Modélisation des grandes bases de données statistiques : application aux enquêtes alimentaires INSEE et aux panels SECODIP

Jean-Claude Poupa

INRA

LABORATOIRE DE RECHERCHE SUR LA CONSOMMATION

MODELISATION DES GRANDES BASES DE DONNEES STATISTIQUES : APPLICATION AUX ENQUETES ALIMENTAIRES INSEE ET AUX PANELS SECODIP

J.C. POUPA
INRA - Station d'économie
et sociologie rurales de Rennes

Juin 1991

Etude informatique réalisée pour le Laboratoire de Recherche sur la Consommation dans le cadre du programme relatif à la mise en place de l'Observatoire des Consommations Alimentaires financé par la DGAL (Direction Générale de l'Alimentation), la DGCCRF (Direction Générale de la Concurrence de la Consommation et de la Répression des Fraudes), la DGS (Direction Générale de la Santé), et le Ministère de la Recherche et de la Technologie (DIST).

Table des matières

Ce document est un rapport d'étape. Il présente les résultats informatiques disponibles en juin 1991, dans le cadre du projet de mise en place de l'Observatoire des Consommations Alimentaires. La phase actuelle de ce projet consiste à évaluer les possibilités d'utilisation des données primaires des principales enquêtes réalisées en France sur la consommation alimentaire des ménages.

Le passage de l'environnement de production industrielle, mis en oeuvre dans les grands centres de calcul où sont traitées habituellement ces données, à une informatique décentralisée sur station de travail, avec un système d'exploitation et des logiciels conçus dans un environnement universitaire, n'a pas été sans poser un certain nombre de problèmes. Toutes ces difficultés techniques, qui passent par la création et la traduction de "primitives systèmes" de manipulation de l'information au niveau binaire, ne sont pas abordées dans ce rapport(1). Les procédures adaptées sont opérationnelles : elles sont à l'heure actuelle assemblées par les informaticiens à la demande, mais sont destinées à être intégrées ultérieurement comme composantes d'une chaîne de traitement automatisée, qui reçoit en entrée les fichiers d'enquêtes et rend l'information dans une base de données(2).

Le premier chapitre de ce rapport restitue les sources d'information et la problématique informatique du projet.

Les deux chapitres suivants présentent les bases relationnelles associées à l'enquête sur la consommation alimentaire de l'INSEE et au panel P2 de SECODIP.

Le dernier chapitre décrit le processus de vectorisation qui permet de transmettre les données sous forme de tableaux, de la base de données relationnelle vers les logiciels de traitement statistique.

La solution présentée dans ce rapport a été mise au point sur une station de travail UNIX SUN 4/65, équipée d'un disque dur de 770 mégaoctets. Le système de gestion de bases de données choisi au terme d'une analyse comparative avec expérimentation en situation réelle est INGRES. Le logiciel cible retenu pour les traitements statistiques est SAS. Les programmes associés sont développés en langage C et seront vraisemblablement intégrés dans le moteur INGRES dans une étape ultérieure.

---

(1) Des notes techniques spécifiques décrivent ces opérations (N. Calchera, J.C. Poupa)

(2) La réalisation d'un automate de conversion du standard LEDA de l'INSEE est en cours, en collaboration avec l'Institut de Formation Supérieur en Informatique et Communication (IFSIC) de l'Université de Rennes 1.

Le volume final de la base SECODIP, avec les deux panels P1 et P2, devrait être inférieur à la centaine de mégaoctets ; le volume de la base INSEE pour l'enquête 1987 sera vraisemblablement de l'ordre de la vingtaine de mégaoctets.

1.L'analyse de la consommation alimentaire des ménages : sources d'information et modèles de données

Les enquêtes sur la consommation alimentaire des ménages de l'INSEE et de SECODIP recueillent des informations relatives aux actes d'achats, tels qu'ils pourraient figurer sur les lignes d'un carnet de comptes.

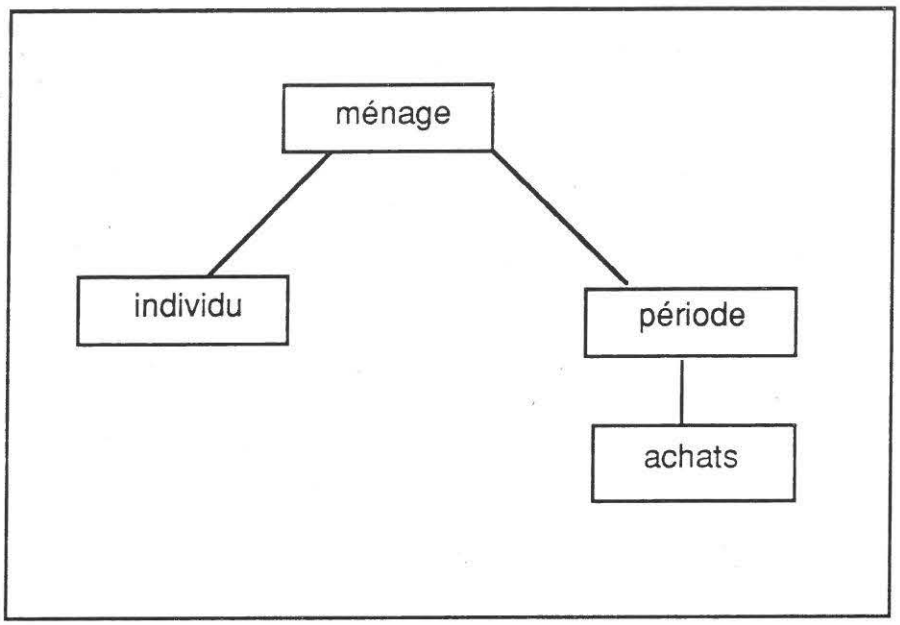
L'enquête INSEE 1987 concerne un échantillon de 6938 ménages, avec une seule période d'observation limitée à une semaine donnée dans l'année, et un ensemble de 327 produits répertoriés dans une nomenclature.

