Contribution à la réalisation d'une chaine d'analyse bibliométrique exploratoire : Tests de modules de classification

Loïc Vinet

Unité Mixte de Recherche
Economie – Droit Rural et Agroalimentaire
Rue de la Géraudière
BP 71627
44316 NANTES CEDEX 03

Institut de Mathématiques Appliquées
44, rue Rabelais
BP 808
49008 ANGERS CEDEX 01

CONTRIBUTION A LA REALISATION D'UNE CHAINE D'ANALYSE BIBLIOMETRIQUE EXPLORATOIRE : TESTS DE MODULES DE CLASSIFICATION

Rapport de stage

Loïc VINET
sous la direction de
Elise et Michel ZITT

IMA IV
Année 2000-2001

Avertissement

Tous les mots suivis d'un astérisque (*) renvoient au glossaire.

Remerciements

Je remercie tout le personnel du Laboratoire de Recherches et d'Etudes Economiques (LERECO) pour leur accueil chaleureux, et plus particulièrement :

- Elise BASSECOULARD-ZITT pour ses nombreux conseils,

- Michel ZITT pour ses nombreux conseils, et notamment les petites astuces du logiciel SAS,

- Valérie JACQUERIE et Maurice QUINQU pour l'approfondissement de mes connaissances des logiciels SPAD et SAS,

- Catherine VASSY, secrétaire du LERECO, pour ses petits dépannages informatiques.

Sommaire

Introduction

Après avoir présenté succinctement l'INRA (Institut National de la Recherche Agronomique), et plus particulièrement le service (ou laboratoire) au sein duquel s'est déroulé mon stage, je présenterai les différentes études réalisées. La première, qui poursuivait une étude préalablement commencée, explique comment réduire l'espace du champ d'étude en imposant certaines contraintes aux éléments traités ; la seconde étude consistait, sur un échantillon donné, à appliquer un algorithme de partitionnement rapide et de comparer les résultats obtenus avec les classifications selon les quatre critères les plus souvent utilisés : saut minimum, diamètre, distance moyenne, critère de Ward (*).

Contexte du stage

I. Présentation de l'entreprise

1.1. L'INRA

1.1.1. Historique

En 1921, l'Etat crée l'Institut des Recherches Agronomiques (IRA) chargé de développer les recherches scientifiques appliquées à l'agriculture. Il y a alors création de plusieurs centres de recherche un peu partout en France. En 1934, l'IRA disparaît : les centres passent sous autorité directe du Ministère de l'Agriculture. Au sortir de la Seconde Guerre mondiale, un rapport puis un projet de loi proposent de créer l'Institut National de la Recherche Agronomique afin de concilier la recherche fondamentale désintéressée et le souci d'application pratique des résultats tenant compte des milieux en s'appuyant largement sur l'expérimentation. L'INRA est créé le 18 mai 1946.

1.1.2. Statut et missions

A sa création, l'INRA est un établissement public doté de la personnalité civile et de l'autonomie financière dirigé par un scientifique sous l'autorité du Ministère de l'Agriculture. Il a pour mission l'organisation, l'exécution et la publication de tous les travaux de recherches portant sur l'amélioration et le développement de la production végétale et animale et sur la conservation et la transformation des produits agricoles.

En 1985, l'INRA devient un Etablissement Public à caractère Scientifique et Technique (EPST), au même titre que le CNRS (Centre National de la Recherche Scientifique) ou l'INSERM (Institut National de la Santé Et de la Recherche Médicale). L'INRA est placé sous la tutelle commune du Ministère de l'Agriculture, de l'Alimentation et de la Pêche et du Ministère de l'Education, de l'Enseignement Supérieur et de la Recherche. Il doit contribuer à l'élaboration de la politique nationale de recherche et doit organiser et exécuter toute recherche scientifique intéressant l'agriculture et les industries qui y sont liées, publier et diffuser les résultats de ses travaux, participer à la valorisation de ses recherches et de son savoir-faire.

1.1.3. Partenariats et coopérations nationaux et internationaux

Au niveau national, l'INRA a de nombreux partenariats de recherche avec :

- le CNRS
- l'INSERM
- l'INRIA (Institut National de Recherche Informatique et en Automatique)
- l'ONF (Office National des Forêts)
- l'IFREMER (Institut Français de Recherche et d'Exploitation de la MER)
- le CEMAGREF (Recherche pour l'Ingénierie de l'Agriculture et de l'Environnement)
- le CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement)
- l'IRD (Institut de Recherche pour le Développement)

L'INRA a également plusieurs coopérations internationales bilatérales ou multilatérales avec les pays d'Europe (Programmes d'Actions Intégrées), le Maghreb (Actions Intégrées), le Liban (CEDRE : Coopération pour l'Evaluation et le Développement de la REcherche), l'Afrique du Sud, la Chine, le Venezuela, le Brésil, l'Europe de l'Est, l'Amérique du Nord, etc. : en tout 90 pays étrangers sont en relation avec l'INRA.

1.1.4. Quelques chiffres

L'INRA dispose d'un budget annuel de 3,4 milliards de FF, provenant principalement de subventions des différents ministères (environ 85%), et détient actuellement 150 brevets et 500 obtentions végétales. Plus de 10000 personnes travaillent, de façon permanente ou non, à l'INRA, dont 3800 chercheurs et ingénieurs, 4800 techniciens et personnels d'appui à la recherche, 1000 thésards et 1000 stagiaires d'écoles d'ingénieurs, de DEA, de DESS ou chercheurs étrangers. Ces 10000 personnes sont réparties dans 280 unités de recherche, 80 unités expérimentales et une centaine d'unités de service ou d'appuis à la recherche. Les unités de recherche et d'expérimentation dépendent de 17 départements scientifiques. Les diverses unités de l'INRA sont regroupées dans 21 centres de recherche répartis sur tout le territoire.

1.2. Le centre de Nantes

Ce centre fut créé en 1976 et il est le seul centre dont les activités concernent principalement les transformations des produits agricoles. Il est implanté sur le site de la Géraudière, qui accueille des établissements comme l'ENITIAA (Ecole Nationale d'Ingénieurs des Industries Agricoles et Alimentaires), la Chambre d'agriculture de Loire-Atlantique ou IQUABIAN (centre d'ingénierie de la qualité en biotechnologie et industries alimentaires accueillant 3 DESS et certaines ressources documentaires de l'INRA et de l'ENITIAA). Le centre est en relation avec l'INSERM, le CNRS, les universités de Nantes, du Maine et d'Angers, l'INH et l'ESA d'Angers, ...

Le centre concentre ses activités dans trois grandes thématiques :

- deux thématiques concernent les sciences de la vie :

① approfondissement des connaissances sur la structure, les interactions et les fonctions des macromolécules pour améliorer la compréhension des propriétés des aliments et pour proposer de nouvelles valorisations non alimentaires

② élaboration de nouvelles méthodologies de caractérisation et d'amélioration de la qualité des aliments

- la troisième thématique concerne les sciences économiques et sociales.

Le centre rassemble 14 unités de recherche dépendant de 6 départements scientifiques :

- transformation des produits végétaux (4 unités)
- transformation des produits animaux (2 unités)
- nutrition, alimentation et sécurité alimentaire (1 unité)
- santé animale (4 unités)
- hydrobiologie et faune sauvage (1 unité)
- économie et sociologie rurales (2 unités : l'UREQUA : Unité de Recherche Economie des Qualifications Agroalimentaires, associée à l'université du Mans et l'UMR EDRA : Unité Mixte de Recherche Economie – Droit Rural et Agroalimentaire où j'ai effectué mon stage.)

250 personnes travaillent sur le centre, dont 110 chercheurs et ingénieurs, 60 techniciens, 30 agents administratifs et 50 thésards. De plus, une cinquantaine d'étudiants effectuent leur stage dans l'une ou l'autre unité.

1.2.1. L'Unité Mixte de Recherche Economie – Droit Rural et Agroalimentaire (UMR EDRA)

Cette UMR a été créée le 1er janvier 2001 et rassemble le CEDRAN (Centre d'Etudes de Droit Agroalimentaire de Nantes) de l'Université de Nantes et le LERECO (Laboratoire d'Etudes et de Recherches Economiques) de l'INRA de Nantes. Elle résulte d'un processus de rapprochement entre les économistes et les juristes de l'INRA et est l'occasion de conforter la volonté régionale de création d'un pôle agroalimentaire dont la première phase a abouti à la mise en place du projet IQUABIAN. L'enjeu de cette unité est le rapprochement de deux disciplines sœurs mais très autonomes que sont l'économie et le droit et elle constitue la première expérience de ce type en France. Il s'agit non seulement d'appréhender les effets croisés entre droit et économie mais aussi d'intégrer les questions spécifiques à chacune des deux disciplines.

L'UMR est structuré autour de trois axes :

Premier axe : " *Orientation des politiques publiques et dynamiques des exploitations agricoles* "

Les recherches menées dans cet axe visent à mieux comprendre l'enjeu des changements de politique publique pour les exploitations agricoles françaises et européennes. Les conséquences économiques de différents scénarios de politique agricole sont testées de façon statique et complétées par l'analyse à la fois économique et juridique des conditions d'adaptation des exploitations.

Deuxième axe : " *Ouverture internationale des marchés agricoles et alimentaires : implications économiques et juridiques.* "

Parallèlement à une libéralisation accrue des échanges, on assiste au niveau international à l'émergence de nouvelles normes et au développement d'accords régionaux. Ces deux éléments peuvent se comprendre comme une entrave à la liberté des échanges. Le deuxième axe de l'UMR se focalise sur les enjeux des négociations internationales pour les échanges agricoles et alimentaires européens et sur le droit alimentaire français et européen.

Troisième axe : *" Propriété intellectuelle, recherche et innovation "*

On associe sur cette thématique les approches bibliométrique (quantification de la production et des échanges de connaissance scientifique et technologique), économique (en lien avec les spécialistes de l'économie du changement technique et de l'économie spatiale, ainsi que l'axe 2 de l'UMR), et juridique (propriété intellectuelle et droit communautaire), autour de deux thèmes majeurs :

- les mécanismes de propriété intellectuelle sous l'angle de la réalisation d'un optimum social (compromis efficace pour la société entre libre circulation et appropriation des connaissances) et de leur interaction avec les stratégies des acteurs publics et privés.

- les systèmes d'innovation et la construction européenne avec une analyse spatiale des activités scientifiques et technologiques et des coopérations à l'échelle des régions européennes, et une analyse de la proposition de règlement sur le brevet communautaire.

Le travail de l'équipe IBIS (Indicateurs Bibliométriques et Scientométrie), au sein de laquelle j'ai effectué mon stage, s'inscrit en partie dans ce troisième axe; il comporte d'autres volets détaillés ci-dessous.

Actuellement, 7 agents de l'INRA (un directeur de recherche, un technicien de recherche, un adjoint technique de recherche, deux chargés de recherche et deux ingénieurs de recherche), 3 thésards et 2 stagiaires travaillent, dans le cadre de cette UMR, sur le site de l'INRA de Nantes.

1.2.2. Le travail de l'équipe IBIS

En dehors des travaux sur les systèmes d'innovation, l'équipe IBIS conduit des recherches sur deux thèmes : les méthodes infométriques pour l'analyse des réseaux scientifiques et les méthodologies d'indicateurs des sciences et des techniques à diverses échelles (pays, institutions, régions, ...). Mon stage s'inscrivait dans la première thématique.

Quel que soit l'objectif, cognitif ou opérationnel, l'étude des univers scientifiques par les méthodes quantitatives suppose la maîtrise d'une série de méthodes. La mise au point au laboratoire de la plate-forme de recherche et d'étude SINDBAD permet de mobiliser pour de nombreux types d'application les ressources statistiques des logiciels SAS. L'effort méthodologique a porté d'abord sur les méthodes de structuration et de cartographie des univers et réseaux scientifiques, en particulier celles reposant sur les relations de citations entre documents, les co-citations et les couplages bibliographiques (ces différentes notions seront abordées dans la prochaine section), puis sur les structurations à base de mots (méthodes lexicales) pour une présentation rapide [Zitt 1998]. Certains travaux en cours concernent la transposition des méthodes citationnistes aux univers du Web [Prime 2001].

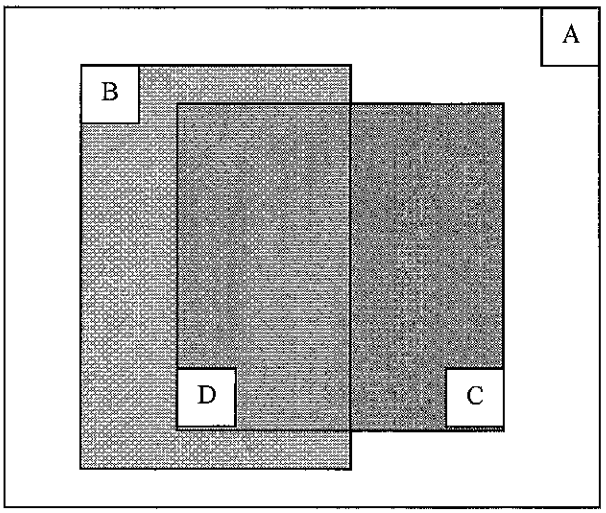
II. Problématique du stage

2.1. La démarche d'une étude bibliométrique

Les études bibliométriques comportent typiquement trois étapes : la recherche des données pertinentes pour le champ de l'étude (constitution d'un corpus), des analyses quantitatives diverses sur ces données (construction d'indicateurs), enfin une interprétation et une validation des résultats (par exemple par dire d'expert). Les études auxquelles j'ai participé sont focalisées sur un domaine, et un aspect important de l'analyse quantitative consiste en général à structurer ce domaine en ses diverses composantes (par exemple « thèmes de recherche »), afin d'établir sur chaque thème une série d'indicateurs, et aussi d'analyser les relations entre ces thèmes. Ces indicateurs servent en général à positionner les différents acteurs (pays, institutions, laboratoires, etc.) sur le domaine, en termes de volume de publication, de spécialisation, de citations reçues (« impact » des publications), et aussi de réseaux de collaboration.

2.1.1. Constitution du corpus

Un corpus est un ensemble limité d'éléments pris en compte pour une recherche bibliométrique sur un champ donné : c'est une population, un recensement exhaustif ou un échantillon de cette population, la population étant l'ensemble de l'information disponible sur le champ. Pour constituer ce corpus, plusieurs systèmes existent : recherche sur Internet, consultation de bases de données bibliographiques, etc. Cependant, il est très rare d'avoir un fichier propre d'entrée car deux phénomènes entrent en jeu : l'apparition d'un « bruit » et d'un « silence ». Représentons ces deux notions sur un schéma.



Supposons une base de données exhaustive sur les publications scientifiques. On s'intéresse à l'ensemble A qui représente tous les articles scientifiques publiés à ce jour. Le sous-ensemble B rassemble les articles pertinents sur les pathologies animales. En lançant une requête sur les pathologies animales, on a obtenu le sous-ensemble E, constitué des sous-ensembles C et D. On constate que certains articles sur les pathologies animales ($B-D$) sont absents, alors que des articles abordant d'autres aspects de la recherche animale ou des articles généraux ont été obtenus (ensemble C). L'ensemble $B-D$ représente le silence (puisque des informations importantes sont tues) et l'ensemble C représente le bruit (des informations sans valeur sont prises en compte) [1]. Nous pouvons résumer ces différentes notions dans le tableau suivant :

	Pertinent	Non pertinent
Sélectionné	D	C : bruit
Non sélectionné	B - D : silence	A - B - C

[1] Ces notions de bruit et de silence proviennent de la théorie documentaire (information retrieval en anglais).

