Fourth International Symposium on
Data Analysis and Informatics
Versailles, October 9-11, 1985

Sponsored by
ADI, ANVAR, ASU, CNET, CNRS,
CEA, ISI, INRA, SFC

Organised by
Institut National de Recherche en Informatique
et en Automatique (INRIA)

Scientific Organisation Committee
E. Diday, Y. Escoufier, L. Lebart,
J. P. Pagès, Y. Schektman, R. Tomassone

Scientific Secretariat
G. Celeux, J. J. Daudin,
Y. Lechevallier, C. Perruchet

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO

DATA ANALYSIS AND INFORMATICS, IV

Proceedings of the Fourth International Symposium on
Data Analysis and Informatics,
organised by the Institut National de Recherche en Informatique
et en Automatique,
Versailles, October 9-11, 1985

edited by

E. DIDAY
Université Paris IX - INRIA-Rocquencourt

Y. ESCOUFIER
Université de Montpellier

L. LEBART
Centre National de la Recherche Scientifique - CREDOC, Paris

J. PAGES
Commissariat á l'Energie Atomique, Fontenay-aux-Roses

Y. SCHEKTMAN
Université Paul Sabatier, Toulouse

R. TOMASSONE
Institut National de la Recherche Agronomique - INA-PG, Paris

1986

NORTH-HOLLAND
AMSTERDAM • NEW YORK • OXFORD • TOKYO

ISBN: 0 444 70061 7

Published by:
ELSEVIER SCIENCE PUBLISHERS B.V.
P.O.Box 1991
1000 BZ Amsterdam
The Netherlands

Sole distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
52 Vanderbilt Avenue
New York, N.Y. 10017
U.S.A.

Library of Congress Cataloging-in-Publication Data

International Symposium on Data Analysis and Informatics
(4th : 1985 : Versailles, France)
Data analysis and informatics, IV.

1. Multivariate analysis--Data processing--Congresses.
2. Factor analysis--Data processing--Congresses.
I. Diday, E. II. Institut national de recherche en informatique et en automatique (France) III. Title.
QA278.I56 1985 519.5'35'02854 86-13459
ISBN 0-444-70061-7 (U.S.)

PRINTED IN THE NETHERLANDS

REFEREES

All the papers included in the Proceedings have been reviewed by at least two referees of the following list ;

AGUILAR-MARTIN	J.	FRANCE
BARTKOWIAK	A.	POLAND
BAUFAYS	P.	BELGIUM
BECKER	M.	FRANCE
BELLACICCO	A.	ITALY
BERTRAND	P.	FRANCE
BLUMENTHAL	S.	FRANCE
BOCK	H.H.	F.R.G.
BONNEFOUS	S.	FRANCE
BRENOT	J.	FRANCE
BROSSIER	G.	FRANCE
BRY	F.	F.R.G.
BURTSCHY	B.	FRANCE
CARLIER	A.	FRANCE
CARROLL	J.D.	U.S.A.
CAUSSINUS	H.	FRANCE
CAZES	P.	FRANCE
CELEUX	G.	FRANCE
CHABANON		FRANCE
CHAH	S.	FRANCE
CHANDON	J.-L.	FRANCE
CHARLES	C.	FRANCE
CLOUTIER	E.	CANADA
DAUDIN	J.J.	FRANCE
DAUXOIS	J.	FRANCE
DAY	W.H.	CANADA
DE FALGUEROLLES	A.	FRANCE
DE SOETE	G.	BELGIUM
DELLA RICCIA		ITALY
DER MEGREDITCHIAN	G.	FRANCE
DEVIJVER	P.	BELGIUM
DEVILLE	J.-C.	FRANCE
DIDAY	E.	FRANCE
DIEBOLT		FRANCE
DUBY	C.	FRANCE
EPITALON	J.-M.	FRANCE
ESCOFIER	B.	FRANCE
ESCOUFIER	Y.	FRANCE
FACY	F.	FRANCE
FENELON	J.-P.	FRANCE
FICHET	B.	FRANCE