L'enquête SECODIP relative à l'année 1989 porte sur deux échantillons de taille équivalente, dits panels P1 et P2, avec une périodicité d'observation hebdomadaire. Pour situer les volumes, le panel P2 regroupe au départ 5 840 ménages ; les achats sont observés pour un ensemble de 19 135 produits, définis selon une nomenclature évolutive dans le temps. Dans la pratique, les données hebdomadaires ne sont pas disponibles pour la totalité des ménages et les observations relatives à un ménage doivent être pondérées par le nombre de semaines d'activités.

Au sein d'une période, pour l'INSEE comme pour SECODIP, un même produit peut être acheté plusieurs fois : l'observation élémentaire est la ligne inscrite au carnet de comptes, à laquelle sont associées systématiquement une quantité achetée et une dépense.

Les ménages sont eux-mêmes décrits par des variables socio-économiques. Ces ménages regroupent des individus, pour lesquels une information minimale est recueillie.

Globalement et schématiquement, toutes ces données peuvent être représentées dans une structure hiérarchisée , illustrée dans le schéma ci-dessous.



La racine de l'arborescence est le niveau "ménage". Une première branche relie les individus aux ménages. Une seconde ramification enchaîne les périodes et les actes d'achats élémentaires pour chacune de ces périodes.

Pour environ 6 000 ménages, on dénombre 20 000 individus, 300 000 achats pour l'enquête INSEE et plus de deux millions d'achats pour chaque panel de l'enquête SECODIP.

1.1.La structuration de l'information dans un environnement de calcul statistique

Les traitements statistiques sont réalisés avec le logiciel SAS, qui gère des tableaux rectangulaires avec en ligne les observations et en colonne les variables. Pour l'enquête INSEE, la construction d'un tel tableau ne soulève pas de difficultés si l'on accepte d'agréger les actes d'achats élémentaires relatifs à un produit pour la période observée : les ménages sont en ligne, les colonnes représentant successivement les variables descriptives des ménages puis les vecteurs des consommations relatives aux produits répertoriés dans la nomenclature. On aboutit finalement à un tableau à 6 938 lignes et moins de 500 colonnes pour l'enquête 1987.

Pour les données SECODIP, la production d'un tel tableau déboucherait sur des impossibilités de traitement. Avec comme critères d'observation le ménage et la période, on aboutirait à un tableau de quelque trois cent mille lignes et vingt mille colonnes !

Dans la pratique, le logiciel SAS est plus difficile d'utilisation dès qu'il s'agit de mettre en correspondance des données observées à plusieurs niveaux de la hiérarchie. C'est en premier lieu un logiciel de traitement statistique, bien adapté à la manipulation de tableaux statistiques, mais ce n'est pas un système de Gestion de Bases de Données.

1.2.La structuration dans un environnement "bases de données"

A l'heure actuelle, les grandes enquêtes nationales utilisent des logiciels qui s'appuient sur le modèle de données hiérarchique. C'est le cas du logiciel LEDA de l'INSEE. Le modèle de données relationnel apparaît comme peu utilisé pour administrer de grandes bases de données statistiques. Les quelques expérimentations effectuées sont souvent décrites comme lourdes et coûteuses.

Le problème posé dans un rapport antérieur était de trouver un logiciel capable de gérer sur le plan logique le modèle relationnel, mais aussi d'autoriser le pilotage du modèle physique de données, de telle sorte que les algorithmes exécutés pour manipuler et archiver les données soient adaptés au contexte applicatif des grandes bases statistiques.

Le logiciel INGRES, conçu et développé par l'Université de Berkeley, répond parfaitement aux besoins qui ont été exprimés dans la première étape de notre analyse. L'expérimentation en vraie grandeur a largement confirmé les espérances que suggérait l'examen du produit d'un strict point de vue fonctionnel.

2.Modélisation sous INGRES des données INSEE : exemple de l'enquête "consommation alimentaire des ménages" en 1987

Les fichiers livrés par l'INSEE au format du logiciel LEDA, sous forme hiérarchisée, sur support bande magnétique, standard IBM, occupent un volume physique d'une trentaine de mégaoctets.

La base INGRES définitive permettra de gérer l'intégralité de l'information disponible dans un espace physique de l'ordre de la vingtaine de mégaoctets. Pour cette première expérimentation, compte tenu des besoins de calcul immédiats, nous avons traité trois ensembles d'informations :

1) les variables descriptives des ménages : 76 variables de type "code" et 35 variables de type "quantité", pour 6 938 ménages, soit le niveau "ménage" de la hiérarchie Leda ;

2) les variables descriptives des individus des ménages : 21 variables observées pour 19 103 individus, soit le niveau "individu" de la hiérarchie Leda ;

3) les actes d'achats élémentaires, enregistrés dans le niveau Leda "inscriptions au carnet de compte" : 267 794 lignes.

Cette base occupe actuellement un volume physique de 11 mégaoctets. La structuration choisie privilégie l'accès au vecteur d'observations : variables descriptives des ménages ou des individus, valeur et quantité achetées des produits. Les délais de réponse pour des questions qui rendent de tels vecteurs sont instantanés, et les ressources consommées sont négligeables. La consommation de ressources en puissance de calcul intervient en amont, au cours du processus d'acquisition, de validation et de restructuration des données.

Une image de cette base est disponible sous forme de fichiers "texte" classiques, dans un volume de 17 mégaoctets.