Pour avoir un corpus aussi propre que possible, il faut minimiser le bruit, comme le silence. Bruit et silence dépendent de la formulation de la requête et de la qualité de l'information interrogeable dans la base. Les requêtes efficaces sont en général le résultat d'un processus itératif. Les logiciels d'interrogation des bases de données ne permettent pas toujours d'obtenir des corpus parfaitement pertinents. Il est alors souvent souhaitable d'opérer une pré-analyse du corpus après déchargement, pour éliminer efficacement le bruit (domaines hors sujet rappelés par exemple par des mots polysémiques dans les requêtes). Il est souvent utile aussi de pratiquer des filtrages (sur certaines catégories de fréquence de termes ou sur certains documents comme les doublons). En fait, de nombreuses techniques de statistique ou d'analyse des données utilisées pour structurer finement l'univers (cf. § 2.1.2.) peuvent aussi être employées dans un premier temps pour « nettoyer » le corpus.

2.1.2. Structuration du corpus

La structuration consiste à identifier des sous-ensembles et les liens qui les relient. Pour expliquer comment s'organise cette recherche de structure, examinons une étude très courante en bibliométrie (*) : l'étude des références bibliographiques. Le corpus est typiquement constitué de notices couvrant un domaine défini. Si ce domaine est vaste, il convient de le découper en sous-ensembles homogènes pour bâtir des indicateurs significatifs. Les notices bibliographiques comportent plusieurs champs, dont voici un exemple.

FN- Science Citation Index (Jan 91 - Dec 91) ***Nom du fichier***

TI- STUDIES ON APPLE PROTOPECTIN .3. CHARACTERIZATION OF THE MATERIAL EXTRACTED BY PURE POLYSACCHARIDASES FROM APPLE CELL-WALLS ***Titre***

LA- ENGLISH ***Langue***

AU- RENARD CMGC; SCHOLS HA; VORAGEN AGJ; THIBAULT JF; PILNIK W ***Auteurs***

CS- INRA, BIOCHIM & TECHNOL GLUCIDES LAB/RUE GERAUDIERE/BP 527/F 44026 NANTES 03//FRANCE; AGR UNIV WAGENINGEN/FOOD CHEM LAB/6700 EV WAGENINGEN//NETHERLANDS ***Adresses***

JN- CARBOHYDRATE POLYMERS, 1991, V15, N1, P13-32 ***Journal***

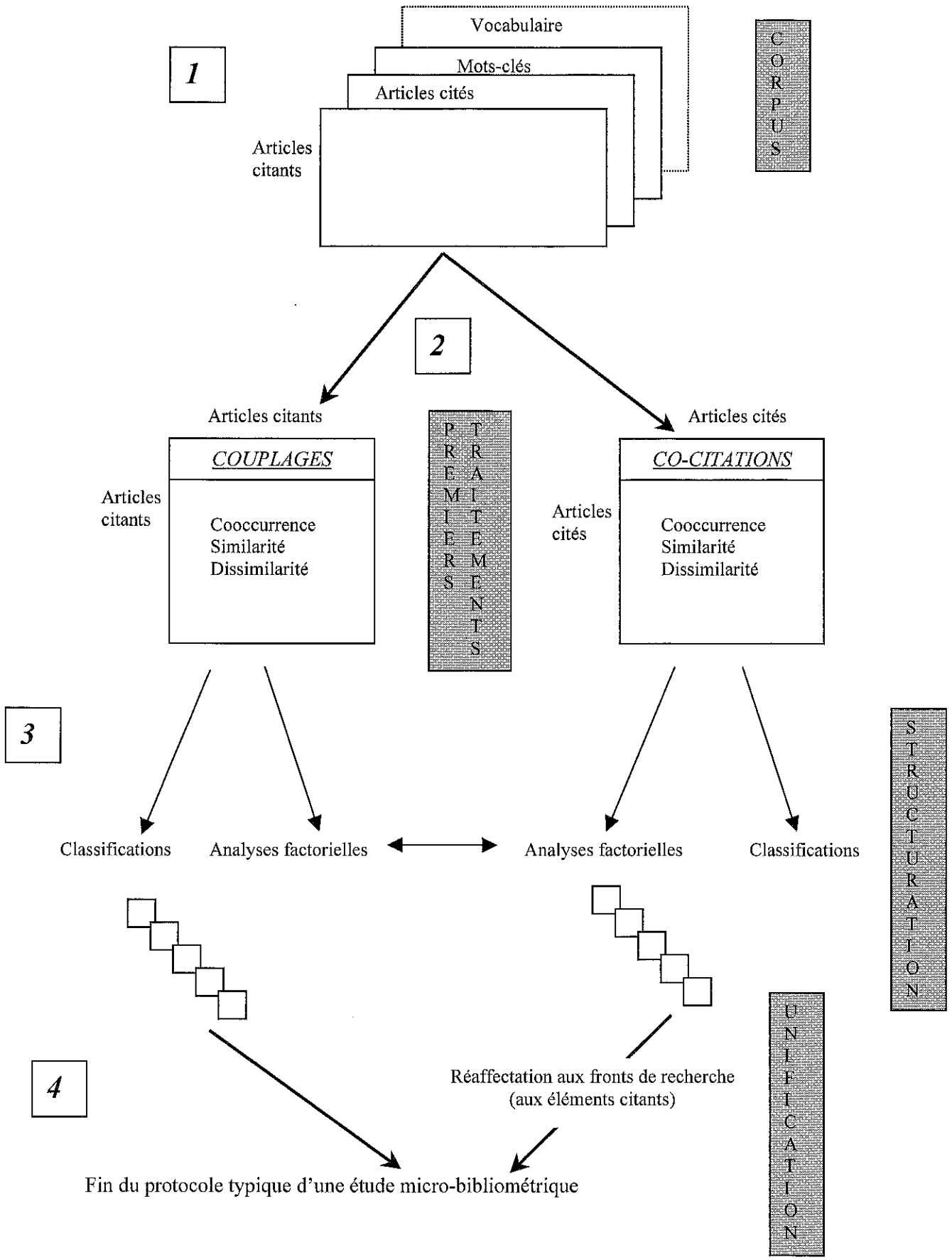
DT- ARTICLE ***Type de document***

NR- 35 ***Nombres de références***

CR- ASPINALL GO, 1984, V4, P193, CARBOHYD POLYM
BARRETT AJ, 1965, V94, P617, BIOCHEM J
BAUER WD, 1973, V51, P174, PLANT PHYSIOL
BELDMAN G, 1985, V146, P301, EUR J BIOCHEM
BELDMAN G, 1988, V31, P160, BIOTECHNOL BIOENG
BENSHALOM N, 1986, V51, P720, J FOOD SCI
DEVRIES JA, 1981, V1, P117, CARBOHYD POLYM
DEVRIES JA, 1982, V2, P25, CARBOHYD POLYM
DEVRIES JA, 1983, V3, P193, CARBOHYD POLYM
DEVRIES JA, 1983, V3, P245, CARBOHYD POLYM
DEVRIES JA, 1983, THESIS AGR U WAGENIN
HAYASHI T, 1984, V75, P605, PLANT PHYSIOL
ISHII S, 1981, V20, P2329, PHYTOCHEMISTRY
ISHII S, 1982, V21, P778, PHYTOCHEMISTRY
JARVIS MC, 1981, V32, P1309, J EXP BOT
KEEGSTRA K, 1973, V51, P188, PLANT PHYSIOL
KNEE M, 1973, V12, P1543, PHYTOCHEMISTRY
KNEE M, 1973, V12, P637, PHYTOCHEMISTRY
KNEE M, 1975, V14, P2213, PHYTOCHEMISTRY
KNEE M, 1978, V17, P1257, PHYTOCHEMISTRY
KNEE M, 1978, V17, P1261, PHYTOCHEMISTRY
KONNO H, 1982, V69, P864, PLANT PHYSIOL
MASSIOT P, 1989, V190, P121, CARBOHYD RES
PILNIK W, 1970, V1, P53, BIOCH FRUITS THEIR P
RENARD CMG, 1990, V14, P295, CARBOHYD POLYM
RENARD CMGC, 1990, V12, P9, CARBOHYD POLYM
ROBERT P, 1985, V5, P501, SCI ALIMENT
ROUAU X, 1984, V4, P111, CARBOHYD POLYM
SAULNIER L, 1987, V7, P329, CARBOHYD POLYM
STEVENS BJH, 1984, V135, P155, CARBOHYD RES
TALMADGE KW, 1973, V51, P158, PLANT PHYSIOL
THIBAULT JF, 1988, V9, P119, CARBOHYD POLYM
VORAGEN AGJ, 1980, V2, P458, J APPL BIOCHEM
VORAGEN AGJ, 1986, V370, P113, J CHROMATOGR
VORAGEN FGJ, 1986, V183, P105, Z LEBENSM UNTERS FOR ***Références bibliographiques***

Exemple de notice bibliographique

Le schéma suivant représente le processus, nous allons l'expliquer par la suite.



5

Indicateurs par thèmes
Relations entre thèmes

INDICATEURS

Etape 1 :

On part d'une matrice rectangulaire à n dimensions, $n \geq 2$, croisant des codes associés à des articles, dits citants, à d'autres codes : références bibliographiques (articles cités), mots-clés, vocabulaire, ... Souvent c'est une matrice booléenne de présence-absence. Par la suite, on ne s'intéressera qu'à la table citants-cités.

Etape 2 :

Ensuite le problème se pose de savoir quel univers on souhaite explorer en priorité, les articles citants ou les articles cités. Si on souhaite une structuration primaire sur les relations entre citants, on emprunte la voie des couplages bibliographiques; si on souhaite une structuration primaire sur les relations entre cités, on emprunte la voie des co-citations.

L'analyse des couplages bibliographiques : Cette méthode a été introduite par Kessler en 1963. Elle part de l'hypothèse suivante : des documents ayant des références bibliographiques communes ont une parenté thématique. Pour déterminer les couplages existants, deux critères sont définis : ① premier critère, noté G_A : Soit P_0 un article scientifique. On définit $G_A(P_0)$ comme l'ensemble des articles ayant au moins une référence commune avec P_0 et $G_A(P_0, n)$ l'ensemble des articles ayant n références communes avec P_0. Ce critère permet d'élargir le nombre d'articles potentiellement intéressants, puisqu'à partir d'un document pertinent, on retrouve l'ensemble des documents couplés avec celui-ci.

② deuxième critère, noté G_B : G_B est un ensemble d'articles qui ont au moins une référence commune (ce critère peut être généralisé à n articles). Ce critère essaie de rapprocher thématiquement les documents d'un corpus.

Un exemple d'utilisation de cette analyse figure dans la section I de la partie « Etudes menées et résultats obtenus ».

La méthode des co-citations : Elle est un renversement du couplage bibliographique. On cherche à déterminer les ensembles (appelés noyaux) d'articles fréquemment cités ensemble de la même façon que l'on a déterminé les couplages bibliographiques. Puis on réaffecte les articles citants à ces noyaux. Un exemple d'utilisation de cette méthode figure dans la section II de la partie « Etudes menées et résultats obtenus ».

En calculant diverses valeurs (fréquences de cooccurrence, fréquences marginales ou d'occurrence), on peut déterminer des indices de similarité (*) ou de dissimilarité (*) entre articles citants ou articles cités. Comme indiqué précédemment, ces techniques peuvent aussi être utilisées pour nettoyer le fichier, par exemple en identifiant des variantes graphiques désignant le même article citant (voir application au § 1.2.1. de la partie « Etudes menées et résultats obtenus »).

Etape 3 :

Une fois notre matrice de co-citation (ou de couplage) définie, on peut utiliser les techniques d'analyses factorielles ou de classifications, ainsi que d'autres méthodes plus sophistiquées, afin de déterminer des groupements d'articles citants ou d'articles cités (ces groupements peuvent être des blocs diagonaux si le domaine est complètement hétérogène).

Etape 4 :

Après structuration, par exemple par les co-citations, il est en général nécessaire de réinjecter l'information sur les articles citants. Des techniques comme les méthodes de réaffectation entrent alors en ligne de compte.

Les résultats trouvés suite à l'étape de structuration permettent souvent d'éliminer le bruit, mais ils peuvent aussi montrer que le corpus n'est pas totalement pertinent. Il faut alors une itération supplémentaire.

2.2. Objectifs du stage

Les méthodes d'analyse des données (analyses factorielles, classifications, ...) peuvent donc être utilisées lors de la phase de constitution du corpus, afin de détecter et d'éliminer le bruit, ou bien lors de la phase de structuration du corpus, pour mettre en avant les groupements d'éléments significatifs. Cependant, outre le problème du choix de la méthode à adopter, se pose le problème de la rapidité et de l'efficacité de la méthode : faut-il prendre des méthodes simples et rapides d'utilisation mais peu efficaces, ou préférer des méthodes plus lourdes en terme de temps de calcul et de manipulation mais plus pertinentes pour les résultats ?

Pour éclairer cette question, il m'a été demandé d'étudier différentes méthodes sur un échantillon (fonctionnement, résultats obtenus) : une méthode de partitionnement rapide et les méthodes de classification ascendante hiérarchique, et de comparer ces méthodes. Cette étude, pour se reporter au schéma de la page 15, se situe à l'étape 3. De plus, afin de bien comprendre comment était mené une recherche bibliométrique, j'ai exécuté quelques travaux se concentrant sur la constitution d'un corpus pertinent.

Etudes menées et résultats obtenus

I. Elimination du bruit

1.1. Données et but de l'étude

Cette étude fait suite à un stage qui eut lieu en juillet 2000, effectué par Camille Prime [2]. Le but du stage était de déterminer, dans un domaine précis, une structure des pages Internet, pour savoir quels sont les différents univers possibles, afin d'accélérer la recherche de sites intéressants lors d'études ultérieures.

Le test a été effectué sur le champ de compétences de l'équipe : scientométrie / bibliométrie, indicateurs, science et technologie.

Le fichier de départ contient une double liste de numéros. La première liste correspond à des adresses URL citantes, la seconde à des adresses URL citées, c'est-à-dire que pour chaque numéro de la première liste, on a toutes les adresses citées caractérisées par les numéros de la deuxième liste. Ce fichier a 704 numéros d'adresses citantes et 385 numéros d'adresses citées.

L'objectif de cette étude est d'obtenir une bonne classification des adresses citées, en effectuant plusieurs tris au préalable, d'abord sur les adresses citantes, ensuite sur les adresses citées.

[2] PRIME Camille, « Transposition des méthodes bibliométriques pour caractériser les univers du Web », mémoire de DEA DISIC (Document Multimédia, Images et Systèmes d'Information Communicants), juillet 2000.

1.2. Déroulement de l'étude

Une étude préalable avait montré que, en associant un rang à chaque page citante selon le nombre décroissant de liens, la fonction log(nombre de liens)*log(rang) suivait une loi de Zipf, c'est-à-dire que la fonction pouvait être approximée par une droite. Toutefois, on pouvait observer l'apparition de trois types de pages citantes :

① Les portails généraux ou scientifiques : ils sont caractérisés par le fait qu'ils ont beaucoup de liens, mais ils représentent une faible part des pages citantes

② Les portails spécifiques à la bibliométrie et à l'infométrie : ceux-ci sont plus nombreux et ont moins de liens mais ils peuvent tout de même donner des résultats peu pertinents

③ Les autres pages citantes, caractérisées par des contenus très pertinents, un faible nombre de liens et qui représente la grande majorité des pages étudiées.

