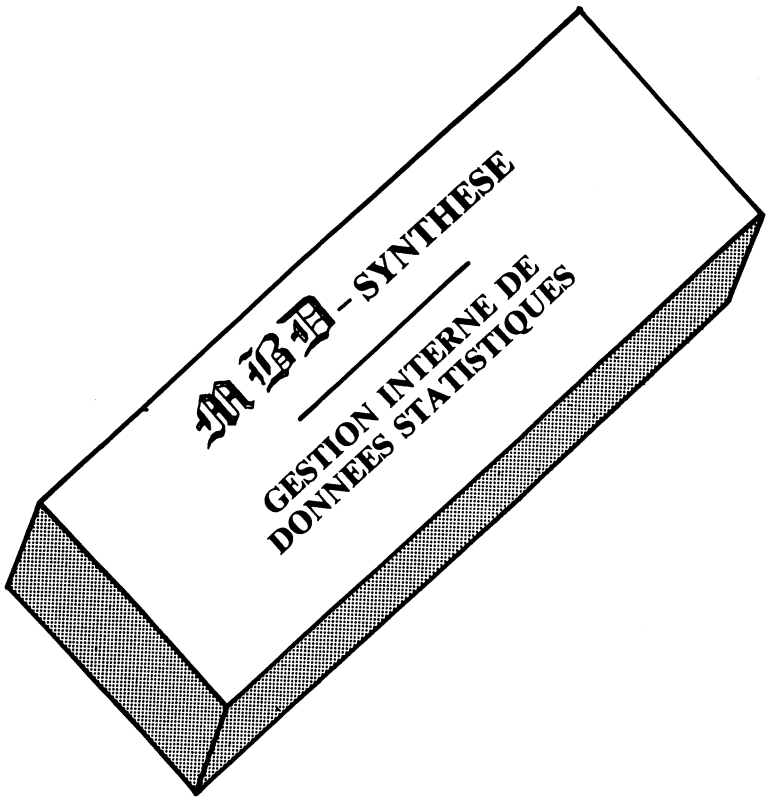
N° 4 - SEPTEMBRE 1986 PUBLICATION TRIMESTRIELLE

MBD

MODÈLES ET BASES DE DONNÉES

AFCET INFORMATIQUE

ISSN 0767-3639



François Bry
ECRC, Arabellastr. 17
D-8000 München 81
Allemagne Fédérale
Tél. : (089) 92 69 91 48

Gérard Thauront
INRETS
2, avenue du Général Malleret-Joinville
94114 Arcueil
Tél. : (1) 45 81 12 12

MOTS-CLES :

Bases de Données Statistiques,
Gestion de Fichiers Statistiques.

• **François BRY** *a travaillé de 1983 à 1984 à l'Institut de recherche des Transports (aujourd'hui INRETS) sur les bases de données statistiques. Depuis le début de l'année 1985, il travaille à l'ECRC à la définition et à la réalisation d'un système d'aide à la conception et à la gestion de schémas de bases de données déductives.*

• **Gérard THAURONT** *est responsable des bases de données au centre informatique (CIR) de l'INRETS. A ce titre il dirige l'équipe d'exploitation de la base de données statistiques SITRAM du Ministère des Transports. De part sa formation, il est statisticien et informaticien.*

RESUME :

Les bases de données statistiques sont pratiquement toutes gérées par des systèmes conçus sur mesure. Cela vient de ce que les SGBD conventionnels, destinés essentiellement aux applications de gestion, ne conviennent pas aux bases statistiques. C'est avant tout en ce qui concerne le stockage des données, la gestion interne de la base, que les problèmes posés et donc les solutions à apporter sont différents. Les bases statistiques nécessitent en particulier des structures de fichier spéciales, permettant des lectures rapides de très nombreuses données ainsi que des compactages spécifiques. Cet article est une synthèse critique des organisations internes proposées pour les bases de données statistiques. Des propositions nouvelles sont également décrites.

TABLE DES MATIERES page

GESTION INTERNE DE DONNEES STATISTIQUES

1. INTRODUCTION

Bien qu'il existe de très nombreuses bases de données statistiques, relativement peu d'attention leur a été consacré jusqu'à maintenant. Leur étude est une direction de recherches pleine d'avenir, d'une part d'un point de vue économique, et d'autre part d'un point de vue technique. La demande en matière de gestion de données statistiques est en effet très forte et se trouve accrue depuis que l'accès aux bases de données est possible par Minitel. D'un point de vue technique, les méthodes développées pour les bases de gestion (notamment en matière de langages d'interrogation et d'accès concurrents) sont adaptables aux bases statistiques. Ces bases sont pratiquement toutes gérées par des systèmes conçus *sur mesure,* pour une application particulière. L'approche consistant à développer un système utilisable par de nombreuses applications, qui a conduit aux SGBD actuels, n'a encore que très partiellement été suivie pour ce type de bases. Or, leurs particularités excluent de les gérer à l'aide de SGBD conventionnels. Au niveau conceptuel, les bases statistiques sont difficiles à décrire par les modèles de données (hiérarchique, réseau, relationnel, entité-association, etc...) conçus principalement pour les applications de gestion : cette question, qui demanderait à elle seule un article, n'est volontairement pas étudiée ici. Au niveau interne, le stockage physique de données statistiques pose des problèmes totalement différents de ceux habituellement résolus par les SGBD. Les bases de données statistiques rassemblent en effet des volumes de données incomparablement plus grands que les bases de gestion. Elles connaissent des interrogations et des traitements différents. Elles demandent donc des structures de fichier spécifiques, permettant des compactages appropriés aux données statistiques et adaptés à leurs interrogations assez particulières.

Après avoir rapidement rappelé les particularités des bases statistiques (section 2), cet article fait une synthèse critique des organisations de fichiers statistiques qui sont proposées dans la littérature (section 3). Des propositions nouvelles sont faites, notamment pour la gestion des « méta-données statistiques », données particulières aux applications statistiques dont le rôle est de décrire la signification des autres données. Cette étude est issue d'un travail mené de 1983 à 1985 à l'Institut de Recherche des Transports (aujourd'hui INRETS) qui assume l'exploitation d'une importante base de données statistiques, la base SITRAM sur les transports de marchandises en France [BODI 82].

2. PARTICULARITES DES BASES DE DONNEES STATISTIQUES

Il est possible de distinguer deux types de bases de données statistiques : les *serveurs d'informations* et les *bases d'analyse* [BRY 85]. Un serveur d'informations peut être défini comme une base de données conçue avant tout pour diffuser de l'information statistique relative à un domaine donné. La base AGRISTAT du Ministère de l'Agriculture qui fournit des informations socio-économiques sur l'agriculture française [RUHL 82], la base SITRAM, du Ministère des Transports qui rassemble des données sur les transports de marchandises en France, la base SEEDIS, issue du recensement américain [McCA 82] en sont des exemples. Une présentation des principaux serveurs d'informations statistiques des administrations publiques françaises est donnée dans [BODI 82]. Une base d'analyse est au contraire une base de données construite pour les besoins d'une étude statistique particulière. Elle sera généralement utilisée par une seule personne, durant une durée limitée, celle de l'étude (de quelques semaines à quelques mois).