2.1.Le modèle logique

Le modèle retenu est un ensemble de cinq relations : MENAGE_CODE_INSEE, MENAGE_QUANTITE_INSEE, INDIVIDU_INSEE, IDENTIFIANT_INSEE, ACHATS_INSEE.

2.1.1. *La relation MENAGE_CODE_INSEE (521 379 lignes)*

MENAGE_CODE_INSEE (*ménage*, *nv*, *x*)

C'est pratiquement une table à trois colonnes, les colonnes étant appelées attributs. La notation adoptée pour décrire les relations souligne les attributs qui composent la clé primaire.

L'attribut *ménage* est un numéro d'observation, qui identifie un ménage et un seul. Ce numéro est repris dans la relation IDENTIFIANT_INSEE, qui établit la correspondance avec les variables d'identification des ménages (champ identifiant de l'arborescence Leda). Il est dit "clé étrangère" dans la terminologie de l'algèbre relationnelle.

L'attribut *nv* est un numéro de variable, affecté par programme en fonction du rang de la variable dans l'enregistrement décrit dans le dictionnaire Leda. Ce dictionnaire sera ultérieurement géré sous forme d'une relation INGRES, ce qui permettra de remplacer ce numéro logique par le nom Leda de la variable dans les requêtes d'interrogation, sans modifier le modèle de données. L'attribut *x* représente la valeur de la variable nv pour l'observation ménage. Le couple (*ménage*, *nv*) identifie une valeur et une seule : il constitue une clé d'identification unique dite "clé primaire".

La distinction entre "codes" et "quantités" reprend la définition des types de données des enregistrements Leda. Elle est maintenue pour souligner le fait que le calcul numérique n'est autorisé que sur des quantités, et non sur des codes, même s'ils sont représentés par des chiffres.

2.1.2. *La relation ménage_quantité_insee (101 611 lignes)*

MENAGE_QUANTITE_INSEE (*ménage*, *nv*, *x*)

La structuration est rigoureusement identique à celle de la relation précédente.

2.1.3. *La relation identifiant_insee (6 938 lignes)*

IDENTIFIANT_INSEE (*ménage*, *lot*, *région*, *vague*, *numéro*, *logement*)

Elle reprend la variable Leda dite "identifiant". La valeur de l'attribut *ménage* est un numéro logique attribué par algorithme , cet attribut constituant une clé étrangère d'autres relations. Au terme du processus de construction de la base, cette relation peut être supprimée s'il n'y a plus lieu de retourner aux fichiers originaux. C'est un moyen

d'apporter des garanties supplémentaires dans le cadre de la législation sur le secret statistique.

2.1.4. *La relation individu_insee (401004 lignes)*

INDIVIDU_INSEE (*ménage*, *rang*, *nv*, *x*)

L'attribut *rang* a pour valeur le numéro d'ordre de l'individu dans la famille, calculé lors du parcours de l'arborescence Leda.

L'attribut *nv* est un numéro logique de variable, affecté par programme. La clé primaire unique est composée des trois attributs (*ménage*, *rang*, *nv*).

2.1.5. *La relation achat insee (267 794 lignes)*

ACHAT_INSEE (*ménage*, *produit*, *tag*, *quantité*, *valeur*)

L'attribut *produit* est le code produit défini dans la nomenclature analytique de l'INSEE.

L'attribut *tag* a dû être introduit au terme d'une suite d'échecs pour exhiber une clé primaire unique à partir des informations disponibles sur l'enregistrement élémentaire "inscription au carnet de comptes" de l'arborescence Leda. L'idée initiale était de retenir comme attribut supplémentaire le numéro d'ordre absolu n de la ligne dans le carnet, mais le triplet (*ménage*, *produit*, *n*) ne constituait pas dans les faits une clé primaire.

Nous avons choisi de différer l'étude de ce problème (les relations utiles seront créées à la demande en temps opportun) et mis en oeuvre une convention indépendante de la sémantique des données.

Un programme externe dit "distributeur de tags", reçoit en entrée la liste des enregistrements, ordonnée selon le couple (*ménage*, *produit*). Il affecte à chaque enregistrement un numéro logique, dit tag, distribué en parcourant séquentiellement une suite de n nombres entiers, pratiquement les 99 premiers nombres. Le triplet (*ménage*, *produit*, *tag*) est une clé primaire unique.

L'attribut *tag* joue un rôle fondamental dans la définition de cette relation, puisqu'on dénombre 18 799 quadruplets (ménage, produit, quantité, valeur) présents au moins deux fois, avec un maximum de 16 occurrences pour 2 quadruplets.

Cette relation contient également 25 lignes avec simultanément un code produit nul et des valeurs et quantités nulles.

Pour 47 573 lignes, la quantité est nulle alors que la valeur est positive.

2.2.Le modèle physique

La structuration choisie est adaptée aux besoins d'économistes qui souhaitent extraire des vecteurs d'observations, en vue de réaliser des traitements statistiques. Cela signifie pratiquement que toutes les observations relatives à une variable devront de préférence être rangées de façon contiguë, la pire des situations étant celle où la reconstitution d'un vecteur de dimension n génère n opérations de recherche.

Par ailleurs, les données sont apurées, validées et consultées uniquement en lecture.

La structuration séquentielle indexée, dite ISAM, répond bien à ce besoin. Les données sont rangées physiquement selon l'ordre des clés d'indexation, une table globale des index permettant de se positionner directement sur une zone du fichier associé.

Elle n'est pas pénalisante étant donné qu'il n'y a pas de mise à jour et qu'il n'y a donc pas lieu de réserver de l'espace libre pour d'éventuels ajouts aléatoires.

Un ordonnancement total est réalisé sur la clé primaire des relations, l'ordre des index étant choisi pour optimiser les temps d'accès pour une interrogation sur les valeurs d'un code variable ou d'un code produit.