GHERMANI		FRANCE
GOVAERT	G.	FRANCE
GOWER	J.-C.	U.K.
GUENOCHE	A.	FRANCE
HORBER	E.	SWITZERLAND
JOANNES		FRANCE
JOMIER	C.	FRANCE
JUNCA	S.	FRANCE
KEZOUIT	O.	FRANCE
KLOESGEN	W.	F.R.G.
LAURO	N.	ITALY
LE CALVE	G.	FRANCE
LE FOLL	Y.	FRANCE
LEBART	L.	FRANCE
LECHEVALLIER	Y.	FRANCE
LEFEVRE	F.	FRANCE
LEMAIRE	J.	FRANCE
LEONARD	M.	SWITZERLAND
LEREDDE	H.	FRANCE
LERMAN	I.C.	FRANCE
LIBERT	G.	BELGIUM
LINGOES	J.-C.	U.S.A.
MASSON	J.-P.	FRANCE
MEUNIER	Ph.	BELGIUM
MOLLIERE	J.-L.	FRANCE
MOMIROVIC	K.	YOUGOSLAVIA
MONJARDET	B.	FRANCE
MORINEAU	A.	FRANCE
MURTAGH	F.	F.R.G.
NAKACHE	J.-P.	FRANCE
OK	Y.	FRANCE
PAGES	J.-P.	FRANCE
PERRUCHET	C.	FRANCE
PICARD	J.	FRANCE
RALAMBONDRAINY	H.	FRANCE
RASSON	J.-P.	BELGIUM
RIZZI	A.	ITALY
ROUX	Maurice	FRANCE
ROUX	Michel	FRANCE
SAPORTA	G.	FRANCE
SCHEKTMAN	Y.	FRANCE
TENENHAUS	M.	FRANCE
TERRENOIRE	M.	FRANCE
THAURONT	G.	FRANCE
TOMASSONE	R.	FRANCE
VALETTE	N.	FRANCE
VAN CUTSEM	B.	FRANCE
VILLOING	P.	FRANCE

TABLE OF CONTENTS

CHAPTER 2 – FACTOR ANALYSIS

CHAPTER 3 – DISCRIMINATION

CHAPTER 4 – MULTIDIMENSIONAL SCALING

DATA ANALYSIS AND INFORMATICS, IV
E. Diday et al. (editors)

SPECIFICITES DES S.G.B.D. STATISTIQUES

François BRY (1)

E.C.R.C.
Arabellastrasse, 17
D-8000 MUENCHEN 81
tél.: (089) 92 699 148
télex : 521 69 10

Gérard THAURONT

I.R.T.
2, av du Général Malleret-Joinville
BP 34
F-94114 ARCUEIL Cedex
tél.: (1) 581 12 12
télex : IRT 204 454 F

USEnet : ...mcvax!unido!ecrcvax!francois
CSNet : francois%ecrcvax.UUCP @ germany
ARPAnet : francois%ecrcvax.UUCP @ seismo.ARPA

Abstract

The Particularities of Statistical DBMSs.

The Data Base Management Systems (DBMS) that statistical applications require are very particular. Conventional DBMSs like those presently available are not suitable. The particularities and the needs of the statistical applications are generally ignored by the computer scientists working on databases.

Two somewhat different types of Statistical Databases (SDB) can be distinguished. A SDB is a bank of information if it has been designed primarily for providing statistical informations dealing with the same subject. A SDB is an analysis database if it has been designed to achieve a given statistical analysis. However, all SDBs have in common many properties. Especially update and access methods are different from those of conventional DBs. Also all SDBs need some specific statistical metadata and require some specific storage method taking into account the stability of the stored data.

Résumé

Les statisticiens ont des besoins spécifiques en matière de Systèmes de Gestion de Bases de Données (SGBD). Les SGBD conventionnels que l'on trouve sur le marché ne leur conviennent pas. Or les spécificités de leurs besoins sont mal connues des spécialistes en Bases de Données (BD).

On distingue deux pôles parmi les Bases de Données Statistiques (BDS) : les serveurs d'informations statistiques et les bases d'analyses.

Toutes les BDS présentent des particularités en ce qui concerne les mises-à-jour et les accès aux données. Les langages d'interrogation des SGBD Statistiques (SGBDS) doivent y être adaptés, ils doivent gérer des méta-données particulières aux applications statistiques. Des méthodes de stockage des données propres aux BDS doivent être spécifiées.

---

(1) Travail effectué à l'I.R.T. avant de rejoindre l'E.C.R.C.