Un serveur d'informations rassemble toujours des volumes de données gigantesques (de l'ordre du milliard d'octets), alors qu'une base d'analyse peut parfois être beaucoup plus petite, voire même être exploitée à l'aide d'un micro-ordinateur. Un serveur est interrogé par de très nombreux usagers. Un serveur d'informations devra donc stocker et restituer l'information nécessaire à la compréhension des usagers, telles que :

Définition économique retenue pour le « ménage », liste des professions que recouvre l'appelation de « technicien » entre 1950 et 1980, etc...

Une base d'analyse pourra ne pas stocker de telles *méta-données statistiques* (appelées ainsi pour les distinguer des méta-données classiques en bases de données [BRY 84, BRY 85]) supposées connues de ses usagers. Une telle hypothèse est rarement possible pour un serveur. Un serveur rassemble en effet presque toujours des données aux définitions semblables mais néanmoins différentes. Des données sur le chômage collectées dans les différents pays de la CEE conduiront, par exemple, à plusieurs définitions du chômeur.

Cette étude est consacrée à la gestion, au niveau physique, des données d'une base de type serveur. Certaines des méthodes décrites dans cet article sont adaptables aux bases d'analyse. Les propriétés caractéristiques des serveurs sont rapidement rappelées ci-dessous (pour une discussion plus complète, voir [BRY 84]).

GESTION INTERNE DE DONNEES STATISTIQUES

Les données statistiques sont stables

Les bases de données statistiques présentent un mouvement des données qui est très différent de celui d'une base conventionnelle. Les mises à jour de données statistiques sont rares et de nature différente de celles des applications classiques. Elles ont lieu uniquement en phase d'insertion de nouvelles données dans la base. Une fois stockées, les données statistiques sont validées. Elles ne sont alors plus accédées qu'en lecture. Lors de l'arrivée de nouvelles données, si une modification est nécessaire, elle porte toujours sur de grands ensembles de données. Les modifications au niveau d'un article sont exceptionnelles. Alors qu'une base conventionnelle présente le plus souvent des taux élevés de mises à jour, les données d'une base statistique sont très stables.

Les interrogations d'une base statistique retournent de très grands ensembles

Bien que de nombreux SGBD conventionnels effectuent à des coûts acceptables d'autres requêtes, la plupart d'entre eux est principalement destinée à la recherche d'un petit nombre de tuples. Des exemples classiques, bien qu'extrêmes, sont ceux des systèmes bancaires (« Quel est l'état du compte courant de M. Dupont ? ») et des systèmes de réservation de places (« Reste-t-il une place sur le vol n° 7214 ? »). Les consultations d'une base statistique conduisent au contraire à de très grands ensembles de données (« Revenus, nombre d'enfants, nombre de voitures et caractéristiques des résidences des ménages de telles catégories socio-professionnelles de l'Ile-de-France »). Ces ensembles peuvent représenter des volumes de quelques milliers à quelques millions d'octets. Bien souvent, les fichiers doivent être lu exhaustivement : l'organisation de fichiers doit prendre en compte cette particularité. Des ensembles de données aussi grands ne sont pas demandés pour une lecture humaine immédiate. Ils sont d'abord soumis à des traitements tels que des analyses statistiques, des interprétations graphiques ou des éditions sophistiquées, du type tableaux croisés (cf. Fig. 1).

Les résultats d'interrogations doivent être triés

Les données extraites d'une base statistique sont demandées triées, dans un ordre qui n'est pas forcément celui de lecture. Les programmes d'analyse statistique, d'interprétation graphique ou d'édition manipulent en effet des ensembles triés. De très substanciels gains de temps peuvent être réalisés en intégrant les processus de lecture et de tri (paragraphe 3.3.4).

Des méta-données volumineuses

Une base statistique référence un très grand nombre de termes [CHAN 81, BATE 82, SHOS 82, WONG 82]. Elle peut réunir un très grand nombre de relations (des serveurs d'informations en comprennent souvent plusieurs centaines) et chaque relation peut avoir quelques dizaines d'attributs. L'usager d'une base de données statistiques doit donc manipuler quelques milliers de termes. A chacun de ces termes sont attachées des méta-données statistiques, informations indispensables à la compréhension des données [WONG 82, CHAN 81, ANDE 82, BATE 82]. Notons que la structure de méta-données est très différente de celle des données proprement dites [McCA 82]. En fait, les organisations internes devront être différentes (sous-section 3.2).

Des redondances nécessaires

Dernière particularité, les interrogations statistiques demandent des ensembles de données redondants en

	population adulte masculine	population adulte féminine	population adulte	population non adulte	toutes populations
région parisienne					
province					
France entière					

Figure 1.

un certain sens. Cette redondance est celle d'un ensemble accompagné de totaux partiels, de marges ou de valeurs de fonctions (moyenne, maximum, etc...) comme dans l'exemple d'interrogation suivant :

« Ensemble des salaires des techniciens, moyennes (de ces salaires) par tranches d'âges, moyenne générale et valeurs extrêmes. »

Les valeurs extrêmes sont des données redondantes, car elles sont implicitement retournées par la première partie de l'interrogation. Une moyenne est également redondante : elle correspond à une interrogation sur l'ensemble des salaires. Une telle demande simultanée de plusieurs ensembles de données dérivés d'un premier ensemble est typique en statistique. Cette simultanéité doit être prise en compte car elle permet une optimisation des temps de traitements.

3. D'AUTRES ORGANISATIONS PHYSIQUES

La gestion interne des données statistiques pose essentiellement trois questions : (1) comment optimiser les temps de réponse en consultation compte tenu des caractéristiques des interrogations, (2) comment structurer les méta-données statistiques et enfin (3) comment tirer partie de la nature statistique des données pour les compacter, et réaliser les indispensables gains en espace mémoire imposés par les volumes gigantesques à stocker. Deux premières sous-sections (3.1 et 3.2) proposent des réponses aux deux premières questions. Des organisations de fichier spécifiques aux données statistiques sont ensuite décrites dans une troisième et dernière sous-section (3.3). Cette sous-section traite tout d'abord des ensembles de données particuliers que sont les *séries chronologiques* (paragraphe 3.3.1). Deux types de données statistiques, les caractères qualitatifs et les caractères quantitatifs, sont définis. Il est montré comment réaliser d'importants gains de place par des *codages* appropriés des caractères qualitatifs (paragraphe 3.3.2) et par *compactage* des caractères quantitatifs (paragraphe 3.3.3). Un dernier paragraphe (3.3.4) propose un moyen simple d'optimisation des temps de traitement des requêtes statistiques.

