1392

enssib
Ecole Nationale Supérieure
des Sciences de l'Information
et des Bibliothèques

Université
Claude Bernard
Lyon I

DESS Informatique Documentaire
Rapport de recherche bibliographique

Analyse des données : Méthodes multivariées (factorielles) sur tableaux longitudinaux

GUIRAO Henri

Sous la direction de

M. NORMAND

CREUSET – Université Jean MONNET

BIBLIOTHEQUE DE L'ENSSIB
8028793

Année 1997-1998

enssib
Ecole Nationale Supérieure
des Sciences de l'Information
et des Bibliothèques

Université
Claude Bernard
Lyon I

DESS Informatique Documentaire
Rapport de recherche bibliographique

Analyse des données : Méthodes multivariées (factorielles) sur tableaux longitudinaux

GUIRAO Henri

ENSSIB Bibliothèque Villeurbanne

Sous la direction de

M. NORMAND

CREUSET – Université Jean MONNET

Année 1997-1998

1998
11
12

Analyse des données : méthodes multivariées (factorielles) sur tableaux longitudinaux

Henri GUIRAO

Résumé :
Ce rapport de recherche bibliographique traite des méthodes d'analyse de données appliquées à des données longitudinales (techniques d'analyse de tableaux ternaires). Après avoir présenté notre méthodologie de recherche, nous offrons une synthèse des principales références de notre bibliographie.

Descripteurs :
Analyse des données, Méthodes multivariées, tableaux longitudinaux.

Abstract :
The following bibliographical research deals with longitudinal data analysis methods (methods of analysing 3-way data tables). After introducing the method used for my research, I give a summary of the main bibliographical references of my search.

Keywords :
Data analysis, Multivariate statistical methods, longitudinal data.

SOMMAIRE

INTRODUCTION

L'objet de ce rapport de recherche bibliographique, suite aux demandes de mon commanditaire, Mme Myriam NORMAND, Maître de conférences en statistiques à l'université Jean Monnet, est la réactualisation d'une bibliographie relative au champ de l'analyse des données. L'intitulé exact de notre recherche est : **Méthodes multivariées (factorielles) sur tableaux longitudinaux**.

Répondant aux souhaits de mon commanditaire, les références bibliographiques retenues sont de deux formes : formalisations (de méthodes), et/ou applications des méthodes factorielles sur tableaux longitudinaux dans le domaine des sciences sociales (économie, gestion et sociologie).

Trois difficultés majeures sont apparues pour la réalisation de cette bibliographie.

La première tient au champ particulier dans lequel nous travaillons : l'économie. En effet, s'il existe de nombreuses références de formalisations et en même temps d'application des méthodes factorielles sur tableaux longitudinaux, celles-ci sont en grandes parties dans les domaines de la médecine, de la biologie, de la biochimie ou de la psychologie.

La seconde difficulté résulte de l'inexistence de bases de données spécifique au champ des statistiques.

La troisième difficulté est liée aux différentes terminologies existantes. Souvent, il n'y a pas de correspondances entre les appellations anglo-saxonnes et françaises. Une même méthode statistique est nommée différemment.

Essayant de dépasser ces difficultés, le rapport de recherche suivant s'organise en trois parties. La première résume la méthodologie adoptée pour la recherche des références bibliographiques, ainsi que l'estimation de son coût. La seconde partie fournit au lecteur une synthèse sur les méthodes factorielles sur tableaux longitudinaux. Enfin, la troisième partie, suivant les normes de rédaction en cours, est la liste de nos références bibliographiques.

PREMIERE PARTIE : METHODOLOGIE

La première étape, pour effectuer une recherche bibliographique pertinente, consiste à saisir correctement la requête du demandeur et donc à étudier précisément le sujet donné. Sachant que notre recherche va nous conduire à interroger des bases de données en ligne, des cédéroms, l'Internet, nous devons traduire l'intitulé fourni par le commanditaire en un vocabulaire de recherche (on passe d'un langage naturel à un langage indexé ou documentaire). A partir de ce vocabulaire, suivant le support de recherche (bases de données, cédéroms, Internet, catalogue de bibliothèques), nous devrons établir différentes équations de recherches (chaque support possède une syntaxe d'interrogation qui lui est propre).

Notre recherche nous a conduit à utiliser plus particulièrement 4 types de supports de recherche. Les bases de données en ligne sous DIALOG (1), l'Internet (2), le cédérom FRANCIS produit par l'INIST (3) et enfin les documents papiers, périodiques et monographies (4). Pour chacun de ces quatre types de support, nous commencerons comme précisé ci-dessus, par citer le vocabulaire de recherche que nous avons retenu ainsi que les équations de recherche établies ou la stratégie de recherche. Pour chacune de ces interrogations nous donnerons une estimation du coût (5)

Nous tenons à préciser ici que nous n'avons utilisé aucune limitation par les champs *dates* dans nos différentes équations. Ceci se justifie par le peu de références pertinentes obtenues à partir de nos équations de départ. Le tri sur ce champs (date), dans la mesure ou il avait lieu d'être puisque notre travail consiste en une réactualisation de bibliographie, a été réalisé manuellement à partir des références trouvées.

1. Interrogation en Ligne sous DIALOG.

Le service DIALOG fournit ces utilisateurs en informations depuis 1972. Avec plus de 450 bases de données, couvrant de larges champs disciplinaires, le serveur DIALOG est une importante source d'information. Un choix aussi vaste de bases de données ne permet pas d'offrir, il est bien évident, une syntaxe d'interrogation harmonisée des bases. Pour chacune d'entre elles, il existe donc des pages d'aides (appelées *Bluesheet*).