Introduction

La conception de Systèmes de Gestion de Base de Données (SGBD) dédiés aux applications statistiques est présentée par les spécialistes en bases de données (BD) comme une des directions de recherche privilégiées pour les années à venir [BD3 83], [Brodie 84]. Les concepteurs de SGBD se sont surtout interessés jusqu'à maintenant à des applications très particulières, (gestion du personnel, gestion bancaire, systèmes de réservation de places, etc...), bien différentes de ce que l'on rencontre en statistique. Les bases ou banques de données statistiques sont telles que les SGBD conventionnels, les seuls dont on dispose actuellement, ne conviennent pas à leur gestion. Or, leurs particularités sont mal connues.

Depuis plus d'une dizaine d'années l'utilisation de systèmes de gestion de bases de données (SGBD) dans les entreprises et les administrations s'étend : l'interêt de ces systèmes n'est plus à démontrer. Les applications statistiques, qui furent l'un des moteurs de l'informatique naissante, produisent et utilisent de nombreuses bases de données. Certaines de ces bases couvrent de longues périodes et ont du mal à survivre aux changements d'ordinateurs et de systèmes d'exploitation. Beaucoup d'entre elles ont une grande importance stratégique, car elles sont à l'origine d'analyses prévisionnelles et décisionnelles. Bien peu sont gérées par des systèmes modernes. Cette situation est paradoxale, car les besoins des statisticiens sont criants. Les bases de données statistiques (BDS) forment une classe d'applications assez "typée" pour que la définition de SGBD dédiés ou SGBDS (Système de Gestion de Bases de Données Statistiques) soit possible.

Cet exposé se propose de dégager les caractéristiques communes aux BDS et de montrer en quoi elles se distinguent des bases de données conventionnelles.

Un premier paragraphe distingue entre outils statistiques et outils de gestion des données. Deux tendances extrêmes complémentaires parmi les BDS sont ensuite décrites.

Un second paragraphe montre les spécificités des accès aux données statistiques. Ceux-ci sont très différents de ceux que connaissent les bases conventionnelles, par exemple de gestion.

Une particularité essentielle des BDS vient des méta-données statistiques qu'elles demandent. Un troisième paragraphe leur est consacré.

Finalement, des propositions d'organisations spécifiques des données statistiques sont faites dans un quatrième paragraphe.

1 Quels systèmes et quelles bases ?

Une première particularité des BDS vient de ce que l'introduction de nouvelles données nécessite généralement des traitements de validation et d'apurement beaucoup plus complexes que pour une base de gestion. Nous n'entrerons pas ici dans le détail de ces traitements qui sont bien connus des statisticiens et qui sont décrits, dans une optique SGBD, dans [Bates 82]. Elles sont souvent réalisées à l'aide d'outils fournis par des paquetages (SAS, BMDP, LEDA, MODULAD, etc...).

Les opérations préalables au chargement de données sont longues, complexes et peuvent demander des retours en arrière. C'est pourquoi presque tous les paquetages offrent des services plus ou moins élaborés de gestion de fichiers. Il est ainsi souvent possible de conserver (dans des "lexiques") les définitions des fichiers (type du fichier, structure des articles, nom des champs, etc...). Certains paquetages permettent des regroupements logiques de fichiers. Ces services de gestion de fichiers, pour utiles qu'ils soient, n'en sont pas moins très différents, et en deçà, de ceux offerts par les SGBD.

L'objectif essentiel des SGBD est de rendre invisible à l'utilisateur d'une BD la structure physique des données, en automatisant les opérations de manipulation de fichiers. Un usager d'une BD bancaire (gérée par un SGBD) pourra par exemple ne connaître que les objets, et les traitements logiques de l'application : compte courant, compte d'épargne, crédit, débit, etc... Il n'a pas à savoir quels sont les fichiers accédés par une de ces opérations. Il ne connaitra pas la structure des articles, ni selon quelles clés les données sont organisées physiquement. Ne pas avoir à connaître l'organisation physique des données dégage l'application de ce qui est sans signification pour elle.

Une BDS gérée par un SGBDS, intégrant des outils statistiques, permettra par exemple de demander l'application de telle méthode d'analyse statistique à tel ensemble d'individus et de variables sans que cet ensemble ne constitue un fichier, plutôt que de construire un fichier, puis d'appliquer tel programme à ce fichier. Bien entendu "l'administrateur" de la base connaitra toujours la structure interne des données.

