Data Mining et Statistique

Philippe Besse[*], Caroline Le Gall[†],
Nathalie Raimbault[‡] & Sophie Sarpy[§]

Résumé Cet article propose une introduction au *Data Mining*. Celle-ci prend la forme d'une réflexion sur les interactions entre deux disciplines, Informatique et Statistique, collaborant à l'analyse de grands jeux de données dans une perspective d'aide à la décision. Sans prétention d'exhaustivité, différents exemples sont exposés issus du marketing bancaire, de la détection de défaillances dans un procédé industriel ainsi que des problématiques aéronautiques pour l'aide au pilotage. Ils permettent de tirer quelques enseignements sur les pratiques du *data mining* : choix d'une méthode, compétences de l'utilisateur, fiabilité des résultats.

Mots clefs Data mining, modélisation statistique, discrimination, arbres de décision, réseaux de neurones.

Abstract This article gives an introduction to Data Mining in the form of a reflection about interactions between two disciplines, Data processing and Statistics, collaborating in the analysis of large sets of data. Without any claim to complete coverage, various examples are given, dealing with banking marketing, failure detection in an industrial process, and aeronautical issues on aircraft control. They allow us to draw several lessons about the practical experience of Data Mining : choice of the method, user's skills and reliability of the results.

Key words Data mining, statistical modeling, discrimination, decision trees, neural netwoks.

[*]Laboratoire de Statistique et Probabilités, UMR CNRS C5583 — Université Paul Sabatier 31062 Toulouse cedex 4 — besse@cict.fr

[†]Motorola — Toulouse

[‡]Airbus EADS — Toulouse

[§]CARSO Informatique — Balma

1 Introduction

Dans ses applications marketing ou de *gestion de la relation client*, le *Data Mining*, qui peut se traduire par la *prospection* ou la *fouille* de données et qui sera plus précisément défini en section 2.3, connaît un développement commercial très important et offre de nombreuses opportunités d'emplois pour les étudiants formés à la Statistique. La motivation principale de cette démarche est la valorisation d'une grande base ou entrepôt de données (*data warehouse*) par la recherche d'informations pertinentes pour l'aide à la décision. Les applications les plus répandues concernent la recherche d'une typologie de la clientèle ou encore celle de *scores* mesurant l'adéquation des clients aux produits à promouvoir en vue, par exemple, d'un publipostage. D'autres applications voient le jour dans un contexte industriel comme l'aide au diagnostic de défaillances dans un procédé de fabrication complexe (Mieno et coll, 1999. ; Gardner et coll., 2000) ou encore dans des disciplines scientifiques confrontées à la détection de motifs ou formes (patterns) dans des volumes de données considérables : génomique, astrophysique...

Comme l'*analyse des données* des années 70/80, le data mining se présente principalement comme un assemblage de techniques au sein d'un progiciel. Ces techniques sont issues de différents domaines dépendant de la Statistique ou de l'Intelligence Artificielle. On y rencontre ainsi des outils de statistique élémentaire et multidimensionnelle pour l'exploration et la classification mais également des techniques de modélisation avec l'arsenal des généralisations de modèle linéaire classique (régression multiple, logistique) et de l'analyse discriminante. Ces dernières sont alors en concurrence avec d'autres approches plus algorithmiques visant à la construction d'arbres de décision, de règles d'induction ou de réseaux de neurones. L'objectif principal est présenté comme une *quête de sens* : extraire l'information pertinente, ou *pépite (nugget) de connaissance*, en vue d'analyses puis de prises de décision.

Beaucoup de logiciels sont consacrés à la fouille de données, une bonne soixantaine sont répertoriés dans les sites[1] consacrés au sujet et comparés (Goebell & Le Gruenwald, 1999) ; la plupart mettent en avant des interfaces graphiques sophistiquées, un accès intégré aux bases de données et insistent sur une automatisation poussée des traitements. L'accroche publicitaire souvent citée est d'ailleurs : *Comment trouver un diamant dans un tas de charbon sans se salir les mains ?* Certains sont axés sur des familles de techniques (par exemple les réseaux neuronaux), d'autres se veulent généralistes et offrent un large choix, d'autres encore se spécialisent sur un domaine particulier comme l'analyse de textes appliquée à la veille technologique.

L'utilisation de ces logiciels suscite rapidement quelques questions :

– une technique de modélisation est-elle préférable à une autre ?

[1] `www.kdnuggets.com` est un bon point d'entrée pour surfer sur ce thème.

– quelle confiance accorder aux procédures automatiques et quelles compétences sont nécessaires au prospecteur ?
– quelle significativité accorder aux résultats ?

Il serait présomptueux de vouloir apporter des réponses définitives. Dans cet article, nous souhaitons simplement donner quelques éclairages à travers des expériences concrètes et susciter une certaine prudence. Contrairement au prospecteur minier qui ne trouve pas de pépite d'or là où il n'y en a pas, le prospecteur de données, en fouinant (*data snooping*) suffisamment, en s'acharnant, finit par révéler une pépite de connaissance, c'est-à-dire une liaison ou un motif jugé significatif sur les données présentes, mais qui risque de se révéler sans capacité de généralisation, donc sans intérêt. Mal utilisé ou mal contrôlé, le *data mining* peut conduire le prospecteur dans des pièges grossiers. Friedman (1997) suggère même qu'il pourrait, comme beaucoup de ruées vers l'or, enrichir principalement les fournisseurs d'outils plus que les prospecteurs eux-mêmes.

La seconde section de cet article est consacrée à une description plus détaillée des pratiques relevant de la fouille de données, des relations qu'elles entretiennent avec les outils informatiques, des rapprochements et oppositions qui peuvent être décelés avec la démarche statistique en terme d'objectifs ou de méthodes. Il s'agit d'alimenter le débat déjà amorcé par Elder & Pregibon (1996), Friedman (1997) ou Hand (1998, 1999). Le présent article privilégie les applications pour une approche comparative ; les techniques principales utilisées (régression, arbres, réseaux de neurones) ne sont pas explicitées, des détails sont donc à rechercher dans la bibliographie citée en référence ou dans des exposés plus systématiques (Besse, 2000 ; Jambu 2000 ; Hébrail & Lechevallier, 2002). Trois domaines d'application sont abordés dans les sections 3 à 5. Le premier concerne une application traditionnelle en marketing bancaire : évaluer un *score d'appétence* de la carte Visa Premier afin de mieux cibler une opération promotionnelle ; le deuxième a pour objectif l'aide à la détection d'une défaillance dans un procédé industriel produisant des circuits intégrés ; le troisième concerne des applications aéronautiques pour l'aide au pilotage. Enfin une dernière section tire quelques enseignements de ces exemples quant à la pratique du *data mining*, sur les comparaisons et choix de méthodes, les compétences requises, les difficultés d'évaluer la fiabilité des résultats, la nécessaire implication des statisticiens dans un domaine en pleine expansion.