Le fichier utilisé pour l'étude ici détaillée contenait des adresses des pages citantes avec les adresses de leurs pages citées (représentés par des numéros), en ayant pris soin au préalable d'éliminer les portails généraux et en ne conservant que les pages citantes ayant une fréquence de citations supérieure à 5.

1.2.1. Traitement des citants

Nous avons d'abord essayé de savoir si, dans un souci de bonne interprétation de la classification finale, il ne fallait pas éliminer quelques pages citantes. Pour cela, nous avons déterminé tous les couplages de numéros d'adresses citantes qui ont des liens en commun : 13 962 couplages ont ainsi été trouvés sur les 704*703 = 494 912 possibles (la part des couplages empiriques sur les couplages possibles n'est que d'environ 3 %).

Nous avons alors calculé un indice de couplage en déterminant les valeurs suivantes :

n_i le nombre de liens de l'adresse n° i

n_j le nombre de liens de l'adresse n° j

n_{ij} est le nombre de liens communs aux adresses n° i et n° j.

Nous avons calculé l'indice de similarité (*) suivant : $Cocit_1(i,j) = \frac{n_{ij}}{\sqrt{n_i n_j}}$ (indice d'Ochiai).

Une fois cela fait, nous avons déterminé quels étaient les couplages d'adresses dont l'indice était égal à 1 (c'est-à-dire quand les deux pages ont strictement les mêmes liens). Nous avons alors constaté que cela représentait 525 numéros sur les 704 de départ. En regardant attentivement les intitulés des adresses associées, nous remarquâmes que toutes ces pages étaient simplement des pages particulières de quelques sites, ou bien elles représentaient des portails. Nous n'avons alors conservé que 35 adresses. Le nouveau fichier contenait alors 214 numéros d'adresses citantes (704-525+35) et 372 numéros d'adresses citées. Comme le nombre d'adresses citées a peu baissé (-4 %), les liens des pages supprimées devaient être en grande majorité internes.

1.2.2. Traitement des cités

Le but de l'étude est de classer les adresses citées. Pour cela nous avons déterminé tous les couples d'adresses qui sont référencés par la même page. 28 158 couples ont ainsi été déterminés sur les 138 012 possibles (372*371), soit un rapport empirique/théorique de l'ordre de 2 %.

Nous avons alors calculé le nombre d'occurrences (ou de citation) de chaque page citée. Nous n'avons pris en compte que celles qui étaient citées plus d'une fois pour une raison toute simple : en effectuant une CAH (Classification Ascendante Hiérarchique) avec le critère de la moyenne selon l'indice défini ci-après sur toutes les adresses, les résultats obtenus étaient difficilement interprétables. En outre, une page citée une seule fois ne crée pas de lien entre deux pages citantes. De cette manière, 230 adresses ont été conservées pour la suite de l'étude.

Nous avons alors calculé un deuxième indice de similarité (*) de la façon suivante :

Soient n_k le nombre de liens de l'adresse n° k et $Cit(i)$ l'ensemble des pages référençant l'adresse n° i.

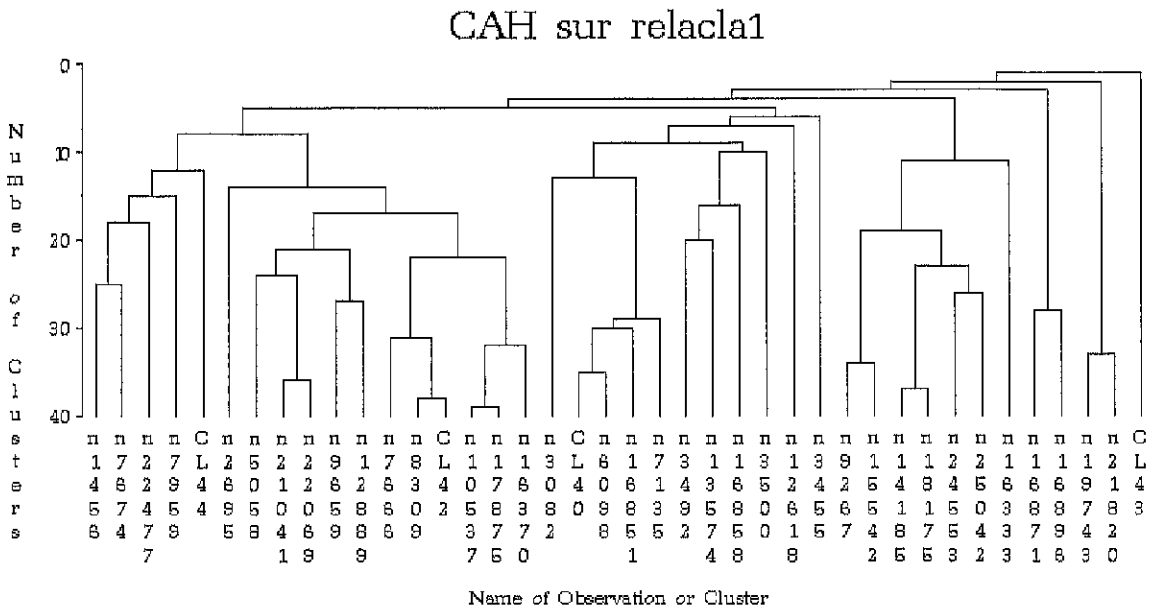
$$\text{Posons } \begin{cases} Sn_i = \sum_{k \in Cit(i)} \frac{1}{n_k} \\ Sn_j = \sum_{k \in Cit(j)} \frac{1}{n_k} \\ Sn_{ij} = \sum_{k \in (Cit(i) \cap Cit(j))} \frac{1}{n_k} \end{cases}.$$

L'indice calculé est le suivant : $Cocit_2(i,j) = \frac{Sn_{ij}}{\sqrt{Sn_i Sn_j}}$ (c'est un indice d'Ochiai pondéré par le nombre de liens des pages citantes).

Nous avons alors procédé à une CAH selon le critère de la distance moyenne sur la matrice de l'indice de co-citation $Cocit_2$[3]. Signalons que, parmi les 52 670 couples, seuls 6 050 – soit 11,5 % – avaient un indice non nul. La matrice utilisée est donc creuse (*).

[3] En fait, sur le complémentaire à 1 de $Cocit_2$, pour pouvoir assimiler la matrice à une matrice de distance.

1.3. Résultats



Ce dendogramme représente la décomposition en 40 classes des pages citées. Après validation des résultats par Michel Zitt et Camille Prime, actuellement en thèse à l'Ecole des Mines de Saint Etienne, une partition des pages en 26 classes et 4 singletons a été retenue. Les résultats ont été exposés dans [Prime 2001].

II. Classification de documents

Une fois tous les éléments disponibles, il faut identifier les liens existant entre ces éléments. Il est donc souvent utile d'utiliser une méthode de classification. Bien sûr, on souhaite avoir une méthode rapide et efficace mais deux problèmes peuvent occasionner une perte de temps : d'abord, le grand nombre d'éléments à classer (il n'est pas rare, en bibliométrie, d'en avoir une bonne dizaine de milliers, voire plus), et ensuite la matrice utilisée pour ces méthodes est très souvent creuse (*). C'est pourquoi je devais étudier un algorithme qui accepte les données sous forme de liste et comparer les résultats obtenus avec les classifications usuelles. C'est ce travail qui est maintenant exposé.

2.1. Jeu de données et premiers traitements

Les données utilisées étaient un tableau recensant, pour un article quelconque, les thèmes abordés. Le fichier comportait 4 906 articles traitant de 154 grands thèmes codés dans la base de données. Nous avons d'abord calculé les fréquences de citations de ces différents codes; nous avons alors constaté que quatre codes – LL600, LL820, LL860 et LL880 – représentaient 54% des citations. Afin de ne pas perturber la suite des calculs, ces quatre codes ont été retirés [4] [5]. Le nouveau fichier ne comportait alors que 2 678 articles et 150 thèmes.

[4] Le tableau des fréquences d'occurrence de chaque code figure en annexe, à la page 26.

[5] Les intitulés précis de ces quatre codes sont :
Animal Science: Animal Physiology and Biochemistry (excluding Nutrition)
Animal Science: Animal Health and Hygiene (General): Parasites, Vectors, Pathogens and Biogenic Diseases of Animals
Animal Science: Animal Health and Hygiene (General): Animal Disorders (not caused by Organisms)
Animal Science: Animal Health and Hygiene (General): Animal Treatment and Diagnosis (non Drug).

Le tableau qui suit représente une notice dont un des champs correspond aux codes traités dans cette partie.

AN 20000404223 ***Code d'accès dans la base de données***

AU Ochirkhuyag, B. Chobert, J. M. Dalgalarrondo, M. Haertle, T ***Auteurs***

IN Institute of Chemistry, Academy of Sciences, Ulanbator, Mongolia ***Instituts***

TI Characterization of mare caseins. Identification of alpha S1- and alpha S2-caseins. ***Titre***

SO Lait. 2000. 80: 2, 223-235. 31 ref. ***Source (nom du journal, année, volume, numéro, pages, nombre de références)***

AB Whole mare milk casein (from Mongolian and French breeds) was obtained from skim milk by isoelectric precipitation (pH 4.6) at 22 deg C. In another series of experiments, mare milk caseins were fractionated after isoelectric precipitation at 4 deg C, according to their sensitivity to temperature. In one group of experiments, the pH of the resulting pellet was adjusted to 6.5 before centrifugation (45 000 g for 30 min) at 4 deg C and the resulting casein fraction precipitated was labelled the cold-precipitated (CP) fraction. In another group of experiments, the supernatant obtained from the centrifugation of skim milk at pH 4.6 and 4 deg C was warmed to 30 deg C before being centrifuged at 30 deg C, 45 000 g, for 30 min. The resulting casein fraction was labelled the thermally-precipitated (TP) fraction. Mare milk caseins were then purified by high resolution gel chromatography on an anion-exchange column followed by reversed phase-HPLC. The casein fractions were analysed by urea- and SDS-PAGE, their amino acid compositions were determined and their first 15 N-terminal amino acids were sequenced. This analysis showed the presence of alpha s-like caseins isolated from the CP fraction. alpha s1-Like-casein showed 4 major double bands by electrofocusing with a pI range of 4.3-4.8. Under the same conditions, alpha s2-like casein showed 2 major bands with a pI range of 4.3-5.1. The TP fraction revealed the existence in mare milk of 6 sub-fractions of beta -like caseins, occurring in the following ratios: 1:5:18:20:15:10, differing in their degree of phosphorylation (as shown also by the action of acid phosphatase). Since no evidence of the presence of kappa -casein was found in mare milk, it is proposed that part of its functions in mare milk may be assumed by the population of less phosphorylated beta -caseins ***Résumé***

RN 9001-77-8 ***Numéro des molécules***

CC Food Science and Food Products (Human): Milk and Dairy Produce [QQ010], Food Science and Food Products (Human): Food Composition and Quality [QQ500] ***Cabicode (codes traités dans l'etude)***

DE mares. casein. milk. phosphorylation. precipitation. ratios. skim milk. mare milk. centrifugation. heat treatment. HPLC. PAGE. acid phosphatase. analytical methods. ***Descripteurs***

GL France, Mongolia ***Localisation géographique***

BT Western Europe, Europe, Mediterranean Region, Developed Countries, European Union Countries, OECD Countries, East Asia, Asia, Developing Countries ***Termes associés***

ID kappa -casein, alpha s-casein, alpha s1-casein, alpha s2-casein, beta -casein ***Identificateurs***

LG English ***Langue***

SL French ***Langue du résumé***

PT Journal article ***Type de publication***

IS 0023-7302 ***Code ISSN***

YR 2000 ***Année***

UP 200005 ***Date d'entrée dans la base***

2.2. La méthode des co-items

La méthode des co-items est très utilisée en bibliométrie (*). Les items classés peuvent être des articles cités (méthode des co-citations), des mots-clés (méthodes des mots associés), des codes de classement (méthode de co-classement),... Par exemple, la méthode des co-citations part d'une table article citant – article cité où pour chaque article citant, on a ses références bibliographiques [6]. On cherche à déterminer les liens qui existent entre les références bibliographiques, notamment quels sont les articles qui sont souvent cités ensemble et quels sont les articles qui apparaissent souvent (souvent des articles de base, généraux) et ceux qui sont mentionnés plus rarement (articles plus pointus). Pour cela, on calcule les fréquences d'occurrence de chaque article et les fréquences de cooccurrence de deux articles (c'est-à-dire le nombre de fois où deux articles sont cités ensemble. Ensuite, on calcule un indice de cooccurrence (qui est souvent un indice de similarité (*)) entre les couples d'articles et on effectue une CAH (Classification Ascendante Hiérarchique) sur la matrice associée à cet indice. Nous allons appliquer ce type de méthode dans une étude de co-classement.

2.2.1. Calcul de l'indice de co-occurrence

Avant de mentionner l'indice utilisé, définissons le tableau de présence-absence de deux thèmes i et j.

		Thème j		
		Présence	Absence	
Thème i	Présence	n_{ij}	$n_{i\bar{j}}$	n_i
	Absence	$n_{\bar{i}j}$	$n_{\bar{i}\bar{j}}$	$n_{\bar{i}}$
		n_j	$n_{\bar{j}}$	n

où n_{ij} désigne le nombre de fois où les thèmes i et j sont abordés par le même article (c'est la fréquence de cooccurrence),

$n_{i\bar{j}}$ désigne le nombre de fois où le thème i est abordé, mais pas le thème j,

[6] Dans notre étude, les articles cités sont les thèmes abordés par les articles.

n_i désigne le nombre de fois où le thème i est abordé (c'est la fréquence d'occurrence du thème i),

$n_{\bar{i}j}$ désigne le nombre de fois où le thème j est abordé, mais pas le thème i,

$n_{\bar{i}\bar{j}}$ désigne le nombre de fois où les thèmes i et j ne sont pas abordés,

$n_{\bar{i}}$ désigne le nombre de fois où le thème i n'est pas abordé,

n_j désigne le nombre de fois où le thème j est abordé (c'est la fréquence d'occurrence du thème j),

$n_{\bar{j}}$ désigne le nombre de fois où le thème j n'est pas abordé,

n désigne le nombre total de fois où un thème est abordé.

En calculant ces différentes valeurs, nous avons constitué 1 852 couples de thèmes différents, ce qui est très loin des 150*149 = 22 350 couples possibles (rapport de 8,3 %).

Une fois cette démarche effectuée, nous pouvons calculer l'indice de cooccurrence. Deux indices sont couramment utilisés :

① L'indice de Jaccard $\left(Jacc(i,j) = \frac{n_{ij}}{n_{ij} + n_{\bar{i}j} + n_{i\bar{j}}} \right)$ (rapport entre l'apparition simultanée de deux thèmes et l'apparition d'au moins un des deux thèmes)

② L'indice d'Ochiai $\left(Ochi(i,j) = \frac{n_{ij}}{\sqrt{n_i n_j}} \right)$ (rapport entre l'apparition simultanée de deux thèmes et la moyenne géométrique d'apparition d'un des deux thèmes).

Un troisième indice est quelquefois utilisé, c'est l'indice d'équivalence qui est tout simplement le carré de l'indice d'Ochiai.