Ces pages d'aides (ou fiches techniques) existent sur support papier, mais elles sont également disponibles sur Internet (URL : http://www.dialog.com). Ces pages fournissent à la fois tous les renseignements nécessaire relatifs à la syntaxe d'interrogation de la base, ses origines, ses producteurs, ses sources, ses tarifs de consultation, les types de données qu'elle propose.

Nos deux sessions d'interrogation du serveur DIALOG ont nécessité trois étapes. La première étape sélectionne le vocabulaire d'interrogation, la seconde étape sélectionne les bases de données a priori intéressantes et la troisième étape constitue les équations de recherches ainsi que leurs applications.

1.1 Sélection du vocabulaire de recherche

La plus grande partie des bases de données dans le monde est anglo-saxonne. Pour effectuer des recherches en ligne, il faut donc commencer par sélectionné un vocabulaire anglo-saxon traduisant l'intitulé du sujet de recherche et s'assurer ensuite que ce vocabulaire est référencé dans la base de donnée (un dictionnaire-index est présent dans chaque base). Notre sujet étant très précis, la recherche d'un vocabulaire correspondant a simplement consisté à traduire en anglais des termes relatifs aux méthodes statistiques et au champ de l'analyse des données. Nous avons donc formé nos équations de recherche à partir des termes suivants :

Data analysis, Statistical analysis, multivariate analysis, correspondence analysis, multivariate methods, longitudinal data, longitudinal analysis.

1.2 Sélection des bases de données

Une fois le vocabulaire trouvé, il s'agit de repérer les bases de données qu'il semble pertinent d'interroger. A partir du catalogue des bases de données sous DIALOG (DataBase Catalogue – Spring 1996, publié par Knight-Ridder Information), nous avons sélectionné les 33 bases de données suivantes :

Interrogation sous DIALINDEX
9: Business & Industry(R) Jul_1994-1997/Nov 27
12: IAC Industry Express (TM)_1995-1997/Dec 01
13: BAMP_1997/Nov W4
15: ABI/INFORM(R)_1971-1997/Nov W4
16: IAC PROMT(R)_1972-1997/Dec 01
20: World Reporter_1997-1997/Nov 26
30: AsiaPacific_1985-1997/Oct B1
75: IAC Management Contents(R)_86-1997/Nov W3
111: Natl.Newspaper Index(SM)_1979-1997/Dec 01-
139: Econ. Lit. Index_1969-1997/Oct
148: IAC Trade & Industry Database_1976-1997/Nov 26
211: IAC Newsearch(TM)_1997-1997/Dec 01
466: Info Latino America_1988-1995/Dec W1
481: Delphes Eur Bus_1980-1997/NOV W1
484: Periodical Abstracts Plustext_1986-1997/Nov W2
485: Accounting and Tax Database_1971-1997/Nov W4
565: Econbase:Time Series & Forecasts_1997/Sep
583: IAC Globalbase(TM)_1986-1997/Nov W4
624: McGraw-Hill Publications_1985-1997/Nov 25
620: EIU Viewswire_1997/Nov W4
627: EIU: Country Analysis_1997/Nov W4
628: Ctry Risk & Forecasts_1997/Nov W4
629: EIU:BUS. NEWSLETTERS_1997/Nov W4
636: IAC Newsletter DB(TM)_1987-1997/Dec 01
637: Journal of Commerce_1986-1997/Nov 26
799: Textline Curr.Glob.News_1995-1997/Oct 12

Interrogation « simple »
239:MathSci(R) 1940-1997/Dec (c) 1997 American Mathematical Society

Interrogation sous - DIALOG OneSearch
137:Book Review Index 1969-1997/Q3 (c) 1997 Gale Research Inc.
470:Books In Print(r) 1997/Nov (c) 1997 R.R.Bowker,Reed Elsevier Inc.
430:British Books in Print 1997/Nov (c) 1997 J. Whitaker & Sons Ltd.
426:LCMARC-Books 1968-1997/Oct W4 (c) format only 1997 Knight Ridder Info.
102:ASI 1973-1997/Oct (c) 1997 Congressional Information Service
122:Harvard Business Review 1971-1997/Dec (c) 1997 Harvard Business Review

Trois critères ont joué dans la sélection des bases :

- contenu économique (dans la mesure où nous recherchons des références de méthodes longitudinales de traitement des données appliquées à l'économie) ;
- formalisme (recherche de formalisations statistiques et mathématiques) ;
- publication (recherche des bases de données couvrant les parutions scientifiques d'articles et/ou de monographies).

Le premier critère nous a mené à retenir la métabase SF BUSECON qui contient 26 bases de données. Ce groupe de 26 bases de données a été interrogé à l'aide du DIALINDEX (code d'appel : B 411). Ce dernier consiste en un regroupement de plusieurs bases de données présélectionnées couvrant un même champ scientifique (l'économie). L'équation de recherche va porter sur ces 26 bases à la fois.

Le second critère, de formalisme mathématiques nous a conduit à retenir la base MathSci(R) (code d'appel correspondant : B 239). Cette dernière a été questionnée en recherche simple. C'est à dire que l'équation de recherche que nous avons établi n'a porté que sur cette seule base.