De nombreux paquetages statistiques tendent depuis peu à fournir des services du types de ceux des SGBD (citons en particulier SAS [SAS 79]). La séparation entre les opérations qui nécessitent la connaissance de la structure interne des données, tel le chargement, et celles qui ne la demandent pas n'est pas clairement faite. L'ajout de services semblables à ceux des SGBD, à des outils par nature différents, ne conduit pas à des SGBD statistiques.

Toute base statistique se situe entre deux pôles que nous appelons "serveur d'informations statistiques" et "base d'analyses".

Le rôle d'un serveur est de rassembler, de pérenniser, et de restituer de l'information dans un domaine particulier. Citons pour la France AGRISTAT, du Ministère de l'Agriculture, [Ruhlman 82], SITRAM, du Ministère des Transports, la Base de Données Macro-économiques (BDM) de l'INSEE. Les principaux serveurs d'informations statistiques des administrations publiques françaises sont présentés dans [Bodin 82].

Un serveur d'information a une durée de vie indéfinie, il est conduit à fournir des données sur de longues périodes (quelques décennies).

Les serveurs ont des volumes de données considérables, de l'ordre de quelques giga-octets (BDM : 2 Go, SITRAM : 8 Go) [Turner 79], [Shoshani 82]. Ce sont de bons candidats pour les disques optiques. Ils ont des volumes très supérieurs à ceux des BD conventionnelles que savent gérer les SGDB disponibles aujourd'hui.

A l'autre pôle, les bases d'analyses sont de tailles et de durées de vie beaucoup plus modestes. Souvent destinées à un seul usager, elles sont construites pour une étude et peuvent, à la limite être implantées sur un micro-ordinateur.

Toutes les BDS, plus ou moins proches de l'un de ces deux pôles, présentent des caractéristiques communes qui les distinguent des BD non statistiques.

2 Les accès aux données statistiques

Une particularité des données statistiques est qu'une fois entrées dans la base elles ne sont pas déchargées. Les données statistiques sont stables. Elles ne connaissent que très rarement des mises-à-jour, et demandent le plus souvent à être conservées dans la base aussi longtemps que possible.

Cette particularité distingue les BDS des BD conventionnelles. La recherche en BD et les développements de systèmes ont tendu, dès le départ, à rendre performantes des modifications très fréquentes de petits ensembles de données. L'exemple classique de système de réservation de d'avion illustre la fréquence possible des mises-à-jour et la taille des données concernées.

Les mises-à-jour que demandent les BDS sont autres : elles ont essentiellement lieu lors de la validation des données arrivant dans la base. Elles ne sont pas ponctuelles mais portent sur de grands ensembles de données : à de très rares exeptions près, si un fichier statistique nécessite des corrections ou des redressements, ils porteront sur un ensemble important d'articles.

Les interrogations que connaissent les BDS diffèrent aussi beaucoup de celles des BD conventionnelles. Les SGBD conventionnels ont été construits avant tout pour effectuer à un moindre coût la recherche d'un petit nombre d'articles. (Ex : état du compte courant de tel client ?) Il est bien rare qu'un statisticien consultant une BDS puisse se contenter d'aussi petits ensembles de données. Ce ne sont en effet pas des unités statistiques isolées qui l'intéressent, mais des ensembles qui peuvent représenter plusieurs millions d'octets. Ces ensembles, afin d'être traités par des logiciels statistiques, doivent souvent être triés. Il est souhaitable qu'un SGBDS réalise au meilleur coût ces lectures et ces tris : des algorithmes différents de ceux utilisés par les SGBD conventionnels sont nécessaires ainsi que des organisations spécifiques de fichiers [Turner 79], [Eggers 80], [Eggers 81], [Shoshani 82].

Différentes familles de "langages d'interrogation" ont été définies, et sont maintenant utilisées de façon classique pour interroger les BD. Ces langages d'interrogation conviennent très mal à la définition d'ensembles statistiques [Anderson 82], [Wong 82]. Il est difficile de faire admettre à un informaticien que même le langage SEQUEL [Chamberlin 74], [Chamberlin 76] (décrit dans [Gardarin 83] de manière didactique) est, malgré ses possibilités de calcul de moyennes, cardinaux, et définition d'ensembles par des fonctions (opérateur "group by"), tout à fait insuffisant pour le statisticien [Anderson 82].