Par manque de temps, seuls les résultats obtenus à partir de l'indice de Jaccard ont été établis et étudiés.

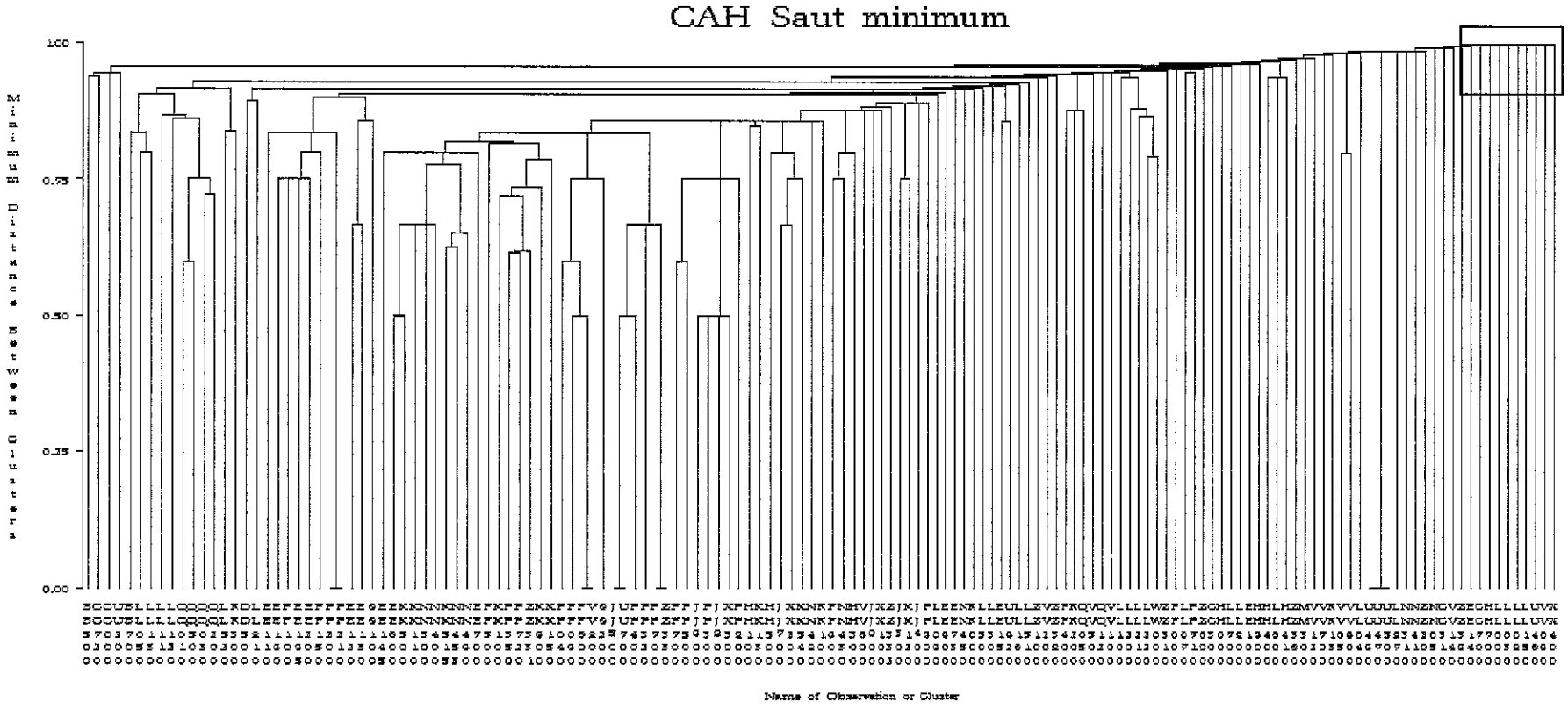
2.2.2. Classifications

Maintenant que la matrice est définie, nous allons pouvoir étudier les résultats donnés par les différentes classifications mises en œuvre. Nous avons utilisé les quatre principaux critères : le saut minimum, le diamètre, la distance moyenne, le critère de Ward (*).[7]

[7] ATTENTION: La matrice utilisée n'est pas la matrice de Jaccard (nommons là MatJacc), mais la matrice complémentaire à 1 de MatJacc (nommons là MatClass). En clair, $\forall(i, j), MatClass(i, j) = 1 - MatJacc(i, j)$.

2.2.2.1. Critère du saut minimum

a) Etude du dendogramme



On remarque que des petites classes se forment au départ mais rapidement, toutes ces classes s'agrègent entre elles, ce qui occasionne un effet de chaîne, puisqu'il y a apparition d'une très grosse classe, qui s'oppose à des classes beaucoup plus petites et à des singletons (encadré).

b) Etude de la taille des classes à un niveau donné

Nous avons effectué un découpage en 38 classes [8]. Voici les résultats obtenus.

CAH Saut minimum

Obs	cluster	frequence
1	CC300	1
2	CC310	1
3	CC700	1
4	CL101	2
5	CL147	3
6	CL38	104
7	CL39	2
8	CL44	4
9	CL50	3
10	EE100	1
11	EE140	1
12	FF000	1
13	HH000	1
14	HH600	1
15	HH700	1
16	LL000	1
17	LL030	1
18	LL040	1
19	LL080	1
20	LL150	1
21	LL700	1
22	LL800	1
23	LL870	1
24	MM300	1
25	NN050	1
26	NN310	1
27	NN410	1
28	RR130	1
29	UU490	1
30	VV060	1
31	VV120	1
32	VV140	1
33	VV700	1
34	XX400	1
35	ZZ200	1
36	ZZ360	1
37	ZZ390	1
38	ZZ900	1

[8] Pour pouvoir comparer les classes ainsi obtenues avec celles obtenues par le « quick clustering » (taille, composition). Bien sûr, le même découpage a eu lieu pour les trois autres critères.

The UNIVARIATE Procedure
Variable: frequence

Frequency Counts

Percents Value	Count	Percents Value	Count	Percents Value	Count	Percents Value	Count
1	32	3	2	4	1	104	1
2	2						

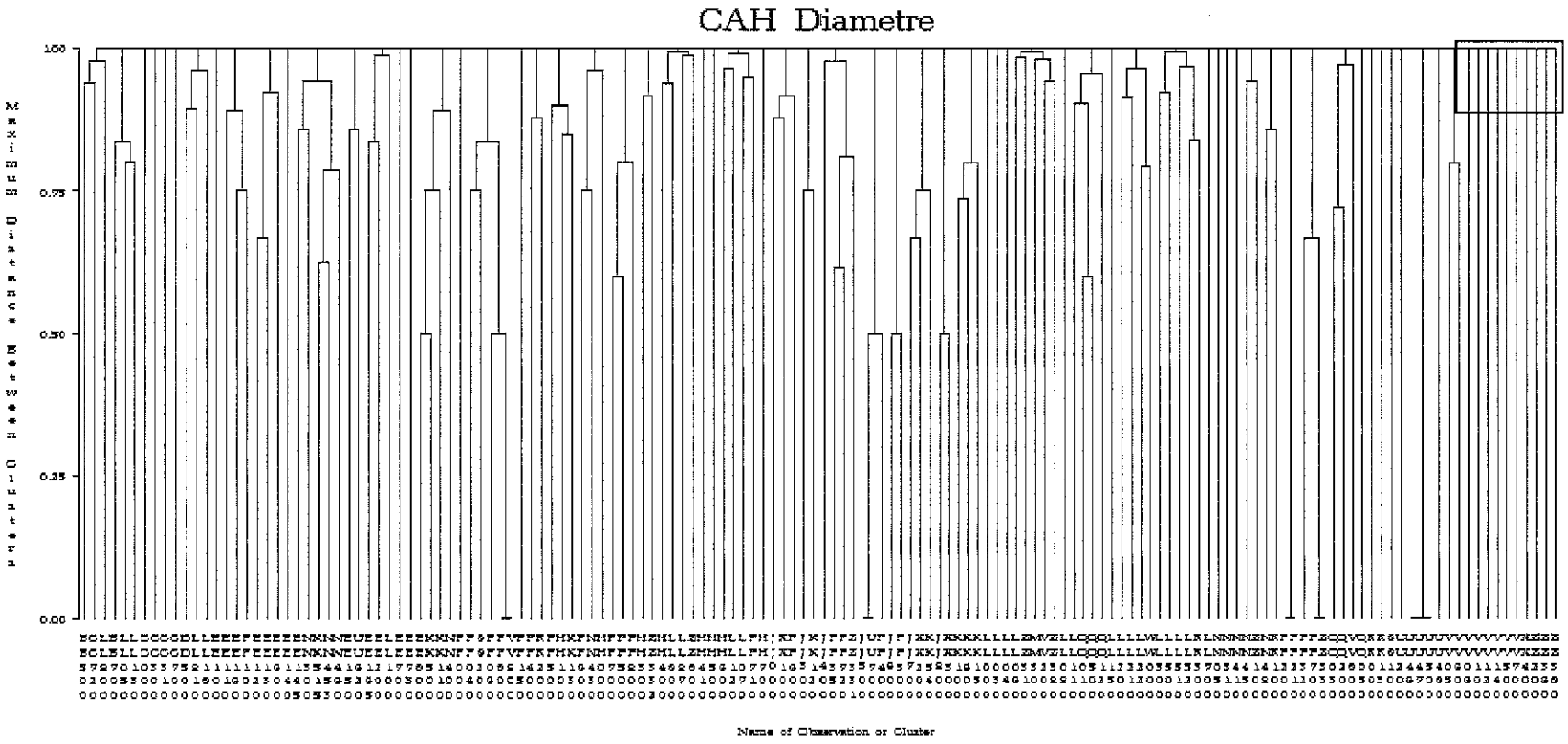
Obs	Nombre	Moyenne	Sigma	Dispersion	Minimum	Maximum	Etendue	Mediane	Mode
1	38	3.94737	16.6830	4.22635	1	104	103	1	1

Comme indiqué au paragraphe a), il y a une très grosse classe de 104 éléments et 29 singletons. Cette classification n'est donc vraiment pas bonne, comme le prouve le coefficient de dispersion[9], qui est de 4,23. Ce critère n'est ainsi pas du tout adéquat pour notre problème.

[9] Le coefficient de dispersion est égal au rapport de l'écart-type sur la moyenne.

2.2.2.2. Critère du diamètre

a) Etude du dendogramme



Nous pouvons constater qu'il y a eu création de plusieurs petites classes qui se sont presque toutes regroupées au niveau 1 avec plusieurs singletons (encadré). On pourrait penser que cette classification est meilleure que celle obtenue avec le critère du saut minimum, mais le découpage en 38 classes infirme cette hypothèse.

b) Etude de la taille des classes à un niveau donné

CAH Diametre

Obs	cluster	frequence
1	CL110	2
2	CL118	2
3	CL121	3
4	CL133	3
5	CL147	3
6	CL149	2
7	CL38	87
8	CL74	5
9	CL75	5
10	CL84	3
11	CL87	4
12	CL90	4
13	CL93	2
14	LL000	1
15	LL030	1
16	LL040	1
17	LL080	1
18	LL150	1
19	LL700	1
20	NN050	1
21	NN310	1
22	NN410	1
23	QQ050	1
24	RR000	1
25	RR130	1
26	SS100	1
27	UU200	1
28	UU490	1
29	VV060	1
30	VV100	1
31	VV120	1
32	VV140	1
33	VV500	1
34	VV700	1
35	XX400	1
36	ZZ200	1
37	ZZ360	1
38	ZZ390	1

The UNIVARIATE Procedure
Variable: frequence

Frequency Counts

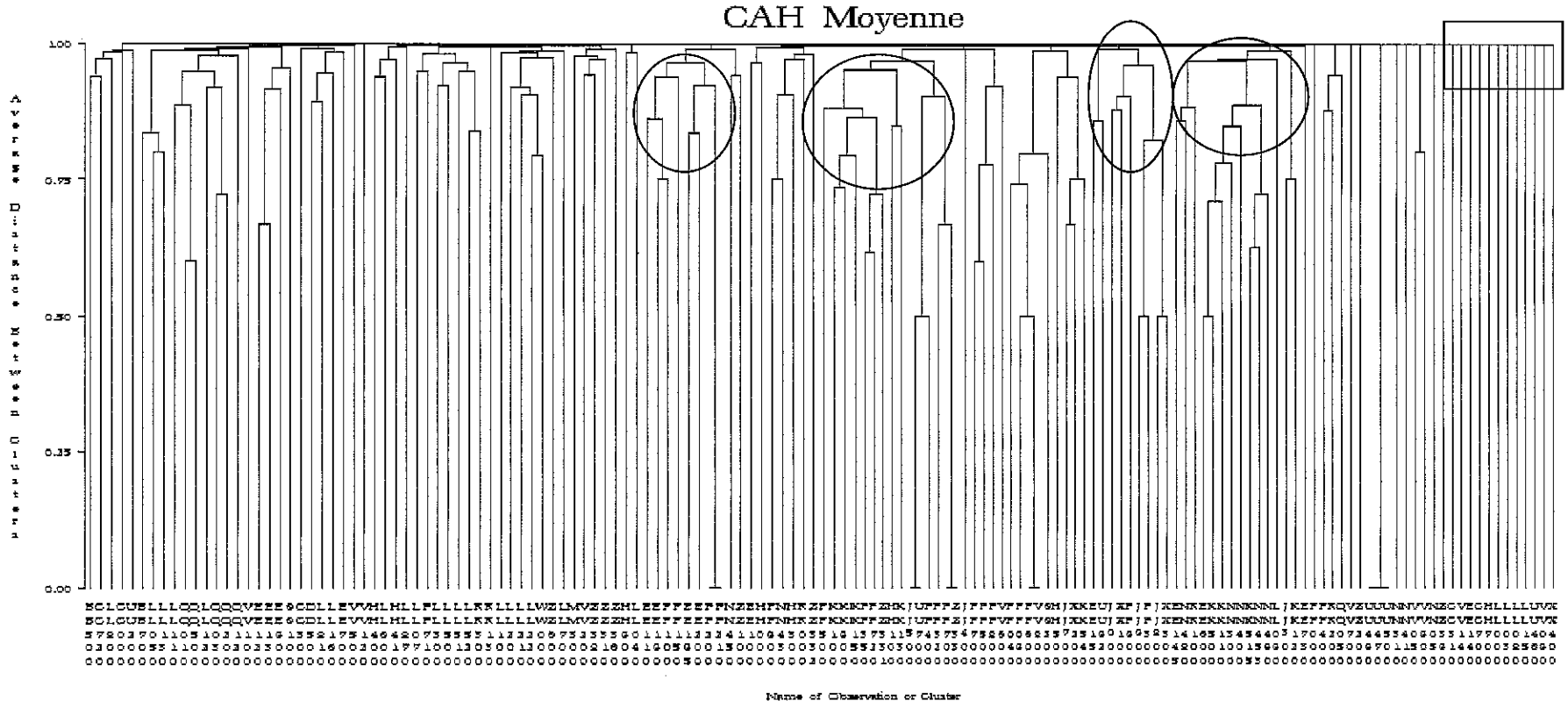
Percents Value	Count	Percents Value	Count	Percents Value	Count	Percents Value	Count
1	25	3	4	5	2	87	1
2	4	4	2				

Obs	Nombre	Moyenne	Sigma	Dispersion	Minimum	Maximum	Etendue	Mediane	Mode
1	38	3.94737	13.8875	3.51816	1	87	86	1	1

Nous constatons qu'il y a une très grosse classe de 87 éléments, 12 classes de 2 à 5 éléments et 25 singletons. Cette coupure donne donc des résultats un peu plus satisfaisants, du point de vue de la taille des classes, que ceux obtenus par le critère du saut minimum. Toutefois, ces résultats ne sont pas acceptables pour le problème puisqu'il y a une classe « mange-tout » et plusieurs singletons (le coefficient de dispersion est de 3,52, ce qui n'est pas très bon).