Avec le troisième critère de publication, nous avons retenus 7 bases relatives aux publications de monographies et/ou périodiques. Ce troisième groupe à été questionné avec le système DIALOG OneSearch. Ce système correspond au DIALINDEX avec une différence : les bases de données interrogées simultanément ont été sélectionnées par l'utilisateur lui-même.

1.3 Formation et application des équations de recherche

Le vocabulaire et les bases de données étant choisis, il ne nous reste alors qu'à former les équations de recherche à appliquer (voir encadré ci-dessous sur la syntaxe des équations) et d'observer les résultats obtenus. Nous ne rendons compte ici que des équations qui nous ont permis d'obtenir des résultats satisfaisants.

Comme indiqué à l'étape précédente, nos premières interrogations ont porté sur la méta-base SF BUSECON (interrogation du DIALINDEX : B 411 demande SF BUSECON).

L'équation la plus efficace est présentée ci-dessous.

SS **(statisti?(w)analys?)** AND **(multivariate?(w)analys?** OR **correspondence(w)analysi?)** AND **longitudi?(w)data**

Syntaxe d'interrogation

SS : pour lancer une recherche,
AND, OR, NOT : opérateurs booléens,
(nW) : opérateur d'adjacence avec contrainte d'ordre,
(nN) : opérateur d'adjacence,
? : troncature.
( ) : notion d'ordre.

Le premier résultat exploitable est : parmi les 26 bases de données, 24 ont une ou plusieurs références qui correspondent à notre équation. Après une première sélection nous ne retenons que les 6 bases offrant le plus de réponses.

15: ABI/INFORM(R)_1971-1997/Nov W4
75: IAC Management Contents(R)_86-1997/Nov W3
148: IAC Trade & Industry Database_1976-1997/Nov 26
211: IAC Newsearch(TM)_1997-1997/Dec 01
484: Periodical Abstracts Plustext_1986-1997/Nov W2
485: Accounting and Tax Database_1971-1997/Nov W4

Parmi ces 6 bases, seule la base Periodical Abstracts Plustext (B 484) est vraiment intéressante. Le DIALINDEX nous a donc permis de faire un tri important. Après cette phase de tri nous avons donc réappliqué une équation de recherche, mais cette fois sur la seule base Periodical Abstracts Plustext (B 484).

SS (statisti?(w)analys? OR **data(w)analys?)** AND **(multivariat?(w)analys?** OR **correspondence(w)analys?)** AND **longitudi?(w)data**

Cette équation nous a permis d'obtenir **75 réponses**. Pour des raisons évidentes de coûts nous nous sommes contentés de visualiser les 25 premières (voir syntaxe de demande ci-dessous). Pour chacune de celles-ci nous avons demandé a voir Titre, Auteur(s), Descripteurs et Résumé, afin de pouvoir évaluer leur pertinence par rapport au sujet. Une seule réponse était vraiment en rapport avec notre sujet. Nous n'avons donc pas visualisé les 50 réponses restantes.

Syntaxe de visualisation
?ts31/ti,au,de,ab/1-25

La seconde interrogation sous DIALOG a été celle de la base MathSci(R). Produite par l' American Mathematical Society, elle couvre une période allant actuellement de 1940 à 1997. L'encadré suivant, relatif à la base MathSci(R), ne fait que retranscrire les informations contenues dans les *Bluesheets* (fiches techniques). Les différents renseignements de l'encadré suivant existent pour chacune des bases sous DIALOG.

MathSci(R) 1940-1997/Dec (c) 1997
produite par l'American Mathematical Society

sources :	600 journals reviewed cover-to-cover, 2500 journals covered selectively, monographs, conference proceedings, theses, technical reports.
dates covered :	mathematics 1959 to the present, statistics 1910 to the present, computer science 1954 to the present,
file size :	1.900.000 records as of April 1997,
update frequency :	monthly (9.000 records per update),
database content :	bibliographic records,
document types indexed :	books and monographs, conferences, symposia, meetings, journal articles, reports,
geographic coverage :	international.

L'équation de recherche utilisée est la même que précédemment.

SS (statisti?(w)analys? OR data(w)analys?) AND (multivariat?(w)analys? OR correspondence(w)analys?) AND longitudi?(w)data

Le résultat obtenu : **12 réponses**, toutes relativement pertinentes. Pour chacune nous avons demandé à voir Titre, Auteur(s), Descripteurs et Résumé.

La troisième et dernière interrogation sous DIALOG à été celle des base Book Review (B137), Books In Print(B 470), British Books in Print (B 430), LCMARC-Books (B 426), ASI (B 102) et Harvard Business Review (B 122). Comme indiqué plus haut, nous avons utilisé le système DIALOG OneSearch pour interroger ces 6 bases.

La stratégie de recherche est toujours identique :

SS (statisti?(w)analys? OR data(w)analys?) AND (multivariat?(w)analys? OR correspondence(w)analys? OR multivariate(w)method ?) and longitudi?(w)data

Sur cette dernière interrogation en mode DIALOG OneSearch, nous n'avons obtenu aucune référence vraiment pertinente.

2. Interrogation en ligne de l'INTERNET

Deux moteurs de recherche ont été testés : ALTA VISTA et INFOSEEK. Nous présentons nos sessions de travail, avec dans l'ordre cité Altavista en premier et Infoseek en second, en indiquant leur syntaxe de recherche, nos équations, l'estimation des résultats obtenus.