Les consultations de BDS conduisent fréquemment à des constructions successives de "sous-bases" afin d'affiner l'analyse. Ces "bases successives", elles-même complexes, sont liées par des relations logiques complexes. Il est nécessaire qu'un SGBDS apporte des services en ce domaine. Il s'agit là encore d'un trait spécifique aux BDS. Les interrogations et consultations de BD conventionnelles sont généralement beaucoup plus rapides et ne nécessitent pas le stockage de résultats intermédiaires volumineux.

Une bonne connaissance d'une BDS (et tout spécialement de type serveur) est longue à acquérir, pour pouvoir l'utiliser valablement. Une base statistique contient en effet de très nombreux types d'objets, aux définitions généralement plus complexes que les objets des BD conventionnelles [Shoshani 82]. Les modèles de données des

SGBD conventionnels ne permettent pas une description satisfaisante des données statistiques [Chan 81], [Bates 82], [Kreps 82], [Bry 84]. Ce point est approfondi au paragraphe suivant.

Une dernière particularité des accès aux données statistiques est que les ensembles de données issus d'une BDS ne sont pas immédiatement interprétables, comme le sont le plus souvent les résultats d'interrogations d'une BD conventionnelle. Ils sont au contraire traités par des programmes et donnent lieu à des éditions raffinées (tableaux croisés, etc...), des représentations graphiques (courbes, histogrammes, cartes,...) ou à des analyses statistiques (analyse de données). L'intégration entre SGBD dédiés aux applications statistiques et outils d'éditions et d'analyse est nécessaire [Anderson 82], [Bates 82]. Elle permet en effet des optimisations de temps de réponse et autorise une plus grande souplesse dans l'expression des interrogations. Une telle tentative d'intégration se trouve dans le système SICLA-PEPIN [Jomier 84]. Mais les "nomenclatures hiérarchiques" et les "méta-données statistiques" ne sont, à notre connaissance, pas prises en compte dans la version actuelle de ce système.

3 Méta-données statistiques

La distinction entre données proprement dites et méta-données est classique en théorie des bases de données. Si l'usager d'une BD (gérée par un SGBD) peut ignorer l'organisation physique des données, c'est parce qu'elle est décrite aux côtés des données de la base et que cette description est manipulée par le système. On donne le nom de méta-données aux données décrivant la structure de la base (cf. par exemple [Gardarin 83]). Le plus souvent les informations "système" (description des fichiers, des index, des clés, etc...) ne sont pas distinguées de nombreuses informations "sémantiques". C'est à ces dernières, et à elles seules, que ce paragraphe est consacré. Nous proposons, pour les distinguer, de les appeler "méta-données statistiques".

Une BDS demande toujours de très nombreuses, méta-données pour être utilisable [Shoshani 82], [Bates 82]. Cela tient d'une part à la nature même des données statistiques et aux méthodes statistiques qui doivent distinguer de nombreux types de variables (quantitatives, qualitatives, valeurs réelles, redressées, agrégées, estimées, etc...). Cela découle d'autre part des volumes de données stockées dans beaucoup de BDS. Un serveur d'informations statistiques rassemble souvent de nombreuses variables aux définitions voisines mais quelque peu différentes (données d'origines ou de périodes différentes). La connaissance de ces différences est souvent essentielle aux statisticiens qui exploitent la base.

Deux types de méta-données statistiques tout à fait spécifiques aux BDS sont les nomenclatures et les systèmes d'unités.

Les nomenclatures sont attachées aux variables qualitatives. Bien que dans la pratique ce terme ne soit utilisé que dans des cas complexes, nous le "généraliserons" aux cas simples comme la nomenclature sexe à deux modalités : féminin, masculin.

De même que deux variables quantitatives peuvent avoir la même unité, deux variables qualitatives peuvent se référer à la même nomenclature. Par exemple, les deux variables qualitatives lieu de

travail et lieu de résidence vont utiliser une unique nomenclature "lieu".

Une nomenclature peut avoir une structure plus complexe qu'un simple ensemble de modalités ; c'est très souvent une hiérarchie. Dans notre exemple, ce peut être une hiérarchie à trois niveaux : communes, départements, régions. On peut imaginer un SGBDS utilisant des structures encore plus complexes pour gérer les variations des nomenclatures (variations dans le temps et/ou suivant la source des données).