2 Data mining, Informatique et Statistique

2.1 Motivation

Historiquement, le développement du *data mining* suit logiquement celui des moyens informatiques de stockage et de calcul. Pour des raisons initialement comptables et de gestion des stocks, les entreprises archivent des masses de données considérables. Il est alors naturel de vouloir valoriser ces données comme support à une stratégie de marketing ou, plus généralement, en les intégrant au processus de prise de décision. Les gestionnaires des bases de données ont multiplié par 10^3 puis 10^6 leur volume, ont complexifié leur architecture en passant d'un fichier unique à des bases réparties dans des environnements hétérogènes et en migrant des systèmes relationnels vers des cubes multidimensionnels. Ils sont donc amenés à utiliser des procédures de consultation plus sophistiquées. Après la simple interrogation : "*Quelles sont les ventes de tels produits à telle période ?*", on peut chercher à connaître les caractéristiques des acheteurs de ce produit, leurs autres préférences d'achat ou plus généralement rechercher des associations client-produits en émergence. Cette évolution nécessite une adaptation des méthodes sous-jacentes. Du langage d'interrogation logique (SQL) on passe aux techniques de recherche d'associations, de classification, de modélisation puis de reconnaissance de formes. Différentes stratégies ont alors été mises en œuvre pour répondre à ces problèmes. Techniquement, elles ont consisté à intégrer ou interfacer des outils statistiques et d'intelligence artificielle à des gestionnaires de bases de données.

2.2 Entrepôts de données

Plus précisément, le contexte informationnel du *data mining* est celui des *data warehouses*. Un entrepôt de données, dont la mise en place est assurée par un gestionnaire de données (*data manager*) est un ensemble de bases relationnelles ou cubes multidimensionnels alimenté par des données brutes et relatif à une problématique :

- gestion des stocks (flux tendu), prévision des ventes afin d'anticiper au mieux les tendances du marché,
- suivi des fichiers clients d'une banque, d'une assurance, associés à des données socio-économiques, à l'annuaire, en vue de la constitution d'une segmentation (typologie) pour cibler des opérations de marketing ou des attributions de crédit ; la *gestion de la relation client* vise à une individualisation ou personnalisation de la production et de la communication afin d'évacuer la notion de client moyen jugée trop globalisante ;
- recherche, spécification, puis ciblage des *niches* de marché les plus profitables ou au contraire les plus risquées (assurance) ;
- suivi en ligne des paramètres de production en contrôle de qualité pour détecter au plus vite l'origine d'une défaillance ;
- prospection textuelle (*text mining*) et veille technologique ;
- *web mining* et comportement des internautes ;
- décryptage d'une image astrophysique, du génome ;
- ...

Un entrepôt de données se caractérise par un environnement informatique hétérogène pouvant faire intervenir des sites distants (Unix, Dos, NT, VM...) à travers le réseau de l'entreprise (intranet) ou même des accès extérieurs (internet). En effet, des contraintes d'efficacité (suivi en "temps réel"), de fiabilité ou de sécurité conduisent à répartir et stocker l'information à la source plutôt qu'à la dupliquer systématiquement ou à la centraliser. Une autre caractéristique est la fréquente incompatibilité logique des informations observées sur des échantillons différents ne présentant pas les mêmes strates, les mêmes codifications. Enfin, ce sont des volumes, chiffrés en téraoctets, et des flux considérables de données qui doivent dans certains cas être pris en compte, surtout lorsque ces données sont issues de saisies automatisées.

2.3 KDD

Toute une terminologie s'est développée autour du *data mining* et de l'intelligence d'affaires (*business intelligence*). On parle ainsi de *Knowledge Discovery in Databases (KDD)* qui se décompose (Fayyad, 1997) en différentes étapes très générales dont certaines sont classiques dans le métier de statisticien :

1. Compréhension du domaine d'application : expliciter la connaissance *a priori* et les buts du commanditaire.
2. Création d'un sous-ensemble cible des données (matrice) à partir de l'entrepôt.
3. Nettoyage : erreurs, données manquantes, valeurs atypiques (*outliers*).
4. Transformation des données : "normalisation", linéarisation, découpage en classes, compression.
5. Explicitation de l'objectif et de la stratégie d'analyse : exploration, association, classification, discrimination, recherche de formes...
6. Choix des méthodes, des algorithmes, en privilégiant interprétabilité ou prédictibilité. Mise en œuvre des outils informatiques appropriés.
7. Test : sur la base de critères à préciser (qualité d'ajustement, de prévision, simplicité, visualisations graphiques...).
8. Exploitation.
9. Diffusion des résultats (intranet) pour prise de décision.

De façon plus précise, le *Data Mining* concerne l'exécution des étapes 3 à 8 ci-dessus. Il nécessite la mise en œuvre, explicite ou non, de méthodes statistiques classiques (graphiques, sondages, composantes principales, correspondances multiples, classification hiérarchique, nuées dynamiques, discriminante, k plus proches voisins, segmentation, régression linéaire, logistique) ou moins classiques (arbres de classification et de régression, modèles graphiques d'indépendance conditionnelle) ou d'intelligence artificielle (perceptron multicouche, réseaux auto-associatif et bayésien, apprentissage et règles d'induction, reconnaissance de formes).

2.4 Convergences

Le développement du *data mining* implique l'association de plusieurs disciplines ne partageant pas les mêmes règles de fonctionnement ou usages. Il est important de faire la distinction entre deux types de points de vue. Les premiers, riches en complémentarités, permettent de confronter puis d'améliorer les techniques, de les adapter au mieux aux données. Des antagonismes plus idéologiques que méthodologiques alimentent les seconds. Citons l'exemple des techniques qui visent à la recherche arborescente de règles de décision[2]. Les méthodes CART (Breiman et coll., 1984) et C4.5 (Quinlan, 1993) sont basées sur les mêmes principes généraux (arbre, critère d'entropie) mais poursuivent des développements indépendants. Cet exemple illustre aussi la remarque de Friedman (1997) : à partir d'une idée originale, un statisticien écrit un article tandis qu'un informaticien crée une entreprise.

Objectifs

Il est clair que les techniques rapidement listées dans la section précédente poursuivent des objectifs similaires et peuvent apparaître comme concurrentes ou plutôt complémentaires. Schématiquement, quatre objectifs non exclusifs sont la cible d'une prospection.

- *Exploration* pour une première approche des données, leur vérification par la recherche d'incohérences, de données atypiques, manquantes ou erronées, leur transformation préalable à d'autres traitements.
- *Classification* (clustering) pour exhiber une typologie ou une segmentation des observations.
- *Modélisation* par un ensemble de variables explicatives d'une variable *cible* quantitative ou qualitative. Il s'agit alors d'une régression ou d'une discrimination (ou classement).
- *Recherche de forme* sans apprentissage. Il s'agit de déceler une configuration (*pattern*) originale se démarquant des données.

Plus radical, Hand (1999) considère la classification comme la recherche d'un modèle et n'oppose donc que deux objectifs : *model building* ou *pattern recognition*. En poussant un peu plus loin la réflexion nous pourrions noter que la découverte d'une particularité ou *pattern* des données doit être exprimée, implicitement ou non, à travers une notion de déviance par rapport à une "norme", c'est-à-dire un modèle qui peut adopter des formes variées. Finalement, quel que soit l'objectif recherché au delà d'une première exploration, la notion de modèle, qu'elle soit de nature statistique, probabiliste,

[2]Consulter Zighed et Rakotomalala (2000) pour une vue d'ensemble sur ces techniques.

logique... reste centrale.