2.1 Moteur de recherche : Alta Vista

<URL : http://www.altavista.digital.com>

Nous avons utilisé le moteur de recherche ALTA VISTA en mode de recherche avancée (*Advanced Mode*). Ce mode permet l'utilisation d'opérateurs Booléens (AND, OR, NOT), d'opérateurs d'adjacence (NEAR), de troncatures (*), et l'utilisation du langage naturel entre guillemets (¨ ¨). Ce mode permet également d'affiner ses recherches (fonction *refine*) en excluant ou en privilégiant certains champs suite à une première série de réponses.

Le vocabulaire retenu pour l'interrogation de l'Internet est le même que celui retenu précédemment, à savoir :

Data analysis, Statistical analysis, multivariate analysis, correspondence analysis, multivariate methods, longitudinal data, longitudinal analysis.

A partir de ce vocabulaire, nous avons établi plusieurs équations de recherche. Nous ne reproduisons ici, que les équations qui ont permis d'obtenir les références les plus pertinentes.

Equation 1 :

('STATISTICAL ANALYSIS' OR 'DATA ANALYSIS') NEAR **LONGITUDIN***

Equation 2 :

'MULTIVARIATE ANALYSIS NEAR **LONGITUDIN***

Equation 3 :

'MULTIVARIATE METHODS' NEAR **LONGITUDIN***

Equation 4 :

'LONGITUDINAL RESEARCH' OR **'LONGITUDINAL METHOD'**

Selon le jour, ou même l'heure à laquelle on applique une même équation, on obtient, dans la plupart des cas, un nombre de références différent.

Nous devons également préciser que le moteur de recherche ALTA VISTA en mode de recherche avancé, permet de choisir la langue d'interrogation. Pensant trouver des sites Web ou références bibliographiques en langue française, nous avons essayer de modifier le critère langage de recherche en remplaçant la valeur apparaissant par défaut 'ANY LANGUAGE' par la valeur 'FRENCH'. L'équation utilisée, ci dessous, ou des variantes de celle-ci n'a permis de trouver aucune référence réellement pertinente.

('ANALYSE DES DONNEES' OR 'METHODES STATISTIQUES' OR 'METHODES MULTIVARIES' OR 'ANALYSE FACTORIELLE' OR 'METHODES FACTORIELLES') NEAR **LONGITUDIN***

2.2 Moteur de recherche : Infoseek

<URL : http://www.infoseek.com>

La syntaxe d'interrogation sur Infoseek est relativement différente de celle d'Altavista. Les opérateurs booléens AND, OR, NOT ne sont pas employés. La structure utilisé se sert des opérateurs arithmétiques (+ , -), tient compte des majuscules pour la recherche sur noms propres et utilise les guillemets pour indiquer l'adjacence de deux termes. Aussi, avant toute recherche il est conseillé de prendre connaissance de l'aide en ligne, que l'on peut trouver à la même adresse.

Stratégie 1 :
Recherche sur **LONGITUDINAL** : réponses 27022
Parmi ces 27022 réponses
Recherche sur « **DATA ANALYSIS** » : réponses 694
Parmi ces 694 réponses
Recherche sur « **MULTIVARIATE METHODS** » : réponses 19

Stratégie 2 :
Recherche sur « **LONGITUDINAL DATA ANALYSIS** » : 171 réponses
Parmi ces 171 réponses
Recherche sur « **CORRESPONDENCE ANALYSIS** » : 11 réponses

2.3 Pertinence des résultats

Le problème de la recherche sur l'Internet, selon le moteur de recherche utilisé tient au bruit qui entoure les réponses. A côté de réponses qualifiées de pertinentes, nombreuses sont celles qui sont soit totalement hors du champs recherché, soit marginales. La recherche induit donc une perte de temps importante à trier les résultats obtenus. Ceci a été notre cas en utilisant AltaVista : en moyenne, sur 100 sites visités, 8 sont en relation avec notre sujet. L'utilisation d'Infoseek ne nous a pas conduit à ce genre de problèmes :

nous avons en effet obtenu des réponses très ciblées (sur une trentaine de sites, la moitié sont pertinents).

Parmi les nombreux sites visités, qui nous ont permis de trouver plusieurs références bibliographiques intéressantes nous tenons à signaler deux sites particulièrement en relation avec notre sujet.

Le premier est un cours en ligne sur les méthodes d'analyse de données longitudinales. Son adresse URL est : http://corelan.bih.harvard.edu/Ida/home.html.

Le second, intitulé *A guide to the web for statisticians* est un site consacré au monde des statistiques. Son adresse URL est : http://www.maths.uq.oz.au/~gks/webguide.

3. Interrogation du cédérom FRANCIS.

Nous avons interrogés plusieurs cédéroms : Docthèses (315 000 références de thèses de doctorat soutenues en France), Bibliographie Nationale Française (1 1,2 millions de références entrés par dépôt légal) et FRANCIS.

Seul le dernier nous a permis de trouver plusieurs références (une référence avec Docthèses). Notre étude est donc ciblée sur ce dernier. le cédérom FRANCIS produit par l'INIST qui couvre le domaine des lettres et sciences. Ce cédérom contient 9000 titres de périodiques, rapports scientifiques, thèses universitaires, comptes-rendus de congrès et monographies. Environ 1,3 millions de références pour une période de 1984 à nos jours.

