La représentation relationnelle des données statistiques: application au traitement des données de panel

Jean-Claude POUPA

Cahiers d'économie et sociologie rurales, n° 26, 1993

Jean-Claude POUPA*

Relational representation of statistical data : an application to panel data

Summary *– The statistical languages and softwares used in economics represent numerical data in multidimensional tables by means of the usual conventions of classical algebra. This logical model, when associated with numerical computations, is not very effective in reorganizing major sets of heterogeneous and complex data, e.g. in harmonizing surveys, monitoring panels, managing multiple nomenclatures, constructing time series, recognizing hierarchical structures, etc.*

The theory of relational algebra makes it possible to define an alternative logical model of representation of multidimensional data which relies on the index notation of the statistician. The data are represented in the form of sets of n-tuples, each element being identified by a single identification "key" which corresponds to the index values for an observation. This model makes it possible, through the use of simple mathematical rules, to formalize cumbersome and apparently complex operations, such as the handling of surveys over a long period or the use of a single nomenclature to deal with data obtained from differing institutional contexts. The reorganisation of the data is carried out by means of the so-called "relational algebra" operations derived from set theory. A "vectorisation" function makes it possible to generate classical data structures that are recognized by statistical languages and softwares.

The application of this theoretical approach raises some practical difficulties with some relational data base management softwares, and this entails analysing the minimum functionalities needed to manage effectively the major statistical data bases. Market softwares appear to have been developed in a high-flow transactional framework, in which a substantial number of users make regular day to day changes from their personal computers. However, the statistician needs a system that is able to extract and reorganize a large flow of factual informations, and to manage the hierarchical relations in these major sets. This paper attempts to define a "benchmark" whose function is to give an account of the behaviour of a system concerning limiting relational calculation operations.

This theoretical analysis is completed by a description of the findings obtained when modelling feed consumption. By means of this example, we are able to make more explicit the methods used in order to carry out matricial calculation with relational algebra operators. To conclude, the simple relational model, as it was defined and used in the seventies is sufficient at the present time to move beyond proprietary hardware and software, and to define solutions to deal with present and future computing configurations in research laboratory environments.

Key-words : *data modelling, statistical data bases, relational calculation, abstract type of data, complexity, benchmark, panel*

La représentation relationnelle des données statistiques : application au traitement des données de panel

Résumé – Les logiciels statistiques représentent les données numériques dans des tables avec les conventions usuelles de l'algèbre classique. Ce modèle logique est peu efficace pour restructurer de grands ensembles de données hétérogènes et complexes : harmonisation d'enquêtes, suivi de panels, gestion de nomenclatures, construction de séries chronologiques, reconnaissance de structures hiérarchisées, etc.

La théorie de l'algèbre relationnelle permet de définir un modèle alternatif de représentation de ces données, qui s'appuie sur la notation indiciaire usuelle. La réorganisation des données s'exprime au moyen d'opérations ensemblistes dites de "calcul relationnel". Une fonction de "vectorisation" permet de générer les tables reconnues par les logiciels statistiques. L'application de ce modèle théorique pose des problèmes avec les logiciels du commerce, ce qui conduit à analyser les fonctionnalités minimales requises pour gérer efficacement les grandes bases statistiques, et à proposer un test d'évaluation des performances.

Mots-clés :
modélisation des données, base de données statistiques, calcul relationnel, type abstrait de données, complexité.

* *Station d'économie et sociologie rurales de l'INRA, 65, rue de Saint-Brieuc, 35042 Rennes cedex.*

Les ressources informatiques nécessaires pour effectuer les traitements statistiques en sciences sociales ont pendant longtemps été fournies exclusivement par les Centres de calcul. Dans le courant des années 80, le micro-ordinateur a pris place à côté du terminal classique. L'accroissement exponentiel de la puissance de calcul offerte et l'explosion de l'offre logicielle ont entraîné un transfert progressif des traitements statistiques usuels vers des équipements individuels. Les grands fonds d'informations statistiques, telles les enquêtes nationales et les statistiques internationales, demeurent néanmoins sur des serveurs centraux. Ils sont régulièrement interrogés pour extraire des jeux de données, transmis sous forme de fichiers pour des traitements locaux.

Les années 90 voient l'arrivée sur le marché des stations de travail et serveurs de laboratoires, qui offrent en local les ressources matérielles et logicielles et la puissance de calcul numérique des grands serveurs classiques. Le coût de la ressource disque décroît régulièrement et il semble désormais possible de disposer des grands fonds d'informations statistiques sur des équipements de laboratoires. Or, cette migration se réalise lentement: la gestion des grandes bases de données statistiques soulève des problèmes spécifiques tant au niveau des volumes que de la complexité (Shoshani, 1982). Les données des grandes enquêtes statistiques restent gérées de façon traditionnelle et les logiciels utilisés, souvent développés sur mesure, ne sont pas portables dans l'environnement des stations de travail et serveurs de laboratoire. Les équipes de recherches en sciences sociales sont de ce fait confrontées à deux informatiques discordantes: bien que l'ensemble des traitements numériques puisse se faire sur des équipements propres, le recours au serveur central demeure indispensable pour extraire les données.

Dans le domaine des sciences économiques, l'intérêt des systèmes de gestion de bases de données relationnelles a été plus récemment évoqué à propos de la gestion des données de panel (David, 1989). Un tel système est décrit comme apportant des fonctionnalités nouvelles significatives: suivi chronologique de vagues d'enquêtes, partage des données, disponibilité d'une méthode de structuration des informations, maîtrise de la complexité, portabilité entre systèmes informatiques ...

L'INRA a engagé en 1990 un programme de recherche lié à la mise en place d'un Observatoire des consommations alimentaires en France[1]. Cet observatoire doit permettre l'étude de la dispersion des

[1] Ce programme de recherche est financé par la Direction générale de la concurrence, de la consommation et de la répression des fraudes (DGCCRF), par la Direction générale de l'alimentation (DGAL), par la Direction générale de la santé (DGS), et par le ministère de la Recherche et de la Technologie (MRT).

consommations ainsi que l'analyse de l'évolution des comportements dans le temps. Pour cela, il doit rassembler les données élémentaires de l'ensemble des enquêtes disponibles. C'est dans cette perspective que l'INRA a élaboré une méthode de gestion des bases de données statistiques. Face aux difficultés rencontrées pour mettre en œuvre une solution efficace avec les systèmes commercialisés, à vocation dite "universelle", nous avons choisi plus prosaïquement d'examiner les représentations mathématiques élémentaires utilisées par les statisticiens et les informaticiens: espaces vectoriels, relations, arbres, treillis, graphes. Ce retour aux définitions s'est poursuivi par la formalisation des règles qui permettent de passer d'une représentation à l'autre. Cet article décrit les méthodes adoptées et propose une traduction concrète, actuellement utilisée par les équipes scientifiques de l'INRA sur un réseau de stations de travail UNIX et de micro-ordinateurs.

La première section de ce document situe la dimension du problème dans le cas des statistiques de consommation alimentaire. La seconde section introduit les fondements mathématiques du modèle relationnel et propose un mode de représentation des vecteurs au moyen de ce modèle. Les fonctions fondamentales qui permettent d'utiliser le modèle relationnel pour "naviguer" dans des espaces vectoriels multiples sont décrites dans la section 3. La section 4 analyse la faisabilité théorique de la solution au moyen de fonctions dites de complexité. Elle se prolonge par la présentation des fonctionnalités minimales nécessaires dans un système de gestion de bases de données statistiques et met en évidence les carences actuelles de l'offre du marché. La dernière section présente les résultats obtenus pour les données de consommation alimentaire. La méthode présentée ici est relative à la structuration des données factuelles en amont des procédures de traitement statistique: elle ne traite pas l'acquisition des données et le chiffrement.

Un premier résultat est la démonstration théorique et pratique du fait que la complexité des données en sciences sociales n'est pas un facteur limitant sur une station de travail d'équipe. Un second résultat est la disponibilité d'une méthode de calcul non numérique pour régler des problèmes jusqu'alors mal résolus: gestion des données de panels, mise en correspondance de sources hétérogènes, gestion de nomenclatures multiples et évolutives, etc. La solution est en outre portable sur toute une gamme de matériel, de la station de travail au serveur central.

LES DONNÉES STATISTIQUES EN SCIENCES SOCIALES: L'EXEMPLE DES ENQUÊTES DE CONSOMMATION ALIMENTAIRE

Le Laboratoire de recherche sur la consommation de l'INRA, en collaboration avec la Division "Conditions de vie des ménages" de l'INSEE, élabore la méthodologie de construction d'une base de données de consommation alimentaire, combinant des données issues de panels et d'enquêtes classiques. L'objectif est d'améliorer l'information statistique disponible. Ces travaux s'appuient sur deux sources principales de données: les enquêtes bisannuelles de l'INSEE et les panels hebdomadaires SECODIP qui sont suivis annuellement.

Description des données

Les enquêtes de consommation alimentaire observent principalement les achats et les repas d'un échantillon de ménages. Un ménage regroupe un ensemble d'individus. Les achats et autres approvisionnements sont notés pour chaque produit pendant sept jours consécutifs, sous la forme d'inscriptions sur les lignes d'un "carnet de comptes". L'INSEE répertorie les repas pris à l'extérieur pour les individus du ménage, pendant la période d'observation. Les caractéristiques des ménages sont établies annuellement. Les enquêtes sont restreintes à une semaine de référence, l'échantillon de base étant éclaté en sous-échantillons observés sur des périodes différentes de l'année. Les panels SECODIP sont suivis sur toutes les semaines d'une période annuelle: on dénombre donc 52 vagues d'observations.

D'un point de vue statistique, la grandeur élémentaire inscrite sur la ligne d'un carnet de comptes est de la forme x_{it}^{pk}. Cette notation désigne l'observation d'une variable x pour la $k^{ième}$ occurrence d'acquisition du produit p, par le ménage i à la période t. Une fourniture élémentaire est en pratique décrite par plusieurs variables: valeur, quantités exprimées avec plusieurs unités, nombre d'articles, ...