3.1. OPTIMISATION DES TEMPS DE REPONSE

Le problème des temps de réponse ne se pose pas de la même manière pour une base de données statistiques que pour une autre base, en raison des tailles tant des ensembles de données consultés que des ensembles retournés par les requêtes d'interrogation. Nous avons pu par exemple observer des temps d'UC de quelques minutes à quelques dizaines de minutes pour l'exécution d'interrogations de la base SITRAM.

L'usager ne peut pas interpréter sans traitements informatiques des ensembles de données aussi grands que ceux retournés par une interrogation. Il demandera, du plus simple au plus compliqué, soit une édition sophistiquée (par exemple sous forme de tableaux croisés avec plusieurs attributs en ligne et en colonne), soit une ou plusieurs interprétations graphiques des données extraites de la base, soit enfin entreprendra la mise en œuvre d'une étude statistique impliquant une ou plusieurs méthodes d'analyse statistique fournies par les logiciels d'analyse (ou « statistical packages »). Si la définition d'une édition ou d'une interprétation graphique peut être rapide, il n'en est pas de même de la mise en œuvre d'une méthode d'analyse statistique. Elle peut prendre de quelques dizaines de minutes à quelques jours, le traitement informatique lui-même pouvant demander plusieurs dizaines de minutes de temps d'UC. L'extraction des données d'une base n'est donc, pour un usager, qu'une première étape dans un processus qui peut être long. Même l'expression d'une interrogation simple (sans traitement statistique) est souvent un exercice délicat et long, car l'usager doit consulter de nombreuses descriptions de données. Bien souvent le processus de définition d'une interrogation, de traitement de cette interrogation suivi éventuellement d'une vérification de la pertinence de la requête puis du lancement d'une méthode d'analyse statistique est décomposé dans le temps par l'usager lui-même. Un certain nombre d'interrogations d'un serveur d'informations pourra donc être traité par lots, sans pénaliser les usagers.

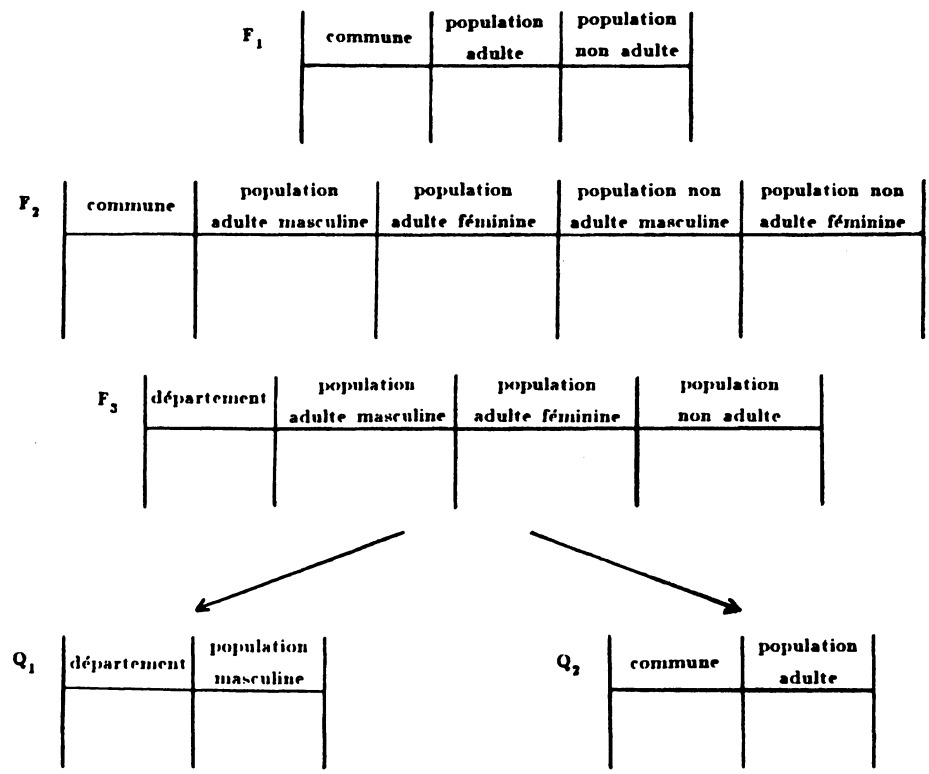
Les volumes de données contenus dans une base de type serveur peuvent être tels que la totalité de la base de données ne puisse pas être stockée totalement sur disque, et qu'un stockage partiel sur bandes soit nécessaire. La base SITRAM représente un volume de 2 milliards d'octets (méta-données non comprises), la base SEEDIS représente près de dix milliards d'octets [McCA 82]. L'interrogation de données sur bandes conduit à des temps de réponse de l'ordre d'une à quelques heures. Une exécution différée de l'interrogation s'impose alors. En pratique les administrateurs de telles bases essaient de ne conserver sur disque que les données les plus fréquemment consultées.

A défaut d'utiliser des disques optiques, un SGBD

statistique destiné à des serveurs doit donc offrir des services de gestion de bandothèques. Idéalement, il devrait gérer automatiquement les copies de bandes à disques et les libérations d'espace disque. Il devrait permettre une expression identique des interrogations, que les données lues soit sur bande ou résidentes sur disque. Bien entendu, les méta-données (statistiques ou conventionnelles) doivent rester constamment disponible sur disque.

Un moyen pour obtenir des temps de traitements corrects consiste à stocker les « mêmes » données à différents niveaux d'agrégation. Précisons cela sur un exemple. Si une variable indique un « lieu en France », on pourra stocker aux côtés des données relatives au « communes », les données agrégées portant sur les « départements » et celles portant sur les « régions ». En stockant différents ensembles ainsi dérivés des données initiales (relatives aux communes), choisis en fonction des consultations les plus fréquentes, il est possible de réduire considérablement l'écart entre les niveaux d'agrégation des données lues et celui des données demandées par l'usager. De plus il est préférable d'accéder aux données relatives aux départements pour une interrogation demandant par exemple un regroupement particulier de départements, plutôt que de lire les données par communes. Bien entendu, le gain en temps est obtenu au détriment de l'espace mémoire.

Une stratégie d'optimisation s'impose alors : plutôt que de chercher à optimiser les temps de traitement de chaque requête, il peut être préférable de minimiser le coût global d'exécution de l'ensemble des requêtes d'une période de temps (demi-journée, par exemple). Ceci est possible en particulier si certaines interrogations sont exécutées en différé. Si deux requêtes doivent lire deux ensembles dérivés tous deux sur bandes, des gains de temps très substanciels sont possibles en ne chargeant sur disque qu'un seul ensemble assez « fin » pour satisfaire les deux requêtes. C'est ainsi par exemple qu'il pourra être préférable de lire des données relatives aux communes pour une requête qui pourrait être traitée en consultant les données agrégées par départements si d'autres interrogations demandent, durant la même période, la lecture sur bande des données portant sur les communes (cf. Fig. 2).



