-à-dire un résumé statistique pertinent pour la question spécifique qu'exprimé la demande d'analyse. Un mot-clé de la demande d'analyse détermine, quant à lui, la ou les procédures à appliquer sur ce protocole dérivé.

Le langage LID est présenté Ici de façon succincte. Pour une présentation plus détaillée, on se reportera au manuel d'utilisation d'EyeLID-1 [30].

2. LE LANGAGE D'INTERROGATION DES DONNÉES: GÉNÉRALITÉS

2.1. Protocole de base: unités, facteurs, variables

La base de données statistiques que le langage LID permet d'interroger est constituée d'un ensemble d'unités. Chaque unité est indexée par les modalités d'un certain nombre de facteurs. Pour chaque unité, plusieurs variables ont été mesurées.

Ces notions seront illustrées sur un exemple schématique inspiré de l'enquête "INOP" ([23], voir aussi les deux interventions: Rouanet H. (1985), "Some aspects of Bayesian multivariate analysis". Communication à la Multivariate section of the Royal Statistical society; Bernard J.-M., Baldy R. (1986), "EyeLID-1: A new program for graphical inspection of multivariate data", Seminar of the Biometrics Unit, Institute of Psychiatry, University of London) dans lequel 20 élèves sont répartis selon la méthode d'enseignement (moderne ou traditionnelle) et selon leur milieu social (défavorisé ou favorisé), ce qui définit 4 groupes (5 élèves par groupe). Pour chaque sujet 3 variables numériques sont mesurées, ceci au cours de 2 sessions (milieu et fin d'année scolaire). On pourra se représenter ces variables, dénommées ci-après V1, V2 et V3 soit comme des résultats à des tests avec notes numériques, soit comme les coordonnées de 3 variables factorielles extraites d'une analyse en composantes principales (ACP).

La description de ce protocole consiste tout d'abord dans l'identification de ses facteurs et de leurs modalités:

- facteur "sujets" S à 20 modalités: (s1 à s20);
- facteur "méthode d'enseignement" A à 2 modalités: moderne (a1), traditionnel (a2);
- facteur "milieu social" B à 2 modalités: favorisé (b1), défavorisé (b2);
- facteur "temps" T à 2 modalités: début (t1), fin (t2) d'année.

Chaque combinaison observée des modalités des facteurs du protocole définit une unité statistique. On a ici 40 unités correspondant à 20 sujets croisés avec 2 sessions. A chaque unité est associée:

- sa description selon les modalités des facteurs; par exemple, "s12 a1 b2 t1" pour la session 1 du sujet 12 (du groupe "moderne" et "favorisé").
- les valeurs pour chacune des 3 variables.

- son poids, ici égal à 1 pour chaque unité car il s'agit d'un protocole élémentaire.

Pour un protocole non-élémentaire, obtenu par exemple par moyennage, ce poids correspondra à un effectif ou nombre d'individus ayant servi à calculer chaque moyenne.

Remarquez que si les données sont structurées (définition de facteurs), cette structure est extrêmement souple puisque ces facteurs peuvent être en relations tout à fait quelconques. Cette souplesse permet aussi bien des structures faibles comportant des lacunes (comme dans les données d'observation) que des structures plus fortes (comme dans l'expérimentation).

2.2. Les demandes d'analyse de LID

Une demande d'analyse de LID permet de définir un protocole dérivé, les opérations pour y parvenir et les procédures à lui appliquer. Une demande LID est composée des trois éléments suivants:

- La formule caractérise le protocole dérivé: d'une part sa structure (unités, facteurs), d'autre part le type de dérivation (moyennage, dérivations résiduelles, etc.) Il existe deux types de formules: ensemblistes ou linéaires.

- La partie **sélection de variables** spécifie quelles sont les variables du protocole de base retenues pour le protocole dérivé.

- Le mot-clé spécifie la procédure à appliquer au protocole dérivé (représentation graphique, tableau, calculs de statistiques, etc.).