Les ménages sont décrits de façon simple par un ensemble de vecteurs définis sur l'espace de l'échantillon: la grandeur notée y_i désigne l'observation de la variable y pour le ménage i. Une suite de m variables sur un échantillon de taille n est représentée dans une matrice A de type $(n \times m)$, le terme a_{ij} désignant la valeur observée de la variable j pour le ménage i.

Dans le cas des panels SECODIP, les ordres de grandeur des intervalles de définition sont les suivants:

$i \in [1, \ldots, 5000]$: échantillon des ménages,

$p \in [1, \ldots, 32000]$: nomenclature détaillée,

$t \in [1, \ldots, 52]$: observations hebdomadaires sur une période annuelle,

$k \in [1, \ldots, 7]$: restriction à un achat quotidien par produit.

Pour l'INSEE, l'indice p est défini sur un intervalle [1,..., 400], l'indice t pour une seule période par ménage. Les nomenclatures INSEE et SECODIP sont élaborées indépendamment l'une de l'autre. Les deux panels SECODIP regroupent respectivement 2 millions et 2,5 millions d'inscriptions élémentaires. Les enquêtes INSEE sont de taille plus modeste avec quelques centaines de milliers d'inscriptions.

La représentation vectorielle

De la statistique descriptive à la modélisation économétrique, les façons de voir les données sont multiples: elles s'appuient cependant sur la représentation vectorielle. Les fichiers d'enquêtes regroupent les valeurs observées pour une entité du monde réel dans une zone de longueur constante, structurée en champs, décrite par un format, appelée enregistrement ou **article**. Les entités de même nature (ménages, achats ...) sont regroupées dans des ensembles homogènes. Avec n articles et m champs par article, un tel ensemble est habituellement représenté par une matrice X de type $(n \times m)$: le terme x_{ij}, toujours défini, désigne la valeur lue dans le $j^{ième}$ champ du $i^{ième}$ article. Les logiciels statistiques utilisés en sciences économiques gèrent les données sous cette forme.

La définition des espaces vectoriels

En se situant au niveau de l'entité ménage, la matrice X contient les vecteurs d'observations associés aux variables descriptives des ménages. Avec l'entité achat, cette représentation est difficilement utilisable par le statisticien: le nombre de lignes par ménage est variable et dépend du nombre d'achats; un ménage qui n'a rien acheté n'est pas représenté. Le problème est de même nature pour l'entité individu, un ménage étant toutefois composé d'au moins un individu. La complexité est maximale dans les enquêtes réelles, qui sont constituées de plusieurs dizaines d'entités réparties dans des générations distinctes[(2)].

[(2)] L'enquête "Consommation alimentaire" de l'INSEE regroupe 17 entités réparties dans 3 générations. L'enquête "Budget des familles" répertorie 33 entités réparties dans 4 générations.

Les relations entre espaces vectoriels

Sur la base de l'information brute acquise sont calculées des variables nouvelles, définies par exemple sur l'espace des ménages: nombre d'individus, nombre d'enfants, nombre d'inscriptions sur le carnet de comptes, valeur totale des achats, moyenne des achats, valeur maximale d'un achat, ... Cette opération, définie d'un espace vectoriel dans un autre, utilise les opérateurs classiques d'agrégation: dénombrement, somme, moyenne, maximum, minimum, ... Elle est appelée **agrégation**.

A l'inverse il peut être utile de disposer, par exemple dans l'espace des individus, de variables qui qualifient le ménage, répétées pour chaque individu d'un même ménage. Cette opération, définie d'un espace vectoriel dans un autre, est appelée **généralisation**.

Les combinaisons linéaires dans un espace vectoriel

Le traitement des données SECODIP au moyen de la nomenclature INSEE nécessite la définition d'une application qui associe aux 32 000 produits SECODIP un code INSEE. La construction des variables descriptives des achats sur l'échantillon des ménages est une combinaison linéaire de vecteurs: le produit p de la nomenclature cible (INSEE) est l'image d'un ensemble de codes de la nomenclature source (SECODIP); la quantité totale achetée est une somme d'achats élémentaires, pondérée en fonction des unités de mesure propres à chaque produit.

La représentation arborescente

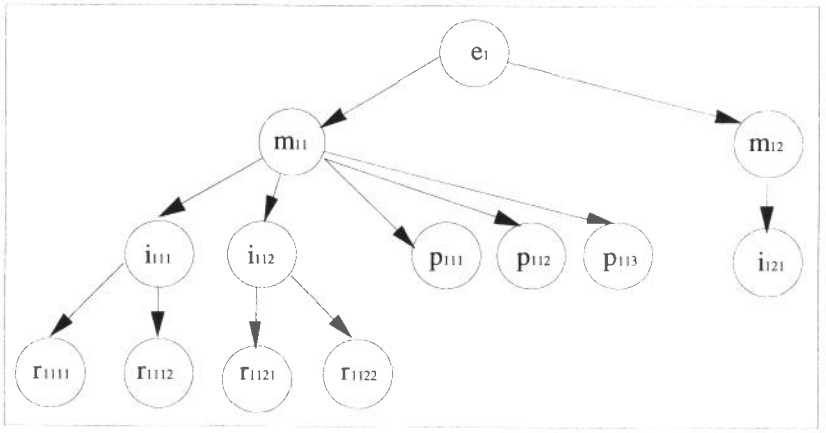
Ce mode de représentation, appelé **modèle de données hiérarchique**, reste très présent sur les sites informatiques centraux pour gérer les grandes enquêtes. Une base de données hiérarchique se représente dans un graphe: les sommets sont les articles, et les arcs représentent des liens de type père-fils. La figure 1 illustre les règles de construction de ce graphe. Dans l'enquête e_1, le ménage m_{11}, qui regroupe deux individus, a effectué trois achats, pour les produits p_{111}, p_{112}, p_{113}. Les individus i_{111} et i_{112} ont consommé chacun deux repas. Le ménage m_{12}, constitué d'un seul individu, n'a pas effectué d'achats.

Le graphe d'une structure de données hiérarchique de ce type est par construction sans circuit et en tout point arrive une flèche et une seule[3]. Le logiciel LEDA de l'INSEE est construit sur la base de ce modèle. Il gère des bases représentées par des graphes dont le nombre de sommets se mesure en millions. L'opération d'agrégation revient à exa-

[3] Cette propriété est propre au modèle hiérarchique et n'est plus vérifiée pour le modèle réseau.

miner les "fils" d'un sommet, et à calculer des statistiques qui "remontent" vers la racine de l'arborescence. Inversement, la généralisation fait "hériter" les fils des caractères du père.

Figure 1.
Représentation d'un graphe dans le modèle hiérarchique



Les limites des solutions classiques

Les représentations hiérarchique et vectorielle ont été valorisées par les équipes de recherche dans le contexte d'une informatique centralisée. Ces solutions ont des limites qui font que certains traitements ne sont pas réalisés du fait de la difficulté, voire l'impossibilité, de restructurer convenablement l'information disponible. Par exemple, bien que la correspondance entre nomenclatures de produits puisse s'exprimer par simple combinaison linéaire de vecteurs, la mise en relation de sources de données hétérogènes au moyen de ces seuls outils apparaît comme une gageure. Ces modèles permettent difficilement de représenter simplement les données de panels sur de vrais échantillons, pour lesquels la composition et la taille varient d'une vague à l'autre, du fait de l'usure et des modifications des caractéristiques socio-professionnelles dans le temps.

LA REPRÉSENTATION RELATIONNELLE DES DONNÉES

Le mode de représentation des données statistiques proposé dans cette section découle naturellement de la notation indiciaire: la $i^{ième}$ occurrence d'une variable x est notée x_i. Dans une enquête simple, la notation x_{ij} désigne la valeur x de la variable numéro j pour l'individu numéro i. Avec

un panel suivi sur T périodes, l'observation élémentaire est notée x^t_{ij}, pour désigner la valeur x de la $j^{ième}$ variable pour l'individu i à la période t. L'ajout de la dimension spatiale introduit un quatrième indice k, pour identifier la région, ce qui donne x^{tk}_{ij}. Ce processus peut se prolonger avec l'introduction d'autres dimensions. La notion de relation permet de redéfinir précisément les bases mathématiques de cette convention de notation.

Définition et propriétés du modèle relationnel

Origine du modèle: les relations

La théorie des ensembles introduit la notion de **relation binaire**. Une relation binaire entre un ensemble E et un ensemble F est une partie du produit cartésien $E \times F$. Dans un espace vectoriel de dimension n, un vecteur peut être décrit comme une relation binaire entre l'ensemble N des entiers naturels et l'ensemble R des réels: c'est un ensemble de couples $\{(i, x) \mid i \in N, x \in R\}$, un élément étant noté x_i. Par définition du produit cartésien, il n'y a pas de couples identiques dans une relation binaire.

Cette notion se généralise à un produit cartésien de p ensembles: une **relation p-aire** entre les ensembles $E_1, E_2, ..., E_p$ est une partie du produit cartésien $E_1 \times E_2 \times ... \times E_p$, p étant le **degré** de la relation. Un élément d'une relation p-aire est appelé **p-uple** ou **p-uplet.** La théorie du modèle relationnel (Codd, 1970) s'appuie sur ces fondements théoriques.

Le modèle relationnel de Codd

Les ensembles de départ, opérandes dans l'expression du produit cartésien, sont appelés **domaines.** Les éléments d'une relation, quel que soit son degré, sont appelés **tuples.** Une **base de données relationnelle** est un ensemble de relations dont les valeurs évoluent dans le temps, par **ajout, modification** ou **destruction** de tuples: la notation $B(t)$ désigne la base à l'instant t.

Les relations sont usuellement visualisées sous forme de tableaux à deux dimensions, auxquels sont associées les propriétés suivantes:

(i) une ligne représente un tuple,

(ii) il n'y a pas de lignes identiques,

(iii) l'ordre des lignes est quelconque,

(iv) chaque colonne est identifiée par un nom, appelé **attribut**, qui la désigne de façon non ambiguë.