F_3 est le fichier le mieux adapté à la question Q_1, mais il ne permet pas de traiter la question Q_2. Symétriquement, F_1 est le mieux adapté à Q_2, mais ne permet pas de répondre à Q_1. Il peut être préférable de calculer Q_1 et Q_2 à partir de F_2 : cela demande plus de calculs, mais moins d'entrées-sorties.

Figure 2.

Le temps de traitement d'une requête peut être allongé, mais le coût collectif s'en trouvera réduit. Bernard Schnetzler, de l'INRETS, a validé pour la base SITRAM les principes d'un tel gestionnaire de files d'attente. Les requêtes en attente ne spécifient pas un ensemble à lire, mais une « famille », unité rassemblant les ensembles dérivables les uns des autres ou d'un même ensemble « père » [BRY 84]. Le système gère automatiquement les chargements de bande à disque, met à jour la méta-base lors de création de nouveaux ensembles, et libère l'espace disque. Le système consulte la méta-base pour sélectionner un ensemble dans une famille. Des ensembles peuvent être maintenus sur disque à la demande de l'administrateur de la base et des priorités d'exécution peuvent être définies.

3.2. GESTION INTERNE DES META-DONNEES STATISTIQUES

Les méta-données statistiques sont de structures très différentes des données proprement dites. Elles peuvent être décrites en *dictionnaires* et *nomenclatures.* Les dictionnaires sont des ensembles de méta-données fournissant les *synonymies* et les descriptions (textes) attachées aux objets du schéma conceptuel ou des vues (cf. exemple 1), les nomenclatures sont des objets spécifiques aux applications statistiques, que nous décrivons plus bas.

Exemple 1

Synonymies : « Val-de-Marne, 94, etc... . »

Texte descriptif : « Le département du Val-de-Marne a été créé en telle année lors de telle réorganisation administrative. Il rassemble telles communes de tel ancien département à l'exception de ... etc ... . »

Au contraire des données proprement dites qui sont stables, les méta-données statistiques sont sujettes à de fréquentes mises à jour. Les accès à ces méta-données sont semblables à ceux que connaissent les bases conventionnelles. Ils doivent être possibles de manière interactive. Les consultations des méta-données sont fréquentes et des taux de concurrence élevés doivent être permis, au contraire des données proprement dites [SHOS 82]. Ces caractéristiques conduisent à proposer que la gestion des méta-données soit complètement séparée de celle des données, et soit assurée par un système semblable aux SGBD conventionnels, par exemple par l'un d'entre eux. Ce système gèrerait non seulement les méta-données statistiques mais aussi les schémas (le schéma conceptuel et les vues) de la base statistique. La gestion de dictionnaires serait relativement aisée à l'aide d'un SGBD relationnel. Les particularités d'une base statistique donnée ne seraient exprimées que dans l'extension de la base relationnelle et non dans son schéma. Le schéma de la base relationnelle des méta-données ne serait construit qu'une seule fois pour les différentes bases statistiques. En d'autres termes, une base statistique particulière (et toutes ses extensions possibles) serait une extension possible de cette base des méta-données.

Le concept de nomenclature renvoie à celui de *caractère quantitatif,* que nous décrivons d'abord. Les applications statistiques référencent deux types d'attributs, appelés caractères ou variables par les statisticiens. Un caractère peut être *quantitatif* ou *qualitatif.* Un caractère est quantitatif lorsqu'il prend des valeurs numériques et que des opérations arithmétiques sur ces valeurs ont un sens. Un caractère exprimant un revenu annuel est quantitatif, un autre exprimant un département de naissance ne l'est pas.

Un caractère qualitatif prend des valeurs (appelées modalités) non numériques (lieu, profession, type de marchandise, sexe, nationalité, etc...). Les modalités d'un caractère qualitatif respectent une organisation bien définie dont la structure est celle d'arbre, appelée nomenclature. Plus précisément, une nomenclature est un arbre dont les sommets terminaux sont les différentes modalités que peut prendre un caractère qualitatif. Les sommets non terminaux expriment des regroupements possibles [1]. Des exemples de nomenclatures bien connues des non-statisticiens sont la nomenclature des lieux en France (communes-départements-régions) et celle des catégories socio-professionnelles (csp).

Les nomenclatures tiennent une place particulière, entre les méta-données statistiques et les données. Elles forment en effet des ensembles stables dont les mises à jour sont semblables à celles des ensembles de données statistiques. Leurs volumes peuvent être d'ampleurs comparables : si une nomenclature est un arbre de faible profondeur (les distances des sommets terminaux à la racine ne dépassent généralement pas cinq arêtes), le nombre de sommets terminaux peut être très grand (quelques dizaines de milliers dans le cas des communes). Un système de gestion spécifique aux nomenclatures est nécessaire. Les synonymies ainsi que les textes descriptifs attachés aux sommets d'une nomenclature pourrait être gérés indépendamment de la structure d'arbre, par le système de gestion de la méta-base.

[1]. Les sommets terminaux forment le domaine de l'attribut auquel est attaché la nomenclature.

GESTION INTERNE DE DONNEES STATISTIQUES

3.3. GESTION INTERNE DES DONNEES

Il est souvent mentionné dans la littérature que les organisations de fichiers des SGBD conventionnels ne conviennent pas aux bases statistiques [TURN 79, EGGE 80, EGGE 81, SHOS 82]. Dans cette sous-section, les problèmes de stockage interne des *séries chronologiques* (ensembles de données dont un des attributs est le temps) sont tout d'abord étudiés (paragraphe 3.3.1), puis il est montré comment compacter les données statistiques selon leur nature qualitative ou quantitative (paragraphes 3.3.2 et 3.3.3). Un dernier paragraphe (3.3.4) propose de réaliser tri et agrégation des données au cours des lectures, afin d'optimiser les temps de réponse.

3.3.1. Séries chronologiques

Le stockage en mémoire des séries chronologiques pose essentiellement deux questions. Tout d'abord, faut-il stocker une série multiple comme un ensemble unique, ou convient-il de l'éclater en plusieurs séries stockées séparément (cf. Fig. 3) ?

Cette question, posée dans [DELO 83], n'est en fait pas particulière aux séries chronologiques, ni aux bases de données statistiques. Si des données sont le plus souvent accédées ensembles, il est préférable de les stocker ensembles, séparément sinon. Notons que certains traitements statistiques demandent de réunir en un seul fichier plusieurs séries, qui peuvent être stockées séparément. Les logiciels d'analyse statistique offrent généralement un outil de fusion de séries [DEKK 82].