Conformément à l'objectif de cacher les particularités techniques du système pour mieux laisser voir la sémantique de l'application, les nomenclatures doivent exprimer la correspondance entre noms en clair et codes internes, l'usager pouvant alors ignorer ces derniers. Le SGBDS devra offrir des fonctionnalités de regroupements de modalités (avec recodage des données selon ces regroupements), ainsi que de structurations selon ces regroupements par exemple pour l'édition de tris croisés avec des sommes partielles.

Le SGBDS devra également effectuer le rapprochement de statistiques de sources différentes selon une même nomenclature. C'est un cas particulier de ce que l'on appelle une jointure dans le langage des BD relationnelles. Finalement, c'est grâce aux nomenclatures que le langage d'interrogation pourra donner un sens à l'expression : "Les départements de la region Centre".

La gestion des nomenclatures pose des problèmes généralement liés à leurs modifications dans le temps, et pour lesquels un SGBDS doit apporter, sinon une solution, du moins une aide.

Il existe des nomenclatures ouvertes, croissant avec le temps (une maison d'édition faisant des statistiques mensuelles sur des livres aura chaque mois de nouveaux titres).

Les nomenclatures officielles subissent souvent des variations ponctuelles : éclatement d'une modalité (département Corse coupé en deux départements), regroupement (fusion de communes), changement de regroupement (telle île, territoire d'un Etat, devient territoire d'un autre Etat), changement de niveau (telle autre île, territoire d'un Etat devient un Etat indépendant).

De même qu'aux variables qualitatives est attachée la notion de nomenclature, aux variables quantitatives est attachée celle de système d'unité. Les changements d'unités et d'échelles, qu'ils soient linéaires ou non, doivent être pris en charge par le SGBDS, rendant ainsi plus aisée la manipulation de francs courants / francs constants, le passage au logarithme, la conversion d'une mesure de consommation exprimée en litres aux 100 km en une consommation en milles par galon, etc...

Les SGBD de gestion connaissent des "valeurs nulles", qui recouvrent nos deux notions de "données manquantes" et de "valeurs sans signification". Leur prise en compte dans un SGBD pose des problèmes nombreux et complexes [Gallaire 83], que nous n'aborderons pas ici. Notons également que des applications statistiques demandent à distinguer entre "vrai zéro" et "epsilon" (i.e. inférieur à la précision), et qu'il serait souhaitable de pouvoir attacher à des valeurs numériques, des appréciations sur leurs fiabilité (valeur non fiable, estimation de l'erreur).

Les applications graphiques, et en particulier cartographiques, sont courantes en statistique. Il parait souhaitable qu'un SGBDS puisse, soit inclure des fonds de cartes qui sont des méta-données statistiques sur des nomenclatures géographiques, soit communiquer avec une BD cartographique. Les recherches et les développements

sont nombreux dans ce domaine [BD3 83] et les retombées pour les applications statistiques prometteuses. A notre connaissance, aucune étude en ce sens n'a encore été publiée.

Nous terminerons ce paragraphe par un type de méta-données fondamental : les données documentaires.

Une BDS est accompagnée d'une très volumineuse documentation. Il existe certes des SGBD documentaires, mais ils ne sont pas utilisables pour gérer la documentation d'une BDS pour deux raisons fondamentales : la documentation doit pouvoir être liée à tous les aspects de la BDS et les fonctionnalités sont différentes.

Nous ne connaissons aucune analyse de la documentation des BDS et nos reflexions se bornent à quelques recommandations.

Une documentation (des textes) doit pouvoir être attachée à tous les objets manipulés dans la base (données, méta-données, etc...) et à tous les niveaux (d'un texte général sur la base, à une information anecdotique sur tel chiffre).

La documentation attachée à une enquête annuelle varie légèrement chaque année. Des historiques doivent être conservés afin de pouvoir fournir, à la demande, l'état de la documentation pour une année donnée, ou ses variations entre telle et telle année.

Lors d'une interrogation de la base, des avertissements utilisant les méta-données statistiques (dont la documentation) doivent pouvoir être édités automatiquement. Ce peut être un astérisque dans une case de tableau, avec une note en bas de page, lorsque l'on utilise une valeur à laquelle est attachée un avertissement.