Une **clé primaire** désigne un ensemble d'attributs d'une relation dont les valeurs identifient de façon unique tout tuple de la relation. Une **clé étrangère** est un ensemble d'attributs d'une relation dont les éléments sont clé primaire dans une autre relation. Les domaines sont des ensembles quelconques. Un **domaine simple** est un ensemble d'éléments indivisibles, dits **atomiques**: les exemples classiques sont les nombres et les dictionnaires de termes. Un **domaine complexe** est composé d'éléments non atomiques: vecteurs, matrices, liste de termes, phrases, etc... Les éléments d'un domaine complexe sont construits avec des éléments atomiques. La **théorie de la normalisation**, introduite par Codd, décrit un algorithme qui décompose une relation définie sur des domaines complexes en un ensemble de relations définies sur des domaines simples, donc représentables dans des tableaux à deux dimensions. Le processus de normalisation se poursuit par l'application d'algorithmes qui éliminent progressivement les redondances dans la représentation logique de l'information.

Codd définit plusieurs opérations sur l'ensemble des relations. Les opérateurs ensemblistes usuels s'appliquent: **union**, **intersection**, **différence**, **produit cartésien**. Les trois opérations fondamentales sont la **projection**, la **restriction** et la **jointure**. Globalement, la projection supprime des attributs de la relation opérande. La restriction supprime des tuples et rend un sous-ensemble d'une relation. La jointure est une combinaison de deux relations qui possèdent un ou plusieurs domaines communs. Ces opérations sur l'ensemble des relations définissent une algèbre: l'**algèbre relationnelle**.

Représentation des données statistiques dans une base de données relationnelle

L'image tabulaire associée au modèle relationnel suggère naturellement de représenter une enquête simple sous forme d'une relation dans laquelle les variables définissent les attributs et les observations les tuples. Un attribut supplémentaire, clé primaire de la relation, est ajouté pour identifier de façon unique toute observation: à défaut, le modèle supprimerait les observations identiques pour des entités différentes. Les enquêtes simples, non hiérarchisées, peuvent facilement être gérées sous cette forme.

Les grandes enquêtes statistiques décrivent souvent plusieurs dizaines d'entités regroupant plusieurs centaines de variables[(4)]. La manipulation de relations de degré élevé soulève des difficultés pratiques importantes: il y a lieu de nommer explicitement un grand nombre d'attributs répartis

[(4)] L'enquête "Budget de famille" de l'INSEE regroupe, en 1989, 1 035 variables, dont 568 pour l'enregistrement "ménage".

dans plusieurs relations, sans possibilité de paramétrage. Du point de vue des performances, l'optimisation est complexe dans le sens où le nombre de stratégies est quasi infini[5], et le choix de ces stratégies très lié au contexte d'utilisation. Pour de grandes enquêtes, cette forme est inefficace et peu performante, sauf accessoirement sur un sous-ensemble restreint.

Représentation élémentaire des structures numériques multidimensionnelles

Un vecteur V est représentable dans une relation binaire $R_1(\underline{i} : N, x : R)$. Cette notation signifie que l'attribut i est défini sur le domaine des entiers naturels et l'attribut x sur l'ensemble des nombres réels. La relation est un ensemble d'éléments (i, x), un tuple (i, x) désignant la composante x_i du vecteur V. L'attribut i est clé primaire de la relation, ce qui exprime l'unicité de la composante x_i; par convention, les attributs de la clé primaire sont soulignés dans la notation de la relation.

De la même façon, une matrice M peut être représentée dans une relation ternaire $R_2(\underline{i} : N, \underline{j} : N, x : R)$. Un tuple (i, j, x) désigne cette fois le terme x_{ij} de la matrice M. Les attributs i et j forment la clé primaire de la relation, ce qui exprime l'unicité du terme x_{ij}.

Ce mode de représentation se généralise sans difficulté à l'ordre p. Une structure numérique à p dimensions est vue comme une relation de degré $p + 1$, notée $R_p(\underline{d_1} : N, \underline{d_2} : N, ..., \underline{d_p} : N, x : R)$. Les attributs d_1, d_2, ..., d_p sont définis sur le domaine des entiers naturels et forment la clé primaire de la relation.

Représentation des données statistiques dans les relations

Dans l'univers du calcul numérique, toutes les composantes d'un vecteur sont définies. La prise en compte du concept de donnée manquante nécessite la définition d'une variable auxiliaire pour réaliser les traitements statistiques. Dans ces conditions, on peut choisir de ne pas représenter les valeurs nulles: la composante x_i du vecteur V aura pour valeur celle de l'attribut x si le tuple (i, x) est présent dans la relation R_1, sinon zéro. Avec cette convention, une matrice diagonale de dimension n est représentée dans une relation contenant au plus n éléments, et non n^2 éléments. Dans l'ensemble des achats en valeur pour un panel SECODIP, la relation 5-aire $A(\underline{i} : N, \underline{p} : N, \underline{t} : N, \underline{k} : N, x : R)$ contient quelque 2,5

[5] Pour une relation n-aire, il est possible de décrire n relations d'ordre total sur une base à un instant donné. Avec $n = 10$, on dénombre $10! = 3\ 628\ 800$ façons d'indexer la relation.

millions d'éléments, au lieu de 36,4 milliards s'il fallait représenter les valeurs nulles.

Cette convention traduit fidèlement le processus réel d'observation: si le ménage i n'a pas acheté le produit p, on n'introduit pas un enregistrement spécifique dans le fichier d'enquête. La valeur nulle est introduite ultérieurement pour les besoins de la représentation vectorielle, dans les procédures de traitement statistique. S'il est nécessaire de représenter explicitement le fait qu'un ménage i n'a pas acheté un produit p à la période t, une solution alternative est de créer la relation $B(\underline{i} : N, \underline{p} : N, \underline{t} : N)$ qui contient les triplets (i, p, t) vérifiant cette propriété.

Application à la représentation relationnelle des données d'enquêtes

La notation indiciaire x_i désigne l'élément de rang i dans une suite de n objets notée $(x_1, x_2, ..., x_i, ..., x_n)$. Cet élément fait référence dans une enquête à une entité du monde réel, par exemple un ménage dans un échantillon, repéré par un identifiant unique. Soit m_i l'identifiant du ménage de rang i. Un vecteur d'observations d'une variable de l'échantillon peut être décrit comme l'ensemble des couples (m_i, x_i) d'une relation $S_1(\underline{m} : A_1, x : \mathrm{R})$. De façon analogue, si p_j est l'identifiant de la variable numéro j, l'ensemble des triplets (m_i, p_j, x_{ij}) de la relation $S_2(\underline{m} : A_1, \underline{p} : A_2, x : R)$ représente une matrice.

Pratiquement, les domaines de définition des identifiants, A_1 et A_2, sont confondus avec l'ensemble N des entiers naturels. La codification utilise des tables de correspondances avec des lexiques de termes qui sont des relations: ce volet, qui relève des applications classiques de l'algèbre relationnelle, n'est pas traité dans cet article.

Exemple

Soit un ensemble de 5 ménages $\{m_1, m_2, m_3, m_4, m_5\}$ pour lesquels sont observés les achats en quantité de 3 produits $\{p_1, p_2, p_3\}$ sur 2 périodes $\{t_1, t_2\}$. La figure 2 traduit la représentation matricielle usuelle. La figure 3 illustre une représentation relationnelle sous forme d'une image tabulaire, avec les attributs ménage, produit, période et valeur: l'ordre des lignes est quelconque. Les ensembles d'identifiants R_1, R_2 et R_3 sont des relations à un seul attribut, dites unaires. La relation R_4 contient les valeurs observées pour chaque triplet d'identifiants.

Figure 2. Représentation matricielle d'un tableau tridimensionnel

	p1	p2	p3
m1	0	0	1000
m2	0	1000	0
m3	0	500	0
m4	2000	0	0
m5	0	0	100

période *t1*

	p1	p2	p3
m1	500	0	0
m2	0	0	1000
m3	500	0	500
m4	0	1000	0
m5	0	500	0

période *t2*

Figure 3. Représentation relationnelle d'un tableau tridimensionnel

ménage
m1
m2
m3
m4
m5

relation R_1

produit
p1
p2
p3

relation R_2

période
t1
t2

relation R_3

ménage	produit	période	valeur
m1	p3	t1	1000
m2	p2	t1	1000
m3	p2	t1	500
m 4	p1	t1	2000
m5	p3	t1	1000
m1	p1	t2	500
m2	p3	t2	1000
m3	p1	t2	500
m3	p3	t2	500
m4	p2	t2	1000
m5	p2	t2	500

relation R_4

Dynamique de la base

Les équipes scientifiques accèdent à une base de données statique qui contient un fonds d'informations factuelles à une période donnée. Cette base évolue par ajout de vagues successives, sous forme de flots de données apurées. La périodicité des mises à jour est souvent annuelle.

L'accès à cette base factuelle pour les utilisateurs n'est autorisé qu'en lecture: les équipes extraient les données pour des traitements statistiques, ou construisent des sous-bases pour leurs besoins propres. Les méthodes d'accès sont présentées dans les sections suivantes.

ALGÈBRE RELATIONNELLE ET ESPACES VECTORIELS

Le calcul scientifique en sciences économiques s'effectue traditionnellement sur des espaces vectoriels. Le modèle relationnel est une représentation logique de l'information étrangère à l'univers du statisticien. Il permet cependant de résoudre une série de problèmes insolubles avec les méthodes classiques de l'algèbre linéaire. Cette section décrit des fonctions élémentaires, définies sur des ensembles de relations, avec comme ensembles d'arrivée des espaces vectoriels.

La fonction de vectorisation

La fonction de vectorisation admet comme ensemble de départ une base de données relationnelle et rend une suite de vecteurs dans un espace vectoriel. Les variables sont des relations qui possèdent une clé primaire unique, notée en soulignant les attributs qui la composent.