Choix de modèle

Dans l'optique fréquente de rechercher un modèle prédictif, quelle que soit la méthode utilisée, celle-ci nécessite d'optimiser certains paramètres : la liste des variables explicatives retenues, d'éventuelles interactions, le nombre de neurones dans une couche cachée et le temps d'apprentissage, le nombre k de plus proches voisins en analyse discriminante, le nombre de feuilles d'un arbre de décision... Les critères de comparaison employés sont communs aux approches et bien connus en Statistique. Il s'agit de toute façon d'estimer et de minimiser des erreurs quadratiques de prévision ou encore des taux de mal classés, éventuellement un risque bayésien si des informations *a priori* et des coûts de mauvais classement sont connus. Estimation sur un échantillon de validation (n'ayant pas participé à l'estimation), validation croisée et critères (C_p de Mallow, AIC, BIC...) impliquant une pénalisation de la complexité du modèle sont couramment proposés. L'objectif central est la sélection d'un modèle *parcimonieux* réalisant un meilleur compromis entre l'ajustement à l'échantillon ou aux données dites d'*apprentissage* et la variance des estimations de ses paramètres pour aboutir à une amélioration des qualités de prédiction.

Comparaison de méthodes

De nombreux travaux sont consacrés à la comparaison de méthodes sur des données simulées ou réelles. Chaque année, des concours[3] de fouilles de données sont organisés. Le projet Statlog (Michie et coll., 1994) propose une comparaison systématique d'une vingtaine de méthodes de discrimination sur une vingtaine de jeux de données issus de problématiques différentes. Les critères employés à cette fin sont à nouveau classiques puisqu'il s'agit toujours d'estimer une erreur soit avec un échantillon test (qui n'a participé ni à l'estimation, ni à la procédure de choix de modèle) soit, dans le cas d'un échantillon réduit, par validation croisée ou rééchantillonnage (bootstrap).

La leçon de bon sens que l'on peut tirer de ce travail méticuleux est qu'il n'y a pas de meilleure méthode ; leurs propriétés intrinsèques et les hypothèses requises s'adaptent plus ou moins bien au problème posé. Ainsi, dans un problème de discrimination, les propriétés topologiques d'une technique permettent ou non de séparer des régions par des frontières linéaires, quadratiques, convexes, fermées... En premier lieu, il apparaît raisonnable de tester chaque grande famille de méthodes par l'un de ses représentants ; en effet, à l'intérieur d'une même famille (par exemple les arbres de classification), les différences observées semblent, en première lecture, peu significatives. Enfin, il s'agit de comparer les résultats obtenus à travers la

[3] www.kdnuggets.com/publications/kddcup.html.

mise en place d'un protocole rigoureux. Les dernières propositions tendent même à proposer des combinaisons pondérées de modèles ou *bagging* (bootstrap aggregating ; Breiman, 1996) pour s'affranchir d'un éventuel manque de robustesse des méthodes. Dans le cas des arbres, à la suite de travaux de Shlien (1990) et Ho (1998), Breiman (2001) propose la constitution de forêts aléatoires (voir aussi Ghattas, 2000) pour réduire la variance d'une prévision. Chaque arbre est estimé sur un échantillon bootstrap. Le résultat d'une classification est alors celui du vote de chacun de ces modèles tandis qu'une simple moyenne fournit la prévision d'une variable quantitative.

2.5 Divergences

Plusieurs raisons essentielles, fondamentalement liées aux données, expliquent les différences et ruptures méthodologiques observées entre Statistique et *Data Mining*.

Données *a priori*

Dans la plupart des problèmes de *data mining*, les données sont *préalables* à l'étude ; elles sont même souvent recueillies à d'autres fins. En revanche, la planification expérimentale ou le sondage, c'est-à-dire la saisie organisée des données, sont partie intégrante d'une démarche statistique traditionnelle qui cherche à en optimiser les caractéristiques, par exemple en tâchant de minimiser simultanément coûts de mesure et variance des estimateurs. Même en marketing, l'impact d'une campagne publicitaire pourrait être préalablement testé sur des échantillons représentatifs des clients potentiels. Bien sûr une telle démarche a un coût qu'il faut mettre en balance avec la qualité des données recueillies et donc la précision et la fiabilité du contrôle de la décision qui en découle.

Nous touchons un sujet sensible, point d'achoppement entre représentants de différentes disciplines. Le statisticien traditionnel, fidèle aux procédés de la planification expérimentale, s'attachera à organiser l'observation de façon à se mettre dans les conditions optimales le conduisant à des modèles fiables, à des tests d'hypothèses et donc à des décisions sous le contrôle des modèles probabilistes inférés. À l'autre extrémité, la tentation est grande, compte tenu des possibilités des moyens de stockage et de calcul, des coûts très lourds engendrés par une démarche expérimentale, de vouloir prétendre à une exhaustivité de l'exploration, de vouloir tester toutes les relations ou modèles possibles, mais sans contrôle de leur significativité.

Taille des données

La principale différence qui est mentionnée entre Statistique et *Data Mining* concerne le volume ou le flux de données qui sont analysées. Il est évident que tous les algorithmes classiques ne sont pas à même de traiter des millions d'observations décrites par des milliers de variables et la pratique du *sondage* est l'objet d'un débat contradictoire. Techniquement, rien n'empêche de pratiquer un sondage dans l'entrepôt de données afin de mettre en œuvre des méthodes manipulant des matrices de taille acceptable. Le risque est de laisser échapper à travers le crible la pépite d'information pertinente : les groupes de faible effectif mais à fort impact économique, la série de transactions frauduleuses par carte bancaire, la cause d'une défaillance exceptionnelle. C'est l'argument avancé pour justifier d'un traitement *exhaustif* mais il n'oppose pas Statistique et *Data mining* à condition d'avoir clairement défini l'objectif : recherche de caractéristiques générales de la population ou de spécificités. Ce serait faire un mauvais procès à la Statistique que de faire croire qu'elle se limite à la recherche de généralités et négliger ainsi, par exemple, toutes les procédures de recherche de valeurs atypiques (*outliers*). Néanmoins, face aux contraintes générées par le volume des données, la réflexion portant sur les structures de données ou les algorithmes utilisés, c'est-à-dire une réflexion de nature informatique, l'emporte sur une réflexion plus statisticienne portant sur les modèles sous-jacents ou sur leur validité. Ce point se renforce lorsque des résultats doivent être fournis au fur et à mesure d'une saisie automatique des données. Les propriétés adaptatives des méthodes deviennent prioritaires.

Il est important de noter qu'à travers la procédure de choix de modèle, la taille des données (nombre d'observations, nombre de variables) est un paramètre qui influe fortement sur le choix des méthodes. Naturellement, plus le nombre d'observations, ou la taille de l'échantillon, est grand et plus il est possible d'estimer précisément un grand nombre de paramètres d'un modèle : plus de variables explicatives et d'interactions en régression, plus de neurones dans un réseau, plus de feuilles dans un arbre, estimation non paramétrique des densités en discrimination. En pratique (cf. l'exemple de la section 5.1), on se rend compte qu'il est finalement plus simple et plus efficace de construire un réseau de neurones flexible avec de nombreuses entrées que de vouloir sélectionner les bonnes interactions pour un modèle polynomial dans un ensemble d'effectif explosif. Dans cet exemple, la légitimité des modèles connexionnistes croît avec la taille des données. C'est d'autant plus vrai que le caractère explicatif d'un modèle de régression trop complexe se vide d'intérêt.