L'optimisation du modèle physique passe également par un ajustement méthodique des types de données déclarés. Les quantités sont déclarées en entier ou flottant sur 4 octets. Pour les codes, les types peuvent être ajustés sur mesure, après examen du dictionnaire Leda. Dans la pratique, les codes sont représentés sur un seul octet (255 valeurs possibles) ou sur 2 octets (65 535 valeurs possibles).

La taille de la base associée à cette enquête reste modeste et il y a lieu de trouver un compromis entre complexité induite par le pilotage du modèle et gain dû à l'optimisation.

3.Modélisation sous INGRES des données secodip exemple du panel P2

Les fichiers livrés par SECODIP, pour le panel P2, dans un format compressé propre aux systèmes IBM, occupent un volume total de 270 mégaoctets. Pour le panel P1, le volume des fichiers est de l'ordre de 200 mégaoctets.

Comme pour l'enquête INSEE, les informations sont globalement divisées en trois ensembles :

1) les variables descriptives des ménages : 56 variables, pour 5 840 ménages, dont 53 variables extraites du fichier SECODIP "central", 2 variables de pondération et un code sous-panel extraits du fichier SECODIP "acti" ;

2) Les variables descriptives des individus des ménages : 6 variables pour 17 338 individus, les valeurs étant extraites du fichier Secodip "central" ;

3) les actes d'achats élémentaires, enregistrés dans le fichier SECODIP "prodpan", soit un ensemble de 2 189 345 inscriptions au carnet de comptes, pour 5 037 ménages, 52 périodes et 19 135 produits.

Cette base occupe actuellement un volume physique inférieur à 50 mégaoctets, le processus d'optimisation du modèle physique n'étant pas encore conduit à son terme. Il sera vraisemblablement difficile de descendre en dessous de 40 mégaoctets, pour une base qui conserve évidemment l'intégralité des informations fournies, avec une structuration différente.

Comme pour l'INSEE, le modèle physique adopté privilégie l'accès aux vecteurs d'observations : variables descriptives des ménages ou des individus, nombre d'unités, quantité et valeur achetées des produits.

Les délais de réponse sont instantanés pour l'extraction des variables descriptives des ménages ou des individus. Pour l'extraction de vecteurs de consommation de produits courants (les tests ont été réalisés pour la viande), l'ordre de grandeur du temps CPU consommé est de 5 secondes pour 10 000 observations, le temps total d'attente étant par ailleurs lié à la charge de la machine. Si en revanche on extrait les observations pour des produits décrits avec une précision maximale, le délai de réponse est instantané dès lors que la requête rend peu de lignes. Enfin, pour une requête qui nécessite un balayage complet de la table (par exemple recherche des achats de valeur nulle), la ressource CPU consommée est inférieure à 2 minutes.

Si dans la phase d'interrogation la consommation de ressources est relativement modeste, le processus de création de la base est quant à lui coûteux. Pour des restructurations complexes sur une ou plusieurs tables de quelques millions de lignes, le temps CPU d'exécution de certaines requêtes est de l'ordre de l'heure, et l'espace disque temporaire utilisé atteint trois fois le volume des tables initiales. Pratiquement cela signifie que la mise en production d'une base d'une taille finale de 50 mégaoctets demande la disponibilité temporaire d'un espace physique de 200 mégaoctets pour travailler dans de bonnes conditions. Ces coûts en ressource disque temporaire sont liés d'une part aux besoins des algorithmes de tris, d'autre part au fait que le système sait défaire tout ce qu'il a fait en cas d'incident, donc assure la sécurité des données.

3.1.Le modèle logique

Ce modèle regroupe cinq relations et une vue logique.

3.1.1. *La relation panel_p2 (276 879 lignes)*

PANEL_P2 (*no*, *nv*, *x*)

L'attribut *no* est un numéro d'observation affecté par programme à chaque ménage. L'attribut *nv* est un numéro logique de variable. Le couple (*no*, *nv*) forme une clé primaire unique.

3.1.2. *La relation individu_p2 (89 340 lignes)*

INDIVIDU_P2 (*no*, *rang*, *nv*, *x*)

L'attribut *rang* est le numéro d'ordre de l'individu dans la famille, affecté par programme. L'attribut *nv* est un numéro logique de variable. Le triplet (*ménage*, *rang*, *nv*) forme une clé primaire unique.

3.1.3. *La relation artif_p2 (6612 lignes lignes)*

ACTIF_P2 (*no*, *panel*, *sous-panel*, *periode*, *duree*, *poids*)

Les attributs *panel*, *sous-panel*, *periode*, *duree*, *poids* sont les valeurs lues dans le fichier secodip.

3.1.4. *La relation achat_p2 (2 126 987 lignes)*

ACHAT_P2 (*no*, *période*, *produit*, *tag*, *u*, *v*, *q1*, *q2*)

C'est une relation définie à titre provisoire, dans la mesure où elle semble contenir des redondances d'informations.

L'attribut "*période*" identifie un numéro de semaine dans l'année.

L'attribut "*produit*" désigne un produit au sens de la nomenclature SECODIP (voir la définition au paragraphe 3.1.5)

Les attributs *u* (nombre d'unités), *v* (valeur de l'achat), *q1* (quantité en fonction du coefficient 1), *q2* (quantité en fonction du coefficient 2) reprennent les variables correspondantes du fichier SECODIP "prodpan".

L'attribut *tag* a été introduit pour générer une clé primaire, indispensable pour la poursuite des opérations de calcul relationnel. On dénombre précisément 12 555 7_uplets (*no*, *période*, *produit*, *u*, *v*, *q1*, *q2*) présents au moins deux fois, avec un maximum de 12 occurrences identiques. Le triplet (*no*, *période*, *produit*), qui sémantiquement était candidat pour constituer une clé primaire, identifie au moins deux lignes pour 93 350 valeurs distinctes.