Restitution d'une matrice

Soit un échantillon M de n ménages $\{m_1, m_2, \ldots, m_i, \ldots, m_n\}$ et un ensemble P de q variables $\{p_1, p_2, \ldots, p_j, \ldots, p_q\}$ qui désignent q produits répertoriés dans une nomenclature. L'ensemble des achats de ces produits par les ménages, mesurés avec la variable x (valeur des achats par exemple), est représenté par le groupe de trois relations (I). Le couple (m, p) est clé primaire de la relation C; l'attribut m, clé étrangère dans C, est clé primaire de M; l'attribut p, clé étrangère dans C, est clé primaire de P.

$$\begin{aligned} &M(\underline{m} : N), \\ &P(\underline{p} : N), \qquad\qquad (I) \\ &C(\underline{m} : N, \underline{p} : N, x : R). \end{aligned}$$

La fonction de vectorisation $V(M, P, C)$, définie de la base de données vers l'espace vectoriel de l'échantillon, admet comme paramètres trois relations et rend une matrice X construite en appliquant la règle suivante:

$$\forall i, \forall j \text{ si } (m_i, p_j, x) \in C \text{ alors } x_{ij} = x \text{ sinon } x_{ij} = 0$$

La propriété d'unicité des clés primaires fait que la définition d'une telle matrice est **complète** et **non ambiguë**: tous les termes sont définis et il n'y a pas plusieurs candidats à l'intersection de la ligne i et de la colonne j.

On remarquera que la fonction $V(P, M, C)$ rend la matrice X', transposée de X, vue comme une suite de vecteurs dans l'espace des produits.

Restitution d'un panel sous forme d'une suite de matrices

L'ajout d'une dimension temporelle s'effectue en définissant un ensemble T de r périodes $\{t_1, t_2, \ldots, t_k, \ldots, t_r\}$. Le panel des achats est alors représenté par le groupe de quatre relations (II). Le triplet (m, p, t) est clé primaire de la relation C. Les attributs m, p et t, clés étrangères dans C, sont respectivement clés primaires des relations M, P et T.

$$\begin{array}{|l|r|}\hline M(\underline{m} : N), & \\ P(\underline{p} : N), & \text{(II)} \\ T(\underline{t} : N), & \\ C(\underline{m} : N, \underline{p} : N, \underline{t} : N, x : R). & \\ \hline\end{array}$$

Il existe plusieurs modes de vectorisation d'une telle base. Si l'analyse économique traite des vecteurs de ménages sur un échantillon constant de taille n, le panel est décrit comme une suite de r matrices X_k. Pour générer ces matrices, il suffit de définir dans la base la relation supplémentaire $J(\underline{p} : N, \underline{t} : N)$, produit cartésien de relations unaires P et T. Le couple (p, t) est clé primaire de la relation J, qui décrit les colonnes de la matrice globale. La fonction de vectorisation associée $V(M, J, C)$, définie de la base de données vers l'espace vectoriel de l'échantillon, rend une suite de matrices X_k construites en appliquant la règle suivante :

$$\forall i, \forall j, \forall k \text{ si } (m_i, p_j, t_k, x) \in C \text{ alors } x^k_{ij} = x \text{ sinon } x^k_{ij} = 0$$

Si l'analyse économique traite maintenant des séries temporelles individuelles sur une durée de r périodes, le panel est vu comme une suite de n matrices X_i. Les colonnes de la matrice globale sont alors décrites par la relation $J(\underline{m} : N, \underline{p} : N)$. La fonction de vectorisation associée $V(T, J, C)$, définie de la base de données vers l'espace vectoriel des périodes, rend une suite de matrices X_i construites en appliquant la règle suivante :

$$\forall i, \forall j, \forall k \text{ si } (m_i, p_j, t_k, x) \in C \text{ alors } x^i_{kj} = x \text{ sinon } x^i_{kj} = 0$$

On remarquera qu'il est ainsi possible de définir six fonctions de vectorisation, par permutation des trois indices utilisés.

Restitution d'une matrice structurée en blocs

La généralisation des enquêtes à des groupes de pays, par exemple la Communauté économique européenne, introduit une dimension spatiale supplémentaire, décrite par un ensemble S de u régions $\{s_1, s_2, \ldots, s_l, \ldots, s_u\}$. L'ensemble des panels des achats est représenté par un groupe de cinq relations (III). Le 4-uplet (m, s, p, t) est clé primaire de la relation C. Les attributs m, s, p et t, clés étrangères dans C, sont respectivement clés primaires des relations M, S, P, T.

$$
\begin{array}{|l|r|}
\hline
M(\underline{m} : N), & \\
S(\underline{s} : N), & \\
P(\underline{p} : N), & \text{(III)} \\
T(\underline{t} : N), & \\
C(\underline{m} : N, \underline{s} : N, \underline{p} : N, \underline{t} : N, x : k). & \\
\hline
\end{array}
$$

Si les vagues sont regroupées en lignes et les régions en colonnes , il est nécessaire de définir deux nouvelles relations $I(\underline{m} : N , \underline{s} : N)$ et $J(\underline{p} : N, \underline{t} : N)$, qui décrivent respectivement les lignes et les colonnes de la matrice globale. La fonction de vectorisation associée $V(I, J, C)$, définie sur la base de données, rend une matrice structurée en blocs en appliquant la règle suivante:

$$\forall i, \forall l, \forall j, \forall k \text{ si } (m_i, s_l, p_j, t_k, x) \in C \text{ alors } x_{ij}^{kl} = x \text{ sinon } x_{ij}^{kl} = 0$$

On remarquera que par permutation des quatre indices, on dénombre 24 fonctions de vectorisation.

Généralisation

Dans les exemples précédents, la relation globale C contient un seul attribut, x,qui n'appartient pas à la clé primaire. Cet attribut désigne les observations élémentaires de la variable X, par exemple des achats en valeur. En introduisant une nouvelle variable Y (par exemple les achats mesurés en volume), le panel géographique précédent est représenté dans la relation $C(\underline{m} : N, \underline{s} : N, \underline{p} : N, \underline{t} : N, x : R, y : R)$. La fonction de vectorisation $V(I, J, C)$ rend deux matrices structurées en blocs, respectivement pour les variables X et Y. Les attributs x et y de la relation C sont dits en **dépendance fonctionnelle totale** des attributs de la clé primaire: cette notion exprime simplement dans ce cas de figure la fonction explicitée par la notation indiciaire du statisticien, qui associe à chaque combinaison d'indices une valeur de variable.

De façon générale, la fonction de vectorisation fait correspondre à chaque attribut qui n'appartient pas à la clé primaire et dépend totalement de cette clé une matrice. Les attributs de la clé primaire sont répartis dans deux relations I et J, qui décrivent respectivement les lignes et les colonnes des matrices. Les algorithmes de vectorisation associés à ces structures et en cours de développement s'appuient sur ces formes génériques.

Gestion des classes d'équivalence

Les vagues des panels sont définies sur des échantillons constants. Implicitement, cela signifie que les ménages enquêtés sont équivalents d'une vague à l'autre. La notion d'équivalence est à définir. Un échantillon, choisi à l'instant t_0, évolue dans le temps: il est difficile de traiter le couple sans enfant de l'année t_0 comme le même individu statistique que la famille nombreuse de l'année t_k. La modification des caractéristiques et du comportement des ménages pérennes s'accompagne d'un processus d'érosion par disparition de ménages. Il faut aussi tenir compte des évolutions de la population étudiée. Cette composante dynamique est prise en compte par la définition de classes d'équivalence: les ménages sont des entités statistiques qui représentent des groupes, et sont choisis au moyen de techniques d'échantillonnage. Il faut alors gérer les correspondances entre vagues, par un mécanisme d'identification des observations.

Un second problème de même nature est lié à la définition des variables. Dans le cas de la consommation alimentaire, l'ensemble des produits évolue dans le temps. Les nomenclatures sont hétérogènes: elles sont élaborées en fonction des objectifs des institutions qui produisent les enquêtes, et définies pour une période donnée. Il faut pouvoir passer d'une nomenclature à l'autre, en fonction de la vague.

Construction des classes d'équivalence

La démarche est explicitée sur l'exemple de l'exploitation des données SECODIP, décrite dans une nomenclature de 32 000 éléments, au moyen de la nomenclature INSEE qui contient moins de 400 éléments.

Soit l'ensemble P des u produits SECODIP $\{p_1, p_2, ..., p_u\}$ et l'ensemble Q des v produits INSEE $\{q_1, q_2,..., q_v\}$. On définit une application f de P dans Q qui associe à tout produit SECODIP un code INSEE unique. Un produit INSEE est généralement l'image de plusieurs produits SECODIP, éventuellement d'aucun. En conséquence, il n'est pas possible de définir une application inverse qui décrirait les produits INSEE à travers la nomenclature SECODIP. L'application f est représen-

tée dans la relation binaire F ($\underline{p}$: N, q : N), qui contient tous les couples (p_i, q_j) tels que $f(p_i) = q_j$. L'attribut p est clé primaire de la relation F.

Représentation des relations d'équivalence

A partir des relations $C(\underline{m} : N, \underline{p} : N, \underline{t} : N, x : R)$ et $F(\underline{p} : N, q : N)$, on définit la relation $D(\underline{m} : N, \underline{p} : N, q : N, \underline{t} : N, x : R)$ comme une application de $N^3 \times R$ dans $N^4 \times R$ qui associe à tout tuple (m_i, p_j, t_k, x) de C le tuple (m_i, p_j, $f(p_j)$, t_k, x) de D.

Cette application se calcule en algèbre relationnelle par une jointure des relations C et F avec l'attribut de jointure p. Elle est totalement définie seulement si toute valeur de l'attribut p de la relation C est présente dans la relation F. Cette propriété est exprimée par le fait que l'attribut p est clé étrangère de la relation C et clé primaire de la relation F: elle est usuellement appelée **intégrité de référence**.

Cette transformation de la base introduit une redondance dans la relation D: l'attribut q est déterminé par la connaissance de l'attribut p. La normalisation supprime habituellement cette redondance en décomposant la relation D en deux relations, qui sont les relations C et F initiales. Nous introduisons donc un processus de "dénormalisation", justifié par le besoin de rapprocher le modèle logique du besoin applicatif.