2.3. Protocoles dérivés: Structure du protocole dérivé et Calcul des valeurs et des poids associées à chaque unité dérivée

Comme tout protocole, un protocole dérivé est à nouveau constitué d'un ensemble d'unités. Ces unités sont décrites par les modalités d'un sous-ensemble des facteurs du protocole de base. Les valeurs associées à ces unités dérivées correspondent également à un sous-ensemble des variables du protocole de base.

La structure d'un protocole dérivé est ainsi obtenue par deux opérations élémentaires: restriction de l'ensemble des unités de base et regroupement des unités de base restantes pour chaque unité dérivée.

Par exemple, la formule LID "al&B −> V2" va provoquer la construction d'un protocole dérivé constitué de 2 unités correspondant aux diverses combinaisons possibles de la modalité al du facteur A et des modalités du facteur B: a1b1 et a1b2. Ce protocole dérivé est univarié puisque seule la variable V2 est considérée ("−>V2").

A l'unité "a1b1" de ce protocole dérivé sont associées toutes les unités de base indexées par a1 et b1: 10 unités de base à savoir, s1a1b1t1, s1a1b1t2, s2a1b1t1, ... , s5a1b1t1, s5a1b1t2.

Pour les formules ensembliste du langage LID la valeur associée à une unité dérivée pour une variable donnée est la moyenne pondérée des valeurs des unités de base correspondantes. Le poids associé à une unité dérivée sera dans ce cas la somme des poids associés aux unités de base correspondantes.

Pour les formules linéaires d'autres dérivations que le moyennage interviennent: dérivations résiduelles, écarts entre moyennes, etc.

3. LE LANGAGE D'INTERROGATION DES DONNÉES: DESCRIPTION

3.1. Les symboles du langage LID

Une demande de LID est constituée de 5 types de symboles:

Opérandes:

fact = facteur désigné par une lettre majuscule par exemple "A".
mod = modalité élémentaire d'un facteur du protocole de base, composée d'une lettre minuscule et d'un numéro de modalité, par exemple "a1".
var = variable désignée par la lettre V et le numéro de la variable, par exemple "V2".

Opérateurs ensemblistes:

^ = concaténation ("et") de modalités; par définition ce symbole ne s'écrit pas dans une formule effective.
_ = regroupement ("ou") de modalités.
, = séparateur de parties d'une famille de parties.
& = composition de 2 familles de parties.
* = croisement de 2 familles de parties.
<> = emboîtement d'un facteur dans une famille de parties.

Opérateurs linéaires:

() = dérivation intra
. = dérivation d'interaction
: = dérivation par contraste

Opérateur de sélection:

−> = sélection de variables

Mots-clés (détaillés au 4. de ce chapitre)

3.2. Formules ensemblistes

Une formule ensembliste comprend comme seuls opérateurs les opérateurs ensemblistes indiqués précédemment. Ce langage ensembliste permet la construction de parties et de familles de parties de l'ensemble des unités de base. Chaque partie définit une unité dérivée où les valeurs sont calculées par moyennage:

- **s1** désigne une modalité d'un facteur; cette formule fournit la moyenne du sujet s1.

- **s1t1** l'opération de concaténation permet de construire des modalités composées, obtenues par la composition de facteurs différents: cette formule fournit la moyenne du sujet s1 au temps t1.

- **s1_s2** l'opération "_" permet le regroupement de modalités élémentaires ou composées comportant les mêmes facteurs; cette formule fournit la moyenne des sujets s1 et s2 regroupés.

- **s1,s2_s3** l'opération "," permet la séparation de parties et définit une famille de parties, ici de 2 parties; cette formule fournit la moyenne de s1, d'une part, et celle de s2 et s3 regroupés d'autre part.

Le reste des éléments ensemblistes du langage LID permet de définir de façon plus concise des familles de parties:

A désigne la famille de parties constituée par l'énumération des modalités du facteur A; cette formule est équivalente à "al,a2" et fournit donc les moyennes pour

chacune des modalités du facteur A.

- sl, s2&T l'opération "&" permet la composition de familles de parties; cette opération engendre une nouvelle famille de parties où chacune est obtenue en combinant une partie de la première famille et une partie de la seconde; ainsi la formule précédente se réécrit: "s1t1,s2t1,s3t1,s1t2,s2t2,s3t2"; seules les parties non vides sont conservées lors de cette réécriture.