La valeur du *tag* est affectée par programme, comme pour l'INSEE, la liste des enregistrements étant ordonnée selon le triplet (*no*, *période*, *produit*), et les valeurs distribuées sur une suite séquentielle d'au moins 12 nombres arbitraires mais distincts.

Le nombre d'attributs de cette relation pourra vraisemblablement être réduit. Les valeurs des attributs *q1* et *q2* sont calculées et on peut supposer qu'elles dépendent du nombre d'unités du produit et du ménage à travers son poids. Si tel était le cas, ces deux attributs pourraient disparaître de cette relation ACHAT_P2, sans perte d'information dans la mesure où ces quantités peuvent être évaluées à partir des valeurs d'autres attributs, sur d'autres relations.

Ce type d'investigation exploratoire fait partie d'un ensemble d'opérations dites de calcul relationnel, effectuées par l'administrateur de la base de données. Leur complexité et les temps d'exécution associés font qu'elles n'ont pas à ce stade toutes été réalisées.

3.1.5. *La relation produit_p2 (19 135 lignes)*

PRODUIT_P2 (*marché, e1, e2, e3, e4, e5, e6, s, produit*)

La clé primaire unique est le 7_uplet (*marché, e1 e2, e3, e4 e5, e6*) utilisé par SECODIP pour identifier un produit dans les fichiers "prodpan" et "reflist".

L'attribut *produit*, clé étrangère de la relation ACHAT_P2, est un numéro logique attribué par programme qui identifie un produit SECODIP et un seul. L'attribut *s* est une statistique issue de la procédure de calcul de cette relation, qui contient finalement le nombre de lignes associées au produit dans la relation ACHAT_P2.

3.1.6. *La vue Secodip (2 126 987 lignes)*

SECODIP_P2 (*no, période, marché, e1, e2, e3, e4, e5, e6, tag, u, v, q1, q2*).

C'est une relation virtuelle, le système exécutant dynamiquement la mise en correspondance, par une opération de jointure, des relations ACHAT_P2 et PRODUIT_P2. Elle est équivalente, du point de vue fonctionnel, à la requête qui exprime l'opération de jointure, ici l'égalité des 7_uplets (*marché, e1, e2, e3, e4, e5, e6*) dans les deux opérandes concernées.

3.2.Le modèle physique

La structuration séquentielle indexée (ISAM) s'impose de la même façon que pour la base INSEE, avec les mêmes critères d'ordonnancement des clés d'index.

Vu la cardinalité des relations, les types des attributs doivent être ajustés aux domaines de définition.

Dès lors que les opérations de jointure utilisent les algorithmes adaptés au modèle physique, il peut être intéressant de remplacer des suites d'attributs par un numéro logique affecté par un programme. C'est ce qui est réalisé par exemple dans la vue logique SECODIP_P2, à travers les relations intermédiaires PRODUIT_P2 et ACHAT_P2. Cela revient à remplacer un 7_uplet codé sur 10 octets par un attribut unique représenté sur 2 octets : le gain est de plus de 16 mégaoctets sur l'exemple du panel P2. Quant au temps d'exécution, il demeure du même ordre de grandeur et peut être réduit, du fait de la réduction des transferts d'information. Il est ainsi possible de mettre en production des relations virtuelles qui mettront en oeuvre des opérations de

jointure générées à travers plusieurs vues logiques, sur des relations réelles ou virtuelles.

Le logiciel INGRES offre les fonctionnalités "système" pour ajuster un modèle physique à un contexte applicatif donné : l'administrateur de la base a la possibilité de suivre le cheminement algorithmique et d'apporter les modifications qui peuvent s'imposer. Dès lors que les modèles logiques et physiques sont figés, les plans d'exécution des requêtes soumises sont évalués puis mémorisés tant que la base INGRES reste active, pour des utilisations ultérieures.

Dans le contexte de cette première expérimentation sur une base statistique, le processus d'optimisation n'a pas été conduit à son terme. Au delà de la recherche des structurations logique et physique qui garantissent les temps de réponse optimaux pour les besoins d'interrogation des équipes de recherche, INGRES offre la possibilité de construire des tables de statistiques utilisées par le système pour évaluer les ressources nécessaires associées à un scénario d'exécution. Si les performances observées donnent actuellement satisfaction, il apparaît qu'elles peuvent encore être améliorées.

4.Base de données relationnelle et traitement statistique : le processus de vectorisation

La solution décrite dans ce rapport traite le volet "bases de données" et fournit aux logiciels statistiques les données sous les formes standard usuelles imposées par les contraintes du calcul numérique. Vu la dimension des enquêtes, le logiciel SAS répond bien aux besoins des économistes. Les méthodes mises en oeuvre s'appliquent à d'autres produits, jusqu'aux tableurs sous MSDOS sous réserve d'éditer des objets dont la dimension est adaptée aux possibilités des logiciels. Dans la pratique, pour l'INRA, les traitements seront répartis entre INGRES et SAS.

4.1.De l'algèbre relationnelle à l'algèbre linéaire

Les méthodes et algorithmes utilisés sont illustrés sur les relations des bases INSEE et SECODIP décrites dans les chapitres précédents. Pour simplifier la notation, nous parlerons des relations MENAGE, INDIVIDU et ACHAT, la structure étant globalement identique pour ces deux bases.

4.1.1. *La relation ménage (no, nv, x)*

Si i désigne la valeur du numéro d'observation *no* et j la valeur du numéro de variable *nv*, l'attribut *x* contient la valeur du terme d'une matrice M, situé à l'intersection de la ligne numéro i et de la colonne numéro j.

Puisque les attributs qui identifient le couple (observation, variable) forment une clé primaire unique, à un couple (i, j) correspond un terme et un seul.

