JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

GÉRARD D'AUBIGNY

Discussion et commentaires. Data mining et statistique

Journal de la société française de statistique, tome 142, n° 1 (2001), p. 37-52

<http://www.numdam.org/item?id=JSFS_2001__142_1_37_0>

DISCUSSION ET COMMENTAIRES

Data Mining et Statistique

Gérard D'AUBIGNY *

RÉSUMÉ

Ce texte commente la présentation par Besse *et al.*, des liens entre exploitation de gisements de données et statistique. Il présente le data mining comme un nouveau paradigme, qui relève des sciences de l'ingénieur et vient enrichir les métiers de la statistique. Présentée comme un outil d'extraction de l'information dans les grandes bases de données, cette approche est contrainte aux analyses secondaires de données de simple observation. Les conséquences de ces caractéristiques sont abordées et l'absence de définitions d'un certain nombre de concepts du data mining est soulignée.

Mots-clés — Analyse des données, analyse exploratoire des données, data mining, exploitation de gisements de données, méthodologie, modélisation, prévision, statistique, validation.

ABSTRACT

This paper comments the presentation, by Besse *et al.*, of the close bonds joining data mining with statistics. Data mining is introduced as a new paradigma, whose arrival among engineering sciences enriches the catalogue of the statistical craft. Known as an information extraction tool from very large data bases, this approach has been restricted by its developpers to the secondary data analysis of observational data. The consequences of these distinctive attributes are discussed and the lack of accurate enough definitions of data minig concepts is pointed out.

Key words — Data mining, Exploratory data analysis, forecasting, methodology, modelling, Multivariate data analysis, statistics, validation.

1. Introduction

Besse *et al.* [3] ouvrent une discussion sur les évolutions actuelles des métiers de la statistique en entreprise. L'accueil réservé par les entreprises aux logiciels de Data Mining montre à l'évidence que l'action puissante de marketing menée autour de ce nouveau paradigme a su identifier un besoin, proposer juste à temps des solutions adaptées et le faire savoir. Effet de mode ou vague de fond, cette offre remodèle l'expression des attentes des entreprises, les

* LabSad, Université Pierre Mendès-France, Grenoble, France.
E-mail : Gerard.dAubigny@upmf grenoble fr

profils de compétences recherchés et les missions confiées aux chargés d'études spécialistes du traitement analytique de l'information.

Cet article marque une date importante, car contrairement à leurs homologues de langue anglaise, les revues françaises ont tardé à ouvrir leurs pages à un débat déjà largement engagé dans les milieux professionnels, des bureaux en conseil aux services d'études des entreprises. Il a suscité un nombre non négligeable de colloques, fait l'objet d'une pression médiatique et d'innombrabes actions de propagande. En France aussi, les offres d'emploi sont au rendez-vous et poussent les milieux académiques à proposer la création de cursus universitaires. La réponse des autorités de tutelles a pu leur paraître un peu déconcertante, car non nécessairement fondée sur une évaluation ou un repérage des pôles de compétences et sur une doctrine générale.

De plus, comme les auteurs s'en font l'écho, en France comme partout, les attentes suscitées par le data mining ont ouvert un débat sur la culture attendue des statisticiens, sur les champs de compétence reconnus à la statistique, ses articulations - en particulier institutionnelles - avec les mathématiques, les sciences de l'information et de la cognition.

Réagir à l'article de Besse *et al.* [3] nécessite de s'interroger sur son économie générale. Je commencerai, dans la section suivante par une brève mise en perspective historique, afin de distinguer la situation française des réalités anglaise et américaine. Je discuterai dans la section trois des aspects méthodologiques et des rapports entre l'exploitation de gisements de données, la statistique, l'informatique et les sciences cognitives. Enfin, la présentation d'exemples par les auteurs sera juste évoquée dans ma conclusion.

2. Le Data Mining : naissance d'un paradigme

Peut-on, comme Besse *et al.* [3] à la suite de beaucoup d'auteurs, situer le data mining dans la continuité des débats entre approches exploratoire et inférentielle de la statistique ? Je plaide pour une réponse négative, en m'appuyant sur les théories économiques de la recherche et du développement. Pour elles, le data mining a toutes les caractéristiques d'un paradigme, né des initiatives des acteurs du marché du logiciel. En cela il se distingue de disciplines établies telles que l'analyse des données et plus généralement la statistique.

2.1. L'Analyse des Données s'intéresse aux méthodologies

Tout d'abord, la difficulté à trouver un équivalent français au terme anglais de data mining rappelle à chacun les origines anglo-américaines de ce paradigme. Quand les auteurs non français rappellent sa filiation à l'*Analyse des Données* ils font référence aux travaux de Tukey [1] [35] et non à *l'école française d'analyse des données* [2] née de ceux de Benzécri. Mais présenter l'EDA ou

1. L'Analyse Exploratoire des Données (EDA : Exploratory Data Analysis)
2. (AEM : Analyse Exploratoire Multidimensionnelle pour reprendre le titre de Lebart [28])

l'AEM comme *un assemblage de techniques au sein d'un logiciel* me paraît un contresens. Bien au contraire, l'EDA comme l'AEM, trouvent leur fondement dans des principes philosophiques et méthodologiques en rupture avec la pensée statistique dominante de l'époque. Toutes deux ont certes utilisé les progrès technologiques pour traduire en logiciel les méthodes induites, mais en tant que moyen d'action au service d'une méthodologie.

Plus précisément, l'apport de Tukey m'apparaît comme une rupture avec la prédominance des questions d'inférence en statistique. Il traduit une volonté de rééquilibrage d'une discipline qu'il contribue à développer de façon experte, pour faire une place aux méthodes exploratoires dans un pays dépourvu de culture en statistique descriptive. En cela, le data mining peut se présenter comme un effort de propagation des idées de Tukey, en particulier le recours aux graphiques, bien que la place centrale accordée à l'œil par Tukey, suppose implicitement des petits jeux de données. Cette filiation est plus floue pour le choix des méthodologies privilégiées.