2.2.2.3. Critère de la distance moyenne

a) Etude du dendogramme



Si l'aspect général du dendogramme ressemble à celui obtenu avec le critère du diamètre (plusieurs petites classes agrégées au niveau 1), on peut voir quelques classes bien distinctes (ellipses). Nous pouvons aussi remarquer que les mêmes codes sont agrégés à la fin, comme lors des classifications obtenues par les critères précédents. Ces codes caractérisent des isolats (*) (encadré).

b) Etude de la taille des classes à un niveau donné

CAH Moyenne

Obs	cluster	frequence
1	CC310	1
2	CC700	1
3	CL115	2
4	CL147	3
5	CL38	15
6	CL39	4
7	CL40	15
8	CL41	7
9	CL43	9
10	CL44	5
11	CL45	5
12	CL46	10
13	CL48	20
14	CL51	9
15	CL52	5
16	CL53	14
17	CL57	2
18	CL83	3
19	EE140	1
20	EE730	1
21	FF000	1
22	HH700	1
23	LL000	1
24	LL030	1
25	LL080	1
26	LL150	1
27	NN050	1
28	NN310	1
29	NN410	1
30	UU490	1
31	VV060	1
32	VV120	1
33	VV140	1
34	VV500	1
35	VV700	1
36	XX400	1
37	ZZ200	1
38	ZZ360	1

The UNIVARIATE Procedure
Variable: frequence

Frequency Counts

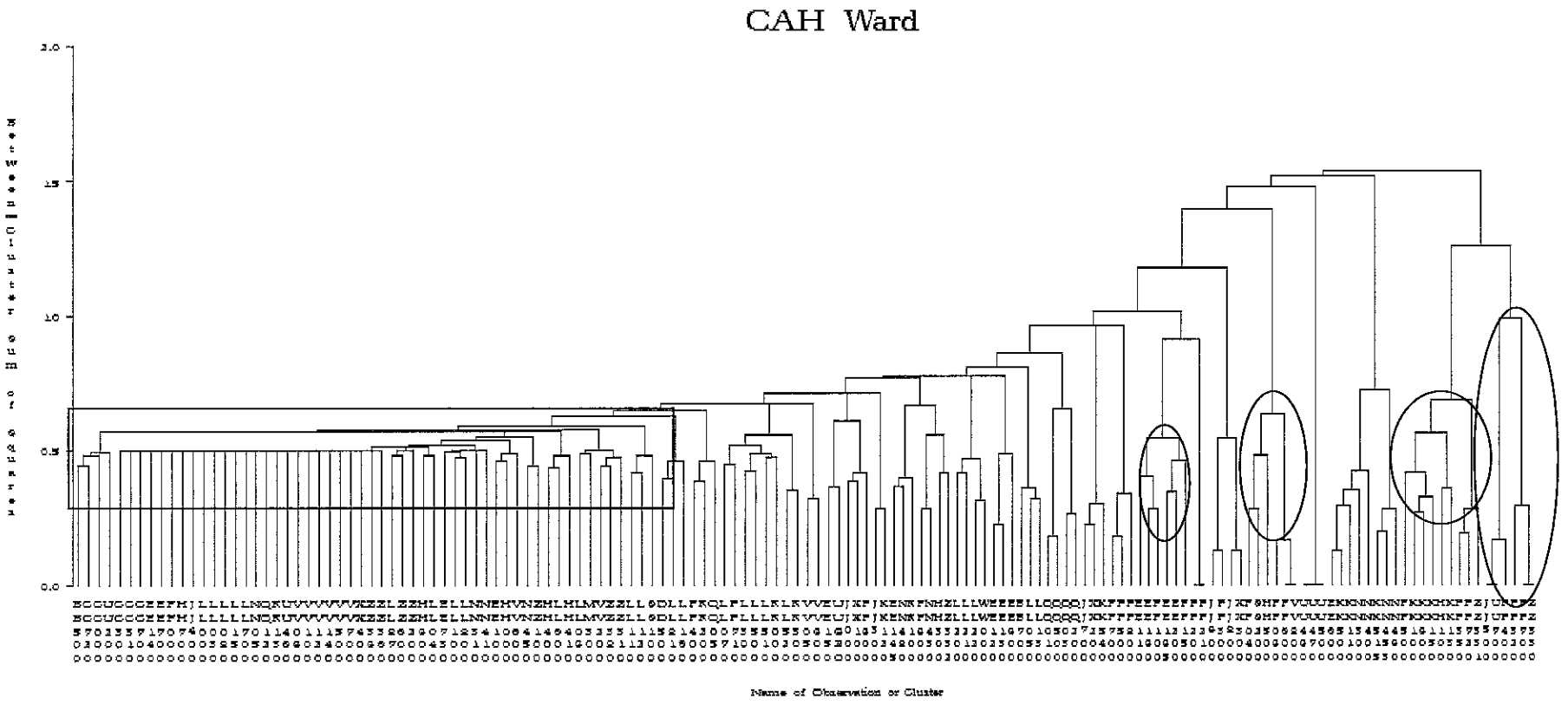
Percents Value	Count	Percents Value	Count	Percents Value	Count	Percents Value	Count
1	22	4	1	9	2	15	2
2	2	5	3	10	1	20	1
3	2	7	1	14	1		

Obs	Nombre	Moyenne	Sigma	Dispersion	Minimum	Maximum	Etendue	Mediane	Mode
1	38	3.94737	4.94263	1.25213	1	20	19	1	1

La dispersion de la taille des classes est beaucoup plus homogène puisqu'il y a des classes à 2, 3, 4, 5, 7, 9, 10, 14, 15 et 20 éléments, le coefficient de dispersion étant de 1,25. La présence de 22 singletons, soit plus d'une classe sur deux, montre que certains thèmes sont soit des isolats soit des thèmes rares apparaissant avec des thèmes fréquents. En effet, comme ils sont agrégés à la fin, ils ont des indices de co-classement très faibles (ne pas oublier que l'on travaille sur la matrice complémentaire).

2.2.2.4. Critère de Ward

a) Etude du dendogramme



Ici, il apparaît très nettement la constitution de classes distinctes (ellipses). On remarque également qu'il n'y a pas de singletons; ceci est peut-être dû au fait que la distance utilisée n'est pas la même que pour les trois autres critères, puisqu'on minimise une perte d'inertie (pour les autres critères, on agrège les éléments les plus proches selon une certaine distance). Par contre, on voit qu'il se forme une classe importante (encadré).

b) Etude de la taille des classes à un niveau donné

CAH Ward

Obs	cluster	frequence
1	CL100	2
2	CL101	4
3	CL102	3
4	CL106	3
5	CL110	2
6	CL112	2
7	CL113	3
8	CL114	2
9	CL117	3
10	CL119	2
11	CL122	3
12	CL123	3
13	CL125	3
14	CL126	3
15	CL128	2
16	CL129	2
17	CL132	2
18	CL137	2
19	CL139	3
20	CL140	3
21	CL141	2
22	CL142	2
23	CL147	3
24	CL149	2
25	CL38	6
26	CL39	39
27	CL41	6
28	CL44	5
29	CL74	4
30	CL78	3
31	CL80	3
32	CL83	3
33	CL84	3
34	CL88	4
35	CL91	3
36	CL92	3
37	CL95	2
38	CL99	5

The UNIVARIATE Procedure
Variable: frequence

Frequency Counts

Percents Value	Count	Percents Value	Count	Percents Value	Count	Percents Value	Count
2	13	4	3	6	2	39	1
3	17	5	2				

Obs	Nombre	Moyenne	Sigma	Dispersion	Minimum	Maximum	Etendue	Mediane	Mode
1	38	3.94737	5.93636	1.50388	2	39	37	3	3

Comme il était dit précédemment, il n'y a aucun singleton. Par contre, trente classes ont au plus 3 éléments et une classe en a 39. Il y a donc une sorte d'effet de chaîne que l'on ne devrait a priori ne pas trouver. Cependant le coefficient de dispersion étant assez bon (1,50), il y a une bonne homogénéité des tailles des classes.

2.2.2.5. Conclusion sur les classifications

Après étude des résultats obtenus (la signification des classes sera étudiée avec les résultats obtenus par le « quick clustering »), on retrouve bien les caractéristiques de ces différents critères : effet de chaîne pour le saut minimum, classes de même taille pour la distance moyenne et le critère de Ward. Il semblerait que l'utilisation de ces deux derniers critères donne de bons résultats.

2.3. La méthode de « quick clustering »

Cette méthode n'est pas une méthode de classification hiérarchique, mais une méthode de partitionnement d'un espace. Cette méthode a été introduite par [Kamen 1970], à partir des travaux de [McQuitty 1964].

2.3.1. Présentation de la méthode

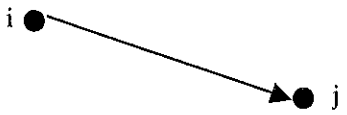
La méthode de « quick clustering » réunit dans une même composante les plus proches voisins successifs du couple de voisins réciproques qui constitue le noyau de départ. Nous allons définir ces deux notions au paragraphe suivant.

2.3.1.1. Algorithme

Afin de mieux comprendre comment fonctionne cette méthode, définissons un graphe G=(X,U) où X est l'ensemble des éléments pris en compte, autrement dit le corpus, et U l'ensemble des arcs reliant ces éléments, arcs valués par l'indice de Jaccard. Notons cet indice d.

Soient i et j deux éléments de X. On dit que j est le plus proche voisin de i (qu'on peut noter $j = ppv(i)$) si et seulement si $d(i,j) = \max_{j \in X - \{i\}} d(i,k)$. On dit que i et j sont voisins réciproques si et seulement si $j = ppv(i)$ et $i = ppv(j)$.

Dans toutes les représentations graphiques qui vont suivre le schéma



signifie que j est le plus proche voisin de i.

L'algorithme du « quick clustering » peut se décomposer en deux étapes :

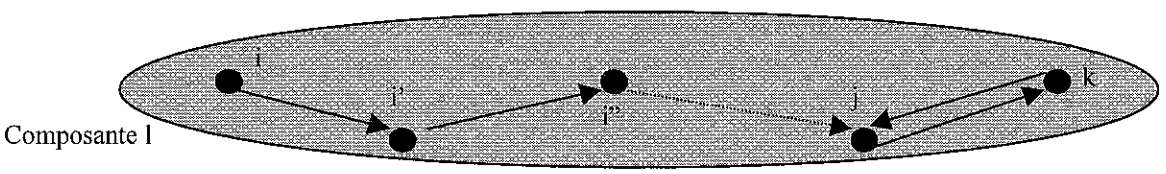
① Il détermine le plus proche voisin de chaque élément i de X. En cas d'ex-æquo, il prend le premier élément par ordre alphabétique. Cette sélection, qui est un tri, est possible en $O(n\log n)$, où n est le nombre d'éléments [10]. Nous avons ainsi un sous-graphe $G'=(X',U')$ de G tel que $X'=X$ et $|U'|\leq n$ ($|U'|$ représente le nombre d'arcs). En effet, un point du graphe G' peut ne pas avoir de plus proche voisin (c'est un isolat (*)); dans le cas contraire, par construction (les ex-æquos sont éliminés), il n'a qu'un seul successeur, son plus proche voisin. Donc $\forall i\in X, d^+(i)\leq 1$ ($d^+(i)$ représente le nombre d'arcs incident extérieurement à i) et $|U'|=\sum_{i\in X} d^+(i)\leq \sum_{i\in X} 1\leq n$.

② Sur ce graphe G', l'algorithme recherche les composantes connexes, de la façon suivante :

~ on part d'un point qui n'a pas encore été affecté à une composante (notons ce point i)

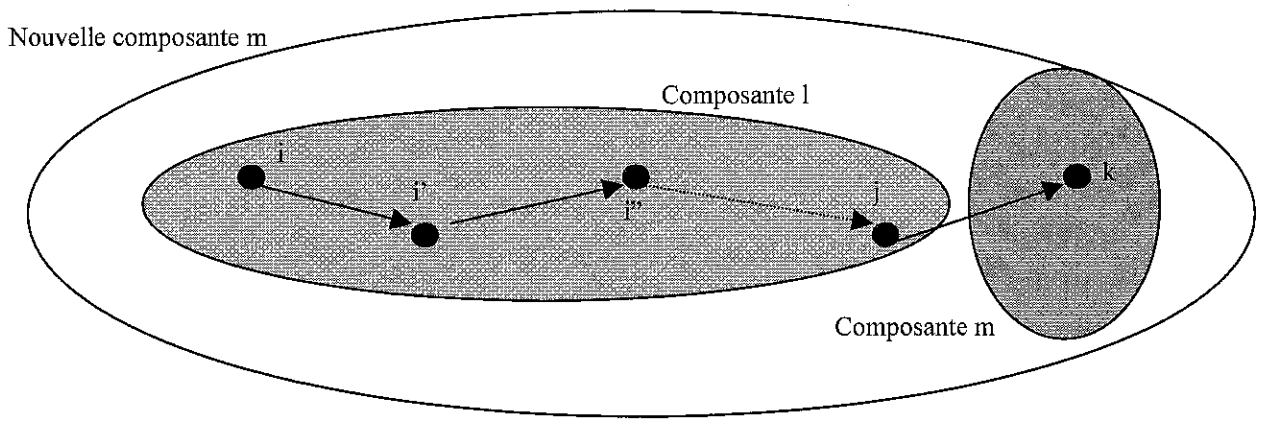
~ on constitue le chemin des plus proches voisins partant de ce point i en l'initialisant à un numéro de composante l jusqu'à ce que :

Soit les deux derniers points j et k de la chaîne sont voisins réciproques; on a alors détecté une nouvelle composante.



[10] Nous reviendrons un peu plus tard sur ce problème des ex-æquos, au § 2.3.2.1.

Soit le dernier point k du chemin a déjà été affecté à une composante m; on remonte alors la chaîne pour réaffecter chaque point à cette composante m.



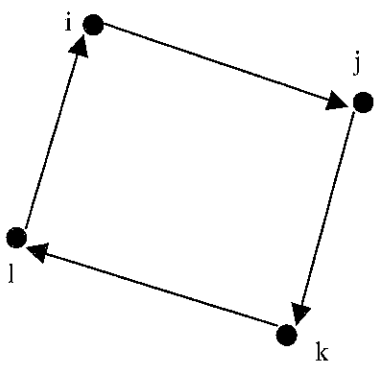
~ on s'arrête quand tous les points ont été traités.

Comme il y a n arcs, et que chaque arc est exploré au plus deux fois, cette deuxième partie de l'algorithme est de complexité au pire cas en $O(n)$. L'algorithme est donc en $O(n \log n)$, il est plus rapide que les algorithmes de classifications, qui sont de l'ordre de $O(n^2)$.

2.3.1.2. Quelques remarques sur cette méthode

Précisons pour commencer que, lors de la sélection du plus proche voisin, les éléments sont triés par ordre alphabétique.