Les regroupements

Le processus de vectorisation décrit plus haut ne peut pas s'appliquer sur la relation D avec comme paramètre la nomenclature Q, l'attribut q n'appartenant pas à la clé primaire de D. Il faut construire une autre relation $E(\underline{m} : N, \underline{q} : N, \underline{t} : N, x : R)$ admettant le triplet (m, q, t) comme clé primaire. L'objectif étant de générer les vecteurs d'achats dans la nomenclature Q, il s'agit d'exprimer la somme des vecteurs élémentaires qui composent un produit q. L'algèbre relationnelle permet de réaliser cette transformation. Une opération de projection supprime l'attribut p dans la relation D et regroupe dans l'attribut a l'ensemble des valeurs de l'attribut x associées à une même valeur du triplet (m, q, t). On obtient la relation $W(\underline{m} : N, \underline{q} : N, \underline{t} : N, a : P(R))$. La notation $P(R)$ désigne l'ensemble des parties de R, donc un domaine complexe au sens de Codd.

L'attribut a contenant des valeurs non atomiques, on dit que la relation W n'est pas en **première forme normale**. Soit une fonction ϕ, dite d'agrégation, définie de $P(R)$ vers R. La relation E, normalisée, se déduit de la relation W en associant à tout tuple (m_i, q_j, t_k, a) de W le tuple (m_i, q_j, t_k, $\phi(a)$) de E. Pour exprimer la somme des valeurs des achats

$\{x_1, x_2, ...x_i, ...x_k\}$, on choisit la fonction $\phi(a) = \sum_{i=1}^{k} x_i$. Pour compter le nombre d'achats élémentaires associé à tout triplet (m, q, t), on utilise la fonction $\phi(a)$ = Card(a). Les autres fonctions d'agrégation classiques sont la moyenne, le maximum et le minimum: elles sont appelées **fonctions ensemblistes.**

SPÉCIFICATION D'UN SYSTÈME DE GESTION DE BASES DE DONNÉES STATISTIQUES

Les sections précédentes décrivent, d'un point de vue théorique, des structures de données et des opérations élémentaires sur ces structures. La question immédiate qui se pose est de savoir s'il est possible de traduire efficacement cette construction logique sur une machine: est-ce calculable et avec quelles ressources? Ce point est abordé en introduisant quelques notions empruntées à la **théorie de la calculabilité**, qui permettent d'évaluer les ressources informatiques nécessaires en fonction de la "taille" du problème.

A l'issue de cette analyse de faisabilité se pose la question du comment. Concrètement, il ne suffit pas de produire une description logique pour une machine abstraite, mais il faut s'assurer que la solution peut être traduite sur une machine réelle. Ce constat conduit à évoquer aussi succintement que possible des contraintes matérielles liées à l'organisation physique des données en mémoire secondaire, dans l'état actuel des technologies commercialisées. Nous examinons enfin ce qu'apportent les langages disponibles, et dans quelle mesure ils permettent de réaliser efficacement les calculs requis pour restituer simplement les structures numériques utilisées par le statisticien.

Calculabilité et complexité

Les consommations de ressources informatiques sont mesurées pour l'essentiel à travers les temps d'exécution, les opérations d'entrées-sorties et l'espace disque occupé. Les deux premiers facteurs sont limités par le temps. La ressource disque est finie et mesurée en octets. Ces consommations sont souvent connues a posteriori. **Les fonctions de complexité** introduisent une méthode d'évaluation des ressources nécessaires : ces méthodes sont décrites dans des manuels récents (Wolper, 1991 ; Beauquier *et al.*, 1992).

Volumétrie

La complexité d'un problème est perçue intuitivement comme liée à la taille des objets manipulés. Avec la représentation proposée plus haut, les cardinalités dénombrent indirectement les informations atomiques effectivement collectées au cours des enquêtes : elles sont bornées et n'excèdent pas en pratique les quelques millions.

S'il fallait représenter simplement les tuples dans des structures de données gérées par un langage de programmation classique[6], il faudrait prévoir un espace de cent mégaoctets pour stocker une relation de dix millions de lignes, soit cent variables pour dix enquêtes annuelles sur un échantillon de dix mille ménages, ou encore dix millions d'achats élémentaires. Dans l'état actuel des technologies disponibles, la ressource disque requise apparaît alors comme relativement modeste.

Classement

Les données sont physiquement représentées en mémoire sous forme d'une séquence d'informations élémentaires contiguës, qui traduit une relation d'ordre implicite. Cette remarque suggère de faire en sorte que cet ordre soit celui des numérotations canoniques usuelles dans les représentations indiciaires, le coût de stockage d'un ensemble ordonné étant identique à celui d'un ensemble sans structure.

Cette organisation simple permet à la fois de transmettre des données sous forme d'un flot ordonné et d'extraire des valeurs élémentaires au moyen d'une recherche dichotomique qui utilise la relation d'ordre existante. Le problème de sa mise en œuvre est abordé plus loin dans la discussion du choix du modèle physique.

Evaluation théorique des temps d'exécution

L'estimation du temps d'exécution d'un programme est basée sur l'étude aux limites d'une fonction $F(n)$ qui exprime le nombre d'instructions élémentaires exécutées en fonction de la "taille" des données n. Soit deux relations de n éléments notées de façon simplifiée $A_1(\underline{i}, \underline{j}, \underline{k}, x)$ et $A_2(\underline{i}, \underline{j}, \underline{k}, y)$. Le triplet (i, j, k) est clé primaire dans chacune des relations. On se propose de construire la relation $B(\underline{i}, \underline{j}, \underline{k}, x, y)$, ce qui s'exprime par une jointure naturelle sur la clé primaire en calcul relationnel.

Soit un premier algorithme P_1 qui ignore l'existence d'une relation d'ordre et l'unicité de la clé primaire. Pour chaque élément de A_1, il examine tous les éléments de A_2, et exécute n^2 comparaisons. Avec $n = 10^6$,

[6] Les indices peuvent être codés sur deux octets, les valeurs sur quatre octets.

on aboutit à 10^{12} tests. Sur une machine hypothétique qui exécuterait une comparaison en 10^{-6} seconde, le temps d'exécution serait de 278 heures. Avec $n = 10^3$, le délai de réponse serait néanmoins instantané. On dit que l'algorithme P_1 est de complexité $O(n^2)$, la notation O() signifiant que le temps d'exécution est une fonction de n^2. Ce formalisme permet de s'abstraire des machines utilisées.

Soit un second algorithme P_2 qui exécute les tris sur les relations opérandes puis opère la fusion. Il existe plusieurs algorithmes de tri classiques de complexité $O(n \log_2 n)$. La fusion étant de complexité $O(n)$, avec $n = 10^6$ ($n \approx 2^{20}$), l'algorithme P_2 s'exécuterait en 41 secondes sur la machine hypothétique précédente.

Supposons maintenant que la cardinalité des ensembles soit 16 fois plus élevée ($n = 2^{24}$). Pour une opération binaire, le temps d'exécution de l'algorithme P_1 serait multiplié par 256 (ce qui ferait plus de huit années de calcul !), alors que le temps d'exécution de P_2 serait multiplié par 18 (soit un temps final de l'ordre de 13 minutes).

Evaluation expérimentale des temps d'exécution : simulations d'applications

Dès lors que l'on manipule de grands ensembles de données, la puissance de calcul relationnel est liée à l'efficacité des algorithmes de tri. Les algorithmes de tri de complexité polynomiale ($O(n^p)$ avec $p \geq 2$) sont inutilisables dans ce contexte. Parmi les algorithmes connus de complexité $O(n \log_2 n)$, le choix de l'algorithme optimal passe par l'évaluation du nombre d'instructions élémentaires exécutées pour réaliser le traitement associé à une comparaison : une analyse très fine dénombre les cycles de calcul de la machine, mesurés en nanosecondes[7].

Une évaluation théorique de type analytique nécessite la connaissance des algorithmes utilisés, voire du code et des techniques de compilation. L'étude des fonctions de complexité peut aussi se faire expérimentalement en réalisant des simulations, usuellement appelées *benchmarks*. Ces simulations représentent l'activité d'un utilisateur qui fait du calcul relationnel pour extraire un flot d'informations ordonnées, par exemple une matrice pour le statisticien, sans devoir au préalable expliciter les méthodes d'accès à travers des mécanismes complexes d'indexation et de regroupement.

[7] Un gain de temps important serait obtenu sur une machine qui disposerait d'un ou plusieurs processeurs spécialisés de tri.

Résultats expérimentaux

Ce test a été réalisé sur une station de travail Sun Sparc 1+ avec le logiciel INGRES (version 6.4). Il simule un utilisateur qui effectue la somme de deux matrices carrées de type (p, p), représentées dans des relations de la forme $M(i : N, j : N, x : R)$. Les attributs i et j contiennent les valeurs des indices, c'est-à-dire un élément de l'ensemble des p premiers nombres[8]. Les valeurs de p sont les puissances successives de $2(p = 2^k)$. L'attribut x contient une valeur réelle quelconque: nous avons retenu sur cet exemple des combinaisons arithmétiques d'indices, ij pour le premier opérande et i/j pour le second. Les relations sont traitées dans ce test comme des ensembles non ordonnés. Le tableau 1 présente quelques résultats, ainsi que les valeurs estimées par extrapolation d'une fonction de complexité $O(n \log_2 n)$.

Tableau 1.
Résultats d'une simulation destinée à mesurer les performances d'un système de calcul relationnel

Dimension de la matrice (p)	256	512	1024	2048	4096
Nombre de termes (n)	65 536	262 144	1 048 576	4 194 304	16 777 216
Temps d'exécution observé (en secondes)	115	508	2 088	8 654	35 282
Temps d'exécution estimé (en secondes)	–	473	2 088	8 561	35 595

Le choix du modèle physique

D'un point de vue technique, les données sont stockées en mémoire secondaire sous forme de blocs indivisibles rangés séquentiellement sur les pistes physiques d'un disque. Relativement au temps d'exécution d'une instruction en mémoire principale, le temps moyen d'accès à un bloc est infiniment grand: le premier s'exprime en nanosecondes et le second en millisecondes, soit un rapport de l'unité au million. Lorsque le processus est initialisé, le débit du transfert de blocs contigus vers la mémoire principale est de l'ordre du mégaoctet par seconde. La performance d'un système pour une application est très liée à son aptitude à prendre en compte ces contraintes technologiques.