L'AEM marquait, elle aussi, une rupture culturelle, mais de nature distincte. Son apport est celui de la géométrie et de la découverte si fondamentale des relations mathématiques de dualité qui sous-tendent la relation entre espace de représentation des individus (entités microscopiques) et espace de représentation des variables (entités macroscopiques) instrumentée de façon canonique par le tableau de données [7], [5]. L'apport méthodologique de ces préoccupations mathématiques est très profond, mais il induit un enrichissement des méthodes, pas une rupture.

La doctrine de Benzécri sur la place du modèle me semble relativement découplée de cet apport. Très proche de la pensée de Tukey et de beaucoup de tenants actuels du data mining, elle fut sans doute soumise aux mêmes influences. De plus, comme Benzécri, au départ plus géomètre que statisticien, le data mining porte un regard différent de celui des statisticiens sur les modèles stochastiques, parce que marqué par sa culture informatique, plutôt déterministe.

2.2. Le data mining s'intéresse aux méthodes

L'époque a changé et la naissance du data mining me semble relever de réalités tout autres, de nature économique et directement induites par le progrès technique. Trois facteurs ont modifié le panorama de l'analyse des données.

1. La mise à disposition de grandes capacités de calcul à faible coût. La plupart des logiciels majeurs de traitement statistique sont nés dans les années 60 ou 70. La puissance de calcul des mainframes de l'époque était alors équivalente à celle d'un PC de base actuel. Cette réalité est de plus couplée aux très fortes réductions du coût et du volume physique des outils de stockage de l'information. L'information ainsi stockée fut alors présentée comme un actif dormant de l'entreprise, gisement potentiel de richesses inexploitées et de profits.

2. Une indéniable explosion de l'accumulation de données - très souvent collectées sans plan de sondage et sans référence à une population ou base de sondage, de simple observation plutôt qu'expérimentales - due aux faibles coûts de stockage, relayée par un effort réel d'automatisation des expériences, ou l'utilisation de procédures de collecte électronique de données. L'informatique de cette époque était de plus marquée par l'évolution des Systèmes de Gestion de Bases de Données [3], vers les Systèmes Informatiques d'Aide à la Décision [4] dont les diffuseurs utilisèrent comme argument de vente la possibilité offerte aux clients de rentabiliser leur achat en exploitant ces gisements dormants d'actions rentables. La mise à disposition de fonctionnalités de traitement de données réputées puissantes, intégrant en particulier des méthodes d'analyse statistique et graphique, devint un enjeu de pénétration du marché.
3. L'introduction de nouvelles méthodes logiques développées par la communauté des chercheurs en Intelligence Artificielle [5], les statisticiens (dont l'*apprentissage statistique* [36, 37] et l'*analyse des données symboliques*) et certains physiciens (travaillant sur les systèmes dynamiques non-linéaires). Dès cette époque, certaines retombées technologiques des modèles de cognition, en particulier d'apprentissage, développées parallèlement dans le cadre de travaux conceptuels allant de la psychologie à l'IA, fournirent à moindre coût des composants logiciels de data mining, telles les techniques de représentation des connaissances, ou encore des méthodes et techniques d'apprentissage offrant de nombreuses possibilités d'application pour la représentation et l'analyse des données, qu'elles renouvellent en complétant les méthodes statistiques traditionnelles.

Les conditions de naissance d'un paradigme étaient alors réunies, et une logique de marché promut alors le data mining, créant et vendant des produits logiciels aux entreprises pour occuper les nouveaux espaces de développement économique ainsi créés.

Le succès de cette opération de promotion est incontestable et ses effets sur ce secteur en expansion de l'informatique se traduisent classiquement par le grand nombre d'acteurs qui proposent une offre diversifiée de logiciels de data mining, apparus dans une courte période de temps et comptent parmi eux tous les grands du marché du logiciel statistique.

Comme tout paradigme tourné vers l'action, le data mining fut dès l'origine conçu comme outil d'exploitation de gisements de données et identifié à un produit logiciel, comme *un assemblage de techniques au sein d'un logiciel*, dont la cohérence n'est pas prioritaire. Il s'agit alors de gagner des parts de marché : plutôt que de les présenter du point de vue méthodologique, ces outils

3. SGBD = Système de Gestion de Bases de Données, en particulier les VLDB = Very Large Data Bases

4. SIAD : Systèmes Informatiques d'Aide à la Décision. En anglais, DSS : Decision Support Systems

5. IA : Intelligence Artificielle

logiciels furent vantés par des arguments d'efficacité compétitive, dont l'offre s'évalue en termes de richesse des fonctionnalités implémentées, de convivialité des interfaces, de puissance des algorithmes et de taille des corpus de données traités.