4 Structuration des données

Les BDS ont besoin d'un type de redondance assez particulier, découlant de la présence dans les bases de données de "résumées". La pratique en statistique, conduit en effet à manipuler différents ensembles contenant les "mêmes" données à différents niveaux de finesse (par exemple, des effectifs de populations par communes, départements et régions). En outre, la taille énorme des ensembles de données stockés conduit souvent à créer a priori des "fichiers résumés" (i.e. moins "fins" que les fichiers initiaux) afin de minimiser l'encombrement en mémoire centrale, et permettre des temps de réponse acceptables. Un SGBDS doit savoir gérer données initiales et résumées, ainsi que les liens logiques entre ces ensembles. Ce type de redondance est difficile, voire impossible, à prendre en compte par un SGBD conventionnel. Pour une interrogation donnée, l'usager devrait pouvoir se décharger sur le système du choix du résumé adapté à la finesse des résultats souhaités.

Le temps est une variable qui intervient dans pratiquement tous les ensembles de données statistiques. Or le temps ne peut pas être traité, quant à la structure logique des données, comme une autre variable [Delobel 83], [Adiba 85]. C'est en effet selon le temps que de nouveaux ensembles de données s'ajoutent à la base : celle-ci doit donc être "ouverte" selon cette variable. Cette structure logique doit être répercutée au niveau physique.

Le temps pose également des problèmes particuliers de représentation des données : des intervalles variables ne peuvent pas être mémorisés comme des périodes constantes. La gestion de plusieurs calendriers dans la même application est aussi un problème spécifique [Adiba 83].

L'organisation de fichiers d'un SGBD doit favoriser la rapidité des lectures de grands ensembles. Les lectures séquentielles, fréquentes dans les applications statistiques, ne doivent pas être trop coûteuses. Les BDS n'imposent par contre pas certaines des contraintes classiques des BD conventionnelles.

L'importance des volumes de données stockées dans une BDS et les tailles des consultations imposent une organisation physique permettant des compactages pour les ensembles de faible densité (i.e. contenant beaucoup de zéros). Des méthodes raffinées de compactages ont été conçues pour les données statistiques [Eggers 80], [Eggers 81], [Shoshani 82], et des organisations de fichiers particulières ont été proposées [Turner 79], [Burnett 81]. Ajoutons qu'une méthode de stockage économique en espace mémoire a été proposée pour des ensembles dérivés (i.e. résumés ou dérivés par des méthodes d'analyse statistique) [Littelfield 83].

Un SGBD doit s'appuyer sur de telles méthodes.

5 Perspectives

Nous nous sommes efforcés de montrer les principaux traits spécifiques des BDS qui doivent être pris en compte par un SGBDS.

Si de nombreux gestionnaires de BDS ont été developpés ([Kobayashi 82], [McCarthy 82], [Nordbäck 82] et [Ruhlmann 82]), ces systèmes sont généralement spécifiques à une application. Des projets de conception de SGBDS commencent à apparaître. Citons en particulier le système FARANDOLE développé autour de M. Léonard à l'Université de Genève, dans le cadre du club MODULAD. Alors que la conception d'un SGBDS est un problème ouvert, bien peu de définitions de ce que doit être un tel système sont proposées. Bien souvent, les informaticiens spécialistes en bases de données cernent mal les aspects spécifiques des BDS.

Bibliographie

Références

[Adiba 83] M. ADIBA :
La Gestion du Temps dans les SGBD; in "Bases de Données, Nouvelles Perspectives", rapport du Groupe BD3, édité par l'INRIA et l'ADI, 1983, 145-148.

[Adiba 85] M. ADIBA, Q.N. BUI et J. PALAZZO DE OLIVEIRA :
Notion de Temps dans les Bases de Données Généralisées, rapport de recherche TIGRE no 23, Laboratoire de Génie Informatique, IMAG, Université de Grenoble, Janvier 1985.

[Anderson 82] A.J.B. ANDERSON :
Software to Link Database Interrogation and Statistical Analysis; Compte rendu de la Conf.Int. COMPSTAT '82 (toulouse) 1-ère partie, Physica Verlag, 1982, 139-144.

[Bates 82] D. BATES, H. BORAL and D.J. DeWITT :
A Framework for Research in Database Management for Statistical Analysis; Proceedings of the ACM SIGMOD International Conference on the Management of Data, 69-70.

[BD3 83] GROUPE BD3 :
"Bases de Données, Nouvelles Perspectives", rapport du Groupe BD3, édité par l'INRIA et l'ADI, 1983.