En mode expert, la syntaxe d'interrogation sur descripteurs à été la suivante :

LONGITUDINAL et **DATA** et **ANALYSIS**

Cette séquence nous a fourni 126 réponses. Parmi ces 126 réponses, seules 3 étaient pertinentes aux vues du sujet.

En mode assisté, la stratégie de questionnement sur descripteurs a été la suivante :

LONGITUDINAL

Suite à cette phase une liste de sélections portant sur le terme longitudinal nous a été proposée. Parmi cette liste nous avons retenu **LONGITUDINAL STUDIES** (10 réponses) et **LONGITUDINAL STUDY** (8 réponses).

Nous sommes retombés sur les 3 réponses pertinentes obtenues en mode expert. Ces 3 réponses sont des applications des méthodes factorielles sur données longitudinales dans le champ de la sociologie.

4. Recherche sur périodiques spécialisés.

Nous donnons ci-dessous une liste non exhaustive de périodiques concernant le domaine des statistiques et de l'analyse des données. Cette liste a pu être établie en consultant les références bibliographiques citées dans des ouvrages fondamentaux de statistiques et d'analyse de données.

- Journal of the American Statistical Association,
- The American Statistician,
- Biometrika,
- Journal of Educational and Behavioural Statistics,
- Journal of the Royal Statistical Society,
- Les Cahiers de l'analyse des données,
- Econometrika,
- Review of Economics and Statistics,
- Revue de Statistique Appliquée,
- Psychometrika,
- Journal of Official Statistics.

On peut consulter entre autres sur Internet le sommaire du *Journal of the American Statistical Association*, et de *The American Statistician* à l'adresse suivante : <URL : http://www.amstat.org/publications/index.html>.

Pour consulter les sommaires de 150 autres périodiques spécialisés dans le domaine des statistiques, on peut également consulter le site Internet mentionné précédemment : *A guide to the web for statisticians* < URL : http://www.maths.uq.oz.au/~gks/webguide>.

L'autre solution pour obtenir les sommaires consiste, pour la localisation en France, à consulter le céderom Myriade ou le 36–17 CCN sur le minitel qui correspondent au Catalogue Collectif National des Publications en Série (CCNPS). Une fois les documents localisés on peut commander les sommaires par les services de Prêt Entre Bibliothèques.

5. Estimation du coût de la recherche.

Les coûts les plus faciles à chiffrer sont ceux de l'interrogation des bases de données sous DIALOG. Suite à chacune de vos interrogations de base de données, le coût de votre session est estimé (voir encadré en exemple ci-dessous). Notre interrogation s'est effectuée en deux sessions.

```
27nov97 03:44:43 User701931 Session D150.1
     Sub account: HENRI
          $3.24    0.216 Hrs File411
    $3.24  Estimated cost File411
    $1.30  TYMNET
    $4.54  Estimated cost this search
    $4.54  Estimated total session cost   0.216 Hrs.
```

```
27nov97 03:58:21 User701931 Session D150.3
     Sub account: HENRI
          $3.50    0.233 Hrs File484
          $0.00  25 Type(s) in Format  5 (UDF)
          $0.00  25 Types
     $3.50  Estimated cost File484
     $1.40  TYMNET
     $4.90  Estimated cost this search
     $9.72  Estimated total session cost   0.463 Hrs.
```

Temps de connexion DIALOG : 1,253 Hrs (0,79 Hrs + 0.463 Hrs)
Coût d'interrogation DIALOG : 26,26 $
Coût d'interrogation DIALOG (en Fr) : 161,83 Fr (cours du $ du 06.03.98).

Le faible coût de notre interrogation des bases de données sous DIALOG résulte du fait que le déchargement des résultats obtenus s'est fait avec un format gratuit (0 $ par référence consultée).

On peut également estimer le coût de commande d'ouvrages ou d'articles dans la mesure où celui-ci est facturé par les services de Prêt-entre-Bibliothèques.

Coût de commande d'une monographie : 20 fr
Coût de commande d'un article : 29 fr les 10 premières pages, 14.50 fr les 10 suivantes
Nombre d'articles commandés : 1
Nombre de monographies commandés : 4
Coût des articles commandés : 43,50 fr
Coût des monographies commandés : 80 fr
Coût global Prêt Entre Bibliothèque : 123,50 fr

Pour ce qui est du coût de la recherche sur Internet, il est délicat d'avancer un chiffre fiable. On peut l'estimer toutefois en se fiant au temps de connexion. Nous précisons ici que l'importance du temps passé sur l'Internet s'explique en partie par le 'bruit' (c'est-à-dire les références obtenues qui sortent du champ recherché) entourant les résultats de l'application des différentes équations de recherche.

Temps de recherche sur Internet : 430 mn
Nombre de sessions Internet : 14
Coût de connexion 3 première minutes : 0,74 fr
Coût de connexion des minutes suivantes : 0,28 fr
Coût estimé des sessions Internet : 129, 72 fr

Le tableau ci-après résume l'ensemble des coûts associés à notre recherche.

	DIALOG	Internet	Autres *	TOTAL
Estimation du temps de recherche (mn)	73.18	430	135	**638.18**
Estimation du coût (fr.)	161.83	129.72	123.5	**415.05**

La colonne *autres** correspond aux cédéroms et aux supports papier. Le coût lié à cette colonne est celui des commandes d'articles et de monographies par les services de prêt-entre-bibliothèque.

Le coût global de notre recherche est donc de **415 francs** pour **11 heures** de recherches (temps de rédaction non compris).