Le data mining s'intéresse aux méthodes plus qu'aux méthodologies et aux principes philosophiques qui motivaient les initiateurs de l'analyse de données et s'en distingue en cela : *il relève des sciences de l'ingénieur*, en particulier de l'informatique et du calcul scientifique. Il en résulte deux caractéristiques du data mining :

1. La première est un mode nouveau de développement et de diffusion des principes et méthodes statistiques, qui opère selon des principes communs aux sciences de l'ingénieur et inverse l'ordre des termes dans lequel est trop souvent enfermée la statistique. On fournit un produit (logiciel, perçu comme concret), qui répond à une demande d'action profitable, et nécessitera tôt ou tard un accompagnement méthodologique pour en maîtriser les fonctions. Les logiciels servent de support à la diffusion d'idées et de méthodologies, mais on n'est plus en situation de devoir convaincre d'abord des bienfaits méthodologiques possibles d'une mission fonctionnelle de conseil et d'aide à la décision, traditionnellement perçue comme abstraite, difficile à appréhender et à valoriser dans les entreprises. Cela change en conséquence le positionnement des chargés d'études spécialistes de l'exploitation de gisements de données dans l'entreprise.
2. Le fait que le data mining opère comme une science de l'ingénieur inverse aussi les termes d'une recherche de méthodes efficaces. Là où la statistique classique valide d'abord des méthodes nouvelles par la démonstration mathématique de leurs qualités de pertinence ou d'optimalité, puis valide leurs résultats par le traitement de données simulées et enfin par celui de données réelles, le data mining capitalise sur la puissance de l'outil informatique et les pratiques du machine learning. Il commence par expérimenter sur machine toute nouvelle heuristique pour en évaluer l'efficacité pratique, puis éventuellement passe à (ou confie à d'autres) l'étude de ses propriétés formelles.

Cette évolution motive les uns et suscite des réserves chez les autres, parmi lesquels beaucoup de statisticiens académiques, moins habitués à concevoir leur discipline dans un cadre heuristique d'aide à la décision qu'en corps de doctrine et de méthodologie, quand ils ne la perçoivent pas comme une sous discipline des mathématiques.

Pourtant, cette dialectique de démarches complémentaires trouve son analogue dans le dialogue entre approches expérimentale et théorique en physique, où elle est ressentie comme positive. Il suffit simplement d'assurer la pérennité des deux approches (sensibilités) et les conditions de leur dialogue fructueux. D'autre part, s'attacher aux seuls aspects négatifs de ce qui fut une entreprise commerciale paraît peu raisonnable. Il semble préférable d'analyser le phénomène, participer activement à ses développements et former pour limiter

les débordements prévisibles et préparer l'accompagnement méthodologique que nécessite ce nouveau métier de la statistique.

2.3. Conséquence sur les pratiques

Non encore stabilisé, le paradigme de data mining n'a pas encore apporté d'innovations méthodologiques déterminantes. Classiquement, l'urgence était d'implémenter les méthodes de l'analyse des données et de les calibrer pour traiter des données nombreuses et il a créé un espace de liberté, pour l'invention d'heuristiques nouvelles d'EDA, validées selon des critères d'utilité plutôt que mathématiques.

D'autre part, l'optique statistique, focalisée sur le traitement numérique des données, ignore les pratiques héritées des SGBD, des sciences cognitives et du machine learning. Ainsi réductrice, elle déforme notre appréciation du data mining. Pourtant, depuis les débuts de l'AEM, les statisticiens ont utilisé des fonctions d'interrogation/visualisation de la base de données de type query/update et des langages de requètes pour l'indispensable retour aux données. Les *méthodes post factorielles* [2, 8, 38] illustrent ces outils logiciels.

Quand l'objectif final d'une étude est d'acquérir de la connaissance sur le domaine d'intérêt, ces deux classes complémentaires d'outils sont nécessaires pour dégager une voie d'entrée économique organisant l'information pour produire une connaissance, vue comme de l'information mise en cohérence et en perspective. Cela nécessite des *a priori* et des stratégies d'analyse qualitative et quantitative. Pour atteindre cet objectif, la statistique offre un cadre méthodologique fort mais ni unique, ni dominant, et pas d'emploi le plus fréquent.

3. Data Mining et méthodologies de traitement de l'information et des connaissances

Le data mining n'échappe pas à l'ambiguïté des domaines de compétences couverts par tout nouveau paradigme. Les définitions affluent [4, 13, 14, 19, 20, 22, 31], rendant difficile son évaluation. Nous discutons ici deux aspects qui correspondent à la présentation des auteurs.

3.1. Le data mining adapte les techniques d'analyse des données au traitement de grandes bases de données ?

Une définition possible du data mining est la composante de la statistique attachée à traiter des grandes masses de données. Cet objectif relève des sciences de l'ingénieur et le facteur d'échelle pose un problème de *calibrage* (scaling) des techniques au traitement des VLDB, beaucoup plus familier aux informaticiens et numériciens qu'aux statisticiens. Mais, valider les méthodes du point de vue de leur efficacité, de façon différenciée en fonction des tailles de corpus des données, pose aussi des questions méthodologiques, qui ne relèvent pas uniquement de la Gestion des données et de l'algorithmique.

Besse *et al.* [3] arguent de l'intérêt des *entrepôts de données réparties* en termes de fiabilité (mais en quel sens ?) et de sécurité. Ce dernier point me paraît contestable à cause de la fragilité accrue de la base, qui résulte de la répartition. Assurer la *pérennité* de la disponibilité (*accessibilité*) de l'information et de son *intégrité* pose des problèmes redoutables. L'utilisation de la toile comme base de données documentaires réparties, ou comme base de données pédagogique illustre de façon très concrète le problème. Les adresses changent, les auteurs actualisent ou non leur page, les temps de réponse s'allongent, ... de sorte que l'hétérogénéité ne me paraît pas devoir être la difficulté principale.

D'autre part, les *Très Grandes Bases de Données* (VLDB) recouvrent des réalités fort contrastées pour le statisticien. Elles posent donc des problèmes statistiques différenciés. Trois caractéristiques au moins des données, de nature très différente, sont couvertes par ce vocable.

1. La référence première concerne les *Bases de Données Nombreuses* : des enregistrements en très grand nombre décrivent un ensemble important d'unités statistiques observées sur un nombre limité d'attributs. Chaque enregistrement est composé d'un petit nombre de champs. C'est une situation conceptuellement familière aux statisticiens. Faut-il traiter leur totalité ? Améliorer les algorithmes pour traiter de gros corpus de données ? Les algorithmes séquentiels ne sont pas tous efficaces, contrairement à ce que disent les auteurs, les méthodes adaptatives n'existent pas pour tous les problèmes et une réflexion sur l'apport des méthodes d'approximation stochastique dans ce cas, explorée en AEM [28] montre que des alternatives existent.

