Banque de données et méthodes interactives de modélisation économétrique

par

H. BOGAERT,
T. DE BIOLEY,
J.-M. PAUL

Bureau du Plan - BRUXELLES

I. INTRODUCTION

Le 23 mars 1976 une convention a été signée, à l'initiative du Bureau du Plan, entre l'Etat belge et quelques centres universitaires.

Elle avait pour but de permettre aux principaux utilisateurs de modèles économétriques de grande taille en Belgique de mettre en commun leurs données et leurs programmes et surtout de réaliser ensemble une banque de données permettant d'alimenter les différents modèles. Le Centre de Traitement de l'Information du Ministère des Affaires Economiques était choisi comme centre hôte pour ce projet.

Une Commission centrale composée de représentants de divers centres participant à la Convention gère le projet.

L'objet de la présente communication est double et vise :

1.- à décrire les réalisations auxquelles a donné lieu la convention précitée (appelée communément "Convention Minibank" et à fournir une évaluation de ce projet;

2.- à présenter, à partir de cet examen, un schéma d'extension et de généralisation.

II. DESCRIPTION ET EVALUATION DE L'EXPERIENCE MINIBANK

1.- Développement des fichiers

a) Conception technique des fichiers

La base de données Minibank est constituée d'un ensemble de fichiers accessibles dans le langage interactif APL et appelés fichiers T.S.A.R. (Time Series APL Retrieval). Chacun de ces fichiers est un ensemble logique de trois fichiers : l'un contient les séries numériques, l'autre la description du contenu de chacune d'elles, la troisième l'index du fichier. La liaison biunivoque établie entre les deux premiers par le troisième permet une documentation complète et précise de chaque série. Les renseignements stockés pour une série sont :

- dans le fichier de données numériques :
  * un header (périodicité, année et période de départ)
  * les observations
- dans le fichier de description :
  * un libellé : brève identification
  * une carte codée (source, aire géographique couverte, unités, année de base, type d'observation, échelle, . . .)
  * un commentaire facultatif.

Au niveau d'un fichier TSAR, chaque série est déterminée par un nom de 6 caractères alphanumériques, ce qui rend sa dénomination compatible avec l'APL ou le FORTRAN. Une série d'observations est donc complètement identifiée dans l'ensemble de la base de données par un nom de fichier TSAR

et un code de six caractères.

b) Contenu, évaluation et champ couvert par l'information

Actuellement, Minibank regroupe une série de programmes dont il sera question plus loin et environ 20.000 séries statistiques regroupées en une vingtaine de fichiers.

Dès le départ, l'accent a été mis sur la constitution d'une information de qualité en donnant la priorité aux besoins les plus urgents des utilisateurs de modèles. L'option consistant à stocker indifféremment toutes les données disponibles a été catégoriquement rejetée.

Deux principes gouvernent le stockage de l'information dans Minibank :

- le respect des sources utilisées et le stockage sous la forme exacte où l'information est publiée;
- le regroupement et l'organisation en vue d'une utilisation pour l'analyse économique et l'alimentation des modèles.

Le respect simultané de ces deux principes qui peuvent être contradictoires est réalisé grâce à :

- une documentation particularisée pour chaque série stockée dans la banque; la liaison documentation-série permet un très grand degré de précision dans la description du matériel statistique stocké;
- la construction de fichiers spécifiques de séries retravaillées lorsque cela s'avère nécessaire.

L'application de ces règles laisse toujours aux utilisateurs la liberté de disposer des séries primaires pour en faire la transformation qu'ils jugent utile mais leur offre aussi la possibilité de disposer directement de séries adaptées et transformées pour une série de besoins courants et communs à tous.

2.- Développement des programmes de gestion et d'utilisation

a) Programmes de gestion

Les programmes de base des fichiers TSAR (création, mise à jour, back-up, édition) sont écrits en PL1 et prennent comme input des cartes codifiées selon des règles strictes. Ces cartes sont générées automatiquement par un programme de préparation d'input pour la création de séries nouvelles ou calculées, pour la mise à jour de commentaires ou de séries numériques, la destruction de séries, etc . . .

Ce programme permet notamment de mettre en connexion un fichier TSAR et un jeu d'identités définissant des séries. L'ensemble de ces programmes est la clef de voûte de la gestion des données de Minibank.

b) Programmes d'utilisation

Un des objectifs de Minibank est de permettre l'utilisation des données dans le langage interactif APL particulièrement puissant pour les applications mathématiques. Les programmes économétriques

développés le sont donc en APL, ce qui donne toute sa valeur à un système de gestion de données développé dans cette optique.