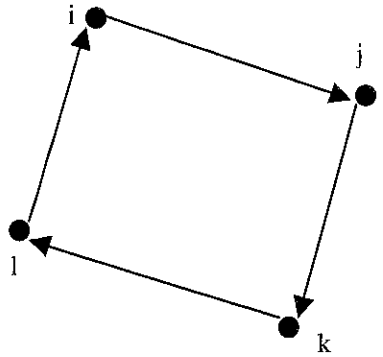
D'abord, il ne peut pas exister de circuit de longueur ≥ 3. Pour le démontrer, raisonnons par l'absurde. Soit un circuit de longueur ≥ 3, où, au moins la valeur d'un arc est différente des autres valeurs des arcs du circuit : prenons par exemple un chemin de longueur 4 en notant les points i, j, k, l.



j est le plus proche voisin de i, k est le plus proche voisin de j, l est le plus proche voisin de k et i est le plus proche voisin de l. Comme la valuation des arcs correspond à un indice de similarité, $\forall (i,j) \in X, d(i,j) = d(j,i)$. Supposons de plus que $d(k,l) > d(i,j)$ donc on a, par définition des plus proches voisins l'inégalité suivante $d(i,j) < d(l,i)$.

Cette inégalité montre que le plus proche voisin de i n'est plus j, mais l, ce qui est impossible par hypothèse.

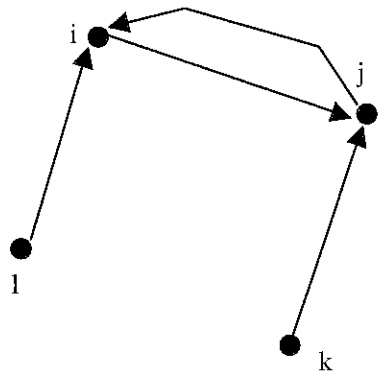
Etudions maintenant le cas d'un circuit de longueur ≥ 3, où toutes les valeurs des arcs sont identiques. Reprenons notre exemple des quatre points i, j, k, l.



On a donc ici $d(i,j)=d(j,k)=d(k,l)=d(l,i)$. Comme d est un indice de similarité, on a également $d(i,j)=d(j,i)=d(k,j)=d(l,k)=d(i,l)$. Cela signifie que nous avons le tableau des plus proches voisins suivant (la colonne de gauche indique le point, la colonne de droite son (ou ses) plus proche(s) voisin(s)) :

i	**j**
i	l
j	**i**
j	k
k	**j**
k	l
l	**i**
l	k

Or, comme on ne prend que le premier plus proche voisin par ordre alphabétique, les lignes sélectionnées dans l'algorithme correspondent à celles en gras dans le tableau, ce qui donne la représentation suivante :



Il ne peut donc pas exister de circuit de longueur ≥ 3.

De plus, par rapport à une CAH par le critère du saut minimum, cette méthode est moins susceptible de créer un effet de chaîne, puisque, une fois arrivée sur une paire de voisins réciproques, elle ne passe pas au subvoisin.

Signalons enfin que cette méthode est stable, en ce sens qu'il n'y a pas de coupure de composantes à un niveau donné. En clair, on construit les composantes à un niveau (ou seuil) S_1. Si on construit les composantes à un niveau $S_2 < S_1$, en plus des liens déjà présents au seuil S_1 apparaissent des liens moins forts en terme de valeur de l'indice de similarité utilisé. Les éléments associés à ces nouveaux liens, soit se rattachent aux composantes créées au niveau S_1, soit occasionnent la création de nouvelles composantes.

2.3.2. Application de la méthode au problème de classification des thèmes

2.3.2.1. Problème des ex-æquos

Lors de l'exécution de la première étape de l'algorithme, quelques codes avaient plusieurs plus proches voisins. On a alors testé manuellement quatre variantes de l'algorithme, notées A, B, C et D. Pour les deux premières, tous les plus proches voisins ont été pris en compte, alors que pour les variantes C et D, nous n'avons considéré que le premier plus proche voisin par ordre alphabétique. Ensuite, pour constituer les composantes, soit on le faisait dans un ordre quelconque (variantes B et D), soit on le faisait par ordre alphabétique (variantes A et C). Nous en avons tiré les conclusions suivantes :

① 9 isolats (*) ont été détectés. En effet, sur les 150 codes de départ, seuls 141 ont été affectés à une composante. Les neuf autres, puisqu'ils n'ont pas été pris en compte, n'avaient pas de plus proche voisin. Par définition, ces neuf codes sont des isolats.

② Les résultats étaient stables pour la conservation du seul premier plus proche voisin (ppv) (constitution des mêmes 29 composantes).

③ Pour la conservation de tous les ppv, sur les 28 composantes trouvées, 2 ont été modifiées lors de l'application de l'option B (un code est passé de la composante n°2 à la composante n°4).

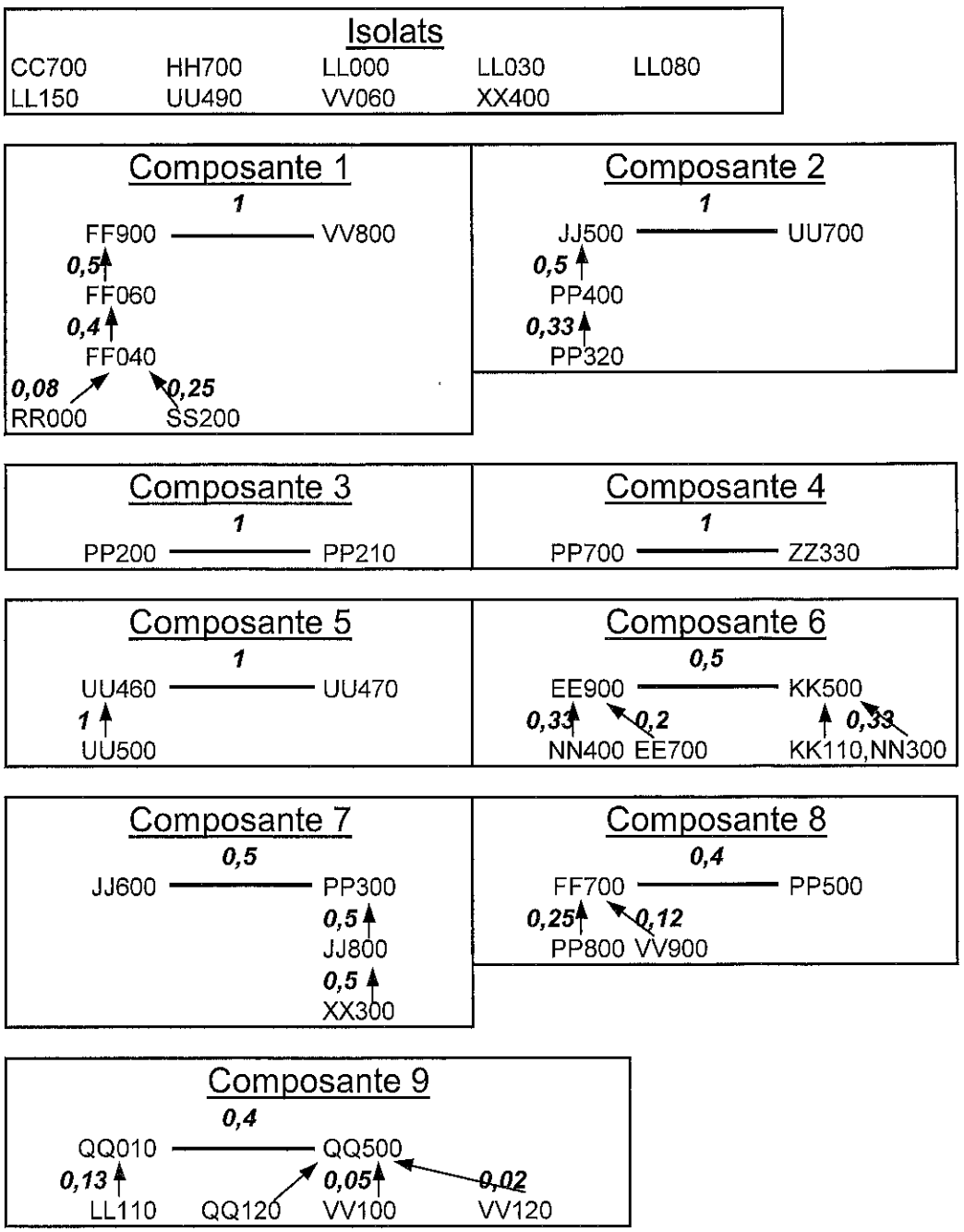
④ Avec les options C et D, ce code ne s'est pas déplacé, par contre, la composante {*JJ600, PP300, JJ800, XX300,* **XX100, JJ000, PP600**} s'est scindée en deux composantes, l'une contenant les codes en italique, l'autre les codes en gras.

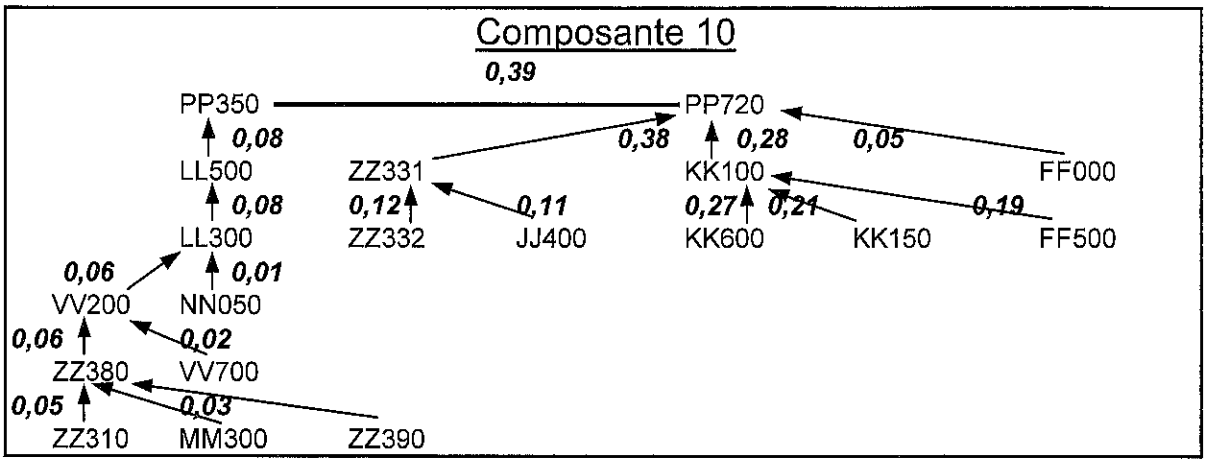
Nous pouvons donc dire que, pour des questions de stabilité des résultats et de rapidité des calculs, il faut appliquer la variante C : sélection du premier ppv par ordre alphabétique et détermination des composantes par ordre alphabétique.

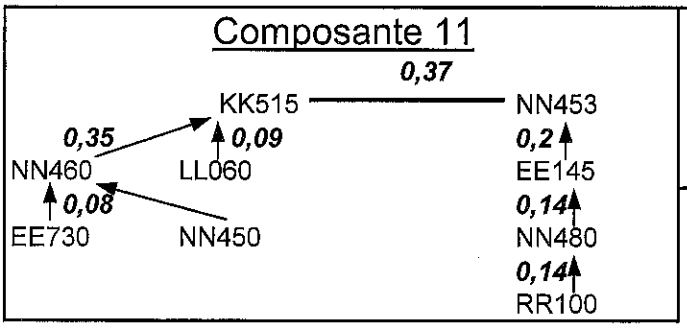
2.3.2.2. Résultat de la méthode

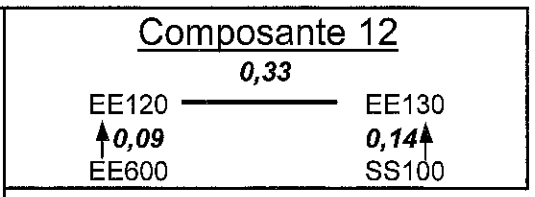
a) Composition des composantes

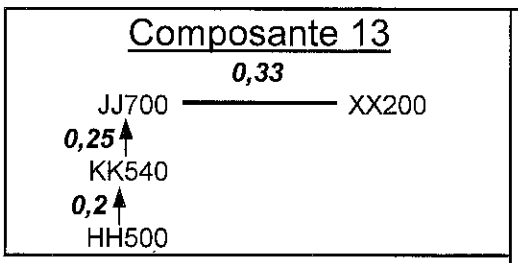
Voici la représentation des 29 composantes, ainsi que les 9 isolats détectés.

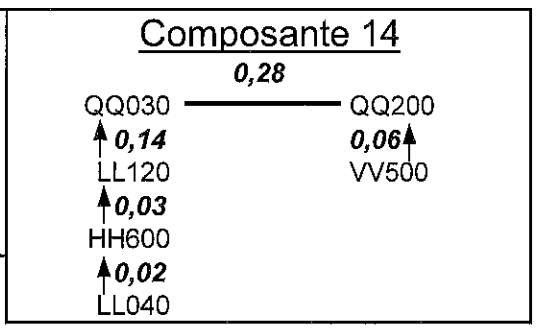


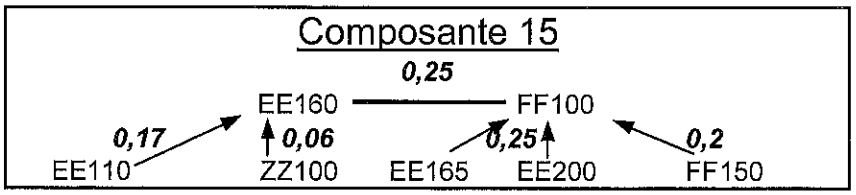

Composante 10
0,39
PP350
PP720
0,08
0,38
0,28
0,05
LL500
ZZ331
KK100
FF000
0,08
0,12
0,11
0,27
0,21
0,19
LL300
ZZ332
JJ400
KK600
KK150
FF500
0,06
0,01
VV200
NN050
0,06
0,02
ZZ380
VV700
0,05
0,03
ZZ310
MM300
ZZ390

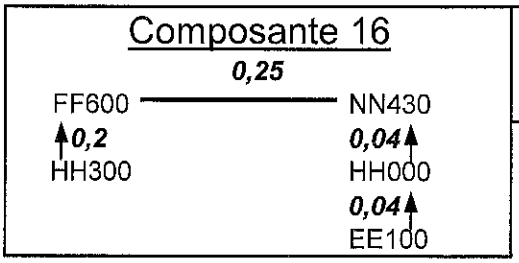

Composante 11
0,37
KK515
NN453
0,35
0,09
0,2
NN460
LL060
EE145
0,08
0,14
EE730
NN450
NN480
0,14
RR100