   L'une consiste à réduire les objectifs d'analyse pour rendre possibles les calculs, mais toutes les méthodes ne sont pas égales sur ce plan. Pour une même taille de fichier, une *Analyse de la Variance* ou l'étude d'un *modèle log-linéaire* sont plus gourmands en mémoire centrale qu'une *régression linéaire*.

   On peut aussi réduire les données pour en permettre une analyse classique : par exemple, extraire un ou des échantillon(s) pour rendre raisonnables les temps de traitement et limiter le poids des erreurs de mesure. C'est ce que font, sans toujours le dire assez nettement, certains logiciels du marché. Mais cette procédure de ré-échantillonnage est un peu étrange, lorsque l'ensemble des enregistrements collectés provient d'une procédure de collecte non maîtrisée (sans plan de sondage), dans une population non définie.

   D'autre part, moins connue des informaticiens, l'adoption de postulats de distribution de probabilité conjointe des variables étudiées permet de façon classique de réduire la taille des fichiers traités, par exemple en utilisant la notion de *résumé exhaustif* (théorique, empirique ou approchée). La statistique fournit alors des outils opérationnels, tel que le concept d'*auto-consistence* (instrumenté sous la forme des *formules de reconstitution* des données en AEM) nécessaire pour assurer l'indispensable retour aux données. Cette deuxième forme de *réduction de données*, peu

familière aux informaticiens spécialistes des SGBD, pourrait enrichir les fonctionnalités d'un logiciel de data mining.

2. Les *Bases de Données de Grande Dimension* se caractérisent par un ensemble d'enregistrements de taille classique caractérisés par un très grand nombre d'attributs. Il offre un challenge plus troublant pour les statisticiens qui considèrent une telle situation comme paradoxale et dangereuse (risques de multicolinéarité, séparation difficile entre pertinent ou non). Pourtant, au plan mathématique, la dualité individus variables offre une symétrie sur laquelle peut s'appuyer une série de propositions méthodologiques raisonnables.
3. Comme l'indiquent Besse *et al.* [3], les *Bases de Données Réparties* collectent des mesures d'indicateurs observés sur des unités statistiques compatibles ou non. Dans le premier cas, on peut les amener à répondre aux contraintes de nomenclatures hiérarchiques par transformation et implémentation d'opérateurs d'agrégation/désagrégation. Dans le second cas, l'utilisation de méthodes statistiques d'imputation, de transfert d'information ou de greffe, est envisageable pour les logiciels de demain. De plus, ces deux questions ne sont solubles que si l'on dispose d'informations réparties, mais répondant à des définitions communes ou tout au moins qu'il est possible de mettre en cohérence, grâce aux méta-données disponibles [10, 11, 12, 15].

Des développements sont nécessaires pour traiter ces sources hétérogènes, car les méthodes actuellement implémentées en statistique comme en *machine learning* supposent en général un seul niveau de mesure. Les méthodes statistiques d'*analyse multiniveau* pourraient trouver un domaine d'application à ce cas, en particulier pour des bases de données nombreuses, où un échantillonnage à chaque niveau se propose naturellement.

Les auteurs ont-ils une expérience de ces problèmes ?

Outre l'implémentation d'opérateurs d'agrégation [1, 9], les bases de données statistiques se différencient de leurs homologues en gestion par leur finalité. Dans le premier cas, le développement de *stratégies d'analyse* nécessite des actualisations du schéma de la base et rompt donc avec le dogme passé des SGBD : la fixité du schéma de la base. Celle-ci est peu compatible avec une pratique usuelle d'analyse : la construction de nouveaux indicateurs synthétiques mesurés sur de nouvelles unités statistiques, avec multiplication des variantes. De ce point de vue, comme pour la gestion des *méta-données* (pour lesquelles la définition des informaticiens est trop restrictice) d'Aubigny *et al.* [10, 11, 12] suggèrent que les *treillis conceptuels* sont mieux adaptés au développement de stratégies d'analyse statistique des données que les SGBD classiques. La gestion dynamique du schéma est-elle possible dans les entrepôts de données modernes ? Gère-t-elle des *données structurées*, mieux adaptées au codage d'une information mixte (chiffres, graphiques, images, textes) ?

Enfin, la place importante accordée dans la communauté du data mining à la notion d'OLAP [6], [33], me semble ignorée à tort dans l'article. Son utilité

6. OLAP · On Lıne Analytic Processıng, et ROLAP Relational OLAP

tient à une faiblesse des SGBD relationnels destinés à la gestion, reconnue initialement par Nelder. La conduite d'analyse statistique (exploratoire) met classiquement en jeu des *tableaux multi-indices*, [18]. Dans ce cas, l'expérience empirique montre que les requêtes *SQL* s'avèrent nettement moins performantes (en temps de réponse) sur de grandes bases que leur équivalent effectué par un logiciel/langage statistique tels que R ou S, cf. d'Aubigny *et al.* [10, 11].

3.2. Le data mining propose des méthodes d'analyse secondaire des données

Les objectifs fixés au data mining par les auteurs me semblent difficiles à délimiter, faute de principe organisateur. De plus, je me suis étonné de la place très réduite accordée au concept de *généralisation* pourtant très présent dans la littérature consacrée au data mining. Tout d'abord, il me semble utile de rappeler que, comme toute étude (statistique), extraire de l'information d'une base de données suppose une question clairement énoncée à laquelle on souhaite répondre. Sans cela le terme même de *motif* utilisé par les auteurs est indéfini. L'objectif courant de réduction des données à une taille gérable n'échappe pas à cette règle. Mais les questions qu'il est possible de poser dépendent des données et se formulent au travers de modèles. Aussi, l'articulation Données-Modèle est-elle centrale pour établir des stratégies d'analyse.