Ils répondent à trois fonctions :

- Interrogation de la base de données.
- Estimation économétrique.
  Un interface entre le Times Series Processor (TSP) et le système TSAR a été implémenté, permettant à un utilisateur APL l'utilisation du TSP et le stockage des résultats en fichiers APL.

  En outre, un système conventionnel (AUTOSAMI) intégré permet à l'utilisateur de se limiter à introduire une ou plusieurs équations et à spécifier les paramètres d'estimation.

  La lecture des données dans un fichier TSAR, le stockage des résultats et la gestion complète du système sont automatisés.
- Création des données et d'une routine de simulation.
  Le passage à la phase de simulation se fait par l'implémentation d'un programme qui, prenant comme input une liste d'équations gérées par AUTOSAMI (cfr supra) génère une routine d'équations et d'identités traduites d'APL en FORTRAN et qui s'insère dans un software de simulation ainsi que le fichier des données de simulation, d'après des noms de séries repris dans les équations spécifiées.

3.- Evaluation

a) Structure des informations

Au niveau conceptuel on peut regrouper les avantages et inconvénients de la structure actuelle des informations selon trois aspects : l'unité d'information, l'identification de ces unités, la documentation de ces unités.

i : l'unité d'information des fichiers TSAR est la série chronologique.
La principale difficulté de cette organisation est qu'une variable peut être observée selon des périodicités différentes, des périodicités aléatoires ou même être sans périodicité dans le cas de séries transversales. Ceci pose donc le problème d'une part qu'avec un même concept de variables on peut référencer plusieurs types de séries selon les différences de périodicité, d'autre part, les périodicités aléatoires entraînent que le vecteur des nombres référencés par une variable pourrait contenir des éléments inexistants.

ii : L'identification des séries statistiques ne peut se faire actuellement que sur base du nom codifié de la variable ou par l'index correspondant à la position relative de celle-ci dans le fichier.
Ceci impose aux utilisateurs de disposer d'un index contenant la codification et la signification en "clair" des codes. En outre, il faut que l'utilisateur sache dans quel fichier TSAR il sera

susceptible de trouver les variables qui correspondent à sa recherche.

Les critiques émises au sujet de l'identification et de l'accès aux séries chronologiques peuvent se résumer en trois points :

- L'absence de transparence entre le niveau d'implémentation physique des données et la manière dont l'utilisateur voit ces données.
- L'accès à l'unité d'information, c'est-à-dire la série : les accès à un ensemble de variables ou à des références temporelles plus précises que celles qui sont données par l'entièreté de la série statistique demande un minimum d'effort de programmation.
- Alors qu'à l'origine, le concept de "fichier TSAR" n'avait pas été conçu délibérément, actuellement on constate que le concept de fichier est nécessaire et même indispensable au bon fonctionnement du système.

iii : Au point de vue de la documentation des séries, l'avantage considérable d'avoir un système qui permet que toutes les séries soient documentées autorise la mise à la disposition des utilisateurs de séries très différentes à la fois dans leur forme et dans leur contenu.

Au niveau de l'implémentation physique des structures d'information, deux critiques ont été mises en avant :

i : La redondance au niveau du stockage de nombreuses informations telles que certains commentaires des caractéristiques de mise à jour : date de mise à jour, etc. . .

ii : A un fichier TSAR correspond une table d'index qui doit se trouver en mémoire centrale, c'est-à-dire en WORKSPACE APL, pour pouvoir accéder aux séries contenues dans le fichier. Etant donné la contrainte d'espace de travail de 120 K on a toujours été obligé de maintenir une taille relativement réduite à chaque fichier TSAR. Cette même contrainte entraîne qu'il est difficile d'accéder à de nombreux fichiers TSAR en même temps; en effet, ceci impliquerait d'avoir chaque index de chaque fichier en WORKSPACE.

Une contrainte d'ordre technique a entraîné la nécessité du concept de fichier qui dans une banque de données, devrait être transparent pour l'utilisateur.

b) Contenu des fichiers

A l'issue des cinq années d'existence de la convention Minibank on peut résumer comme suit ses principaux apports :

- l'organisation des fichiers en thèmes distincts et leur gestion par des spécialistes des questions qu'ils recouvrent conduit à une très bonne qualité de la maintenance;
- la qualité des statistiques reprises dans la banque est sans doute la meilleure possible compte tenu

de l'information disponible. L'avantage majeur de la banque de données est d'offrir aux utilisateurs un ensemble, sans doute unique en Belgique, de statistiques homogénéisées. Le nombre et la qualité des participants a créé une demande qui a justifié, de la part des fournisseurs d'informations, une recherche et un travail important de traitement des données.