Automatisation

Les promoteurs de certains logiciels de fouille de données insistent fortement sur les possibilités d'automatisation des traitements : l'intervention d'un expert deviendrait inutile car le traitement peut être opéré par le commanditaire pilotant les analyses à l'aide d'une interface conviviale. Cette présentation irrite bien sûr tout statisticien qui a pu expérimenter, dans des situations concrètes, combien il est important de s'assurer de l'intégrité et de la cohérence des données avant de se lancer dans une méthodologie sophistiquée. Il s'inquiète également devant la complexité des méthodes mises en jeu dont toutes les options, choisies souvent par défaut, ainsi que leurs conditions d'application et limites, restent opaques à l'utilisateur non formé. Paradoxalement, l'utilisation d'une interface ergonomique rend trop facile ou trop rapide le lancement des analyses. Le temps nécessaire à la réflexion sur le choix des options ou le bien fondé des résultats obtenus peut s'en trouver singulièrement réduit au profit d'une apparente efficacité.

Cette logique d'automatisation est encore renforcée avec le développement de programmes dits *agents intelligents* chargés, en tâche de fond, d'inférer des changements de modèles ou de méthodes parallèlement à la saisie des données.

Validation

En *Data Mining* comme en Statistique l'importance d'exhiber des modèles parcimonieux fait l'unanimité et les stratégies pour y aboutir sont similaires. En revanche, seule une démarche statistique traditionnelle dans le cadre d'un système contraignant d'hypothèses est susceptible d'apporter directement des précisions quand à une majoration de l'erreur ou des intervalles de confiance. Cette remarque n'a sans doute que peu d'importance dans une application de type marketing où la précision sur le nombre de courriers envoyés n'est pas un réel enjeu. En revanche, dans d'autres applications ce point mérite un développement spécifique. En effet, dans certains contextes industriels (pharmaceutique, aéronautique...) soumis à une législation à travers des procédures de certification, il faut pouvoir *prouver* que l'erreur est bien inférieure à la norme fixée. Il est alors difficile d'éviter de s'interroger sur la *représentativité* de l'échantillon et sur les caractéristiques de la loi de l'erreur. Cette dernière est obtenue de façon théorique comme résultant d'un corpus d'hypothèses ou par simulation (Monte Carlo) ; une réflexion de nature statistique redevient nécessaire.

Le quatrième objectif tel qu'il est cité dans la section précédente (recherche de forme), associé à des techniques issues de l'Intelligence Artificielle, marque principalement l'originalité du *data mining* par rapport à la pratique statistique. C'est son plus fort argument commercial pour la recherche d'une pépite de connaissance mais c'est celui qui pose aussi le plus de problèmes de validation : montrer qu'une forme, une relation, une séquence d'observations, est chargée d'information et révèle une structure sous-jacente qui ne soit ni une erreur, ni un artefact de l'échantillon. La collaboration entre plusieurs disciplines, pas seulement informatique et statistique, est sur ce point incontournable.

Statistique et Mathématiques

Un dernier différent est de nature plus académique. Ainsi, Hand (1998) s'interroge sur la place de la Statistique. Considérée comme branche des Mathématiques, elle est nécessairement attachée à la notion de *preuve* pour valider une méthode en s'assurant, par exemple, de ses propriétés et vitesses de convergence. En revanche, au sein de la mouvance informatique, le *Data Mining* échappe à cette contrainte. Il vise à l'efficacité opérationnelle en admettant une approche empirique consistant à comparer performances et précisions des algorithmes en concurrence. Dans le premier cas, une caricature de la démarche conduit à des méthodes théoriquement performantes mais inutilisables ou inadaptées et ne répondant pas aux besoins des praticiens. Dans le deuxième, on assiste à un foisonnement d'adaptations ou de variantes incrémentales des algorithmes censées en améliorer la vitesse ou la précision.

Bien sûr, ces deux extrêmes ne sont pas exclusifs et constituent un schéma simpliste mais la tendance est nette à travers les politiques éditoriales des revues internationales de Statistique. Elles acceptent prioritairement la publication d'articles décrivant des méthodes nécessairement étayées par des preuves mathématiques de convergence ou d'optimalité. La tendance est renforcée en France par la structure cloisonnée du Comité National des Universités qui filtre les candidatures des Enseignants-Chercheurs avec la difficile tâche de définir les frontières entre Mathématiques appliquées, applications des Mathématiques et autres disciplines. Cela contribue au malaise de la Statistique en France[4] en la coupant artificiellement des domaines d'application qui sont une source naturelle d'innovation.

3 Marketing bancaire

Cette section décrit un exemple typique de gestion individualisée de la *relation client*.

3.1 Données et objectif

Les données sont issues de la CARSO Informatique chargée des études pour les Banques Populaires du Sud-Ouest. Un ensemble de 48 variables décrivent les avoirs, les mouvements, les épargnes, les emprunts... d'un échantillon de 1200 clients anonymes. L'objectif principal est la détermination d'un *score d'appétence* de la carte Visa Premier. L'étude a été confiée à une vingtaine d'étudiants (10 binômes) équipés du module SAS Enterprise Miner (SEM, 2001). Chaque binôme a eu pour consigne de mettre en œuvre deux stratégies. La première a consisté en une exploration manuelle et guidée

[4]Consulter à ce sujet le rapport sur la Statistique de l'Académie des Sciences (2000) dont on trouve une présentation par G. Saporta dans ce journal (140, 4).

afin d'aboutir à un même sous-ensemble de variables recodées en classes. Dans leur deuxième tentative, les étudiants ont repris des données initiales en étant libres d'utiliser les outils automatiques (sélection de variable, recodage) disponibles dans SEM (2001). Ces deux stratégies se sont conclues par la comparaison de trois modélisations (régression logistique, arbre de classification, réseau de neurones) estimées et testées chacune sur plusieurs échantillons issus de tirages indépendants : l'un commun à tous les étudiants, tous les autres différents.

3.2 Résultats préliminaires

Il serait fastidieux de lister les résultats[5] issus de l'exploration de ces données, étape très élémentaire mais nécessaire à leur compréhension et leur vérification.

L'étude unidimensionnelle montre que la plupart des variables présentent des distributions très dissymétriques. C'est un phénomène classique avec des variables mesurant des revenus, de distribution voisine d'une log-normale, illustrant la concentration des richesses. Des transformations sont indispensables. Deux stratégies ont été testées. La première utilise des transformations monotones ($f(x) = \log(a+x)$) pour rendre les distributions plus symétriques, la deuxième transforme toutes les variables par découpage en 2 ou plus rarement 3 classes. Cela résume le fait que l'information importante est la présence ou l'absence de tel produit financier plutôt que le nombre ou le montant de ce produit. Nous pourrions penser que le choix de garder quantitative l'information est plus efficace : plus de degrés de liberté, information moins résumée ; néanmoins la deuxième stratégie (tout qualitatif) s'est montrée plus efficace en terme de qualité prédictive des modèles. Elle seule est mentionnée dans la suite de cet article.

L'étude bidimensionnelle montre que même une gestion automatisée des données est source d'erreurs ou d'incohérences. Nous trouvons ainsi quelques clients ne possédant globalement aucune carte de paiement mais titulaires d'une carte Visa Premier, d'autres plus jeunes que l'ancienneté de leur relation avec la banque. Les coquilles rencontrées sont principalement dues à l'agrégation de fichiers d'origines différentes et qui ne se trouvent pas au même niveau de mise à jour. Touchant peu de clients, elles sont d'importance mineure sauf si, non décelées, elles sont considérées dans une fouille automatique comme très informatives car évidemment très improbables.