Un second problème découle des deux types de données que contiennent souvent les séries chronologiques, données observées et données prévisionnelles. Les données observées sont stables et souvent volumineuses. Il est donc souhaitable de disposer d'une organisation de fichiers très efficace en lecture de grands ensembles données, quitte à ce que les mises à jour (très rares) soient plus coûteuses. Les données de prévision, au contraire, connaissent des mises à jour fréquentes et sont moins volumineuses que les données observées. D'autre part, une même série de valeurs observées peut être « prolongée » par plusieurs suites d'estimations. Nous proposons donc de stocker séparément valeurs observées et prévisions. Cela permet d'une part d'offrir à chaque type de données une organisation interne appropriée et d'autre part facilite la gestion interne de prévisions multiples (cf. Fig. 4).

Un stockage « séquentiel selon le temps » des séries chronologiques permet que les ajouts de nouvelles valeurs ne demandent pas de réorganisation des données déjà stockées. Cela peut être réalisé par une organisation de fichiers du type séquentiel, ou encore en stockant dans plusieurs fichiers les données de différentes périodes.

Les données prévisionnelles posent, quant à leur stockage, les mêmes problèmes que les données des bases conventionnelles : elles ne sont pas stables comme le sont les données observées et sont au contraire sujettes à de fréquentes mises à jour portant sur de petits ensembles. Les organisations de fichiers proposées pour les bases de données classiques conviendraient donc si ce n'est que l'intégra-

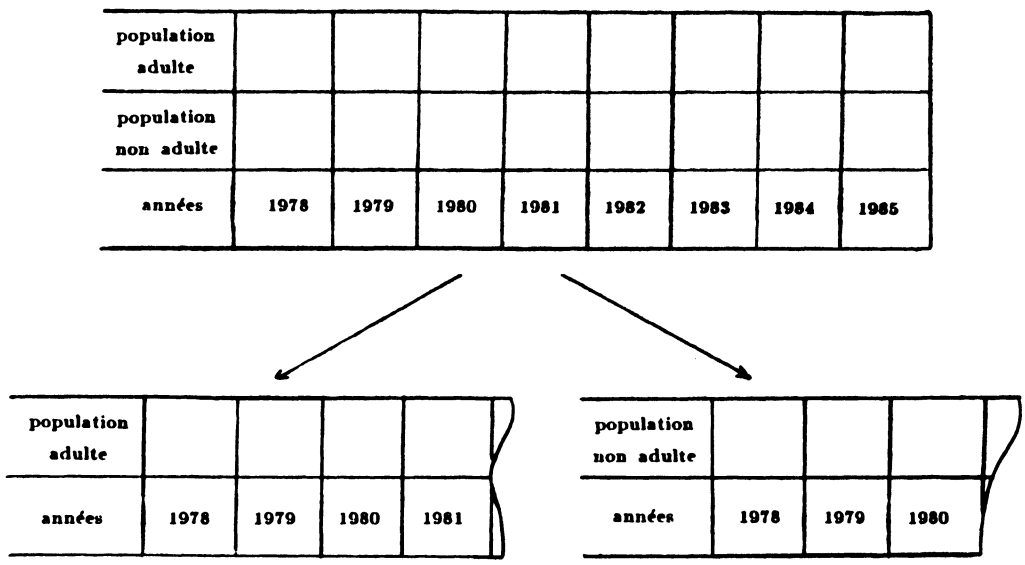


Figure 3.

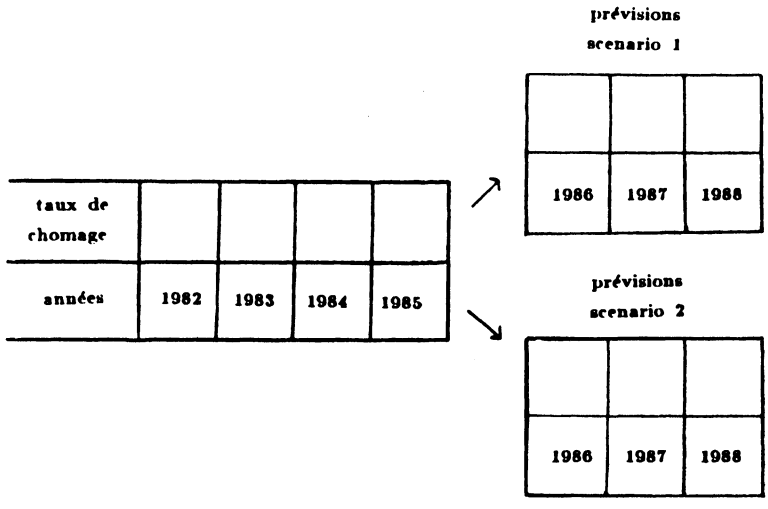


Figure 4.

tion entre outils d'analyse statistique et SGBD peut conduire à préférer une organisation proche de ce que les logiciels d'analyse statistique savent manipuler. Pour cette raison une organisation relative peut être préférable à une organisation indexée.

3.3.2. Caractères qualitatifs : codages

Pout tout type d'ensembles (série chronologique ou autre) de substantiels gains de place sont possibles par des codages. Les modalités d'une nomenclature, et donc les valeurs des attributs des relations d'une base statistique, sont en effet souvent des libellés descriptifs. Il est préférable de faire figurer dans les fichiers des codes de modalités et de stocker dans la méta-base les correspondances entre codes et libellés. La nécessité, pour pratiquement toute application statistique, de référencer divers synonymes d'un même libellé et des textes explicatifs, incline à une telle solution (cf. Fig. 5) [2].

libellé officielle : Ingénieurs et Cadres

abréviation officielle : IC

code officiel : 0012

code de stockage : 31

Figure 5.

Le nombre de bits nécessaires à un codage ne dépend, en toute rigueur que du nombre d'occurences que peut prendre la variable, c'est-à-dire du nombre de modalités d'un niveau de nomenclature. On peut ainsi coder les 95 départements français sur 7 bits. De tels codages sont dits denses : tout code possible (entre deux bornes) correspond à une modalité. Des codages denses sont souhaitables. Des codages non denses demandent en effet des organisations de fichiers assez sophistiquées pour éviter la perte de place en mémoire secondaire provoquée par les valeurs ne correspondant à aucun code. Des organisations séquentielles ou aléatoires ne conviendraient par exemple pas en présence de codage non denses. Nous verrons plus bas, au paragraphe 3.3.3, que des codages non denses peuvent également rendre plus délicats les compactages qui permettent d'éviter le stockage de suites de valeurs nulles (vrais zéros).