On peut conclure que le contenu actuel des fichiers couvre relativement bien la majorité des besoins des principaux utilisateurs. La fiabilité des fichiers est très grande, leur mise à jour rapide et en général ils couvrent de manière très exhaustive, les domaines auxquels ils se rapportent. Au niveau de l'ensemble, il existe toutefois certaines grandes lacunes : statistiques monétaires, statistiques du commerce international.

c) Programmes de gestion et d'utilisation

Les programmes de gestion se caractérisent par leur robustesse et, ce qui est à la fois un avantage et un inconvénient, par la rigidité qu'ils imposent à l'introduction des données. Ils sont en fait les garants de la maintenance correcte des données et, à cet égard, ont donné entière satisfaction.

D'autre part, les programmes exploitant les données Minibank présentés en II.3 sont d'un emploi très aisé. Grâce au développement de programmes conversationnels, l'utilisateur peut, sans être familiarisé à l'APL, ni même à aucun langage de programmation, en tirer un profit maximum.

Si certains points doivent encore être étayés, l'ensemble de ces programmes créés en collaboration avec des économètres, semble rencontrer la plupart de leurs besoins.

III. PROJETS D'EXTENSION ET DE GENERALISATION

1.- Gestion et utilisation de l'information

a) Meilleure intégration et meilleure mise à la disposition des utilisateurs

Sous leur forme actuelle, la plupart des programmes d'utilisation (II.3) présentent un niveau d'intégration avancé. Il reste cependant souhaitable :

i : de développer un programme de recherche de données utilisable sans prérequis informatiques, avec possibilité de réaliser des opérations arithmétiques simples, d'éditer des graphiques, etc . . . Possibilités également de rechercher des séries dans plusieurs fichiers TSAR simultanément, sur base d'un autre critère que le code des séries (recherche par sujet), etc. . .;

ii : d'intégrer l'ensemble des programmes existant dans un programme général présentant à l'utilisateur toutes les possibilités offertes et indiquant les marches à suivre possibles;

iii : de travailler dans le sens des développements d'un plus large choix de programmes économétriques écrits en APL : analyse des données, programmes d'estimation plus spécialisés, etc . . .

b) Elargissement du champ d'application

Un des objects d'un développement futur serait d'offrir la possibilité de gérer, dans Minibank, non seulement des séries vectorielles (en général temporelles), mais aussi des matrices de dimensions quelconques (carrées, cubiques, . . .) comme telles, et non plus comme ensembles de séries vectorielles.

2.- Structure de l'information

Le but d'une révision de la structure d'information est d'améliorer le système dans le sens de l'élaboration d'une véritable banque de données c'est-à-dire d'un ensemble d'informations accessibles par un utilisateur non spécialiste, ces informations ayant pour propriétés d'être intégrées, non ambiguës et constituant le moins de redondance possible au niveau de l'implémentation physique.

En outre, le concept de banque de données laisse supposer que l'utilisateur dispose d'un outil de recherche des informations à l'intérieur de la banque de données, cet outil de recherche étant plus ou moins sophistiqué mais dégagé de toute façon de tout concept d'implémentation physique.

a) Etude théorique de la structuration des informations

L'étude théorique de la structure des informations a déjà été mise en oeuvre et permet de dégager un certain nombre de propriétés fondamentales des données économiques.

i : les caractéristiques relationnelles (1)

L'unité d'information statistique peut être considérée comme la mesure d'un phénomène. Ce phénomène étant rationalisé et décrit par un modèle (implicite ou explicite) de la réalité. Le modèle de perception de la réalité est constitué de façon très abstraite par trois fondements :
- l'identification des entités qui interagissent;
- l'identification des types d'interaction;
- la localisation dans l'espace et dans le temps de ces interactions.

A chacun de ces fondements correspond une terminologie bien connue des utilisateurs de modèles économétriques. Le concept d'agent économique, le concept de variable et de type de variables permettent de modéliser la réalité économique en en décrivant les relations entre agents économiques et en mesurant ces relations au moyen de variables.

Une observation statistique est la mesure d'une relation localisée dans temps et espace entre des entités prises en compte dans un modèle de la réalité.

Les entités et relations reçoivent une acception différente selon le modèle envisagé : il est clair par exemple que les types d'agents économiques recensés ne sont pas les mêmes selon la comptabilité nationale et les statistiques monétaires. A chaque modèle correspondent en fait des entités dont le domaine de définition doit être précisé soit en extension si on donne la liste des entités référencées, soit par un critère de définition.