Besse *et al.* distinguent quatre types de questions posées : exploration, classification (au sens clustering), modélisation, et recherche de forme. Il s'agit dans les quatre cas de construire un modèle *descriptif* au service d'une question relevant du premier aspect du triptyque *Décrire, Expliquer, Prévoir*, [30], faisant du data mining un chapitre de la statistique descriptive. Hand [22] attribue ce fait à l'orientation résolument centrée sur la découverte (discovery : le data mining implémente les méthodes de la statistique exploratoire!) et laisse entendre que cela justifierait la non prise en considération des conditions de production des données. Cela est lourd de conséquences, à cause des conditions contraignantes qui président à l'interprétation causale de relations de corrélation ou d'agrément mises en évidence sur des *données de simple observation*, [32].

Le paradigme du data mining se présente comme consacré à l'*analyse secondaire* [7] de grandes bases de données préexistantes et s'intéresse peu au Processus de Génération des Données (PGD), laissant à d'autres statisticiens des domaines de compétences reconnus tels que la théorie des sondages et celle des plans d'expériences. Le choix des données est une donnée de l'étude, peut-être arbitraire mais non remise en cause, et les modèles spécifiés offrent simplement une réduction des données à l'aide de résumés ou une description de ces données. On parle souvent de modèles phénoménologiques (ou *boîte*

7. **L'analyse** primaire assure la cohérence entre procédures de production et méthodes **d'analyse des** données. L'analyse secondaire concerne des données préexistantes, collectées **dans un but et selon** un protocole non nécessairement liés aux questionnements d'intérêt et **aux méthodes d'analyse** utilisées.

noire, comme les réseaux de Neurones) dans ce cas, par opposition aux modèles *explicatifs* (ou théoriques), qui ont pour ambition de synthétiser une théorie ou un PGD et dont les paramètres reçoivent une interprétation concrète.

Le concept de *généralisation* s'applique à cette deuxième catégorie de stratégies de spécification de modèles et l'un de ses archétypes est l'inférence statistique. Dans ce cas d'analyse confirmatoire, le dogme veut que la spécification du modèle précède la consultation et, en principe même, la collecte des données. Le choix du modèle postulé n'est pas discuté : c'est une donnée du problème, qui ne peut être remis en cause en cours d'analyse, au vu des données. Le point central est l'articulation entre interprétations causales d'associations, rendue possible par l'existence d'un plan de sondage, et le contrôle de variables exogènes potentiellement influentes par randomisation. Ces deux composantes sont nécessaires [8] au passage du descriptif à l'explicatif, car elles assurent une maîtrise raisonnable du PGD.

Cette distinction entre décrire et expliquer est indépendante du type de modèle utilisé. Il peut être *structurel* et décrire des covariations : $g(X^1, ..., X^p) = 0$ (comme en Typologie, en Analyse en Composantes Principales, en Positionnement Multidimensionnel) ou *fonctionnel* et décrire une variation conditionnelle : $E(h(Y)/X^1, ..., X^p) = f(X^1, ..., X^p)$ (comme en Régression, en Analyse Discriminante, ou en Reconnaissance des Formes). Les modèles fonctionnels sont le plus souvent utilisés pour prédire (dans un cadre descriptif) ou prévoir (dans un cadre explicatif). Mais, dans les deux cas, Thom [34] montre que le statut du modèle change du point de vue de la connaissance. Les modèles descriptifs élaborés avec des outils de data mining transforment des données en information utile pour définir des actions, et la thèse de Thom dit que les modèles explicatifs tirent leur valeur de ce qu'ils adjoignent à une analyse quantitative une approche qualitative plus fine et plus décisive pour la connaissance. Certes, mais c'est la place même du modèle par rapport aux données qui a changé.

3.3. Le data mining pose le problème du choix de modèles

Les données étant fixées *a priori*, le problème de *spécification de modèles* pour une analyse secondaire [27] est plus délicat en data mining qu'en statistique classique. La littérature sur le sujet est fournie, car on touche alors à un débat ancien, où le terme même de data mining fut utilisé bien avant son appropriation par les informaticiens. La littérature économétrique donne une connotation négative au data mining pour décrire le risque d'un choix de modèles surparamétrés et donc impropres à la généralisation [29]. Cette question est directement liée au choix de stratégies de sélection en chaîne de modèles *emboîtés* : une approche ascendante, de complexification progressive des modèles, est-elle préférable à une approches descendante ? La doctrine des économètres, due à l'école de Hendry, préconise l'approche allant du général au spécifique, [6, 17, 21, 24, 25, 26].

8. Les travaux de [32] s'attachent à spécifier des modèles explicatifs lorsqu'il n'est pas possible d'appliquer des procédures de randomisation.

Une deuxième facette du problème, plus spécifique à l'analyse des données, tient à la non unicité du modèle. Plus le nombre de variables utilisées pour décrire une structure d'association est grand et plus le nombre de modèles candidats utiles est grand. Cette recherche doit donc être contrainte dans les bases de données de grande dimension. Deux approches au moins sont en concurence. La première, classique pour la spécification de modèles structurels, consiste à utiliser un critère d'optimalité de la solution en termes de motifs intéressants. C'est le cas en Projection Pursuit où est réputée intéressante une répartition projetée sur une droite aussi éloignée que possible d'une loi de Gauss. En analyse linéaire des données comme en analyse en composantes indépendantes, c'est la formule de reconstitution approchée des données qui est jugée intéressante, parce qu'elle permet le retour aux données. La deuxième stratégie, le *statistical learning*, *cf.* Vapnik [36, 37], s'intéresse à la construction de modèles de dépendance fonctionnelle appris (estimés) sur un nombre obligatoirement fini de données. Pour traiter ce problème mal posé, on réduit le nombre de modèles à explorer en introduisant des *termes de régularisation* dans l'expression de la fonction critère.