Ainsi, la recherche imprudente d'une classification conduit aux résultats du tableau 1. Une seule variable, mesurant l'ancienneté du client dans la banque, explique la classification obtenue. Il s'agit en fait d'un artefact dû à un petit sous-ensemble de clients artificiellement anciens. En se démarquant des autres, ils confèrent une variance importante à cette variable. L'expérience

[5] Certains sont explicités à titre d'illustration par Besse (2000).

TAB. 1 – *Statistiques relatives à une classification obtenue par réallocation dynamique (procédure FASTCLUST). La variable exprimant l'ancienneté du client (*RELAT*, liée à l'âge) explique à elle seule la classification obtenue. Toutes les autres ont des influences négligeables.*

Variable	Total STD	Within STD	R-Squared	RSQ/(1-RSQ)
AGER	11.894403	10.526963	0.220380	0.282676
RELAT	156.2459	20.373673	0.983077	58.090326
OPGNBL	0.684994	0.684542	0.005993	0.006029
MOYRVL	1.470973	1.428200	0.061724	0.065784
TAVEPL	4.907649	4.732474	0.074468	0.080460
ENDETL	1.277375	1.266556	0.021471	0.021942
GAGETL	5.468386	5.434921	0.016827	0.017115
GAGECL	3.672788	3.671744	0.005247	0.005275
GAGEML	4.530790	4.508201	0.014582	0.014797
...	...	...	0.0...	0.0...

montre que des étudiants apprentis prospecteurs tombent facilement dans ce piège en se contentant d'une typologie sans intérêt.

3.3 Modélisation

Il s'agit de construire des modèles prédictifs de la variable binaire : possession ou non de la carte Visa Premier. Quatre méthodes sont en concurrence sur cet objectif.

- L'analyse discriminante et ses variantes n'ont pas donné de bons résultats ; absente de la version basique de SEM (2001) elle est laissée de côté.
- La régression logistique associée à un algorithme de choix de modèle pas à pas.
- Le perceptron multicouche qui, compte tenu du nombre de variables en entrée et de la taille de l'échantillon, sera limité à 5 neurones sur la couche cachée. Ils sont munis de fonctions de transfert sigmoïdales.
- Les arbres de classification avec un critère d'homogénéité basé sur l'entropie.

La démarche de choix de modèle et de comparaison de méthode adoptée est celle implicitement proposée par SEM (2001). L'échantillon global est aléatoirement partagé en trois parties : *apprentissage* (60%), *validation* (20%) et *test* (20%). Pour chacune des trois méthodes retenues (logistique, perceptron, arbre), le modèle est estimé sur l'échantillon d'*apprentissage* et optimisé sur celui de *validation* : choix des variables de la régression logistique, limitation de l'apprentissage du perceptron et élagage de l'arbre.

3.4 Comparaison

Finalement, les modèles optimaux de chacune des trois démarches sont comparés en terme de taux de mal classés estimé sur le *seul échantillon test* (précaution nécessaire pour estimer sans biais). Le travail ainsi décrit a été réalisé par 10 binômes d'étudiants après exploration manuelle ou sélection et transformation automatique sur le même échantillon. Puis, pour tenir compte de la source de variation importante due à l'estimation de l'erreur sur l'échantillon test, la procédure a été répétée trois fois par chaque binôme et sur des échantillons différents. C'est très simple avec SEM (2001) puisqu'il suffit de modifier l'initialisation du générateur de nombres aléatoires dans le premier nœud du diagramme de l'interface schématisant l'enchaînement des outils.

Une première analyse de variance élémentaire fournit les résultats du tableau 2. Celui-ci montre l'absence d'effet du facteur binôme. Les étudiants, sans doute trop guidés, n'ont pas fait preuve de beaucoup d'initiatives. Nous ne pouvons donc malheureusement pas nous intéresser à la robustesse d'une méthode ou d'une stratégie vis à vis de l'inexpérience de l'utilisateur. Cette expérience devra donc être reconduite. En revanche, il est possible de s'intéresser aux effets des autres facteurs ; `Stratégie` a deux niveaux : automatique (`Aut`) et manuel (`Man`) tandis que `Méthode` en a trois : arbre de classification (`Arb`), régression logistique (`Log`) et réseau de neurones (`Res`). Les diagrammes boîtes (figure 1) montrent l'importance relative de la variance attachée à l'estimation du taux de mal classés. L'échantillon test de taille modeste (# 200) est une source importante de variation. Le tableau 3 montre le net effet de la stratégie et un effet de la méthode moins marqué tandis que l'interaction stratégie/méthode est négligeable. Ceci est repris dans la figure 2 qui montre la supériorité uniforme de la procédure manuelle et l'absence d'interaction significative même si la régression logistique fait un peu mieux dans le cas automatique.

3.5 Commentaires

Les résultats de cette expérience suggèrent quelques remarques :

1. la procédure manuelle est certes artisanale et plus longue mais elle permet, d'une part de détecter quelques incohérences dans les données et d'autre part de fournir des modèles significativement meilleurs, indépendamment de la technique utilisée. L'expertise humaine s'avère, sur cet exemple, nécessaire, voire incontournable.
2. Pénalisés par un effectif réduit de l'échantillon d'apprentissage, les réseaux de neurones se montrent systématiquement moins performants que la régression logistique (choix de modèle automatique par élimination) et surtout que les arbres de classification (élagage par optimisation

Tab. 2 – *Tous les binômes traitent le même échantillon. Analyse de variance (procédure REG de SAS, 1989) montrant les effets de la stratégie (automatique et manuelle) et, à un moindre degré, de la méthode (régression, arbre, réseau), l'absence d'effet du binôme, sur le taux de mal classés obtenu sur l'échantillon test.*

Type III Tests				
Source	DF	Mean Square	F Stat	Prob > F
BINOME	7	2.8578	1.5425	0.2319
STRATEG	1	460.7222	248.6738	0.0001
METHODE	2	6.8329	3.6880	0.0517
BINOME*STRATEG	7	2.9880	1.6128	0.2113
BINOME*METHODE	14	2.0221	1.0914	0.4362
STRATEG*METHODE	2	9.0747	4.8981	0.0244

Tab. 3 – *Chaque binôme traite trois échantilllons différents. Analyse de variance (procédure REG de SAS, 1989) montrant les effets de la stratégie (automatique et manuelle) et de la méthode (régression, arbre, réseau) mais l'absence d'interaction significative.*

Type III Tests				
Source	DF	Mean Square	F Stat	Prob > F
STRATEG	1	874.6242	146.5821	0.0001
METHODE	2	23.9387	4.0120	0.0197
STRATEG*METHODE	2	10.6399	1.7832	0.1710