De sorte qu'une observation statistique est une occurrence d'une relation entre différents domaines de définition. Un exemple : l'importation belge de pétrole en provenance de Koweit peut être décrite comme une relation R entre les domaines D1 : pays origine
D2 : pays destination
D3 : un bien
D4 : un montant

Chacun de ces domaines de définition sont décrits comme des ensembles :

- en extension, c'est-à-dire par détermination d'une nomenclature : l'ensemble des pays, l'ensemble des biens.
- par un critère : le montant est un chiffre E à R dont l'unité est le dollar dont la valeur d'échelle est le milliard.

La relation R est définie sur un cinquième domaine qui correspond à la localisation du montant dans le temps.

D5 : temps, ici la période allant du 1er janvier 1981 au 31 décembre 1981.

Formellement la relation R : importation est décrite de la façon suivante :

R : importation (D1 : Pays, D2 : Pays, D3 : Biens, D4 : montant, D5 : TEMPS).

Cette relation est parfaitement compatible avec l'algèbre relationnelle dont les opérations s'appliquent bien entendu à cette relation.

Par exemple : une série chronologique des importations de pétrole en provenance de Koweit est une projection de la relation R sur un triplet (D1 = Koweit, D2 = Belgique, D3 = pétrole).

Du point de vue de l'utilisateur, l'unité d'information à prendre en considération est celle qui est définie par les occurrences d'une relation non décomposable. La structure de l'unité d'information correspondant à une observation statistique sera donnée par un nom de relation (correspondant au type de variables économiques) et aux occurrences des domaines de définition que cette relation met en concordance.

Au niveau de la banque de données en général il conviendra donc de déterminer les domaines de définitions, soit en extension par un fichier de nomenclature, soit par des contraintes d'intégrité pour les domaines qui sont définis par une règle.

ii : Les caractéristiques fonctionnelles

Si les caractéristiques relationnelles définissent complètement le concept de variables économiques, des caractéristiques fonctionnelles doivent en outre être ajoutées pour préciser l'état dans lequel

se trouve l'observation statistique définie par les caractéristiques relationnelles.

Les caractéristiques fonctionnelles qui sont traditionnellement reprises dans les fichiers TSAR sont : l'auteur de la mise à jour, ou de la création, la date de la mise à jour de la création, la source bibliographique, le statut de l'observation : définitif, provisoire . . .

b) Utilisation des structures relationnelles dans la conception d'une banque de données statistiques

Une variable économique peut être décrite formellement par les concepts de :

- domaine de définition;
- relation entre ces domaines de définition;
- occurrence d'une relation, à savoir un n-uplet dont les éléments appartiennent au domaine des définitions mis en relation.

Sans modifier de façon substantielle les fichiers TSAR actuels, qui seraient considérés dans une première étape comme l'implantation physique de tous les n-umplets de toutes les relations contenues dans la banque de données, c'est-à-dire comme les occurrences de toutes les observations statistiques, il y aurait moyen d'implanter une superstructure décrivant les modèles de définition et les relations qui sont implicitement contenues dans la banque de données.

Le but de cette superstructure serait double : d'une part, l'élaboration d'un langage de désignation plus général, d'autre part, l'élaboration d'un langage de manipulation plus général pour l'utilisateur.

i : Elaboration d'un langage de désignation

Si l'on reprend l'exemple d'une relation R sur les domaines A, B, C, D que l'on pourrait concrétiser, par exemple, comme étant la relation importation sur les domaines : pays d'origine, pays de destination, biens importés, montant des importations, il y a moyen de concevoir une désignation des variables économiques ou d'un ensemble de données économiques.

En effet, on peut désigner l'ensemble des données d'importation par la relation R sur l'entièreté des domaines A, B, C, D, où on peut désigner un sous-ensemble des importations en sélectionnant les occurrences relatives à des sous-ensembles des domaines de définition. Un cas particulier étant le cas de l'opérateur de projection de la relation R sur des occurrences d'un sous-ensemble des domaines de définition. Exemple : les importations de la Belgique en provenance de l'Allemagne Fédérale constituent un sous-ensemble de la relation importation projetée sur le couple des domaines de définition, pays d'origine-pays de destination dont l'occurrence est (Belgique-Allemagne Fédérale).

ii : Elaboration d'un langage de manipulation pour l'utilisateur

Un utilisateur que nous supposons toujours être un utilisateur APL doit pouvoir disposer d'opérateurs de consultation de la banque de données qui soient cohérents et si possible intégrés au langage APL.