En général, ces stratégies conduisent à retenir plusieurs modèles utiles pour répondre à la question d'intérêt. Ces modèles ne sont pas nécessairement emboîtés, donc difficiles à comparer et ils n'apportent pas nécessairement une réponse identique. Quelles caractéristiques communes aux modèles retenus conserver ? De plus, on répond à la question avec une certaine *incertitude*. Comment la mesurer ? En termes de contraste moyen entre modèles retenus ? Certains auteurs préconisent de recourir à des modèles qualifiés de robustes, car obtenus en moyennant les modèles (model averaging). Cela fournit-il une mesure d'incertitude associée ? Les auteurs ont-ils eu l'opportunité d'expérimenter cette approche sur les études présentées ?

D'autre part, le texte ne définit pas ce qu'on appelle un *motif*, une particularité des données, un pattern. Interpréter les modèles retenus comme un échantillon (en un sens flou) de modèles donne une notion d'*invariant* couplée à celle de *motif intéressant*. Au sens strict, un invariant est une caractéristique que l'on peut déceler (donc d'un point de vue opérationnel, écrire ou reproduire) dans l'ensemble des modèles utiles spécifiés. C'est-à-dire un sous-modèle recouvert (au sens de encompassing) par tous les modèles retenus. Par extension, un invariant généralisé ou motif intéressant est un modèle proche (par exemple pour le contraste de Kullbak-Leibler) d'une composante de tous les modèles retenus.

Cette discussion même est assez difficile à mener, car les auteurs ne précisent pas ce qu'ils appellent un modèle. Prenons l'exemple des structures exponentielles pour fixer le langage. Trois composantes sont alors à considérer :

1. La structure statistique donne la famille paramétrée de lois retenue pour décrire la variabilité des données. Ici, faute de connaître le PGD, l'analyste n'a pas la maîtrise de cette structure statistique et seule une modélisation des erreurs de mesure peut contrebalancer l'arbitraire d'un choix ;

2. La fonction lien occupe une part non négligeable d'une étude, et les logiciels de data mining rivalisent de fonctionnalités de ré-expression des données, nécessaires à la construction de résumés utilisables, compromis entre régularités statistiques et interprétabilité. Mais cette question échappe au propos des auteurs ;
3. Le modèle comportemental (structurel ou fonctionnel) de (co)variation retenu.

Il me semble que c'est sur cette troisième composante que porte la notion de motif utilisée dans le texte ! Si c'est le cas, la détermination d'une structure statistique aussi généralisable que possible, par le principe d'entropie maximum [9] peut être utilisée pour déterminer une structure statistique respectant les motifs suggérés par l'analyse et leur donner un statut d'hypothèses statistiques, éventuellement testables dans une phase confirmatoire ultérieure, sur des données fraîches, collectées dans ce but. On cherche une famille de lois de probabilité, optimale sous contrainte de respecter ces motifs. L'intérêt de cette démarche est double : elle fournit une méthode constructive de choix de modèle et elle donne un contenu à la notion d'estimation robuste avec des données en nombre limité, *cf.* [16].

3.4. Le problème de la complexité des modèles en data mining

Ce problème est loin d'être standard dans le cas de bases de données de grande dimension. On dispose alors d'un ensemble de règles d'associations potentielles riche, qui détermine un espace énorme de combinaisons (scenarii) possibles. Evaluer un paramètre unidimensionnel dans un modèle paramétrique nécessite des échantillons relativement petits, mais générer des règles qui s'influencent mutuellement impose des échantillons très grands pour décrire cet espace de grande dimension engendré par l'ensemble des scenarii possibles. Il faut donc définir des procédures d'évaluation de la complexité d'un modèle. La taille n de l'échantillon donne le nombre de degrés de liberté, c'est-à-dire la dimension algébrique de l'espace des observations, qui n'est pas toujours liée à la dimension intrinsèque ou topologique de la variété décrite par les données. Le découplage provient des non linéarités de cette variété, de sorte que certains auteurs identifient complexité et non linéarité de la composante comportementale du modèle. Mais le degré de curvilinéarité modélisable est lié à la richesse en données. Par exemple, lors de l'étude d'un processus stochastique non stationnaire, distinguer un modèle de tendance déterministe et d'une tendance stochastique est impossible sur des séries courtes, bien que les deux modèles soient décrits par peu de paramètres.

Besse *et al.*, reprennent la distinction faite par Hand [21] entre deux finalités d'une étude qui guiderait le choix d'une stratégie d'analyse. Tout d'abord, lorsque le but assigné est de construire un modèle descriptif ou une prédiction, on recherche un résumé afin d'identifier et décrire les principales caractéristiques de forme de la loi de répartition, des invariants macroscopiques, permettent de décrire les caractéristiques communes à l'ensemble

9. MEP . Maximum Entropy Principle

de l'échantillon. Peu importe alors le type de base de donnée traité. Au contraire, la détection d'événements rares (Hand [21] parle de *subtle patterns or departure from regularity*) nécessite des bases de données nombreuses, pour identifier des petits écarts à la norme c'est-à-dire des détails en terme de variabilité, *a priori* importants du point de vue substantiel, de façon à détecter des comportement inusuels, par exemple des formes sporadiques dans un électrocardiogramme, un patron inhabituel de dépenses par carte banquaire, un indice de tentative de fraudes (modèles de ruptures, les interventions en séries chronologiques), etc... Plus le détail est fin, plus il est associé à un motif difficile à déceler si la variabilité mesurée par la structure statistique n'est pas maîtrisée. Plus donc la place des métainformations et des connaissances exogènes, les *a priori*, est grande.