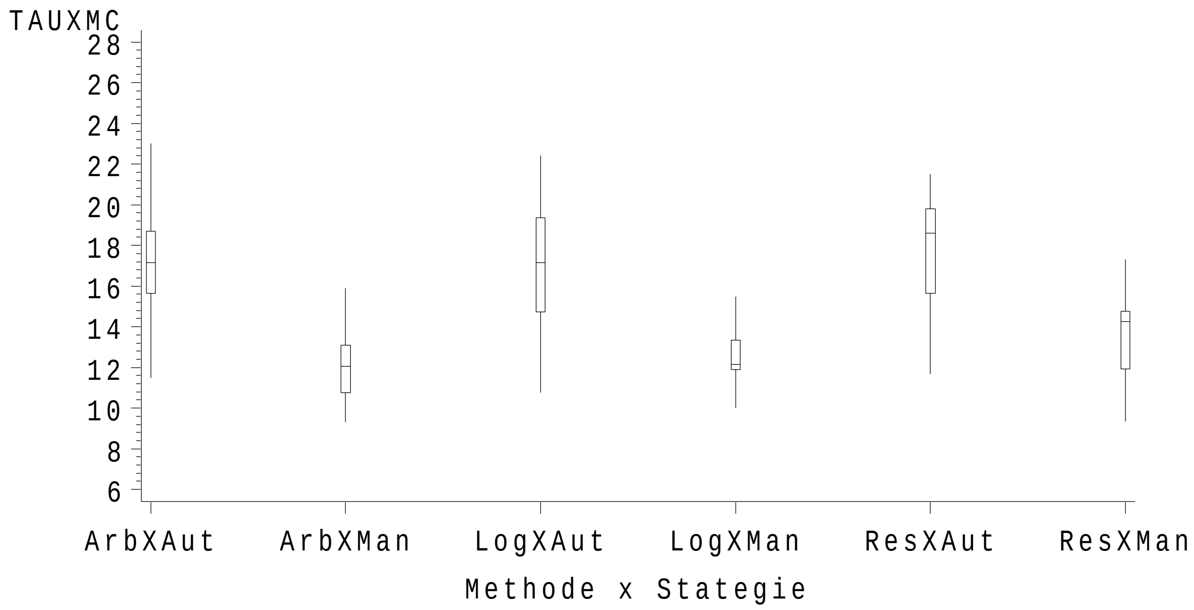

FIG. 1 – *Diagrames boîtes parallèles pour chaque combinaison de méthode et stratégie. Ceci montre l'importance relative de la variance associée à l'estimation du taux d'erreur sur un échantillon test de taille modeste (200).*

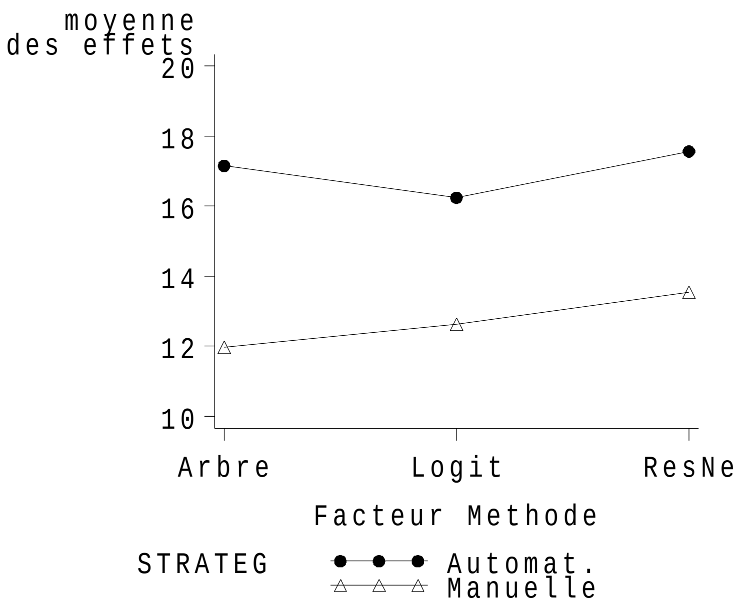

FIG. 2 – *Graphique comparant les effets moyens observés pour chaque combinaison de facteurs : interaction non significative et, globalement, meilleurs résultats de la procédure manuelle. L'arbre de classification semble plus performant sur cet exemple.*

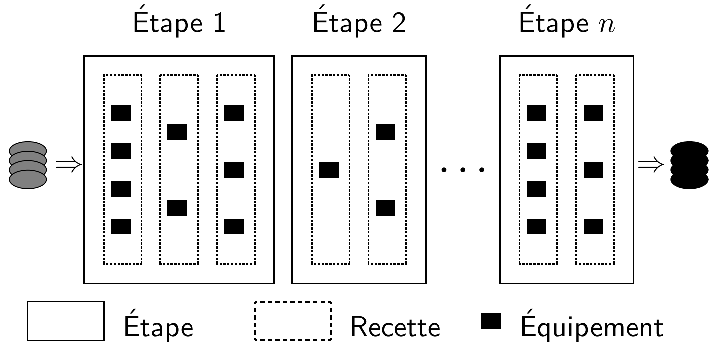

FIG. 3 – *Schématisation du processus de fabrication qui reçoit en entrée des lots de plaquettes vierges de silicium sur lesquelles sont intégrés les circuits au cours de cent à deux cents étapes durant un temps de cycle d'une dizaine de semaines.*

sur un échantillon de validation). De plus, la grande facilité d'interprétation de ces derniers les font nettement préférer sur cet exemple.

3. *Attention*, la stratégie qui consiste à comparer la qualité des modèles par leur performance sur un échantillon test est associée à une variance importante. La variance inter test est même plus importante que la variance inter méthode. Seul un jeu de données beaucoup plus volumineux ou une procédure de validation croisée (pas prévue dans SEM, 2001) permettrait d'améliorer la qualité de l'estimation des erreurs.

4 Procédé industriel

Le Contrôle Statistique des Procédés est largement utilisé dans l'industrie et intégré aux normes de qualité pour optimiser des réglages (plans d'expérience) ou détecter des dérives de procédés (cartes de contrôle). Dans le cas de la fabrication de circuits intégrés, le procédé est long et complexe, il requiert d'autres outils pour détecter d'éventuelles défaillances.

4.1 Données et objectifs

Les circuits intégrés sont fabriqués à partir de plaquettes de silicium, chacune contenant des centaines de circuits. Ces plaquettes sont regroupées par lots passant sur des équipements à chaque étape du procédé (cf. schéma de la figure 3). Une étape est composée de recettes au cours desquelles plusieurs équipements peuvent être utilisés indifféremment. De manière abusive mais pour simplifier, les recettes sont confondues avec les étapes. Les technologies actuelles requièrent plus d'une *centaine* d'étapes différentes. En *fin* *de fabrication*, les circuits sont testés électriquement. Il existe deux grands types de test :

– les tests de contrôle de procédé vérifient si celui-ci est conforme à la technologie. Ils permettent de détecter des dérives du procédé mais ne servent généralement pas à rejeter des circuits. Selon la technologie, on en compte plusieurs centaines.
– Les test sous pointes, ou tests électriques, sont effectués sur chaque circuit de chaque lot. Le test sous pointes consiste en un ensemble de mesures électriques (courants, fréquences, tensions) dont les spécifications ont été déterminées avec les clients. Tout circuit qui n'est pas dans la spécification pour un test est rejeté. La proportion de circuits ayant passé l'ensemble des tests est appelée rendement.