SECONDE PARTIE : SYNTHESE

La synthèse qui suit a pour objet de présenter ce que sont, en analyse des données, les méthodes multivariées sur tableaux longitudinaux. Cette synthèse a été rédigée à partir de quelques ouvrages majeurs, référencés dans notre troisième partie. Nous tenons à préciser que, malgré le dévouement de Mme Nicole FURNON, responsable du service de Prêt-entre-Bibliothèques du Service Commun de la Documentation de l'université Jean Monnet, il ne nous a pas été possible d'obtenir l'ensemble des documents primaires désirés.

Analyse des données : méthodes multivariées (factorielles) sur tableaux longitudinaux.

Pour décrire de façon synthétique des tableaux de données, les économistes disposent depuis plusieurs années de techniques statistiques dont les plus connues sont l'analyse en composantes principales et l'analyse factorielle des correspondances. Ces techniques sont adaptées à des ensembles de données structurés par deux indices : l'analyse en composantes principales s'utilise dans le cas de tableaux croisant des individus et des variables, et l'analyse factorielle des correspondances dans le cas de tableaux croisant les modalités de deux variables qualitatives.

Dans le cas où l'on dispose d'un ensemble de données dépendant cette fois de trois indices, en particulier de tableaux étudiés à plusieurs périodes de temps, il est nécessaire de recourir à d'autres méthodes. Les méthodes factorielles sur tableaux longitudinaux correspondent à ce cas de figure.

Nous commencerons par présenter ce que recouvre la notion de données longitudinales. Nous verrons ensuite les méthodes factorielles sur tableaux longitudinaux.

1. La notion de données longitudinales (longitudinal data)

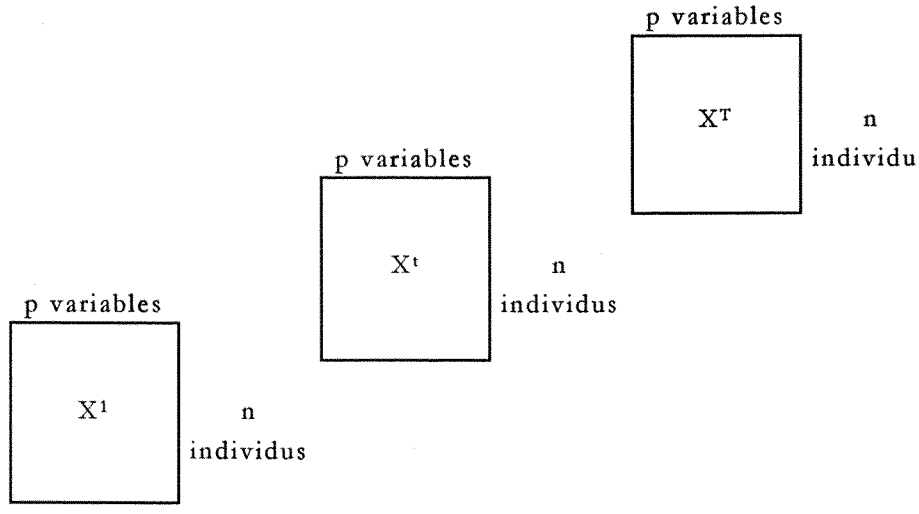
La recherche longitudinale concerne la collecte et l'analyse de données, à différents moments dans le temps (MENARD, S.1991, HEIJDEN VAN DER, PGM. 1987). La recherche longitudinale s'applique à des données pour lesquelles une relation longitudinale, c'est-à-dire une relation entre des même phénomènes étudiés à des périodes distinctes, doit être estimée (HEIJDEN VAN DER, PGM. 1987). Les mêmes variables sont mesurées dans la même unité pour deux périodes au moins. Soit la suite de tableaux X^t, t = 1,..,T donnant les valeurs prises par p variables pour n individus, à T périodes différentes (voir représentation ci-dessous).

La recherche longitudinale doit être définie en termes de types de données et en termes de méthodes spécifiques de traitement de ces données. Ses caractéristiques impliquent que les données soient collectées, pour chaque variable, pour deux ou plusieurs périodes ; les sujets ou individus analysés soient les même ou au moins comparables d'une période à une autre ; les

analyses impliquent des comparaisons de données inter et intra périodes (MENARD, S.,1991).

Le travail sur données longitudinales a deux objectifs principaux : permettre une description des changements sur un échantillon de données, et établir s'il y a indépendance (ou dépendance) entre des variables (MENARD, S.,1991). La prise en compte des mêmes variables appliquées aux mêmes individus à différentes périodes dans le temps permet d'atteindre ces deux objectifs avec rigueur et qualité (PETERSEN, T. 1993, HUBER, GP. VAN DE VEN, AH.,1995).

Représentation des tableaux indexés par le temps



Dans le domaine de l'économie, la collecte de données, sous forme de différents ratios permettant de juger de la santé d'une économie nationale ou d'une entreprise, conduit à l'élaboration de tableaux longitudinaux dès que l'on estime ces différents ratios sur plusieurs périodes (années, trimestre, mois).

En décrivant ce que recouvre la notion de données longitudinales, nous nous situons à un niveau de collecte des données. Nous allons maintenant passer à un niveau d'analyse des données en nous intéressant aux méthodes factorielles de traitement de ce type de données.