[8] Le principe de génération automatique de ces ensembles $E_0, E_1, \ldots, E_k$ est basé sur un enchaînement de produits cartésiens selon la règle $E_{i+1} = E_i \times \{0,1\}$, le processus étant initialisé avec $E_0 = \{0,1\}$.

Relation d'ordre et clé primaire

Dans les bases de données statistiques, les relations d'ordre sont des éléments structuraux fondamentaux. Les observations atomiques sont les composantes de vecteurs, regroupés en matrices. L'introduction d'indices supplémentaires définit des suites de matrices, éventuellement des suites de suites de matrices... Les indices sont hiérarchisés et un ordre canonique d'exploration est imposé: la variable j est observée à la période k pour l'individu i.

La notion de relation d'ordre est totalement étrangère au modèle relationnel qui se limite à définir une relation comme un ensemble sans structure. Elle est introduite ultérieurement à travers l'indexation, pour optimiser les temps d'accès. La notion de clé primaire permet d'identifier les tuples d'une relation mais n'introduit pas de relation d'ordre (sauf implicitement lorsque la clé est composée d'un attribut numérique unique).

Spécification du modèle physique

La relation d'ordre du statisticien doit impérativement être conservée par le modèle physique. L'accès à une matrice X_j nécessite alors un simple positionnement en début de zone puis le transfert d'un flot de blocs contigus qui ne contiennent que les tuples utiles. Si cette relation d'ordre est perdue, les tuples sont éparpillés dans les blocs, ce qui multiplie les accès pour rapatrier des blocs isolés qui ne contiennent qu'une partie des tuples recherchés. Si la ressource en mémoire principale est restreinte, une telle situation détériore considérablement les performances. La définition d'un index logique dans le langage d'interrogation de la base régénère la relation d'ordre du point de vue de l'utilisateur, mais pas nécessairement au niveau du modèle physique.

On tiendra compte du fait que les données sont acquises en flots et non modifiées. Vu la cardinalité des relations, il est utile de pouvoir choisir une représentation physique qui minimise l'espace occupé par un tuple et l'ajuste rigoureusement aux domaines de définition.

Les besoins linguistiques

Codd a énoncé des recommandations pour construire un langage universel de gestion des données (Codd, 1970). Ses propositions s'appuient sur une axiomatique rigoureuse (la logique des prédicats du premier ordre), et esquissent le profil d'un langage dont la finalité est de formaliser les questions complexes indépendamment des aspects d'exécution. Ce langage, noté R, permet de déclarer les relations et leurs domaines, autorise les ajouts, suppressions et modifications d'éléments de relations,

offre la possibilité de gérer les clés primaires et étrangères... Spécifiquement orienté vers les données, le langage R est conçu pour être incorporé *(embedded)* dans des environnements de programmation diversifiés, associés à des langages hôtes, notés H.

Le langage SQL

Parmi un ensemble de langages développés autour du modèle relationnel, le langage SQL *(Structured Query Language)*, d'origine IBM, s'est imposé comme norme de fait. Son succès commercial est sans doute dû à la simplicité et à la puissance du langage d'interrogation de la base. La requête de restriction est en mesure de qualifier de façon très concise des questions complexes, au moyen d'expressions booléennes avec les opérateurs logiques et arithmétiques usuels et les quantificateurs "il existe" ($\exists$) et "quel que soit" ($\forall$).

L'efficacité et la complétude sont maximales sur les domaines numériques. Les dialectes SQL proposent des opérations de manipulation de chaînes de caractères. Le langage permet de définir les relations et d'exprimer les opérations usuelles de l'algèbre relationnelle. Il gère les droits d'accès des utilisateurs sur les relations. La notion de domaine est sommaire et il n'est pas possible de définir des domaines spécifiques, par exemple des intervalles ou des ensembles énumérés. Le lien avec les langages de programmation est assuré par deux structures de données prédéfinies (SQLCA, SQLDA), qui sont utilisées par le programmeur pour générer des requêtes SQL, contrôler leur exécution et récupérer les résultats[(9)].

Le paramétrage du modèle physique et des méthodes d'accès

Dans les recommandations initiales de Codd, la déclaration des directives de représentation des données en mémoire relève du langage hôte H. Ces directives ont pour objectif de fournir les informations au système pour qu'il mette en œuvre une stratégie adaptée au contexte applicatif. Pour des mises à jour fréquentes et aléatoires effectuées simultanément par des utilisateurs indépendants, une structuration arborescente s'impose. Pour des acquisitions de flots de données factuelles ordonnées et non modifiables, la structuration séquentielle classique suffit. Dans le premier cas, les relations d'ordre sur les éléments des ensembles varient sur des cycles dont la période est parfois mesurée en millisecondes. Dans le second cas, l'ordre est invariant, avec accessoirement des mises à jour dont la périodicité est annuelle ou pluriannuelle.

[(9)] Cette partie du langage est appelée *embedded SQL*.

Dès lors qu'un langage permettrait de décrire ces contextes, en d'autres termes la dynamique des ensembles, le choix du modèle physique optimal pourrait être fait par un compilateur. Il appartiendrait au compilateur d'évaluer les fonctions de complexité et d'adapter les stratégies algorithmiques en fonction d'informations déclaratives ou d'indications statistiques acquises en cours d'exécution. Un tel langage n'existe pas sur le marché. Les logiciels du commerce privilégient l'optimisation des performances dans la situation dite "transactionnelle à haut débit", dans laquelle des utilisateurs effectuent en permanence des modifications ponctuelles depuis leur poste de travail.

La fonction de "vectorisation"

L'approche relationnelle est étrangère au monde des logiciels statistiques, qui proposent leurs propres systèmes de gestion des données. Si des interfaces dites SQL existent, elles traitent des relations dont les attributs sont les noms des colonnes du tableau statistique. Comme indiqué plus haut, cette solution devient rapidement inutilisable dès que le nombre de variables augmente.

La restructuration des informations numériques, qui nécessite de passer de la représentation relationnelle des matrices proposée plus haut à la représentation tabulaire classique, est une tâche procédurale qui relève du langage hôte H. Le langage SQL ne permet pas de construire la fonction de vectorisation $V(I, J, C)$, définie de la base de données dans un espace vectoriel: elle doit donc être développée au moyen d'autres langages.

Les fonctions d'agrégation

Les fonctions d'agrégation sont présentes dans le langage SQL. Le résultat du calcul, qui utilise les opérations relationnelles de projection, restriction et éventuellement jointure, est affecté dans une nouvelle relation.

Le calcul matriciel

Le langage SQL permet également d'exprimer les opérations usuelles de calcul matriciel: combinaisons linéaires, produit de matrices. La démarche est illustrée dans la section suivante sur l'exemple de l'ajout de la dimension "nutritionnelle" dans les bases de la consommation alimentaire. Cette façon de procéder peut s'avérer très efficace dans un contexte où les matrices contiennent beaucoup de termes nuls. La notion de **type abstrait de données** (Liskov et Zilles, 1974) définit un cadre théorique

pour l'intégration de ces opérations dans le "moteur" de la base, afin que l'utilisateur puisse calculer directement des expressions matricielles.

APPLICATION AU TRAITEMENT DES PANELS DE CONSOMMATION ALIMENTAIRE

Ces méthodes de structuration de l'information statistique sont actuellement traduites sur des stations de travail UNIX de laboratoire (Sun Sparc station 1$^+$) avec le système INGRES[10], pour les enquêtes INSEE et les panels SECODIP. Les traitements statistiques sont réalisés après "vectorisation" avec les logiciels S et SAS. La fonction de "vectorisation" est traduite en langage C.

Description des bases de données relationnelles

Une base de données relationnelle est définie plus haut comme un ensemble de relations qui évolue dans le temps. Dans le contexte des travaux de recherche sur la consommation de l'INRA, les équipes scientifiques manipulent en fait des bases logiques indépendantes. Dans la phase d'exploration des données statistiques disponibles, les bases logiques correspondent aux enquêtes existantes. L'empilement des enquêtes de consommation alimentaire de l'INSEE en vue de générer des séries temporelles longues est actuellement en cours de réalisation.

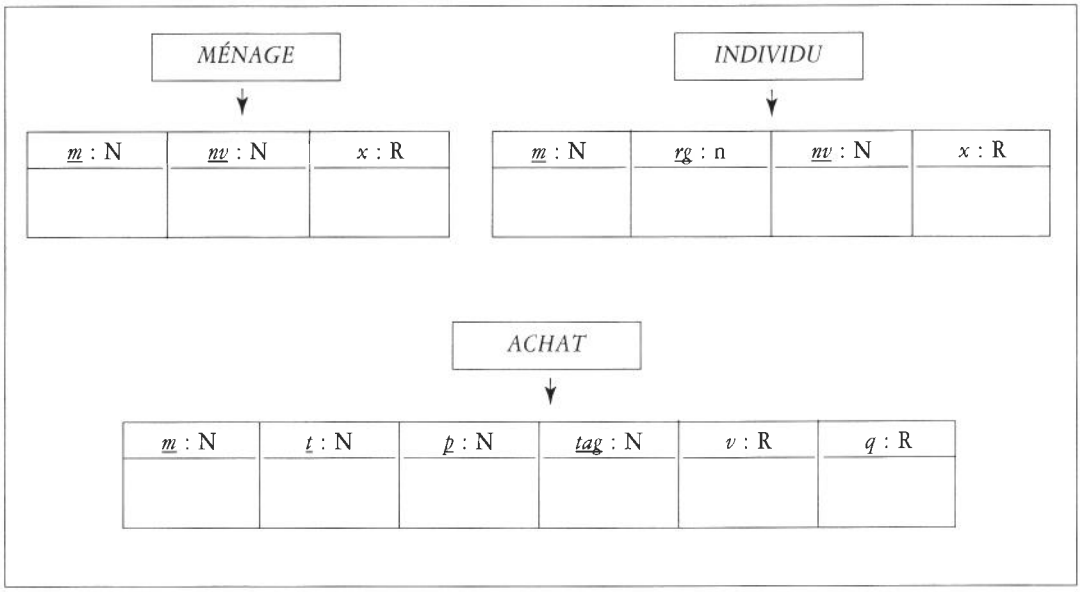
Les bases factuelles d'enquêtes

Les bases logiques associées aux enquêtes forment un ensemble de trois relations: *MÉNAGE* ($\underline{m}$, $\underline{nv}$, x), *INDIVIDU* ($\underline{m}$, $\underline{rg}$, $\underline{nv}$, x) et *ACHAT* ($\underline{m}$, $\underline{t}$, $\underline{p}$, $\underline{tag}$, v, q). Des relations annexes établissent éventuellement les liens entre des codes numériques et les termes des dictionnaires d'identification des produits et des variables. Un élément de la relation *MÉNAGE* est l'observation élémentaire d'une grandeur x associée à la variable nv pour un ménage désigné par son numéro logique m. Un élément de la relation *INDIVIDU* est également l'observation élémentaire d'une grandeur x associée à la variable nv pour l'individu repéré par son rang rg pour le ménage m. Un élément de la relation *ACHAT* identifie