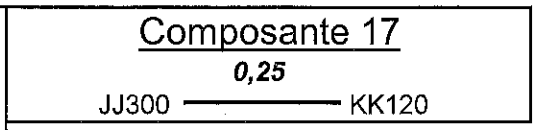

Composante 12
0,33
EE120
EE130
0,09
0,14
EE600
SS100

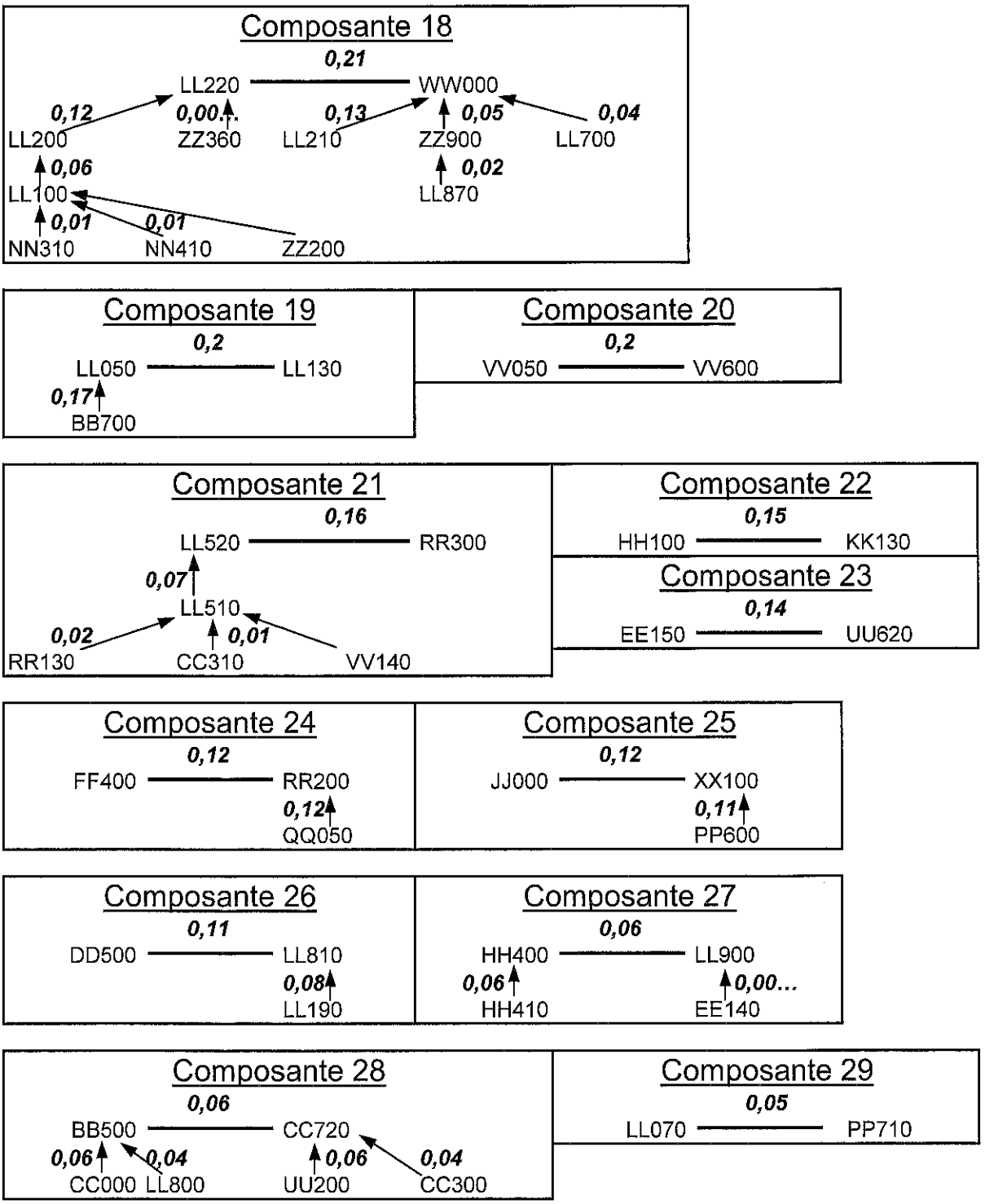

Composante 13
0,33
JJ700
XX200
0,25
KK540
0,2
HH500


Composante 14
0,28
QQ030
QQ200
0,14
0,06
LL120
VV500
0,03
HH600
0,02
LL040


Composante 15
0,25
EE160
FF100
0,17
0,06
0,25
0,2
EE110
ZZ100
EE165
EE200
FF150


Composante 16
0,25
FF600
NN430
0,2
0,04
HH300
HH000
0,04
EE100


Composante 17
0,25
JJ300
KK120



Les traits en gras indiquent les couples de voisins réciproques, les valeurs sont celles prises par l'indice de Jaccard et les composantes sont triées par ordre décroissant de la valeur de l'indice du noyau.

b) Signification des différentes composantes

Composante 1	Plantes terrestres
Composante 2	Sol, érosion ***(marginal)***
Composante 3	Ressources en eau
Composante 4	Ressources biologiques Ecologie générale
Composante 5	Sociologie (communautés et femmes)
Composante 6	Emploi, travail, marchés en sylviculture
Composante 7	Ressources naturelles et pollution
Composante 8	Pathologies végétales et phénomènes climatiques
Composante 9	Qualité alimentaire Nutrition humaine Produits laitiers
Composante 10	Ressources biologiques végétales (divers usages)
Composante 11	Machinisme agricole et agroalimentaire
Composante 12	Commerce des produits animaux
Composante 13	Agrochimie et déchets
Composante 14	Elevage à viande Intoxications alimentaires
Composante 15	Aspects économiques de la production végétale
Composante 16	Pathologie végétale
Composante 17	Gestion des forêts Physique des sols ***(marginal)***
Composante 18	Biotechnologie et génétique animale
Composante 19	Paléontologie Gibiers Œufs ***(marginal)***
Composante 20	Santé humaine (physiologie)
Composante 21	Alimentation animale
Composante 22	Incendies de forêts Contrôle des agents pathogènes ***(marginal)***
Composante 23	Economie environnementale Sport
Composante 24	Plantes cultivées Déchets
Composante 25	Déchets animaux Pollution
Composante 26	Règlements Santé animale Abattage
Composante 27	Toxicologie animale Médicaments

Voici également le tableau des isolats :

CC700	*Professions, education, information and training (general) : professions, practice and service*	*3*
HH700	*Pathogen, pest and parasite management (general) : other control measures*	*1*
LL000	*Animal science*	*1*
LL030	Animal science : culture of other invertebrates (not aquaculture)	2
LL080	Animal science : zoo animals	9
LL150	*Animal science : animal husbandry (general) : animal husbandry (other products)*	*1*
UU490	*Sociology (general) : community development : social psychology and culture*	*1*
VV060	Human health and hygiene (general) : human reproduction and development	1
XX400	Wastes (general) : industrial wastes and effluents	1

La dernière colonne indique la fréquence d'occurrence de chaque code. En étudiant le fichier de départ, il s'avère que cinq codes (ceux en italique) sont les seuls abordés par des articles (ce sont des isolats parfaits) et que les quatre autres codes sont mentionnés avec au moins un des quatre codes les plus fréquents éliminés [11] (ce sont des isolats par construction).

c) Etude de la taille des composantes

Quick clustering

Obs	compo	frequence
1	1	6
2	2	4
3	3	2
4	4	2
5	5	3
6	6	6
7	7	4
8	8	4
9	9	6
10	10	19
11	11	9
12	12	4
13	13	4
14	14	6
15	15	7
16	16	5
17	17	2
18	18	12
19	19	3
20	20	2
21	21	6
22	22	2
23	23	2
24	24	3
25	25	3
26	26	3
27	27	4
28	28	6
29	29	2
30	30	1
31	31	1
32	32	1
33	33	1
34	34	1
35	35	1
36	36	1
37	37	1
38	38	1

[11] Pour rappel, les intitulés précis des quatre codes éliminés sont :
Animal Science: Animal Physiology and Biochemistry (excluding Nutrition)
Animal Science: Animal Health and Hygiene (General): Parasites, Vectors, Pathogens and Biogenic Diseases of Animals
Animal Science: Animal Health and Hygiene (General): Animal Disorders (not caused by Organisms)
Animal Science: Animal Health and Hygiene (General): Animal Treatment and Diagnosis (non Drug).

The UNIVARIATE Procedure
Variable: frequence

Frequency Counts

Percents Value	Count	Percents Value	Count	Percents Value	Count	Percents Value	Count
1	9	4	6	7	1	12	1
2	7	5	1	9	1	19	1
3	5	6	6				

Obs	Nombre	Moyenne	Sigma	Dispersion	Minimum	Maximum	Etendue	Mediane	Mode
1	38	3.94737	3.53322	0.89508	1	19	18	3	1

La méthode de « quick clustering » a construit des composantes sphériques et la fonction des tailles de ces composantes est peu étendue (coefficient de dispersion de 0,90). Toutefois, nous remarquons que 18 composantes ont entre 2 et 4 codes, mais cela s'explique très bien par le fait qu'il y avait peu d'éléments à classer (150, en comptant les isolats). Cette méthode semble bonne d'un point de vue statistique (taille des composantes) et pratique (signification des composantes).

2.4. Comparaison des différentes méthodes

Nous allons voir, pour chaque classification, quelle(s) classe(s) a (ont) le plus grand nombre d'éléments de chaque composante. Ensuite nous verrons quelles sont les classes qui ont « détecté » une composante, en calculant un indice d'équivalence.

2.4.1. Détection des composantes par le nombre d'éléments en commun

2.4.1.1. Critère du saut minimum

Le tableau suivant indique, pour chaque composante, dans quelle classe se trouvent la plus grande partie de ses éléments (le nombre d'éléments est indiqué entre parenthèses).

Numéro de la composante	*Numéro de la classe*
Composante 1	S01 (6)
Composante 2	S01 (4)
Composante 3	S01 (2)
Composante 4	S01 (2)
Composante 5	S02 (3)
Composante 6	S01 (6)
Composante 7	S01 (4)
Composante 8	S01 (4)
Composante 9	S01 (5)
Composante 10	S01 (14)
Composante 11	S01 (9)
Composante 12	S01 (4)
Composante 13	S01 (4)
Composante 14	S01 (4)
Composante 15	S01 (7)
Composante 16	S01 (3)
Composante 17	S01 (2)
Composante 18	S01 (5)
Composante 19	S01 (3)
Composante 20	S03 (2)
Composante 21	S01 (3)
Composante 22	S01 (2)
Composante 23	S01 (2)
Composante 24	S01 (3)
Composante 25	S01 (3)
Composante 26	S01 (3)
Composante 27	S04 (3)
Composante 28	S05 (4)
Composante 29	S06 (2)

Le tableau suivant indique, pour chaque classe citée, le nombre de composantes détectées :

Numéro de la classe	*Nombre de composantes détectées*
S01	24
S02	1
S03	1
S04	1
S05	1
S06	1

Nous voyons donc bien que cette classification n'est vraiment pas satisfaisante, puisque la classe 1 qui englobe la majorité des éléments (104), détecte également la majorité des composantes.

2.4.1.2. Critère du diamètre

Le tableau suivant indique, pour chaque composante, dans quelle(s) classe(s) se trouvent la plus grande partie de ses éléments (le nombre d'éléments est indiqué entre parenthèses).

Numéro de la composante	*Numéro de la classe*
Composante 1	D03 (5)
Composante 2	D03 (3)
Composante 3	D01 (2)
Composante 4	D04 (2)
Composante 5	D02 (3)
Composante 6	D03 (6)
Composante 7	D03 (4)
Composante 8	D03 (3)
Composante 9	D05 (4)
Composante 10	D03 (7)
Composante 11	D03 (5)
Composante 12	D03 (3)
Composante 13	D03 (4)
Composante 14	D03 (2), D06 (2)
Composante 15	D03 (6)
Composante 16	D03 (5)
Composante 17	D03 (2)
Composante 18	D08 (4)
Composante 19	D03 (3)
Composante 20	D09 (2)
Composante 21	D10 (3)
Composante 22	D03 (2)
Composante 23	D03 (2)
Composante 24	D03 (2)
Composante 25	D03 (3)
Composante 26	D03 (3)
Composante 27	D03 (4)
Composante 28	D03 (5)
Composante 29	D03 (2)

Le tableau suivant indique, pour chaque classe citée, le nombre de composantes détectées :

Numéro de la classe	*Nombre de composantes détectées*
D01	1
D02	1
D03	22
D04	1
D05	1
D06	1
D08	1
D09	1
D10	1

Nous voyons ainsi que cette classification est également peu satisfaisante, puisque la classe 3 englobe la majorité des éléments (87), donc la majorité des composantes.

2.4.1.3. Critère de la distance moyenne

Le tableau suivant indique, pour chaque composante, dans quelle(s) classe(s) se trouvent la plus grande partie de ses éléments (le nombre d'éléments est indiqué entre parenthèses).

Numéro de la composante	*Numéro de la classe*
Composante 1	M04 (5)
Composante 2	M03 (4)
Composante 3	M01 (2)
Composante 4	M03 (2)
Composante 5	M02 (3)
Composante 6	M05 (5)
Composante 7	M06 (4)
Composante 8	M03 (4)
Composante 9	M07 (5)
Composante 10	M03 (8)
Composante 11	M05 (7)
Composante 12	M07 (4)
Composante 13	M04 (4)
Composante 14	M07 (3)
Composante 15	M01 (7)
Composante 16	M08 (5)
Composante 17	M05 (2)
Composante 18	M09 (7)
Composante 19	M07 (3)
Composante 20	M10 (2)
Composante 21	M09 (4)
Composante 22	M03 (2)
Composante 23	M06 (2)
Composante 24	M11 (3)
Composante 25	M06 (3)
Composante 26	M12 (3)
Composante 27	M13 (3)
Composante 28	M14 (5)
Composante 29	M09 (2)

Le tableau suivant indique, pour chaque classe citée, le nombre de composantes détectées :

Numéro de la classe	*Nombre de composantes détectées*
M01	2
M02	1
M03	5
M04	2
M05	3
M06	3
M07	4
M08	1
M09	3
M10	1
M11	1
M12	1
M13	1
M14	1

14 classes sont prises en compte, soit une moyenne d'une classe pour deux composantes. Cette classification semble donc meilleure que les précédentes, sans être pour autant la panacée, puisque 2 classes ont détecté 9 composantes (M03 et M07).

2.4.1.4. Critère de Ward

Le tableau suivant indique, pour chaque composante, dans quelle(s) classe(s) se trouvent la plus grande partie de ses éléments (le nombre d'éléments est indiqué entre parenthèses).

Numéro de la composante	*Numéro de la classe*
Composante 1	W04 (3)
Composante 2	W03 (3)
Composante 3	W01 (2)
Composante 4	W05 (2)
Composante 5	W02 (3)
Composante 6	W06 (5)
Composante 7	W07 (2), W08 (2)
Composante 8	W09 (3)
Composante 9	W38 (3)
Composante 10	W38 (5)
Composante 11	W12 (3), W26 (3)
Composante 12	W13 (3)
Composante 13	W14 (3)
Composante 14	W38 (3)
Composante 15	W17 (6)
Composante 16	W19 (2), W38 (2)
Composante 17	W18 (2)
Composante 18	W38 (8)
Composante 19	W22 (3)
Composante 20	W23 (2)
Composante 21	W38 (3)
Composante 22	W25 (2)
Composante 23	W27 (2)
Composante 24	W28 (3)
Composante 25	W29 (3)
Composante 26	W30 (3)
Composante 27	W34 (3)
Composante 28	W35 (4)
Composante 29	W33 (2)

Le tableau suivant indique, pour chaque classe citée, le nombre de composantes détectées :

Numéro de la classe	*Nombre de composantes détectées*
W01	1
W02	1
W03	1
W04	1
W05	1
W06	1
W07	1
W08	1
W09	1
W12	1
W13	1
W14	1
W17	1
W18	1
W19	1
W22	1
W23	1
W25	1
W26	1
W27	1
W28	1
W29	1
W30	1
W33	1
W34	1
W35	1
W38	6

Ici, 27 classes ont détecté les 29 composantes, mais la dernière, qui contient 38 éléments, en a détecté 6 à elle seule. Cette classification détecte donc bien les composantes.