L'analyste traite ici de façon classique une décomposition des données du type signal (comportemental) plus bruit (structure statistique) $(D = S + B)$[10] et le niveau de détail accessible dépend de la qualité initiale des données et d'une bonne description de l'invariant S déjà dégagé du bruit B. La qualité de la description globale S prime. Si le modèle comportemental adopté pour une décomposition $D = S + B$ est insuffisamment complexe (principe de parcimonie) au sens où il ignore des variations non linéaires, il reporte celles-ci dans le bruit B et augmente la variabilité, interdisant le repérage de détails substantiels. Complexifier la partie signal atténue les non linéarités présentes dans les résidus et donc assimilées à du bruit, c'est-à-dire à des motifs non intéressants.

Cette approche, majoritairement adoptée en data mining fonctionne comme une analyse multi résolution : elle part d'un modèle « simple » qu'elle enrichit progressivement de détails, de complexité plus forte et moins perceptibles, contrairement à ce que préconise Hendry. J'aimerais bien connaître le point de vue des auteurs à cet égard.

4. Conclusion

Si le data mining est un paradigme, on doit s'intéresser à son cycle de vie. Va-t-il résolument changer de façon durable les métiers de la statistique, leur articulation avec les sciences de l'information et de la cognition, leur dépendance au langage mathématique ?

La quantité de sources de données, leur volume et les applications liées au traitement de données continue de croître beaucoup plus vite que le nombre de statisticiens. Réduire le temps de traitement informatique des données est un objectif tentant et raisonnable pour automatiser les tâches de routine et conserver un temps suffisant pour interpréter, comprendre et mettre en forme les résultats obtenus. En cela les logiciels de data mining sont un apport.

10. Cette notation abusive mais parlante est illustrée dans les structures exponentielles par la séparation en composante comportementale et structure statistique du modèle. Le rôle d'une réexpression éventuelle par la fonction lien est occulté ici.

Il appartient aux statisticiens d'agir pour former des cadres compétents et prudents à cette technologie, qui porte sa part de dangers. C'est une chance à saisir. Des esprits chagrins ont prétendu qu'on s'enrichissait plus sûrement en vendant des outils de data mining aux prospecteurs qu'en exploitant ces gisements de données. Mais, la vraie richesse cultivable ne résiderait-elle pas dans la compétence méthodologique actualisée du prospecteur ? Il me semble que la responsabilité première des statisticiens consiste à s'attacher à cet aspect de la question.

On a aussi dit que le data mining commençait là où l'on se propose de transformer des données en information. Cet objectif est ambitieux et reste à tenir. Je suis beaucoup plus réservé sur l'ambition de produire de la connaissance, compte tenu de la limitation dans laquelle s'est enfermé le data mining dès ses débuts. Produire de la connaissance par une analyse secondaire de données de simple observation reste une utopie et une promesse intenable à court terme.

La relation liant le data mining aux sciences cognitives est donc plutôt liée à son rapport à la complexité. Celle des modèles statistiques est liée au nombre d'attributs des données retenu et beaucup moins au nombre d'enregistrements. Notre cerveau est apte à manipuler un très petit nombre seulement d'associations entre concepts : 7 au plus disent les psychologues. La machine peut-elle suppléer aux limites naturelles de l'esprit humain sur cet aspect qui pousse à l'humilité et la prudence ? Les modèles utiles à la prise de décision en entreprise nécessitent-ils une telle complexité ? Je ne le crois pas.

Juger le data mining uniquement du point de vue de sa relation aux mathématiques serait aussi une erreur. Science de l'ingénieur, le data mining fournit des réponses efficaces à des demandes ciblées d'actions rentables. Il doit pour cela établir un compromis temporaire entre les optimalités mathématiques et les optimalités opérationnelles. Ce n'est pas forcément la meilleure réponse du point de vue des critères usuels de la statistique, mais plutôt une solution sub-optimale raisonnable qui sait allier efficacité dans la mise en œuvre et optimalité mathématique.

C'est le cas par exemple de l'étude qui nous fut confiée pour construire un classifieur optimal de cellules prostatiques devant fournir la meilleure règle de classement possible sous la contrainte prioritaire de temps de calcul le plus bref possible. La rentabilité tenait alors directement au gain marginal journalier du nombre de plaques analysées par la machine et classées de façon sûre. Le rôle de la statistique est alors de ré-étudier les optimalités comparées des méthodes possibles dans ce cadre contraignant, guidée par les résultats d'une étude de type data mining destinée à trier les méthodes heuristiques admissibles.

La spécificité data mining de la stratégie adoptée et de la réponse apportée ne transparaissent pas forcément dans la présentation des résultats d'une étude. Cela m'amène à poser une dernière question aux auteurs, après lecture des exemples présentés. Quelle est selon eux la spécificité de ces études qui les différencient de l'Analyses Exploratoire de Données pratiquée depuis 25 ans ? En quoi sont-elles spécifiques d'une démarche de type data mining ?

RÉFÉRENCES

[1] AL BOUAZZAOUI A., D'AUBIGNY G., GRAS S. and TASSARD G., Agrégation et modélisation objet dans les SIG. *Revue internationale de Géomatique*, 4 :337-352, 1994.

[2] BERNARD J.M., LE ROUX B., ROUANET H. and SCHILTZ M.A., L'analyse des données multidimensionnelles par le langage d'interrogation de données LID. *Bulletin de méthodologie sociologique*, 23 :3-46, 1989.

[3] BESSE P., LE GALL, RAIMBAULT N. and SARPY S., Data mining et statistique. *Journal de la Société Française de Statistique*, 142, 1 :5-36, 2001.