- s1,s2*t1,t2 l'opérateur de croisement "*" est identique à celui de composition "&", mais il induit de plus la vérification que les deux familles de parties sont croisées, c'est-à-dire qu'aucune partie vide n'est générée.

- S<a1b1,a2b2> l'opération d'emboîtement "<>" d'un facteur (l'emboîté écrit à l'extérieur des <>) dans une famille de parties (l'emboîtant écrit à l'intérieur des <>) est identique à celle de composition "&", mais induit de plus la vérification de la relation d'emboîtement du facteur dans la famille de parties, autrement dit que chaque modalité du facteur n'engendre de partie non vide que pour au plus une seule partie de la famille de parties emboîtante.

3.3. Formules linéaires

Les trois opérateurs linéaires "()". "." et ":" permettent d'introduire d'autres modes de dérivation que le moyennage pondéré. Toute formule linéaire peut être décomposée en deux composantes: une formule ensembliste associée qui définit les unités du protocole dérivé comme précédemment, et un mode de dérivation particulier pour le calcul des valeurs pour chaque unité.

Pour les formules impliquant seulement les opérateurs "()" et ".", le calcul des pondérations associées aux unités dérivées, se fait uniquement en référence à la formule ensembliste associée:

- **S(A)** pour la dérivation intra "()", la structure du protocole dérivé est la même que celle de "S&A". Le calcul des valeurs de "S(A)" se fait par différence entre celles des protocoles dérivés S&A et A; ainsi à l'unité dérivée "s1 et a1" est associée l'écart entre moyenne de "s1a1" et moyenne de "a1".

- **A.T** pour la dérivation d'interaction ".", la structure du protocole dérivé est la même que celle de "A*T". Le calcul des valeurs de "A.T" se fait à partir de celles de "A*T" en construisant le protocole de support "A*T", doublement centré (voir [21] et chapitre IV).

- S:t1,t2 pour la dérivation par contraste ":" , la structure du protocole dérivé est la même que celle correspondant à la formule "S". Le calcul des valeurs se fait par différence entre les valeurs des deux protocoles dérivés "S*t1" et "S*t2": pour le sujet si on calcule l'écart entre les valeurs associées à "s1t1" et "s1t2".

La pondération associée à l'unité dérivée "s1" est la moyenne harmonique des pondérations associées aux 2 unités "s1t1" et "s1t1".

3.4. Sélection de variables

Toute formule est suivie de la sélection des variables à considérer pour le protocole dérivé:

- **−>V2,V3** l'opérateur "−>" sépare une formule quelconque (à gauche) et une liste de variables séparées par des "," (à droite). Pour "S−>V2,V3", on définit le protocole dérivé des moyennes des sujets ("S") pour les deux variables V2 et V3.

4. MOTS-CLÉS D'EYELID-1

Le "mot-clé" apparaissant à gauche d'une formule permet de sélectionner un mode de représentation du protocole dérivé ou certaines opérations à effectuer sur ce protocole dérivé. Nous n'en détaillons ici que les plus importants pour l'analyse des données d'observation. On trouvera l'ensemble des mots-clés dans [5].

4.1. Représentation du protocole dérivé

TABLE - Le protocole dérivé est représenté sous forme de tableau des valeurs des variables sélectionnées indexées par les étiquettes des unités dérivées.

WEIGHTS - Même chose que TABLE avec seule Impression des poids des unités dérivées.

4.2. Représentation graphique

GRAPH - Le protocole dérivé est représenté sous forme graphique avec un grand nombre d'options permettant une modification Interactive du graphique, en particulier sélection d'attributs graphiques en fonction des étiquettes des unités dérivées.

HISTOGRAM - Histogramme des valeurs du protocole dérivé.

4.3. Calculs statistiques divers sur le protocole dérivé

DESC - Minimum, maximum, moyenne, médiane, quartiles, écart-type du protocole dérivé.
VAR - Variance du protocole dérivé.
SS - Somme des carrés centrés du protocole dérivé.
RSS - Somme des carrés non centrés du protocole dérivé.
COV - Matrice des variances-covariances.
SP - Matrice des sommes des carrés et produits centrés.
RSP - Matrice des sommes des carrés et produits non centrés.
CORREL - Matrice des corrélations.