L'algorithme qui restitue une telle matrice est décrit dans l'encadré ci-dessous :

Encadré n°1 : premier algorithme de restitution d'une matrice observation-variable

```
Début
        Pour tout élément i de la liste des observations
        Faire
                pour tout élément de la liste j des variables
                faire
                        - extraire le triplet (i, j, x) ;
                        - affecter la valeur x comme valeur du terme situé à
                          l'intersection des lignes et colonnes associées
                          respectivement aux observations et aux variables ;
                fait ;
        fait ;
Fin ;
```

Nous avons traduit en langage C une procédure qui exécute cet algorithme à partir de trois fichiers produits par le système INGRES :

- la liste ordonnée des numéros de ménage ;
- la liste ordonnée des numéros de variable ;
- la liste ordonnée des triplets (i, j, x).

L'ordonnancement est effectué sous INGRES à travers le choix du modèle physique et une image de la relation résultante est éditée sous forme de fichiers.

A terme, cette procédure pourra être intégrée sous INGRES, avec comme paramètres les deux clés étrangères[1] de la relation à transformer.

La procédure rend

- une matrice (6 938 x 76) pour la relation MENAGE_CODE_INSEE,
- une matrice (6 938 x 35) pour la relation MENAGE_QUANTITE_INSEE,
- une matrice (5 840 x 56) pour la relation PANEL_P2.

Le résultat est correct seulement si la propriété d'unicité de la clé primaire est vérifiée.

---

[1] Dans le modèle logique des bases INSEE et SECODIP, nv peut être vu comme clé étrangère sur une relation de définition des variables, de cordinalité égale au nombre de variables.

4.1.2. *La relation individu (no, rang, nv, x)*

L'algorithme précédent ne s'applique plus, les lignes de la matrice étant des individus, qualifiés par les valeurs des deux attributs *no* (numéro de l'observation au niveau ménage) et *rang*. Cette relation gère un lien hiérarchique, un ménage étant composé d'un nombre variable d'individus.

La méthode mise en oeuvre est cependant généralisable, les éléments des listes utilisées pour décrire les observations et les variables n'étant plus des nombres, mais des n_uplets.

Ce nouvel algorithme est décrit dans l'encadré ci-dessous.

Encadré n°2 : second algorithme de restitution d'une matrice observation-variable

Début

Pour tout i_uplet (i_1, i_2,...i_p) de la liste des i_uplets de la relation descriptive des observations

Faire

pour tout j-uplet (j_1, j_2, ...j_q) de la liste des j_uplets de la relation descriptive des variables

faire

- extraire le n_uplet ;

(i_1, i_2, ..., i_p, j_1, j_2, ..., j_q, x) ;

- affecter la valeur x dans une matrice, pour laquelle la ligne i représente le rang du i_uplet (i_1, i_2, ..., ip) dans la liste des observations, la colonne j le rang du j_uplet (j_1, j_2, ..., j_q) dans la liste des variables ;

fait ;

fait ;

Fin ;

Si l'on applique cet algorithme aux relations "ménage", les relations descriptives des variables et des observations possèdent un seul attribut. Pour les relations "individus", si l'on convient de représenter les personnes du ménage en ligne, la relation descriptive des observations est formée avec les deux attributs (no, rang), la relation descriptive des variables étant réduite au seul attribut nv.

La procédure qui traduit cet algorithme admet comme paramètres deux relations construites avec les clés étrangères de la relation à transformer : I_LISTE (i_1, i_2,...,i_p) et J_LISTE (j_1, j_2 , ..., j_q)

Le résultat est correct seulement si la propriété d'unicité de la clé primaire, qui regroupe tous les attributs des clés étrangères, est vérifiée.

4.1.3. *La relation achat (no, période, produit, tag, u, v, q1, q2)*

La méthode mise en œuvre est présentée sur l'exemple de la relation achat de la base secodip. Elle est transposable sur la relation achat de la base INSEE.

Les relations ACHAT ne contiennent pas l'observation exhaustive de tous les produits, pour tous les ménages et toutes les périodes (la relation ACHAT_P2 contiendrait alors plus de cinq milliards de triplets !). Pratiquement, si un produit n'est pas référencé pour une période et un ménage, cela traduit une absence d'achat du produit, donc des quantités nulles. Ces valeurs nulles (1) peuvent être introduites par modification de l'algorithme précédent, en testant la présence de tout quadruplet (*no*, *periode*, *produit*, *tag*), défini sur le produit cartésien des relations I_LISTE et J_LISTE dans la relation ACHAT_P2.

Par ailleurs, il n'y a plus un seul attribut qui n'appartient pas à la clé primaire, mais quatre pour cet exemple. La solution retenue pour traiter ce problème consiste à éditer un tableau statistique dans lequel ces quatre attributs sont rangés dans des colonnes consécutives.

Les règles de construction sont les suivantes :

- la première colonne contient le numéro d'ordre de la ligne ;
- les seconde et troisième colonne contiennent respectivement les valeurs des attributs *no* et *période*.
- Les colonnes qui suivent correspondent aux valeurs des quatre attributs *u*, *v*, *q1* et *q2*, observés pour les couples (*produit*, *tag*) présents dans la relation ACHAT_P2 ;
- à chaque couple (*no*, *periode*) de la relation ACHAT_P2 correspond une ligne du tableau.

(1) L'application de cette règle peut être liée à la valeur d'autres variables : le problème des données manquantes n'est pas traité dans ce rapport.