[4] BLUNT G., KELLY M.G., ADAMS N.M. and HAND D.J., Data mining for fun and profit. *Statistical Science*, 15 :111-131, 2000.

[5] CAILLEZ F. and PAGES J.P., *Introduction à l'analyse des données*, SMASH, Paris, 1976.

[6] CAMPOS J. and ERICSSON N.R., Constructive data mining : modeling consumer's expenditure in Venezuela. *The econometrics journal*, 2 :226-240, 1999.

[7] CAZES P. *Application de l'analyse des données à l'étude de problèmes géologiques.* PhD thesis, Paris VI, 1970. Thèse de doctorat de troisième cycle.

[8] CHESSEL D. and DODELEC S., ADE Version 3.3 : Hypercard Stacks and Quick Basic Microsoft Programme library for the analysis of environmental Data. Technical report, URA CNRS 1451, Université Lyon 1, 69622 Villeurbanne cedex, 1992.

[9] D'AUBIGNY C. and D'AUBIGNY, Agrégation spatiale et résumés statistiques. *Revue internationale de Géomatique*, 4 :307-336, 1994.

[10] D'AUBIGNY G. and TASSARD G., La modélisation des méta-données dans les systèmes intelligents de traitement de l'information statistique. In EUROSTAT, Luxembourg, editor *The meta-information management systems*, pages 43-63, 1995.

[11] D'AUBIGNY G. and TASSARD G., La modélisation des méta-données dans les systèmes intelligents de traitement de l'information statistique. In Univ. de Liège, Belgique, editor, *Cahiers du centre d'histoire quantitative et du développement économique régional*, pages 1-21, 1995.

[12] D'AUBIGNY G. and TASSARD G., The object oriented approach applied to statistics. In European Communities, Luxembourg, editor, *Proceedings of the Strategic reflexion colloquium on IT issues for Statistics*, pages 76-95, 2000 invited paper.

[13] ELDER J. and PREBIGON D., A statistical perspective on knowledge discovery in data bases, pages 83-113. In Fayyad U.M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy [14], 1996.

[14] FAYYAD U.M., PIATETSKY-SHAPIRO, SMYTH P. and UTHURUSAMY R., editor. *Advances in knowledge discovery and Data mining.* AAAI Press, Menlo Park, CA, 1996.

[15] FROESCL K.A., A metadata approach to statistical query processing. *Statistics and Computing*, 6 :11-29, 1996.

[16] GOLAN A, JUDGE G. and MILLER D., *Maximum entropy econometrics : robust estimation with limited data.* John Wiley & sons, Chichester, 1996.

[17] GRANGER C. and TIMMERMANN A., Data mining with local model specification uncertainty : a discussion of Hoover and Perez. *The econometrics journal*, 2 :220-225, 1999.

[18] GRAY J. Data cube : a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining & Knowledge Discovery*, 1 :29-53, 1998.

[19] FRIEDMAN J.H., Dat mining and statistics : What is the connection ! Technical report, Stanford University, Stanford, CA 94305, 1998.

[20] HAND D.J., Data mining : statistics and more ? *The American statistician*, 52 :112-118, 1998.

[21] HAND D.J., Discussion contribution on 'Data mining reconsidered : encompassing and the general-to-specific approach to specification search' by K.D. Hoover and S.J. Perez. *The econometrics journal*, 2,241-243, 1999.

[22] HAND D.J., Statistics and Data minig : intersecting disciplines. *ACM SIGKDD Exploration* 1 :16-19, 1999.

[23] HAND D.J., MANNILA H. and SMYTH P., *Principles of data mining.* The MIT Press, Cambridge, Mass., 2001.

[24] HANSEN B.E., Discussion of 'Data mining reconsidered'. *The econometrics journal*, 2 :192-201, 1999.

[25] HENDRY D.F and KROLZIG H.M., Improving on 'Data mining reconsidered' by K.D. Hoover and S.J. Perez. *The econometrics journal*, 2 :202-219, 1999.

[26] HOOVER K.D. and PEREZ S.J., Data mining reconsidered ; encompassing and the generalto-specific approach to specification search. *The econometrics journal*, 2 :167-199, 1999.

[27] LEAMER E.E., *Specification search : ad hoc inference with non experimental data.* John Wiley & sons, New york, 1978.

[28] LEBART L., MORINEAU A. and PIRON M., *Statistique Exploratoire Multidimensionnelle.* Dunod, Paris, 1995.

[29] LOVELL M.C., Data mining. *Review of Economics and Statistics*, 65 :1-12, 1983.

[30] MATALON B. *Décrire, Expliquer, Prévoir : démarches expérimentales et terrain.* Armand Colin-Collection U, Paris, 1988.

[31] SMYTH P. Chapter 1 : Data mining at the interface of computer science and statistics. in *Data mining for scientific and engineering applications*, page xxx-XXX, 2001 à paraître.

[32] ROSENBAUM P.R., *Observational studies.* Springer-Verlag, Berlin, 1995.

[33] SHOSHANI A., OLAP and Statistical Databases : similarities and differences. *ACM TODS*, 2 :1-18, 1997.

[34] THOM R., *Prédire n'est pas expliquer.* Editions Eshel, Paris, 1991.

[35] TUKEY J.W., *Exploratory data analysis.* Addison-Wesley, Reading, Mass., 1977.

[36] VAPNIK V.N., *Estimation of dependences based on empirical data.* Springer-Verlarg, Berlin, 1982.

[37] VAPNIK V.N., *Statistical learning theory.* John Wiley & sons, New York, 1998.

[38] YOUNG F.W. VISTA : the VIsual STAtistics System ? Technical report, Psychometric Laboratory Research Report 94-1, UNC Psychometrics Lab, Chapel Hill, NC., 1994.