Cependant, il est difficile en pratique de conserver des codages denses. En effet une nomenclature évolue dans le temps et si l'on souhaite disposer des données de chaque période, les états successifs d'une nomenclature doivent rester accessibles. Plutôt que de stocker toute une nomenclature par période de temps, il est préférable (pour économiser l'espace mémoire) de conserver la nomenclature initiale, et de créer à chaque modification une « règle de calcul » de la nouvelle nomenclature à partir de la précédente (cf. Fig. 6).

Selon les applications un choix devra être fait entre des codages denses compliquant la gestion des nomenclatures, et une gestion plus simple des nomenclatures qui demande des organisations de

[2]. Les exemples de cette figure et des suivantes ne sont pas réels.

avant 1900

20 : Route 21 : Fer 22 : Eau

à partir de 1900

20 : Route

201 : Traction Animale 202 : Traction Mécanique

après 1930

- Le code 201 (Traction Animale disparait).
- Le code 23 (Aérien) apparait.

Figure 6.

fichiers plus complexes ou moins performantes (en temps d'exécution des entrées-sorties ou en utilisation d'espace mémoire).

3.3.3. Caractères quantitatifs : compactages

La représentation des variables arithmétiques pose moins de problèmes que celle des caractères qualitatifs. Notons tout d'abord qu'il est souvent utile, dans les applications statistiques, de *centrer* les caractères arithmétiques. Cela consiste à calculer un paramètre de position (généralement une moyenne) à partir de l'ensemble des valeurs du caractère arithmétique, de la mémoriser et de stocker chaque valeur diminuée de ce paramètre au lieu de la valeur elle-même (cf. Fig. 7). Cela permet de gagner en précision sans accroître l'espace mémoire nécessaire. Cette méthode, utilisée par les statisticiens pour calculer les variances, n'est à notre connaissance pas utilisée en bases de données.

F_1	1230.1	21.725	36.662	72,471
F_2		1258.8	1266.7	1302.6

Les données du fichier F_1 sont plus précises que celles du fichie F_2. avec le même nombre de chiffres significatifs.

Figure 7.

Une organisation de fichiers adaptée à un SGBD statistique doit avant tout permettre un compactage des données, pour les ensembles de faible densité, c'est-à-dire contenant beaucoup de zéros, fréquents en statistique, et permettre aussi des accès rapides en lecture. Les lectures séquentielles, fréquentes dans les applications statistiques, doivent être aussi peu coûteuses que possible. La très grande taille des ensembles de données statistiques (10^5 à 10^6 octets) exclue les structures de fichier demandant des réorganisations périodiques après l'insertion de nouvelles données. Il faut que les modifications de la base de données, c'est-à-dire essentiellement les ajouts de nouvelles données, se fassent en créant de nouveaux secteurs dans le fichier et non en modifiant (par insertion de nouveaux articles et éclatement) des secteurs existants déjà. Des organisations du type ISAM ou « B-tree » sont donc disqualifiées pour des données statistiques.

Certaines des contraintes imposées à l'organisation de fichiers dans les applications générales sont par contre levées dans le cas des bases de données statistiques. La stabilité des données statistiques, et le fait que les modifications affectent le plus souvent des fichiers dans leur totalité, autorise le choix d'organisations de fichiers peu performantes en insertion ou en mise à jour. C'est ainsi, par exemple, que si un placement des articles par hachage des clés présente l'inconvénient pour les applications « classiques » d'entraîner une dégradation sensible des temps de lecture après débordement, ce défaut

n'existe pas pour les bases de données statistiques, pour lesquelles l'ajout de nouvelles données peut être réalisé en créant de nouveaux fichiers.

Certaines particularités des données statistiques peuvent être mises à profit dans la conception d'une organisation de fichiers répondant aux objectifs annoncés plus haut, c'est-à-dire permettre avant tout des compactages et des lectures rapides. Les données statistiques se présentent « naturellement » en articles de taille fixe, si le fichier est plat. On appelle ainsi des fichiers ne contenant pas de *groupes répétitifs,* semblables à des « tableaux » de « structures » PL/1, ni de *groupes hiérarchiques,* correspondant à des « structures imbriquées (les groupes sont définis dans [BRY 84]) : un fichier plat est un fichier dont les articles expriment explicitement les valeurs des différentes variables, comme une relation dans une base relationnelle.

La première technique de compactage est connue des statisticiens. Elle repose sur des fichiers séquentiels et consiste à tirer partie de ce que les clés des articles sont formées d'une suite de valeurs (d'attributs) et peuvent donc être triées selon un ordre lexicographique. Une certaine forme de redondance apparaît alors, car de nombreux articles adjacents donnent des valeurs identiques à certains caractères qualitatifs. Il est possible de ne stocker chaque valeur du caractère qualitatif de poids fort qu'une seule fois, et pour chacune des valeurs de ce premier caractère de ne donner qu'une fois chaque valeur du caractère suivant, etc... (cf. Fig. 8). Les articles étant de taille fixe, il est aisé de calculer la position dans le fichier d'un article de clé donnée.

département	csp	situation de famille	salaire
75	IC	célib.	102
75	IC	célib.	250
75	IC	célib.	150
75	IC	célib.	180
75	IC	marié	120
75	IC	marié	130

(75 (IC (célib. (102. 250. 150. 180). marié (120. 130))))

Figure 8.

Il est souhaitable de chercher à compacter les données non seulement sur les caractères qualitatifs, mais aussi sur les caractères quantitatifs. Une solution simple consiste à enlever purement et simplement les articles pour lesquels tous les caractères quantitatifs prennent des valeurs nulles (vrai zéro). On fait alors la convention que si aucun article n'est trouvé pour une clé donnée, c'est qu'il exprime des valeurs nulles (cf. Fig. 9). Cette méthode est simple si les codages des nomenclatures sont denses. Sinon, toute lecture du fichier doit se faire parallèlement à celle de plusieurs nomenclatures, pour distinguer entre les absences d'articles correspondant aux valeurs nulles et celles résultant de l'absence de signification de la clé. Un compactage par suppression des valeurs nulles permet des adressages relatifs approximatifs dans un fichier séquentiel. Il suffit pour cela que les valeurs nulles soient bien distribuées dans le fichier trié. Notons qu'il y a souvent de bonnes raisons que ce ne soit pas le cas.

commune	code interne commune	information
Arcueil	12	1
Paris	13	0
Créteil	15	12

12	1	15	12

Pour distinguer entre les deux articles possibles [13 | 0] et [14 | 0] celui sans signification, le second car 14 n'est le code d'aucune commune, il faut consulter la liste des codes internes.

Figure 9.

Un autre mode de compactage consiste à remplacer une suite de valeurs identiques par un couple formé de la longueur de cette suite (ou « facteur de répétition ») et de la valeur constante (cf. Fig. 10). Ce mode de compactage présente l'inconvénient de ne plus permettre d'adressage relatif. Un problème intéressant, se rapportant à cette méthode de compactage, a été posé par A. Shoshani [SHOS 82] : trouver un bon algorithme de tri des caractères optimisant la somme des longueurs des séquences constantes en fonction des distributions des valeurs des caractères quantitatifs.