2. Méthodes factorielles sur tableaux longitudinaux

On parle indistinctement de tableaux longitudinaux, évolutifs ou ternaires. Les trois recouvrent le même concept de prise en compte de trois indices (individus, variables et temps). Il existe plusieurs méthodes statistiques permettant d'analyser et de résumer ces tableaux spécifiques, le

plus souvent à l'aide de représentations graphiques sur des axes factorielles, cercles des corrélations. Nous donnons ci-dessous un bref descriptif des principales méthodes.

Les méthodes qui ont fait l'objet de recherches théoriques les plus importantes sont l'analyse canonique généralisée et la méthode STATIS (CASIN, P. 1995 b ; FAHRMEIR, L. and TUTZ, G. 1994). Ces deux méthodes s'appuient sur l'analyse de la proximité entre les tableaux X_T, mais ne permettent pas toujours une interprétation aisée des résultats.

L'analyse en composantes principales généralisée (ACPG) détermine des variables synthétiques résumant les proximités entre des tableaux et décrivant au mieux les tableaux de départ. Cette méthode combine l'analyse en composantes principales (ACP) et l'analyse canonique généralisée (ACG).

L'analyse discriminante de tableaux (ADT) est une variante de l'analyse canonique généralisée (ACG). Elle peut être adaptée au cas de données longitudinales. L'analyse discriminante de tableaux évolutifs (ADTE) est obtenue à partir du critère de généralisation de l'analyse canonique en imposant des contraintes d'orthogonalisation dans chaque espace (CASIN, P. 1995 b).

L'ADTE est relativement proche de la méthode LONGI qui, elle aussi, permet l'étude de données longitudinales. La différence entre les deux méthodes tient à la manière dont sont construites les variables synthétiques du type $Z^j{}_t$ (une variable synthétique résume l'ensemble des tableaux X_T).

LONGI (analyse des données LONGItudinales) calcule des indices multivariés, décrivant les individus indépendamment du temps ou les évolutions temporelles indépendamment des individus. Cette caractéristique en fait une technique bien adaptée aux données économiques évolutives.

Dans le cadre de tableaux ternaires, la méthode STATIS permet d'extraire l'information contenue sous forme de graphiques : résumé global par un nuage de point-tableaux, position compromis des individus dans un système d'axes interprétables à l'aide des variables, évolution de chaque individu autour de sa position compromis dans ce même système d'axes.

Conclusion

Si l'on s'intéresse à l'évolution (dans le temps) d'individus ou d'objets à travers certaines variables (croisement de trois indices), alors on se positionne dans le champ de l'analyse de données sur tableaux longitudinaux. Tout comme en analyse des données standard, il n'existe pas une méthode particulière mais plusieurs méthodes. Le choix de l'application d'une méthode plutôt qu'une autre doit alors être pensé en fonction de la nature des variables (qualitative ou quantitative), du temps pris en compte (discret ou continu) et en fonction des objectifs de l'analyse.

TROISIEME PARTIE : BIBLIOGRAPHIE

AGRESTI, A. (1996), *An introduction to categorical data analysis*, New York : John Wiley and Sons.

Association of GCRC statisticians : On-line course on longitudinal data analysis, [on line], [20.02.98],Available from internet : <URL : http ://corelan.bih.harvard.edu/Ida/home.html>.

BERHANE, K. TIBSHIRANI, R. (1994), *Longitudinal data analysis using varying-coefficient models* [on line],. Available from internet : <URL : http ://lib.stat.cmu.edu/joint94>.

BESSE, P. (1987), Optimal metric in principal components analysis of longitudinal data, In *Data analysis and informatics*, Versailles.

CASIN, P. (1995 a), L'Evolution économique de six pays de 1973 à 1992 décrite par la méthode LONGI, *Economie Appliquée*, vol XLVIII (3), p 151-174.

CASIN, P. (1995 b), L'Analyse discriminante de tableaux évolutifs, *Revue de Statistique Appliquée*, vol XLIII (3), p 73-91.

CASIN, P. (1996), L'Analyse en composantes principales généralisée, *Revue de Statistique Appliquée*, vol XLIV (3), p 63-81.

CIBOIS, P. DEGENNE, A. (1990), L'Analyse des données numériques, *Année sociologique (L')*, vol 40, p 345-349.

CLARK, W.A.V. (1992), Comparing cross-sectional and longitudinal analyses of residential mobility and migration, *Environment and planning*, vol 2(9), p 1291-1302.

COLEMAN, J.S. (1981), *Longitudinal data analysis*, New York : Basic Books.

COREY, P. ESCOBAR, M. (Coord .) Longitudinal data analysis working groups, [20.02.98], Available from internet : <URL : http ://www.utstat.toronto.edu/biostat/groups/long.html>.

DIGGLE, P.J. LIANG, K.Y. ZEGER, S.L. (1994), *Analysis of longitudinal data*, London : Chapman and Hall.

EYE, A. VON (ed.) (1990), *Statistical methods in longitudinal research*, Boston : Academic press, 570 p.

FAHRMEIR, L. and TUTZ, G. (1994), *Multivariate statistical modelling based on general lienar models*, New York : Springer-Verlag, p 204-218.

GREGOIRE, T.G. BRILLINGER, D.R. DIGGLE, P.J. (1997), *Modelling longitudinal and spatially correlated data*, New York : Springer-Verlag, 416 p.

GROSSMAN, W. (1986), Statistical analysis of longitudinal data, In *7th international summer school on problems of model choice and parameter estimation in regression analysis*, Holzau : Nov. 26- Dec. 2, p 79-87.