La présence d'une défaillance, cause d'une chute de rendement, n'est observable qu'en *fin de production*. Compte tenu du temps de cycle qui peut être de plusieurs mois, plus tôt la cause est trouvée, plus tôt l'action corrective peut être mise en place et les plaquettes épargnées. Dans ce contexte, le *data mining* est utilisé pour accélérer le processus de recherche de défaillance. Deux sortes d'analyse sont expérimentées :

– une analyse qui consiste à trouver si les lots mauvais sont passés par une même séquence d'équipements au cours du procédé de fabrication et donc à suspecter ces équipements ;
– une analyse qui consiste à caractériser physiquement des rejets à des tests électriques à l'aide des tests de contrôle de procédé afin de préciser l'origine de la défaillance.

Les données se présentent sous la forme d'un tableau dont les lignes sont les lots et les colonnes des variables précisant : le rendement moyen, l'équipement par lequel est passé le lot pour chacune des étapes, les tests de contrôle. Pour ne pas bruiter les données relatives au moment de la défaillance, il est nécessaire de ne considérer que les lots correspondant à la période incriminée. Ainsi, on travaille rarement sur les données de plus d'une centaine de lots. Pour les deux types d'analyse, compte tenu du nombre d'étapes du procédé de fabrication, ou du nombre de tests de contrôle, il y a plusieurs centaines de variables explicatives pour une seule variable à expliquer : un rendement moyen par lot ou une variable binaire le caractérisant (bon ou mauvais).

L'objectif semble *a priori* classique : expliquer une variable par un ensemble de variables qualitatives (les étapes du procédé) prenant diverses modalités (les équipements) ou encore de variables quantitatives (tests de contrôle). La situation est néanmoins très particulière par le nombre de variables explicatives ainsi que par l'objectif recherché : la construction d'un modèle à une variable désignant l'équipement ou le test de contrôle incriminé, éventuellement un modèle à deux variables dans le cas, semble-t-il rare, d'interactions entre équipements.

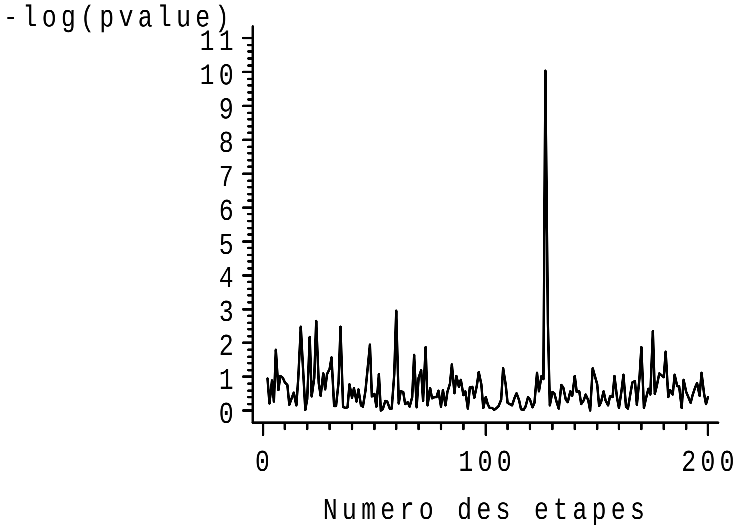

FIG. 4 – *Graphique de* $\log_{10}(1/p)$ *(p-valeurs : probabilités de dépasser la valeur de la statistique de test) des analyses de variance pour chacune des étapes du procédé. Compte tenu du nombre de tests réalisés il est naturel d'obtenir plusieurs valeurs supérieures à 2 mais une étape se distingue nettement des autres.*

Différentes approches sont envisageables. Un outil est déjà installé en production (Bergeret & Chandon, 1999) sous la forme d'un tableau de bord. Il calcule systématiquement, pour chaque étape, les analyses de variances permettant de tester l'influence du facteur équipement sur la variable rendement. Autrement dit, le rendement moyen des lots passés par un même équipement est-il significativement différent de ceux passés par d'autres équipements à une même étape. Les résultats obtenus sont comparés avec ceux d'autres approches, arbres ou régression logistique. Compte tenu du faible nombre de lots au regard du nombre de variables en entrée et à cause de leur difficulté d'interprétation, les réseaux de neurones sont dans ce cas inadaptés.

4.2 Recherche d'une séquence d'équipements défaillants

Dans ce premier exemple, le problème de rendement a été causé par l'utilisation d'un type d'équipement plutôt qu'un autre à une étape du procédé de fabrication. Autrement dit, les lots mauvais sont majoritairement passés par des équipements de type A alors que les lots bons ont été traités principalement par des équipements de type B. Il n'y a pas d'interaction connue avec d'autres équipements à d'autres étapes. Le tableau de bord, comparant systématiquement les rendements moyens des équipements d'une même étape par des analyses de variance, révèlent plusieurs tests significatifs au niveau de 1% (cf. figure 4). Un test se détache nettement associé à une probabilité de l'ordre de 10^{-10}. Chaque étape pointée est alors analysée avec les responsables du procédé de fabrication. Les comparaisons visuelles des histogrammes des rendements par équipement et de leurs séries de taux de rejets en fonction de la date de passage des lots à l'étape permettent de bien identifier les équipements et d'éliminer les fausses alarmes. Elles sont dues à des particularités du procédé de fabrication ou à la présence de certaines valeurs qui peuvent influencer l'analyse de variance ou encore au simple hasard compte tenu du nombre de tests réalisés sur le même échantillon. Les investigations aux étapes retenues par ces analyses ont montré que l'étape nommée IMPD, associée à la plus faible probabilité (10^{-10}), et les équipements suspectés (IMPL1) étaient responsables de la défaillance. Pour information, le gain en rendement obtenu après l'utilisation exclusive des équipements de type IMPL2 est de l'ordre de 7,5 %.

Pour faciliter la diffusion des résultats, un arbre de classification, plus simple à interpréter, a également été estimé à l'aide du logiciel S-plus (1997). Dans ce cas, les lots sont classés en bon ou mauvais selon leur taux de rejet à un test électrique, ou plus simplement selon leur rendement. L'arbre de classification obtenu (figure 5) est structuré en 3 niveaux. Chaque niveau met en cause une étape et des équipements. Le premier niveau sépare les lots selon deux branches : celle des “bons” ne contenant aucun lot mauvais et celle des “mauvais” ne contenant que 7 lots bons sur 78. Les deux autres niveaux permettent d'affiner les résultats sur les lots mauvais, néanmoins ils semblent moins déterminants dans la classification des lots. Un élagage par validation croisée confirme cette observation puisque la première division est seule retenue (figure 6). Les résultats suggèrent donc que les lots mauvais sont passés par les équipements de type IMPL1 à l'étape IMPD, et qu'il n'y a pas d'interaction avec d'autres équipements à d'autres étapes.

Un arbre de régression a également été estimé sur les mêmes données. L'arbre ainsi obtenu est structuré en quatre niveaux. Le premier niveau est aussi basé sur l'étape IMPD. Il distingue clairement les lots bons des lots mauvais selon l'utilisation de l'équipement. Les équipements de type IMPL1 semblent générer les lots qui ont en moyenne les rejets les plus importants (19,4 % de rejets en moyenne contre 0,13 % pour l'équipement de type IMPL2). Les autres niveaux de l'arbre sont différents de ceux de l'arbre de classification ; ils semblent eux aussi très peu informatifs. Quant à l'élagage par validation croisée, il est encore plus favorable à la conservation du seul premier niveau.