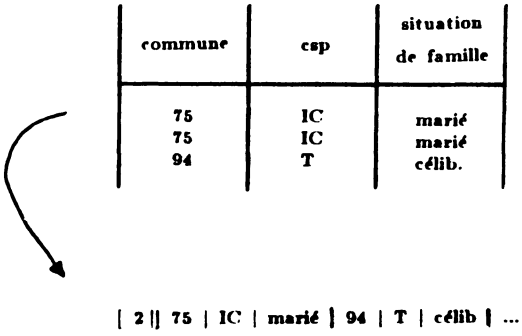
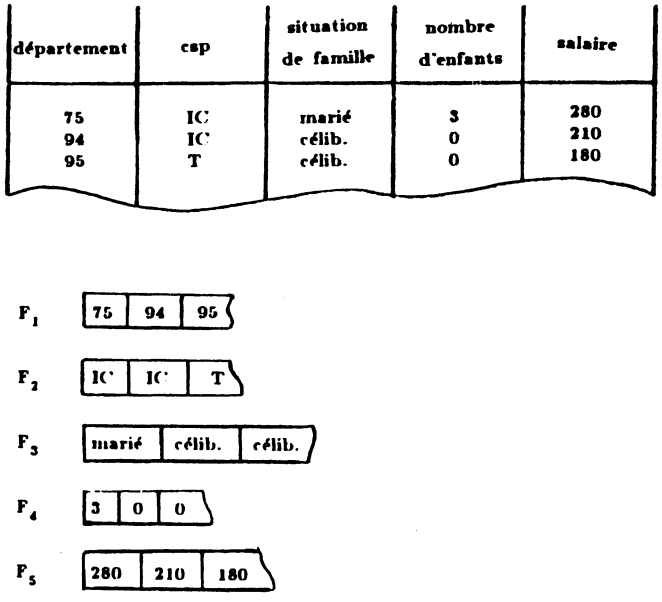


Figure 10.

Des compactages sophistiqués, dérivés de la méthode précédente mais offrant un adressage relatif (en temps logarithmique) ont été proposés [EGG 80, EGGE 81]. Ces compactages présentent toutefois l'inconvénient de sortir du cadre connu des logiciels statistiques, qui ne savent généralement manipuler que des fichiers compactés par suppression d'articles ou par insertion de facteurs de répétition. Des structures de fichiers inconnues des programmes statistiques imposent des réorganisations des données avant le lancement de procédures d'analyse statistique. De telles réorganisation sont souvent coûteuse en temps, car elles demandent en général un tri.

Une dernière structure de fichier proposée dans la littérature et qui semble la plus intéressante, est celles des *fichiers verticaux* [TURN 79, BURN 81]. Elle tire son nom de la représentation usuelle en table des fichiers ou des relations. Elle consiste à stocker les données par attribut (verticalement) plutôt que par tuple (horizontalement) (cf. Fig. 11). Elle fut proposée par les concepteurs du système RAPID, un des premiers SGBD statistiques [TURN 79].

Trois raisons justifient cette méthode. Premièrement, on observe fréquemment que pour un carac-



La question "moyenne des salaires par csp" ne demande la lecture que de deux fichiers. F_2 et F_5. Si les données sont triées par département et csp, F_1 et F_3 peuvent être compactés.

Figure 11.

tère quantitatif, les valeurs nulles ont « naturellement » tendance à être regroupées, car les nomenclatures ordonnent le plus souvent les modalités selon les niveaux supérieurs. Si on considère par exemple un fichier des types de productions exprimées en valeurs marchandes par régions géographiques et qu'une région est peu industrialisée, la séquence des modalités exprimant les production industrielles a de forte chance de contenir de nombreuses suites de zéros. Une seconde raison en faveur de cette méthode est qu'une consultation d'une base statistique demande bien souvent beaucoup moins d'attributs que le fichier n'en contient. Une dernière raison en faveur de cette organisation de fichier est que les techniques de compression sophistiquées décrites plus haut sont plus efficaces sur les « colonnes » que sur les « lignes » des fichiers triés, en raison du peu de différence entre les valeurs des caractères qualitatifs d'articles proches [EGG 81, EGGE 80].

Une autre méthode pour diminuer l'encombrement en mémoire d'une base statistique consiste à tirer partie de la structure algébrique de certains ensembles de données statistiques. Certaines matrices produites par des méthodes d'analyse statistique sont des symétriques ou presque diagonale. Il est possible de ne stocker que les « parties utiles » (c'est-à-dire non nulles) de ces matrices.

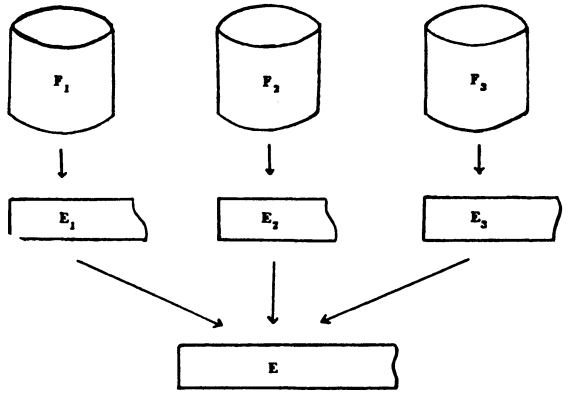
Citons également des propositions de stockages efficaces d'ensembles dérivés (par des méthodes d'analyse statistiques) [LITT 83]. Les méthodes proposées visent à ne mémoriser que les différences entre l'ensemble dérivé et l'ensemble de données initial, lorsque celles-ci sont peu importantes. Il semble difficile, par ces méthodes, de conserver l'indépendance physique des données : l'usager créateur d'ensembles dérivés doit intervenir dans le choix du mode de stockage.

3.3.4. Trier et agréger au cours des lectures

La lecture des fichiers statistiques permet une optimisation originale. Nous avons vu que les consultations d'une base de données statistique produisent de grands ensembles de données, qui doivent être triés. Afin de réduire les temps de réponse, il est souhaitable que les tris soient partiellement effectués au cours de la lecture des données de la base et non *a posteriori* (cf. Fig. 12). Lorsque la requête demande des données agrégées, c'est-à-dire à des niveaux de nomenclatures moins fins que ceux des données lues, un gain très sensible de temps est obtenu en calculant progressivement les cumuls également au cours de la lecture.