[10] Le prototype INGRES a été développé à l'Université de Californie à Berkeley, et conçu au départ comme une couche externe du système UNIX. Les premières versions ont été publiées en 1975 avec le langage QUEL, parallèlement au développement du prototype IBM SYSTEM R. Le langage SQL a été introduit beaucoup plus tard dans les versions commercialisées du logiciel INGRES.

tout achat effectif d'un produit p sur la période t pour un ménage m, tel qu'il pourrait figurer sur les lignes d'un carnet de compte: les achats identiques, par exemple quotidiens sur une période hebdomadaire, sont distingués au moyen de l'attribut instrumental *tag*. Un achat est mesuré par la valeur monétaire v et la quantité q. L'attribut p est clé étrangère dans les relations de définition des nomenclatures de produits. La figure 4 visualise les schémas des relations sous la forme tabulaire simple.

Figure 4.
Schéma des relations statiques *MÉNAGE, INDIVIDU, ACHAT*



La dynamique des bases

A partir des relations précédentes, et d'autres relations externes qui représentent par exemple des nomenclatures ou des classes de ménages, les bases évoluent par création et destruction de relations, en fonction des besoins. Il s'agit principalement de calculs lourds, en vue par exemple de produire une matrice. Les interrogations ponctuelles sont peu fréquentes et restreintes à la phase exploratoire initiale. Les mises à jour par modification d'un élément d'une relation sont normalement interdites.

Il est envisagé de rassembler dans une même base des données factuelles d'enquêtes relatives à plusieurs années: les relations associées sont modifiées avec une périodicité annuelle (SECODIP) ou pluri-annuelle

(INSEE). Cette base est bâtie autour des trois relations principales *MÉNAGE (w, m, nv, x)*, *INDIVIDU (w, m, rg, nv, x)* et *ACHAT (w, m, t, p, v, q)*. L'attribut *w* désigne la vague. Les achats identiques sont regroupés, ce qui se traduit par une projection qui supprime l'attribut *tag* des bases annuelles.

Le paramétrage du modèle physique

Le tableau 2 fournit les cardinalités des relations associées aux enquêtes INSEE 1987 (6 938 ménages) et au panel SECODIP P2 de l'année 1989 (5 840 ménages). A titre indicatif, le stockage des trois relations principales du panel SECODIP P2, totalement ordonnées sur la clé primaire, requiert moins de 40 mégaoctets.

Tableau 2. **Cardinalité des relations associées aux enquêtes annuelles de consommation**

	MÉNAGE	*INDIVIDU*	*ACHAT*
INSEE 1987	622 900	401 004	267 794
SECODIP P2	276 879	89 340	2 138 186

La requête de modification du modèle physique (Modify)[11], commune aux langages QUEL et SQL de INGRES, permet de choisir une structure de stockage adaptée à la gestion d'un ensemble statique muni d'une relation d'ordre, c'est-à-dire un treillis. Si l'équipe de recherche choisit d'analyser la consommation des produits, l'attribut *p* constitue le premier critère d'ordonnancement: les autres critères sont successivement le ménage *m*, la période *t* et la marque *tag*. Le quadruplet *(p, m, t, tag)* est explicitement déclaré comme **clé d'identification unique** et comme expression canonique d'une **relation d'ordre totale**.

Si le projet de recherche d'une autre équipe privilégie le comportement individuel des ménages, il est recommandé de créer une autre relation, qui représente le même ensemble et la même clé primaire, mais avec un modèle physique qui gère la relation d'ordre exprimée par le quadruplet *(m, t, p, tag)*.

[11] Cette requête est étrangère au standard SQL et propre à INGRES. D'autres systèmes proposent un mécanisme de regroupement des données sur le disque en fonction des valeurs d'un ou plusieurs attributs *(cluster)*, et gèrent ainsi une relation d'ordre entre classes.

Les bases SECODIP

Dans la relation *ACHAT* des schémas précédents, l'attribut p, instrumental, assure la correspondance avec les nomenclatures externes. Le lien avec la nomenclature INSEE est géré dans la relation *INSEE* ($\underline{p}$, λ).

Les correspondances entre nomenclatures

La jointure des relations *ACHAT* et *INSEE* sur l'attribut p, clé étrangère de *ACHAT* et clé primaire de *INSEE*, rend une relation $A_1(\underline{m}, \underline{t}, \underline{p}, \underline{tag}, \lambda, v, q)$, de même cardinalité que *ACHAT*. La projection qui supprime les attributs p et *tag* permet de construire la relation $A_2(\underline{m}, \underline{t}, \underline{\lambda}, v, q, n)$, qui décrit les achats des produits *INSEE* par ménage m et période t: les attributs v et q représentent respectivement les valeurs et quantités totales achetées; l'attribut supplémentaire n dénombre les actes d'achats élémentaires.

Le choix de l'échantillon

L'analyse statistique s'appuie sur un échantillon de ménages dont il faut évaluer la qualité des réponses: un moyen simple est de déterminer pour chaque ménage le nombre de périodes K pour lesquelles k achats au moins ont été déclarés. On peut alors choisir de restreindre l'échantillon aux ménages qui ont répondu sur au moins T périodes, ce qui s'exprime par le prédicat $[K \geq T]$. Cette restriction, appliquée à la relation A_2 précédente, rend une relation de même schéma $A_3(\underline{m}, \underline{t}, \underline{\lambda}, v, q, n)$ qui contient les achats des ménages qui répondent aux critères de choix retenus.

Ce processus est paramétrable: il est ainsi possible d'examiner plusieurs critères et de produire les jeux de données associés, repris par les logiciels d'analyse statistique. Cela revient à paramétrer le choix de l'espace vectoriel de l'échantillon en fonction de règles simples exprimées par des formules classiques.

La restitution des matrices

Pour le panel P_2, l'analyse détaillée des achats par période a conduit à retenir un échantillon final de 3136 ménages, pour lesquels sont disponibles des données hebdomadaires. Ces achats sont décrits avec la nomenclature INSEE: on dénombre 170 produits dans la relation $A_3(\underline{m}, \underline{t}, \underline{\lambda}, v, q, n)$, qui correspondent à 19135 produits SECODIP initiaux. Dans un premier temps, l'objectif est de réaliser la description statistique sur l'espace de l'échantillon final, afin de comparer les distributions avec celles de l'INSEE. La méthode va consister à produire, au moyen de la

fonction de vectorisation, des matrices avec en ligne les ménages et en colonne les achats des produits et les variables descriptives des ménages.

Une première opération de calcul relationnel supprime l'attribut t de la relation A_3 (par projection), tout en effectuant la somme des achats: le résultat est la relation $A_4(\underline{m}, \underline{\lambda}, v, q, n)$. Si les relations $I(\underline{m})$ et $J(\underline{\lambda})$ représentent les listes des ménages et des produits, la fonction $V(I, J, A_4)$ rend trois matrices qui regroupent les observations relatives aux variables mesurant les achats en valeur (v), en volume (q) et en nombre d'opérations (n). La traduction a provisoirement été effectuée par l'utilisation conjointe d'une requête de restructuration du modèle physique et d'une fonction externe écrite en langage C.

La matrice décrivant les caractéristiques des ménages est produite de la même façon, à partir des relations $I(\underline{m})$, $J(\underline{nv})$ et $MÉNAGE$ $(\underline{m}, \underline{nv}, x)$. Les fonctions de regroupement permettent d'ajouter à la demande des variables agrégées: nombre d'individus du ménage, nombre d'enfants, indicateurs d'activité professionnelle, etc.

L'analyse des fluctuations hebdomadaires nécessite la prise en compte de la dimension temporelle. Le tableau statistique soumis au logiciel SAS est produit comme précédemment à partir des relations $I(\underline{m}, \underline{t})$, $J(\underline{\lambda})$ et $A_3(\underline{m}, \underline{t}, \underline{\lambda}, v, q, n)$. Il contient 163 072 lignes (52 périodes pour 3 136 ménages).

Les bases INSEE: le passage du modèle hiérarchique au modèle relationnel

Relativement aux bases annuelles SECODIP, les bases INSEE sont de taille plus modeste: avec 6 938 ménages, l'enquête alimentaire 1987 décrit 267 794 achats élémentaires. L'empilement des enquêtes annuelles disponibles depuis 1973 génèrerait une relation $ACHAT$ $(\underline{w}, \underline{m}, \underline{t}, \underline{p}, v, q)$ de moins de cinq millions d'éléments.

Le traitement des enquêtes INSEE pose le problème de la transformation du modèle hiérarchique géré par le logiciel LEDA sur système central, en modèle relationnel sur station de travail UNIX. Dans la mesure où il s'agit de représenter une même information dans deux modèles logiques différents, nous avons choisi de définir une méthode de traduction aussi indépendante que possible des données et des logiciels. L'algorithme est actuellement traduit dans un prototype écrit en langage C, utilisable pour convertir tout fichier au standard LEDA en un ensemble de relations définies sur des domaines simples. Il parcourt le graphe associé à l'arborescence dans un ordre canonique et attribue une clé d'identification unique à chaque sommet du graphe, qui représente un article. Il extrait ensuite les valeurs atomiques des variables codées dans cet article et génère les tuples des relations. Les clés d'identification des som-

mets du graphe sont utilisées pour construire les clés primaires des relations. A ce stade, les valeurs des attributs des clés primaires expriment uniquement la structuration syntaxique initiale de l'arborescence LEDA.