Reprenant l'exemple d'importation décrit plus haut, on peut concrétiser la relation R sur les domaines A, B, C, D dans une workspace APL de deux manières :

* Par un tableau

R :

A	B	C	D
a_1	b_1	c_1	d_1
a_2	b_2	c_2	d_2
a_3	b_3	c_3	d_3
	.		
	.		
	.		
	.		

* Par une matrice à trois dimensions, par exemple les dimensions A, B, C. Cette matrice contient alors les éléments de D classés selon les axes A, B et C mais ceci implique que les triplets (a_1, b_1, c_1) soient non redondants.

Ce passage à la forme matricielle permet d'utiliser tous les opérateurs matriciels disponibles en APL.

Le langage de manipulation peut dès lors comporter les opérateurs suivants :

Opérateur T : T R (A, B, C, D) qui amène une table en WS.

Opérateur M : R (A, B, C, D) M D qui amène en WS la matrice définie sur l'ensemble D.

Opérateur P : R (A, B, C, D) P A = a, B = b qui réalise la projection de la relation R sur les éléments du sous-ensemble a de A et b de B, ou plus généralement sur la relation R' (A, B).

Opérateur J : qui réalise la jointure de deux relations : soit la relation R_1 (A_1, B_1, C_1, D_1) et la relation R_2 (A_1, B_1, C_1, D_2).
On peut définir une jointure des relations sur des domaines communs.

Opérateur A : R (A, B, C, D) A R' (A, B, C, D) réalise la suppression d'adjonction des n-uplet de la relation R et de la relation R'.

Opérateur S : R (A, B, C, D) S R' (A, B, C, D) réalise la suppression des n-uplet contenus

dans R et désignés par les n-uplet de R'.

Avec ces 6 opérateurs plus un opérateur qui permettrait de définir une relation nouvelle, on disposerait des éléments d'un langage de manipulation des relations et donc des données de la banque de données qui seraient compatibles avec le langage APL.

iii : Implantation physique de concepts relationnels

L'implantation des structures relationnelles peut se faire en partant des fichiers actuels et en liant les records contenant les séries chronologiques à une superstructure par un système de pointeurs.

On aurait alors un système à trois étages :

Etage 0 : les occurrences des relations, c'est-à-dire :
- les séries chronologiques,
- les pointeurs vers des relations et des domaines de définition,
- les caractéristiques fonctionnelles,
- un libellé.

Etage 1 : décrivant l'ensemble des relations se trouvant dans la banque de données, les liens de ces relations avec les données de définition et la localisation des relations dans les fichiers TSAR.

Etage 2 : les domaines de définition.

Dans un stade ultérieur, il conviendrait de refaire la structuration de l'information au niveau de l'implantation physique pour tenir compte des problèmes :
- de rapidité d'accès,
- de non redondance de l'information en rejetant une grande partie de la documentation qui se trouve actuellement au niveau des séries vers les structures d'information qui décrivent les domaines de définition des relations ou les caractéristiques des données.

IV. CONCLUSION

En guise de conclusion, on essayera de mettre en exergue ce qu'on peut considérer comme les deux points qui distinguent le projet Minibank d'autres banques de données statistiques et économiques.

a. Le projet ne s'est pas développé à partir d'une vision théorique de son état final, mais a, au contraire, été conçu pour répondre aux besoins existants d'utilisateurs différents. Toutefois, une structuration extrêmement souple a été définie au départ de manière à permettre une évolution vers un système dont le but et les fonctions (répondant aux besoins des utilisateurs de modèles économétriques) ne sont définis que provisoirement.

b. Le système a été d'emblée conçu pour permettre la consultation aisée de l'information statistique y compris sa documentation, et pour servir de support, sans interface particulier aux softwares d'estimation et de construction des modèles économétriques. C'est le choix qui a justifié l'adoption de l'APL comme langage de base.

L'extension prise par le projet et la nécessité d'asseoir l'estimation et la simulation de modèles économétriques sur des ensembles de données de plus en plus larges justifient que, tout en gardant à Minibank sa spécificité propre et ses points forts, on opère un "changement d'état" pour passer d'un ensemble de fichiers à une base de données intégrée et directement utilisable en APL. L'utilisation de structures relationnelles et l'élaboration d'un langage de manipulation intégré à l'APL répond à cet objectif.

NOTE

(1) Les concepts relationnels font référence aux concepts définis dans les bases de données relationnelles. On peut trouver un exposé clair de cette approche dans :

a) CODD,E.F. : "Relationnal Model of Data for Large Shared Data Bank", ACM Communications, 13, 377-387, 1970.

b) MARTIN, James : "Computer Data-Base Organization",Second Edition, Prentice Hall.