4. CONCLUSION

Dans cette synthèse sur la gestion interne des données statistiques, nous avons tout d'abord rappelé les caractéristiques des bases de données statistiques qui influent sur les choix d'organisation de fichiers. Ces caractéristiques distinguent très nettement les bases statistiques des bases de gestion. Les bases



Si plusieurs fichiers, F_1, F_2 et F_3, doivent être lus pour former un résultat E, et si E doit être trié, on peut d'abord former les résultats intermédiaires triés E_1, E_2 et E_3 à partir de chacun des fichiers (les temps de lecture peuvent être mis à profit pour les tris). Le résultat E est obtenu par fusion-tri de E_1, E_2 et E_3.

Figure 12.

statistiques rassemblent en effet des volumes souvent gigantesques de données stables, accompagnées de volumineuses méta-données statistiques nécessaires à la compréhension par l'usager des informations de la base. Nous avons ensuite montré que les organisations internes adaptées aux bases de données statistiques doivent répondre à des objectifs différents de ceux des bases conventionnelles : les mises à jour de données sont en effet d'un tout autre type, et les accès aux bases statistiques sont essentiellement des lectures, souvent exhaustives. Les différentes organisations internes proposées pour les bases statistiques ont été ensuite décrites et discutées. Distinguant d'une part selon la nature des données (données observées ou prévisions) et d'autre part selon le type des attributs (qualitatif ou quantitatif) les diverses organisations sont expliquées, ainsi que les techniques de compactage des données compatibles avec chacune d'entre elles.

Cette synthèse n'est qu'une première tentative de comparaison des organisations internes de données statistiques qu'il serait intéressant de prolonger. Deux directions sont possibles.

Premièrement, il faudrait disposer de mesures pratiques des temps d'accès et des gains en espace mémoire des diverses solutions. Réaliser une telle étude représenterait un travail considérable. En effet, les solutions retenues pour les bases statistiques actuellement exploitées reposent le plus souvent sur des combinaisons empiriques de diverses méthodes. Des mesures (des temps d'accès et des espaces mémoires) effectuées sur les logiciels existants se prêterait donc difficilement à des comparaisons. Il serait très probablement nécessaire d'implanter ou de simuler les diverses méthodes pour les besoins de l'étude.

Répertorier les différentes bases statistiques actuellement exploitées et leurs techniques de gestion interne des données est une deuxième direction possible. A notre connaissance, aucun serveur n'est géré par un SGBD commercialisé. Est-ce bien le cas ? Les critères de comparaisons retenus dans cet article résistent-ils aux contraintes que connaissent d'autres applications que la base SITRAM ?

Pour conclure, nous essaierons de répondre à une question souvent posée (en particulier par le comité de rédaction de MBD) : quel est le devenir de la recherche sur les bases de données statistiques ? Si pendant longtemps les recherches sur les bases de données statistiques ont essentiellement portées sur la confidentialité, ce n'est plus le cas. Depuis quelques années le nombre des publications traitant des bases statistiques s'accroît. Deux principales directions de recherche émergent : la gestion interne des données statistiques et la modélisation des bases statistiques. L'essentiel des travaux est mené aux Etats-Unis. En Europe, très peu d'efforts ont été jusqu'à présent consacrés à l'étude des bases de données statistiques. Il semble cependant que ce sujet de recherche soit prometteur. Très certainement, les années à venir verront l'apparition sur le marché des SGBD dédiés aux grosses applications statistiques.

BIBLIOGRAPHIE

[ANDE 82] A.J.B. Anderson : Software to Link Database Interrogation and Statistical Analysis. In : *Proceedings of the International Conference COMPSTAT,* pages 139-144, 1982, (Toulouse).

[BATE 82] D. Bates, H. Boral and D.J. DeWitt : A Framework for Research in Database Management for Statistical Analysis. In : *Proceedings of the ACM - SIGMOD International Conference on the Management of Data,* pages 69-70, 1982.

[BODI 82] J.-L. Bodin : Les Banques de Données dans le Système Statistique Public. *Le Courrier des Statistiques* (18), 1982.

[BRY 84] F. Bry et G. Thauront : *Bases de Données Statistiques, Bases de Données d'un Autre Type.* Rapport Interne IRT-Groupe SITRAM, Institut de Recherche des Transports (aujourd'hui INRETS), 2, avenue du Général Malleret-Joinville, 94114 Arcueil (France), décembre, 1984.

[BRY 85] F. Bry et G. Thauront : Spécificités des S.G.B.D. Statistiques. In : *Compte-rendu des Quatrièmes Journées Internationales Analyses des Données et Informatique.* INRIA, octobre, 1985, (Versailles).

[BURN 81] R.A. Burnett and J.J. Thomas : Data Management Support for Statistical Data Editing and Subset Selection. In : *Proceedings of the Workshop on Statistical Database Management,* 1981, (Menlo Park).

[CHAN 81] P. Chan and A. Shoshani : SUBJECT : A Directory Driven System for organizing and Accessing Large Statistical Databases. In : *Proceedings of the International Conference on Very Large Data Base,* 1981, (Cannes).

[DEKK 82] A.L. Dekker : Postgraduate Training for Statisticians - Database Methods. In : *Proceedings of the International Conference COMPSTAT,* pages 179-185. 1982, (Toulouse).

[DELO 83] C. Delobel : Les Bases de Données Economiques. *Bases de Données, Nouvelles Perspectives.* INRIA - ADI, 1983, pages 31-34. Rapport du Groupe BD3.

[EGGE 80] S.J. Eggers and A. Shoshani : Efficient Access of Compressed Data. In : *Proceedings of the International Conference on Very Large Data Bases,* 1980, (Montreal).

[EGGE 81] S.J. Eggers, F. Olken and A. Shoshani : A Compression Technique for Large Statistical Database. In : *Proceedings of the International Conference on Very Large Data Bases,* 1981, (Cannes).

[LITT 83] R.J. Littlefield and P.J. Cowley : Some Statistical Data Base Requirement for the Analysis of Large Data Sets. *Computer Science and Statistics : the Interface.* North-Holland, 1983, pages 24-30.

[McCA 86] J.L. McCarthy : Metadata management for Large Statistical Databases. In : *Proceedings of the International Conference on Very Large Data Bases,* 1982, (Mexico).

[RUHL 82] O. Ruhlmann : AGRISTAT : Banque de Données Socio-Economiques et Statistiques sur l'Agriculture française. *Le Courrier des Statistiques* (24) : 21-24, 1982.

[SHOS 82] A. Shoshani : Statistical Databases : Characteristics, Problems, and some Solutions. In : *Proceedings of the International Conference on Very Large Data Bases,* 1982, (Mexico).

[TURN 79] M.J. Turner, R. Hamond and F. Cotton : A DBMS for large Statistical Databases. In : *Proceedings of the International Conference on Very Large Data Bases,* pages 319-327, 1979.

[WONG 82] H.K.T. Wong and I. Kuo : GUIDE : Graphical User Interface for Database Exploration. In : *Proceedings of the International Conference on Very Large Data Bases,* 1982, (Mexico).