Le schéma ci-dessous illustre la structure d'un tel tableau.

numero ligne	n°	periode	produit1, tag1				produit2, tag1				produit2, tag2				etc...			
			u	v	q1	q2	u	v	q1	q2	u	v	q1	q2	u	v	q1	q2

L'algorithme qui produit cette structure de données et édite le résultat dans un fichier texte destiné aux logiciels statistiques est une généralisation des algorithmes précédents. Il est décrit dans l'encadré numéro 3. Sa traduction a été faite en langage C et la procédure est opérationnelle pour "vectoriser" toute relation de la forme :

$$R(\underline{i1}, \underline{i2}, \ldots, \underline{ip}, \underline{j1}, \underline{j2}, \ldots, \underline{jq}, x_1, x_2, \ldots, x_r)$$

Encadré n°3 : algorithme général de la fonction de vectorisation.

Début

Pour tout i_uplet (i1, i1, ..., ip) de la liste des i_uplets de la relation descriptive des observations

Faire

- éditer le numéro de la ligne ;
- éditer les p attributs i1, i2, ..., ip ;
- pour tout j_uplet(j1, j2, ..., jq) de la liste des j_uplets de la relation descriptive des variables

  Faire

  - si le n_uplet (i1, i2, ..., ip, j1, j2, ..., jq, x1, x2, ...,xr)

    existe alors éditer les r attributs x1, x2,..., xr,

    sinon éditer r valeurs nulles ;

  fin si ;

  fait ;
- éditer une marque de fin de ligne ;

fait ;

Fin ;

4.2.La production de relations intermédiaires soumises au processus de vectorisation

La vectorisation de la relation achat associée au panel P2 de la base SECODIP aurait abouti à un tableau de plusieurs centaines de milliers de lignes et colonnes : L'algorithme associé ne peut donc pas être appliqué sur une relation qui contient l'intégralité de l'information.

4.2.1.*Exemple : la consommation annuelle de viande d'après le panel SECODIP*

Supposons qu'il s'agisse d'extraire les achats de viande des ménages, avec une ligne par ménage et agrégation des achats relatifs à toutes les périodes. Après examen du dictionnaire SECODIP, ces produits sont référencés par le code marché "243".

On souhaite récupérer une matrice pour lesquelles les colonnes correspondent aux valeurs distinctes prises par l'attribut e_2, ce qui donne en fait une typologie des consommations de viande, avec 38 rubriques .

La démarche consiste à produire des relations temporaires intermédiaires.

Une première interrogation produit la table MARCHE_243 (*no, période, tag, e2, u, v, q1, q2*). Cette table peut ne pas être une relation dans la mesure où les doublons sont conservés. Elle est créée à partir de la vue secodip_p2, avec le prédicat de sélection (marché = 243).

L'opération relationnelle de projection va supprimer les attributs *période* et *tag* en regroupant et en additionnant les quantités afférentes aux couples distincts (*no, e2*). Au passage, il peut être judicieux de transformer les types des attributs *v, q1 et q2* en mode dit flottant, voire en double précision si les projections sont accompagnées d'extrapolation de valeurs.

Ce processus permet de fabriquer la relation (*no, e2, u, v, q1, q2*). La nature des opérations relationnelles réalisées fait que le couple (*no, e2*) forme une clé primaire unique.

L'algorithme de vectorisation décrit précédemment est appliqué : le prototype actuellement utilisé lit les fichiers images des relations, ordonnés sous INGRES. Il restitue un tableau, avec les 5 037 ménages du fichier prodpan en ligne. On retrouve

en colonne 38 paquets de 4 variables, qui prennent pour valeur celle des 4 attributs *u, v, q1, q2*, pour les 38 produits répertoriés.

4.2.2. *Production d'un tableau avec en ligne les ménages SECODIP, et en colonne les quantités achetées selon la nomenclature de produits de l'INSEE.*

L'enquête INSEE 1987 répertorie 327 produits. Un objectif est d'interroger la relation achat de SECODIP dans une vue logique qui contiendrait comme attribut d'identification du produit le code de la nomenclature analytique de l'INSEE.

Du point de vue du calcul relationnel, il suffit de pouvoir générer une relation qui aurait la structure suivante :

INTERPRETE (*marché, e1, e2, e3, e4, e5, e6, insee*).

Cette relation est utilisée pour générer une table virtuelle (vue logique) :

INSEE_P2 (*no, période, produit, tag, insee, , u, v, q1, q2*).

De cette table virtuelle est extraite une relation réelle, par projection obtenue en supprimant les attributs *produit, tag* et *période* et en effectuant le cumul des quantités. Cette dernière relation possède une clé primaire unique (*no, insee*).

Le processus de vectorisation peut être appliqué et les traitements statistiques effectués , sur une matrice à 5037 lignes et 327 colonnes.

Dans ce processus, l'attribut période aurait pu être conservé, le critère d'observation étant alors le couple (ménage, période). Pratiquement, le tableau résultant regrouperait 261 924 lignes, ce qui peut poser des problèmes de ressource disque avec le logiciel SAS.

Un compromis serait de travailler sur des périodes mensuelles ou trimestrielles. La construction de telles relations intermédiaires avec INGRES ne soulève pas de difficultés.

4.2.3. *Génération de variables calculées sur plusieurs générations d'une arborescence.*

Les méthodes statistiques utilisées mobilisent des variables construites par agrégation d'informations disponibles à différents niveaux d'une arborescence : poids des ménages, nombre de personnes d'un ménage, nombre de semaines pour lesquelles un ménage a rempli un carnet de comptes. Le calcul relationnel s'avère être particulièrement efficace pour construire les vecteurs d'observations associés à ces variables.

Le poids du ménage est extrait de la relation PANEL_P2 par simple interrogation sur le critère de la valeur d'un numéro logique de variable. L'opération rend la relation :

POIDS (*no*, *poids*),

dans laquelle l'attribut *poids* désigne la pondération affectée au ménage identifié par l'attribut *no*.