Etudions maintenant ces méthodes avec l'indice d'équivalence.

2.4.2. Détection des composantes par un indice d'équivalence

2.4.2.1. Calcul de l'indice

On calcule le nombre d'éléments par composante i et par classe j (notons respectivement ces nombre n_i et n_j) et le nombre d'éléments présents à la fois dans la composante i et la composante j n_{ij}.

L'indice calculé est le suivant : $Equi(i,j) = \frac{n_{ij}}{n_i}\frac{n_{ij}}{n_j}$ où i est le numéro d'une composante et j celui d'une classe.

Plus cet indice est grand, plus la composante et la classe sont identiques.

2.4.2.2. Résultats

Le tableau suivant indique les composantes détectées par classification, dont l'indice est supérieur ou égal à 0,5. Les nombres entre parenthèses la valeur de l'indice.

Saut minimum	Diamètre	Moyenne	Ward
Q05 (1)	Q03 (1)	Q05 (1)	Q01 (0,5)
Q20 (1)	Q04 (0,67)	Q15 (0,7)	Q02 (0,75)
Q27 (0,75)	Q05 (1)	Q16 (0,71)	Q03 (1)
Q28 (0,67)	Q09 (0,67)	Q20 (1)	Q04 (0,67)
Q29 (1)		Q24 (1)	Q05 (1)
		Q26 (0,6)	Q06 (0,83)
		Q27 (0,56)	Q07 (0,5+0,5)
		Q28 (0,83)	Q08 (0,75)
			Q12 (0,75)
			Q13 (0,75)
			Q15 (0,86)
			Q17 (1)
			Q19 (1)
			Q20 (1)
			Q22 (1)
			Q23 (1)
			Q24 (1)
			Q25 (1)
			Q26 (1)
			Q27 (0,75)
			Q28 (0,66)

Nous remarquons ainsi que seul le critère de Ward fournit de bons résultats, puisqu'il détecte 21 composantes sur 29 dont 10 entièrement.

2.4.3. Remarques sur ces comparaisons

Ces résultats ne sont donnés qu'a titre indicatif, ils ne peuvent être correctement interprétés pour deux raisons principales :

① Il est très difficile de comparer des méthodes de classifications hiérarchiques avec une méthode de partitionnement, puisqu'il y a une notion de niveau d'agrégation dans les premières que l'on ne retrouve pas dans la méthode de « quick clustering ».

② Une coupure dans une CAH en X classes (ici, le nombre de composantes issues de la méthode de « quick clustering ») est fortement arbitraire, puisqu'on n'est pas du tout certain de trouver un "bon" découpage. Par exemple, pour le critère de Ward, les isolats ont été agrégés ensemble en cours de procédure (et non pas à la fin comme pour les trois autres critères) et la coupure en 38 classes s'est faite après cette agrégation. En modulant le nombre de classes, nous aurions peut-être pu retrouver ces isolats.

Conclusion

La méthode de « quick clustering » détecte très bien les isolats, puisqu'ils ne sont pas pris en compte, et donne une partition homogène de l'ensemble des éléments à classer. Elle donne ainsi des résultats intermédiaires, mais de façon beaucoup plus rapide, entre les classifications ascendantes hiérarchiques selon le critère de la distance moyenne et de Ward. On peut ensuite affiner les résultats avec l'une ou l'autre classification, ou les deux.

Prolongements et remarques

Les points abordés ci-après se rapportent à la deuxième partie du rapport sur les comparaisons des méthodes de classification et de partitionnement.

Nous pouvons déjà dire que le travail a été effectué sur un petit échantillon (150 codes) par manque de temps, il serait donc intéressant de l'étendre à une population plus importante. Sur le même fichier, on a par exemple plus de 6000 codes mots-clés. Cependant, comme il y aura plus d'éléments à traiter, les composantes et les classes seront plus importantes, donc plus difficiles à interpréter : il pourrait être intéressant de fixer une taille maximale des classes, donc introduire un seuil, pour limiter cette croissance. Le tout bien sûr est de déterminer la bonne valeur de ce seuil.

De plus on ne s'est intéressé qu'à l'indice de Jaccard, une étude s'appuyant sur l'indice d'Ochiai, très souvent utilisé en scientométrie, permettrait de compléter l'étude et de voir s'il y a des ressemblances et des différences entre les résultats obtenus à partir de ces deux indices. Il serait également intéressant d'utiliser des indices pondérés, puisqu'il y a moins de chance d'avoir des ex-æquos.

Justement, en parlant d'ex-æquo, comme il y a une sélection des plus proches voisins, des composantes ont été créées alors qu'elles auraient pu être rattachées par des liens supprimés. La résolution de ce problème de fusion de composantes pourrait être éventuellement traité.

Résumé

J'ai réalisé mon stage au sein de l'équipe IBIS (Indicateurs Bibliométriques et Scientométrie) de l'UMR EDRA (Unité Mixte de Recherche Economie – Droit Rural et Agroalimentaire) sur le site nantais de l'INRA (Institut National de la Recherche Agronomique). Les chercheurs de cette équipe ont développé avec le logiciel SAS une suite de programmes statistiques pour la réalisation d'études et de recherches bibliométriques.

De nombreuses méthodes d'analyses bibliométriques reposent sur des mesures de proximité et des classifications appliquées à des éléments de notices bibliographiques ou d'autres textes. J'ai été associé à trois applications : une étude sur la transposition des méthodes citationnistes aux pages Web, une étude de structuration de thèmes et l'étude et l'implémentation d'une méthode de classification rapide.

La première étude m'a permis de me familiariser avec les différentes notions de base et les processus typiquement mis en place pour une famille d'études bibliométriques : le calcul d'indicateurs sur thèmes structurés. L'objectif de l'étude était d'identifier des groupes d'adresses URL centrées sur une même thématique, en utilisant l'analogie entre les hyperliens du Web et les citations de la bibliométrie. Plus particulièrement, j'ai utilisé la méthode bibliométrique dite des couplages bibliographiques pour dédoublonner des adresses URL (« pages miroir ») sur le critère d'identité de leurs hyperliens. Après ce nettoyage du fichier, j'ai appliqué une méthode de classification ascendante hiérarchique (la distance moyenne) pour structurer les adresses URL citées.

La seconde étude visait à l'analyse thématique d'un corpus par la méthode classique des co-classements qui, comme celle des co-citations, des couplages bibliographiques ou des mots associés, est fondée sur un calcul de proximité suivi par un algorithme de structuration. La classification a été ici effectuée par quatre méthodes (le saut minimum, le diamètre, la distance moyenne, le critère de Ward) dont j'ai comparé les résultats. Il s'est avéré que les méthodes de Ward et de la distance moyenne donnaient une structuration homogène des éléments, alors que les critères du saut minimum et du diamètre occasionnaient de grandes disparités dans la taille des groupes.

Enfin le stage comportait la mise en œuvre informatique d'un algorithme de classification rapide. L'équipe IBIS du laboratoire utilise pour l'essentiel les méthodes de classifications élaborées disponibles sous SAS (en particulier celle de la distance moyenne), adapté à un traitement approfondi de tableaux carrés relativement petits (2000*2000). Pour certaines applications, le conditionnement préalable de grands fichiers (élimination du bruit), des algorithmes plus frustes mais capables de traiter rapidement des tables de grande taille sont nécessaires. J'ai implémenté en programmation sous SAS un algorithme particulier, le « quick clustering » de Kamen, qui satisfait ces exigences.

Glossaire

Bibliométrie :

Ensemble des méthodes et techniques quantitatives relatives au traitement d'informations bibliographiques.

Diamètre (critère du) :

$$\begin{array}{l}\forall (s,s') \in I^2, s \neq s', nv(s,s') = \sup_{i \in s, i' \in s'} \{d(i,i')\} \\ \forall (s,s',t) \in I^3, s \neq s', s \neq t, s' \neq t, nv(s \cup s', t) = \sup\{nv(s,t), nv(s',t)\}\end{array}$$

où d est une distance.

Deux classes ne sont reliées entre elles que si tous les liens entre éléments des deux classes sont uniformément petits. Il construit des classes compactes et évite de constituer des classes comprenant des éléments très éloignés les uns des autres.

Dissimilarité (indice de) :

Application d de I*I dans R^+ telle que $\begin{array}{l}\forall (i,j) \in I \times I, d(i,j) = d(j,i) \\ \forall i \in I, d(i,i) = 0\end{array}$.

Distance moyenne (critère de la) :

$$\begin{array}{l}\forall (s,s') \in I^2, s \neq s', nv(s,s') = \dfrac{\sum_{i \in s, i' \in s'} d(i,i')}{Card(s).Card(s')} \\ \forall (s,s',t) \in I^3, s \neq s', s \neq t, s' \neq t, nv(s \cup s', t) = \dfrac{Card(s).nv(s,t) + Card(s').nv(s',t)}{Card(s) + Card(s')}\end{array}$$

où d est une distance.

Ce critère est un compromis entre les critères du diamètre et du saut minimum. Il a tendance à construire des classes sphériques (classes de mêmes tailles).

Isolat :

Un élément i de l'ensemble I est un isolat si et seulement si $s(i,j) = \delta_{ij} = \begin{cases} 1, j = i \\ 0, j \neq i \end{cases}$ où s est un indice de similarité.

Matrice creuse :

Une matrice creuse est une matrice où plus de 80% des termes sont nuls.

Saut minimum (critère du) :

$$\begin{array}{l} \forall (s,s') \in I^2, s \neq s', nv(s,s') = \inf\limits_{i \in s, i' \in s'} \{d(i,i')\} \\ \forall (s,s',t) \in I^3, s \neq s', s \neq t, s' \neq t, nv(s \cup s', t) = \inf\{nv(s,t), nv(s',t)\} \end{array}$$

où d est une distance.

L'utilisation de ce critère permet de traiter un grand nombre d'individus. Sans correction appropriée, il y a souvent création d'un effet de chaîne (s'il existe entre deux points très éloignés i et j une suite de points dont chaque point est très proche du suivant, alors les points i et j sont considérés comme très proches), donc création de classes longilignes (classe ayant beaucoup d'éléments dont les points extrêmes sont très éloignés les uns des autres).

Similarité (indice de) :

Application s de I*I dans R^+ telle que $\begin{array}{l} \forall (i,j) \in I \times I, s(i,j) = s(j,i) \\ \forall (i,j) \in I \times I, i \neq j, s(i,i) = s(j,j) = s_{\max} > s(i,j) \end{array}$.

En posant $d(i,j) = s_{\max} - s(i,j)$, d est un indice de dissimilarité.

Ward (critère de) (ou critère de la perte d'inertie intraclasse minimale) :

$$\forall (s,s') \in I^2, s \neq s', nv(s,s') = \frac{m_s . m_{s'}}{m_s + m_{s'}} d(g_s, g_{s'})$$

$$\forall (s,s',t) \in I^3, s \neq s', s \neq t, s' \neq t, nv(s \cup s', t) = \frac{(m_s + m_t).nv(s,t) + (m_{s'} + m_t).nv(s',t) - m_t.nv(s,s')}{m_s + m_{s'} + m_t}$$

où d est une distance, g_k est le centre de gravité de la classe k et m_k est la masse de la classe k (généralement, $m_k = Card(k)$).

Ce critère a tendance à construire des classes sphériques (classes de mêmes tailles).

Bibliographie

Théorie de la classification

[Celeux 1989] CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y., RALAMBONDRAINY H., « Classification automatique des données », Dunod, 1989, 285 p.

[Lebart 1995] LEBART L., MORINEAU A., PIRON M., « Statistique exploratoire multidimensionnelle », Dunod, 1995, 439 p.

[Nakache 2000] NAKACHE J.-P., CONFAIS J., « Méthodes de classification avec illustrations SPAD et SAS », CISIA-CERESTA, 2000, 185 p.

[Volle 1980] VOLLE M., « Analyse des données », coll. Economie et Statistiques avancées, Economica, 1980, 319 p.

Calculs d'indices

[Cailliez 1976] CAILLIEZ F., PAGES J.-P., « Introduction à l'Analyse des Données », S.M.A.S.H., 1976, 616 p.

[Hennequet 1993] HENNEQUET C., « Exploration et validation de méthodes de classification en vue d'analyses bibliométriques », rapport de stage, 1993, 32 p.

Méthode de « Quick clustering »

[Bertier 1981] BERTIER P., BOUROCHE J.-M, « Analyse des données multi-dimensionnelles », PUF, 1981, 270 p.

[Kamen 1970] KAMEN J. M., « Quick clustering », Journal of Marketing Research, vol. VII, mai 1970, pp. 199-204.

[McQuitty 1964] McQUITTY L. L., « Capabilities ansd improvements of linkage analysis as a clustering method », Educational and Psychological Measurement, vol. 24, automne 1964, pp. 441-456.

Utilisation de SAS

[Le Moigne] LE MOIGNE L., REMOND-TIEDREZ E., SILBERBERG S., « Le langage SAS », Europstat Editions, 204 p.

[Leroux] LEROUX-SCRIBE C., « L'analyse des données », guides SAS, Europstat Editions, 397 p.

[Sautory] SAUTORY O., « La statistique descriptive avec le système SAS », INSEE guides, 280 p.

Méthodes bibliométriques

[Prime 2001] PRIME C., BASSECOULARD E., ZITT M, « Co-citations and co-sitations : a coutionery view on an analogy », Proceedings of the 8th international conference on scientometrics and informetrics, vol. 2, the University of New South Wales faculty of commerce and economics, Sydney, Australie, 16-20 juillet 2001, pp. 529-540.

[Zitt 1998] ZITT M, BASSECOULARD E., « Méthodes de structuration pour l'analyse stratégique des univers scientifiques : les techniques de citation », Veille Stratégique, Scientifique et Technologique, Actes VVSST'98, Toulouse, France, 19-23 octobre 1998, pp.31-41.

Abstract

After a brief presentation of bibliometrics, and in particular structuring methods of scientific universes based on data analysis, three studies are detailed, which concerned the use of the proximity measures and automatic clustering. In the first study bibliographical coupling is applied to clean a set of documents before co-citation analysis. The second study compares the results of four hierarchical clustering techniques for subject code co-classification. Finally, a quick clustering algorithm has been implemented in order to partition large sets of items, complementary to previous clustering methods. Some properties of this algorithm are sketched.

Résumé

Après une brève présentation de la bibliométrie, et en particulier des analyses de structuration des univers scientifiques fondées sur l'analyse des données, le rapport présente trois études qui ont trait à l'utilisation des méthodes de calcul de proximité et de classification automatique. En premier lieu, on décrit une méthode de dédoublonnage de documents dans un contexte d'analyse des co-citations. En second lieu, on compare les résultats de quatre techniques de classification ascendantes hiérarchiques dans une analyse de co-classement. Enfin, un algorithme de partition fruste mais rapide et capable de traiter des ensembles de grande taille, complémentaire des méthodes classiques de classification hiérarchique a été implémenté. Quelques particularités de cet algorithme sont abordées.

Mots-clés

Bibliométrie, indice de similarité, classification ascendante hiérarchique, méthode de « quick clustering », méthode des co-citations, analyses des couplages bibliographiques, logiciel SAS.