HAGENAARS, J.A. (1990), *Categorical longitudinal data : Log-linear panel, trend, and cohort analysis*, London : SAGE Publications.

HECKMAN, J.J. SINGER, B. (eds.) (1992), *Longitudinal analysis of labor market data*, Cambridge : Cambridge university press, 410 p.

HEIJDEN VAN DER, P.G.M. (1987), *Correspondence analysis of longitudinal data*, Leiden : DSWO Press.

HELLER, H. BROWN, A. (1995), Group feedback analysis applied to longitudinal monitoring of the decision making process, *Human relations*, vol 48 (7), p 815-835.

HUBER, G.P. VAN DE VEN, A.H. (1995), *Longitudinal field research methods : studying processes of organizational change*, London : SAGE Publications, 392 p.

JONES, R.H. (1993), *Longitudinal data with serial correlation : a state-space approach*, London : Chapman and Hall.

KASPARIAN, R. (1993), L'Analyse longitudinale de la population active : une typologie de profils de carrière des générations françaises de 1911 à 1935, *Population*, vol 48(3), p 620-653.

KAUFMAN, R.L. (1993), Decomposing longitudinal from cross-unit effects in panel and pooled cross-sectional designs, *Sociological methods & research*, vol 21, p 482-504.

KROONENBERG, P.M. (1985), Three-mode principal component analysis of multivariate longitudinal organizational data, *Sociological Methods and Research*, vol 14, p 99-136.

LIANG, K.Y. ZEGER, S.L. (1986), Longitudinal data analysis using generalised linear models, *Biometrika*, vol 73, p 13-22.

LELIEVRE, E. COURGEAU, D. (1991), Approches longitudinales, In *Interrogations et parcours sociologiques*, STENDLER, F. WATIERS, P. (éds.), Paris : Méridiens Klincksieck, p 105-115.

MENARD, S. (1991), *Longitudinal research*, London : SAGE Publications, 88 p.

MERHAN, F. (1989), Analysis of discrete longitudinal data : infinite-lag Markov models, In *Statistical data analysis and inference*, Neuchatel : August 21-24, p 533-541.

Modelling longitudinal and spatially correlated data : methods, applications, and future directions. Nantucket, [Mass.] 15-18 October 1996 Available from internet : <URL : http ://www.tgg.fw.vt.edu/Nantucket>.

PAPADOPOULOS, S. AMENLYA, Y. *On factor analysis of longitudinal data*, [12.04.97] Available from internet : <URL : http ://www.lib.stat.cmu.edu/joint95/Abstracts/025P02>.

PERNIN, M.O. (1986), *Contribution à la méthodologie d'analyse de données longitudinales*, Thèse Doct. : Univ. Claude Bernard-Lyon 1.

PETERSEN, T. (1993), Recent advances in longitudinal methodology, *Annual review of sociology*, vol 1, p 425-454.

PLEWIS, L. (1981), A Comparison of approaches to the analysis of longitudinal categoric data, *British Journal of Mathematical and Statistical Psychology*, vol 34, p 118-123.

PLEWIS, L. (1985), *Analysing change, measurement and explanation using longitudinal data*, New York : Wiley.

PLEWIS, L. (1995), Statistical analysis of longitudinal data, *In Statistical analysis of longitudinal data*, Leiden, 29 mei-2 juni 1995.

ROGOSA, D.R. And SANER, H.M. (1995), Longitudinal data analysis examples with random coefficient models, *Journal of Educational and Behavioural Statistics*, vol 20, p 149-170.

ROGOSA, D.R. And SANER, H.M.(1995), Longitudinal data analysis examples with random coefficient models : reply to discutants, *Journal of Educational and Behavioural Statistics*, vol 20, p234-238.

ROGOSA, D.R. (1980), Comparisons of some procedures for analyzing longitudinal panel data, *Journal of Economics and Business*, vol 32, p 136-151.

ROGOSA, D.R. (1995), Myths and methods : « Myths about longitudinal research » In *The Analysis of change*, GOTTMAN, JM. (ed.), Hillsdale : Lawrence Erlbaum Associates, p 3-65.

ROVINE, M.J. and EYE, A. VON (eds) (1991), *Applied computational statistics in longitudinal research*, Boston : Academic press, 237 p.

RUNGIE, C. (1993), Planning the data analysis of longitudinal studies of Aging, , *Beijing multidimensional longitudinal study of Aging*, Beijing : UNFPA Project p23.

RUNGIE, C. (1994), Longitudinal data analysis and outcomes, *Beijing multidimensional longitudinal study of Aging*, Beijing : UNFPA Project p23.

RUNGIE, C. (1994), Increased interest in longitudinal studies, Research news, *Market research society of Australia*, vol 11, n° 8, p 6.

SAPORTA, G. (dir.) (1996), *L'Analyse des données évolutives : méthodes et applications*, Paris : éd. Technip, 227 p.

VISSER, R.A. (1985), *The Analysis of longitudinal data in behavioural and social research*, Leiden : DSWO Press.

XIANG ZHANG, C. Multivariate longitudinal data analysis with a family of covariance matrices, [12.04.97] Available from internet : <URL : http ://www.lib.stat.cmu.edu/joint95>.

YOUNG, C.H. SAVOLA, K.L. PHELPS, E. (1991), *Inventory of longitudinal studies in the social sciences*, London : SAGE Publications, 576 p.