[Bodin 82] J.-L. BODIN :
Les Banques de Données dans le Système Statistique Public; Courrier des Statistiques, 18, 1982.

[Brodie 84] M.L. BRODIE :
On the Development of Data Models; in "On Conceptual Modelling", M.L. Brodie, J. Mylopoulos and J.W. Schmidt eds, Springer Verlag, 1984, 19-48.

[Bry 84] F. BRY :
Un Modèle de Données Statistique, rapport interne I.R.T., 1984.

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

[Chamberlin 74] D.D. CHAMBERLIN and R.F. BOYCE :
SEQUEL: A Structured English Query Language, Procceedings of the ACM-SIGMOD Workshop on Data Description Access and Control, 1974.

[Chamberlin 76] D.D. CHAMBERLIN, M.M. ASTRAHAN, K.P. ESWARAN, P.P. GRIFFITHS, R.A. LORIE, J.W. MEHSL, P. REISNER and B.W. WADE :
SEQUEL 2: A Unified Approach to Data Definition, Manipulation and Control, I.B.M. Journal for Research, Vol. 20, No 6, November 1976, pp 560-575.

[Chan 81] P. CHAN and A. SHOSHANI :
SUBJECT: A Directory Driven System for Organizing and Accessing Large Statistical Databases; Proceedings of the International conference on Very Large Data Base (Cannes), 1981.

[Delobel 83] C. DELOBEL :
Les Bases de Données Economiques; in "Bases de Données, Nouvelles Perspectives", rapport du Groupe BD3, édité par l'INRIA et l'ADI, 1983, 31-34.

[Eggers 81] S.J. EGGERS, F. OLKEN and A. SHOSHANI :
A Compression Technique for Large Statistical Database; Proceedings of the International Conference on Very Large Data Bases (Cannes), 1981.

[Eggers 80] S.J. EGGERS and A. SHOSHANI :
Efficient Access of Compressed Data; Proccedings of the International Conference on Very Large Data Bases (Montréal), 1980.

[Gallaire 83] H. GALLAIRE :
Informations Incomplêtes; in "Bases de Données, Nouvelles Perspectives", rapport du Groupe BD3, édité par l'INRIA et l'ADI, 1983, 149-158.

[Gardarin 83] G. GARDARIN :
Bases de Données - les Systèmes et leurs Langages, Editions Eyrolles (Paris), 1983.

[Jomier 84] G. JOMIER, O. KEZOUIT et H. RALAMBONDRAINY :
SICLA-PEPIN: A System Integrating Data Analysis and Relational Data Base Management System, Compte-rendu de la Conf. Int. COMPSTAT '84 (Prague), Physica Verlag, 1984, 263-290.

[Kobayashi 82] Y. KOBAYASHI, K. FUTAGAMI and H. IKEDA :
Implementation of a Statistical Database System: HSDB, Compte-rendu de la Conf. Int. COMPSTAT '82 (Toulouse), 1ère partie, Physica Verlag, 1982, 282-287.

[Kreps 82] P. KREPS :
A Semantic Core Model for Statistical and Scientific Databases; in "A LBL Perspective on Statistical Database Management", Lawrence Berkeley Laboratory (Berkeley), 1982, 129-157.

[Littlefield 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, J.E. Gentle ed., North-Holland, 1983, 24-30.

[McCarthy 82] J.L. McCARTHY :
Metadata Management for Large Statistical Databases; Proceedings of the International Conference on Very Large Data Base (Mexico), 1982.

[Nordbäck 82] L. NORDBACK and A. WIDLUNG :
AXIS - The Manager of Very Large Statistical Databases, Compte-rendu de la Conf. Int. COMPSTAT '82 (Toulouse), Physica Verlag, 1982, 303-204.

[Ruhlmann 82] O. RUHLMANN :
AGRISTAT : Banque de Données Socio-Economiques et Statistiques sur l'Agriculture Française; Courrier des Statistiques, 24, 1982, 21-24.

[SAS 79] SAS Institute Inc. :
SAS User's Guide, Raleigh ed., 1979.

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

[Turner 79] M.J. TURNER, R. HAMOND and F. COTTON :
A DBMS for large Statistical Databases; Proceedings of the International Conference on Very Large Data Bases, 1979, 319-327.

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