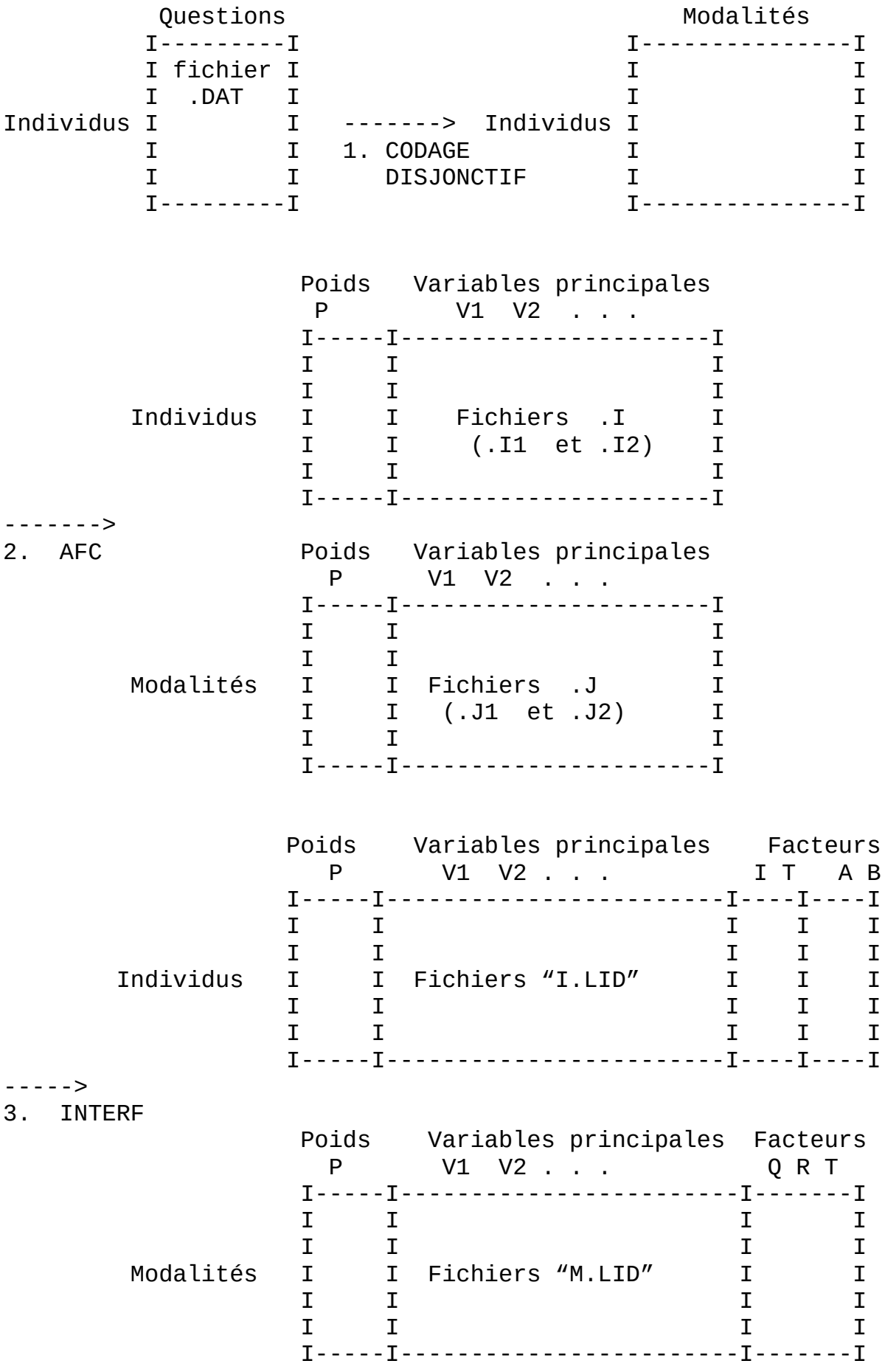
4.4. Réutilisation du protocole dérivé

FILE - Le protocole dérivé est recopié dans un fichier pour une possible réutilisation ultérieure par le programme en tant que nouveau protocole de base.