Le nombre d'individus par ménage est calculé par une projection de la relation INDIVIDU_P2 qui supprime l'attribut *rang*, après dénombrement des lignes pour chaque valeur de l'attribut *no* d'identification des ménages. Le résultat est une relation :

FAMILLE (*no*, *ni*)

qui contient le nombre d'individus *ni* du ménage *no*.

Le calcul du nombre de semaines d'activité des ménages enchaîne plusieurs opérations de calcul relationnel. Une première projection supprime les attributs *tag* et *produit* et rend une relation intermédiaire :

ACTIF (*no*, *période*),

qui répertorie les semaines pour lesquelles les ménages ont rempli au moins une ligne du carnet de comptes.

La projection de l'attribut *période*, avec dénombrement des lignes pour chaque valeur de l'attribut d'identification des ménages no, rend une relation :

ACTIVITE (*no*, *ns*),

qui contient le nombre de semaines *ns* dites d'activité pour le ménage *no*.

Le calcul se poursuit avec l'opérateur relationnel de jointure naturelle, qui met en correspondance les attributs *poids*, *ni* et *ns* pour les valeurs de la clé de jointure *no*. Le résultat est une relation

ECHANTILLON (*no*, *poids*, *ni*, *ns*),

qui contient toute l'information utile.

Un ordonnancement sur la clé primaire *no* est réalisé, et la relation est copiée dans un fichier sous forme d'un tableau statistique à 4 colonnes. Ce tableau peut être lu par le logiciel SAS, qui interprète cette fois la colonne associée à l'attribut *no* comme une variable utilisable pour réaliser des opérations de fusion avec les autres tableaux extraits de la base de données.

Conclusion

L'expression "bases de données" est communément employée pour désigner un système informatique au sein duquel toute personne, quelle que soit sa qualification, accède à des informations. On demande volontiers à un tel système de guider infailliblement un personnage virtuel, dit "utilisateur", dans les arcanes de ses préoccupations.

Dans le contexte de ce projet, le concept de "base de données relationnelle" fait référence à une méthode mathématique de calcul de type ensembliste. Elle a été choisie parce qu'elle s'avère être efficace pour structurer rigoureusement de grands ensembles d'informations statistiques, en amont des calculs numériques.

L'outil élémentaire qui permet d'exprimer les opérations de calcul relationnel est un langage informatique, connu sous le sigle SQL (Structured Query Language), d'origine IBM. D'autres langages sont disponibles (dont le langage QUEL développé pour le prototype INGRES par l'Université de Berkeley), mais SQL est devenu de fait un standard partagé par tous les systèmes de gestion de bases de données, dont INGRES. Nous n'introduirons pas ici un débat sur les qualités et les propriétés de ces langages.

Pour les bases provisoires INSEE et SECODIP décrites dans ce rapport, le moteur relationnel est piloté interactivement avec SQL : le "pilote" est seul responsable de la cohérence et de la complexité des calculs relationnels exécutés. INGRES met cependant à disposition un "tableau de bord" qui fournit sur demande toutes les informations utiles pour évaluer la complexité des algorithmes, mesurer les performances et éventuellement limiter la consommation de ressources[1]

Les langages de développement d'applications dits de "quatrième génération", n'ont pas été étudiés : nous nous sommes restreints, dans un premier temps, aux seules fonctionnalités du moteur relationnel, et en particulier à son comportement aux limites, avec comme opérandes des relations de plusieurs millions de lignes.

Dans la mesure où les résultats obtenus donnent satisfaction et sont conformes aux espérances théoriques, nous prévoyons de poursuivre ces investigations en expérimentant le langage de développement d'applications proposé par INGRES en environnement graphique. L'objectif est d'ouvrir l'accès pour des bases développées sur des serveurs UNIX à des postes de travail MSDOS, OS2 ou MAC dits "clients", qui voient la base avec leurs propres interfaces graphiques.

---

[1] Ces fonctionnalités ne sont pas disponibles dans le module de base, mais fournies dans le module dit de "gestion des connaissances".

Des procédures de niveau "moteur" restent à créer. Nous avons évoqué la production dans la base SECODIP d'une relation, nommée interprète, qui établirait la correspondance entre les codes produits SECODIP et la nomenclature analytique de l'INSEE. Il est possible d'établir par des méthodes empiriques une telle correspondance pour une partie des produits de la base SECODIP, mais l'approche est coûteuse en ressources humaines, partielle et accessoirement ambiguë, dès lors que les règles de décision ne sont pas formalisées dans un système.

En s'appuyant sur la théorie des langages, très utilisée dans les techniques de compilation, il paraît raisonnable d'envisager la production d'un automate qui appliquerait des règles de codification au vu des textes descriptifs associés aux codes SECODIP. Cette voie reste à explorer, la traduction de cet automate pouvant se faire en langage C ou sous forme de règles dans le module de gestion des connaissances de INGRES.

A terme, l'ensemble de ces procédures, existantes ou à créer, est à intégrer au sein d'un système homogène, accessible à travers un langage simple adapté à la culture scientifique des disciplines concernées. Pratiquement ce processus est amorcé pour les bases INSEE avec la production d'une interface entre le logiciel LEDA et le système INGRES. Si les résultats donnent satisfaction, les outils "système" développés pour convertir les représentations binaires élémentaires et parcourir l'arborescence, traduits en langage C, pourront être installés dans le noyau du moteur relationnel de INGRES, à travers une approche dite "type abstrait".

A l'issue de cette première étape, à la fois théorique et pratique, l'architecture informatique pour l'Observatoire des Consommations Alimentaires se profile progressivement : d'un côté le produit INGRES, pour lequel il semblerait que nous innovions dans le domaine de la gestion des grandes bases de données statistiques ; de l'autre côté le produit SAS pour les traitements statistiques.