La prise en compte de la sémantique des données s'effectue dans une seconde étape, sous la responsabilité de l'équipe scientifique et en fonction des besoins, par une séquence d'opérations relationnelles. Pour les enquêtes alimentaires, ce processus aboutit à la production des relations *MÉNAGE, INDIVIDU* et *ACHAT* décrites plus haut.

Les opérations relationnelles les plus coûteuses sont des jointures naturelles, qui font correspondre aux codes instrumentaux ($k^{ième}$ variable du $j^{ième}$ fils du $i^{ième}$ père) les valeurs de champs qui identifient le contenu d'un article ($n^{ième}$ achat d'un produit p pour le ménage m). Les cardinalités des ensembles opérandes se mesurent en millions: les algorithmes de jointure doivent en conséquence être efficaces dans ce contexte.

L'ajout de la dimension "nutritionnelle"

Quelle que soit la source des données, les quantités q achetées annuellement pour des produits λ par des ménages m d'un échantillon sont représentables dans une relation de schéma $B(\underline{m, \lambda}, q)$. La composition des aliments est usuellement décrite dans un espace vectoriel sous forme d'une matrice M, le terme m_{ij} désignant la teneur d'une unité de l'aliment i en nutriment j. Une telle matrice est représentable dans une relation $C(\underline{\lambda, \mu}, k)$, l'attribut k mesurant la quantité du nutriment μ contenu dans une unité de l'aliment λ. Le couple (λ, μ) est clé primaire de la relation C, ce qui exprime le fait qu'il existe une valeur et une seule pour qualifier une teneur dans une table. La jointure naturelle des relations B et C rend une relation $D(\underline{m, \lambda, \mu}, k, q)$, de laquelle on déduit la relation $E(\underline{m, \lambda, \mu}, x)$: la valeur de l'attribut x mesure la quantité consommée du nutriment μ, à travers l'aliment λ pour le ménage m.

En s'appuyant sur l'exemple de répertoire général des aliments du Centre informatique sur la qualité des aliments (CIQUAL), qui décrit la teneur des aliments en 34 nutriments, on peut estimer à environ cinq millions d'éléments la cardinalité de la relation E résultante pour le panel P2 de SECODIP. Les teneurs manquantes sont codées dans une relation annexe $G(\underline{\lambda, \mu})$. Le traitement des données SECODIP ou INSEE à travers la nomenclature du CIQUAL s'effectue au moyen de la relation CIQUAL $(\underline{p}, \lambda)$ dans laquelle l'attribut p est la clé étrangère des relations *ACHAT*. Les calculs ne soulèvent pas de difficultés et l'analyse des fonctions de complexité démontre leur faisabilité. La "vectorisation" de la relation E permet d'effectuer des analyses socio-économiques sur des vecteurs de nutriments.

La projection de la relation E qui supprime l'attribut λ, et calcule la quantité totale de nutriments μ consommée par un ménage m, rend une

relation $H(\underline{m}, \underline{\mu}, x)$. Cette relation H représente la matrice de consommation potentielle des nutriments sur l'échantillon des ménages. C'est le produit des matrices représentées par les relations B (consommation des aliments) et C (composition des aliments).

CONCLUSION

L'algèbre relationnelle offre une alternative aux représentations tabulaires multidimensionnelles usuelles pour coder simplement des structures numériques complexes: vecteurs, matrices, tenseurs... Le calcul relationnel permet d'exprimer les opérations de base du calcul matriciel: transformations linéaires, produit de matrices. La compilation d'un langage qui gère ces relations et traduit les opérations relationnelles ne soulève pas de difficultés théoriques: les domaines utilisés sont restreints aux entiers et aux réels et les algorithmes existent. L'analyse de calculabilité montre qu'une simple station de travail est en mesure d'exécuter ces algorithmes avec des délais de réponse acceptables, pour des opérandes de plusieurs millions de tuples. Ce résultat théorique se vérifie expérimentalement sans difficulté: le modèle logique est donc potentiellement efficace pour structurer les grandes bases de données statistiques.

Or, il apparaît que l'utilisation du modèle relationnel reste aujourd'hui limitée dans les institutions de production de statistiques. Cette réserve apparente des statisticiens pourrait être justifiée par une relative inadéquation de l'offre logicielle du marché. En effet, les systèmes commercialisés sont évalués au moyen de critères qui mesurent principalement la "performance transactionnelle", ou de "commutation", c'est-à-dire l'aptitude d'un système à gérer efficacement des accès simultanés à des objets partagés, consultés en lecture ou mis à jour. La gestion des grandes bases de données statistiques, dès lors qu'elles sont constituées, ne requiert pas ce type de puissance; les extractions sont faites épisodiquement et massivement par les seules équipes scientifiques autorisées.

La réalisation décrite dans cet article, faite au sein d'un département de recherche en sciences sociales pour lequel l'informatique est un instrument et non un objet de recherche, démontre qu'il est possible de mettre en œuvre des solutions efficaces et performantes dans le contexte d'un laboratoire. Nous pouvons aujourd'hui transformer au moyen d'un automate de traduction un fichier hiérarchique LEDA, exploité sur site central IBM, en un ensemble de relations gérées dans un système de gestion de bases de données. Cette interface s'appuie sur le langage SQL et utilise la spécification IBM pour communiquer avec les langages de programmation classique. La fonction de vectorisation, dont le principe est décrit dans cet article, a également été traduite avec ces mêmes outils, et admet comme paramètres des noms de relations.

Nous soulignons ici le fait que la solution adoptée ignore les langages dits "de quatrième génération" promus par les fournisseurs de logiciels. En conséquence, elle n'est pas liée à un produit spécifique et est facilement transposable dans d'autres contextes, dès lors que les standards sont respectés. Ce transfert paraît pouvoir se faire en particulier sur les grands serveurs centraux des institutions statistiques, et notamment avec le système de gestion de bases de données DB2 d'IBM.

Il ne faudrait pas pour autant en déduire que cette application est portable sur tout système qui se prétend relationnel et déclare offrir le langage SQL. Parmi les critères d'évaluation à retenir, il paraît indispensable de mesurer la puissance absolue d'exécution des opérations de calcul relationnel sur de grands ensembles, de la même façon que l'on mesure la puissance de calcul numérique en virgule flottante. Le test proposé au chapitre 4 est suffisant et facilement réalisable: il peut fournir des résultats significatifs avant de franchir le cap du million de tuples, tant pour évaluer l'efficacité de la stratégie algorithmique que la gestion de la ressource disque.

Les problèmes illustrés dans cet article sur l'exemple de la représentation de structures numériques usuelles ne sont pas propres aux sciences sociales. Ils font actuellement l'objet de recherches actives dans les laboratoires informatiques (Silberschatz *et al.*, 1991): optimisation, parallélisme, règles, objets, systèmes de stockage, gestion de mémoire, événements, distribution des données, transactions longues ... Le laboratoire IBM de San José poursuit ses travaux sur le prototype Starburst (Lohman *et al.*, 1991). L'Université de Berkeley travaille sur le prototype Postgres (Stonebraker et Kemnitz, 1991). Cette première approche se traduit par un élargissement du champ d'application du modèle relationnel. Une seconde approche reprend les acquis de la théorie des langages et propose des systèmes à objets qui se substituent aux systèmes de gestion de bases de données relationnelles (Delobel *et al.*, 1991 ; Gardarin, Valduriez, 1991) . On ne peut donc qu'espérer l'arrivée sur la marché de produits commerciaux qui intègreront les fonctionnalités présentées dans cet article pour mieux répondre aux besoins du statisticien. En attendant, le modèle relationnel simple, tel qu'il a été défini et traduit dans les années 70 dans le laboratoire IBM de San José et à l'Université de Berkeley, est aujourd'hui suffisant pour s'abstraire de technologies matérielles et logicielles propriétaires, et définir des solutions ouvertes sur les configurations informatiques actuelles et futures, dans l'environnement des laboratoires de recherche.

BIBLIOGRAPHIE

BEAUQUIER (D.), BERSTEL (J.), CHRÉTIENNE (P.), 1992 — *Eléments d'algorithmique*, Paris, Masson.

CODD (E. F.), 1970 — *A Relational Model of Data for Large Shared Data Banks*, Communications of the ACM, Vol. 13, n° 6.

DAVID (M.), 1989 — *Managing Panel Data for Scientific Analysis: The Role of relational Data Base Management Systems, in*: KASPRZYK (D.) *et al.* (ed.), *Panel Surveys*, Wiley, pp. 226-241.

DELOBEL (C.), LECLUSE (C.), RICHARD (P.), 1991 — *Bases de données: des systèmes relationnels aux systèmes à objets*, Paris, Interéditions.

GARDARIN (G.), VALDURIEZ (P.), 1991 — *SGBD avancés. Bases de données objets, déductives, réparties*, Paris, Eyrolles.

LISKOV (B.), ZILLES (S.), 1974 — *Programming with Abstract Data Types*, SIGPLAN Notices, Vol. 9, n° 10.

LOHMAN (G. M.), LINDSAY (B.), PIRAHESH (H.), SCHIEFER (K. B.), 1991 *Extensions to Starburst: Objects, Types, Functions, and Rules*, Communications of the ACM, Vol. 34, n° 10.

SHOSHANI (A.), 1982 — *Statistical Databases: characteristics, problems and some solutions*, Computer Science Division Publications, Berkeley, University of California, California 94720.

SILBERSCHATZ (A.), STONEBRAKER (M.), ULLMAN (J.), 1991 — *Database Systems: Achievements and Opportunities*, Communications of the ACM, Vol. 34, n° 10.

STONEBRAKER (M.), KEMNITZ (G.), 1991 — *The Postgres Next-Generation Database Management System*, Communications of the ACM, Vol. 34, n° 10.

WOLPER (P.), 1991 — *Introduction à la calculabilité*, Paris, Interéditions.