5. UTILISATION D'EYELID-1 POUR LES DONNÉES D'ENQUÊTE

Dans cet article, le logiciel EyeLID-1 est utilisé comme un outil d'analyse post-factoriel. Il convient donc de préciser ici comment l'analyse factorielle des correspondances et EyeLID-1 peuvent être utilisés successivement et/ou conjointement sur des données d'enquête.

En ce qui concerne des données d'enquête qui se présentent sous la forme d'un tableau "Individus x Questions", nous détaillons Ici de façon plus technique les étapes, déjà Indiquées au chapitre I, qui conduisent des données brutes à la création de 2 fichiers de données pour EyeLID-1 (fichiers .LID).


Questions
fichier .DAT
Individus
Individus
1. CODAGE DISJONCTIF
Modalités
Poids P
Variables principales V1 V2 . . .
Individus
Fichiers .I (.I1 et .I2)
2. AFC
Modalités
Fichiers .J (.J1 et .J2)
Facteurs I T A B
Fichiers “I.LID”
3. INTERF
Facteurs Q R T
Fichiers “M.LID”

1. A partir du tableau des réponses de chaque individu (fichier .DAT), on procède au codage disjonctif complet des données, d'où le tableau "Individus x Modalités".

2. On procède ensuite à l'analyse factorielle des correspondances de ce tableau, avec un logiciel tel que ADDAD. Cette étape conduit entre autres à la création des deux séries de fichiers de résultats factoriels (poids, puis coordonnées selon les axes factoriels:

- fichiers .I pour les lignes (ici les individus): .I1 pour les éléments actifs de l'analyse, .I2 pour les éléments supplémentaires.
- fichiers .J pour les colonnes (ici les modalités): .J1 pour les éléments actifs de l'analyse. .J2 pour les éléments supplémentaires.

Ces deux premières étapes sont les étapes usuelles d'une A.F.C.

3. L'étape suivante fait appel au programme INTERF qui réalise l'interface entre AFC et EyeLID-1. Comme résultats de cette étape deux fichiers analysables par EyeLID-1 sont créés: le fichier des Individus et le fichier des Modalités.

Le fichier des Individus (fichier I.LID):

- chaque unité statistique correspond à un individu.
- le poids de chaque unité est celui fourni par l'AFC.
- les variables (au sens de LID) sont les variables principales fournies par l'AFC dans les fichiers .I1 et .I2.
- les facteurs (au sens de LID) sont des descripteurs des individus et correspondent généralement à des questions initiales extraites du fichier .DAT (le logiciel INTERF permet de récupérer certaines questions initiales pour en faire des facteurs). Le facteur I qui énumère simplement les individus est généré automatiquement par INTERF. Le facteur T. pour "type", également généré automatiquement, distingue les individus actifs (modalité t1) provenant du fichier .I1, des individus supplémentaires (modalité t2) provenant du fichier .I2.

Le fichier des Modalités (fichier M.LID):

- chaque unité statistique correspond à une modalité ;
- le poids de chaque unité est celui fourni par l'AFC;
- les variables (au sens de LID) sont les variables principales fournies par l'AFC dans les fichiers .J1 et .J2;
- plusieurs facteurs sont construits automatiquement: "Q", la question dont chaque modalité est extraite, "R", le numéro de réponse dans la question pour chaque modalité et "T", le facteur "type" comme dans le fichier des individus (t1 pour le fichier .J1, t2 pour le fichier .J2). Les facteurs Q et R sont des facteurs formels qui jouent un rôle d'intermédiaire de calcul.

On a déjà donné des exemples de demandes LID sur chacun de ces fichiers qui permettent de retrouver un certain nombre de résultats de l'AFC (voir chapitre I). La suite de cet article va illustrer la richesse de ce langage pour procurer des résultats, graphiques, non fournis par l'AFC.