En comparant arbre de régression et tableau de bord basé sur le test de Fisher, il n'est pas surprenant de constater que ces deux méthodes désignent la même variable. En effet, la statistique de test correspond au critère d'homogénéité maximale qui s'exprime comme une déviance d'un modèle linéaire à un facteur pour la sélection de la variable à utiliser. Néanmoins, ces deux approches apportent quelques nuances. Dans l'exemple présenté où un type

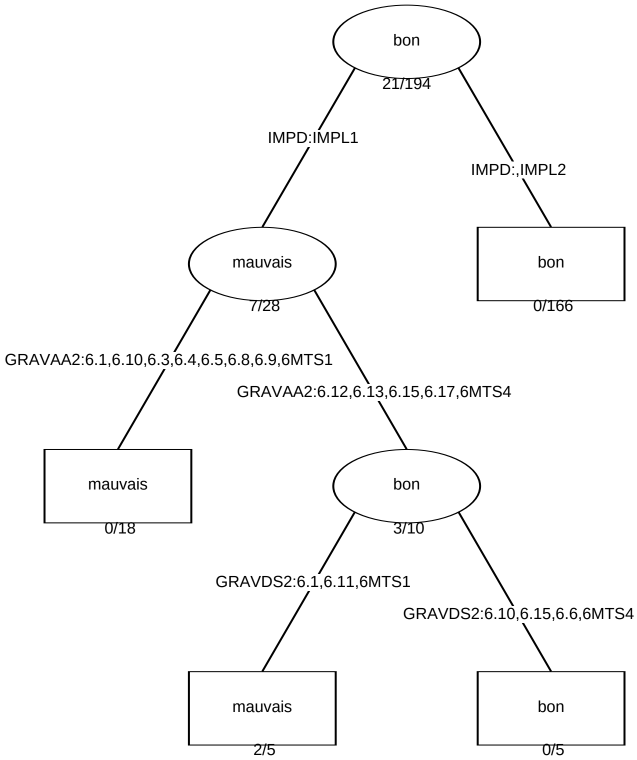

FIG. 5 – *Arbre de classification non élagué pour la recherche d'une séquence défaillante d'équipements. À chaque niveau, sont précisés la variable (l'étape) opérant la meilleure dichotomie entre les lots ainsi que les modalités (équipements concernés). On donne aussi pour chaque nœud le rapport du nombre de mal classés sur l'effectif total.*

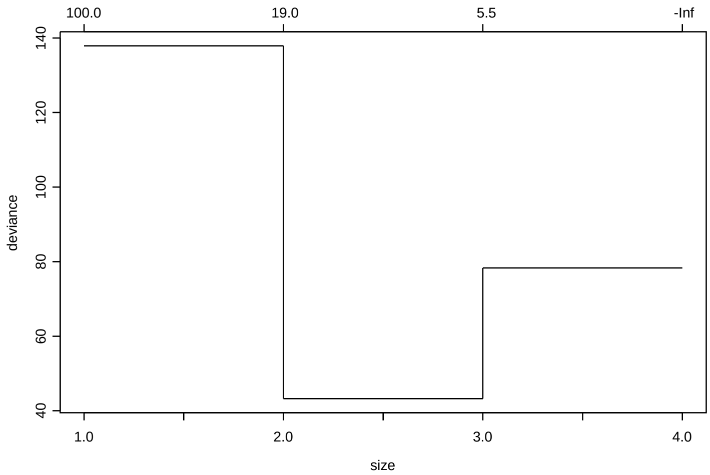

FIG. 6 – *Élagage par validation croisée. Comportement de la déviance du modèle en fonction du nombre de nœuds retenus dans l'arbre. Un seul nœud avec deux feuilles est retenu.*

d'équipement est mis en cause, l'arbre met directement en avant cette information par une partition des modalités, donc des équipements, par types. Dans le cas du tableau de bord, une étape complémentaire de comparaison multiple est nécessaire. De plus, un arbre est susceptible de prendre en compte naturellement d'éventuelles interactions entre équipements, contrairement aux analyses de variance considérées. Néanmoins, le problème des fausses alertes peut s'avérer délicat dans des situations moins contrastées. Si la défaillance d'un équipement n'est pas franche ou encore si l'écart entre les équipements défaillants et les équipements corrects est faible, le choix des divisions de l'arbre est influencé par le nombre de modalités. Autrement dit, entre deux étapes dont l'une présente un équipement épisodiquement défaillant parmi trois et l'autre composée de beaucoup d'équipements, l'algorithme de choix de la variable la plus discriminante peut s'arrêter sur celle présentant le plus de modalités. Il n'est donc pas possible de négliger une analyse manuelle et détaillée des résultats obtenus.

4.3 Caractérisation physique des rejets à un test électrique

Dans l'exemple suivant, la stratégie décrite précédemment ne permet pas de mettre en cause un équipement particulier. Une autre approche consiste à caractériser la défaillance par les tests de contrôle capables de détecter une dérive du procédé de fabrication.

Ici, la chute de rendement est causée par des rejets importants à un test électrique. Les arbres de régression et de classification donnent des résultats différents, et, dans les deux cas, le premier niveau est construit à partir d'une étape utilisant de nombreux équipements. L'analyse manuelle de chacune des étapes laisse supposer qu'on est dans le cas où la cause du problème est plutôt marginale. Peu de crédit peut être accordé à ces résultats. Le tableau de bord, quant à lui, pointe trois étapes différentes de celles des arbres. Deux d'entre elles mettent en cause le même équipement (6203). Une analyse détaillée a montré que ces résultats étaient assez crédibles. Néanmoins, ils ne satisfont pas complétement l'ingénieur responsable du produit. Pour approfondir ses investigations, il souhaite caractériser physiquement les rejets du test électrique. C'est pourquoi, différents modèles sont estimés pour expliquer les rejets par les tests de contrôle de procédé. Avant élagage, l'arbre de régression obtenu (figure 7) est structuré en deux niveaux. Le premier, basé sur la variable quantitative P22351, sépare les lots en deux sous-populations. Clairement selon le nombre moyen de rejets de chaque sous-population, cet arbre sépare en deux branches les lots mauvais (moyenne élevée donnée dans les ellipses qui symbolisent le nœud) des lots bons (moyenne faible). Les autres divisions de l'arbre affinent ces résultats mais elles ne semblent pas apporter d'informations supplémentaires utiles. L'élagage par validation croisée conforte cette observation puisqu'il s'arrête au premier niveau. Le graphique donnant les rejets au test en fonction de la variable P22351 souligne la présence d'un effet seuil (figure 8). En effet, à gauche du seuil donné par l'arbre (-0.84), la proportion de lots mauvais est importante.

À partir de ces résultats, l'ingénieur responsable du produit a identifié trois étapes du procédé de fabrication dont deux mettant en cause l'équipement 6203. En pratiquant des modifications à ces étapes, il a pu augmenter le rendement du produit d'environ 5 %.

L'efficacité de cette approche est évaluée par comparaison à une régression logistique en classant les lots bons et mauvais selon leur taux de rejet au test électrique. La procédure `logit` de SAS (1989) avec sélection pas à pas automatique des variables les plus significatives conduit aux résultats du tableau 4.

La variable P22351 est la première variable entrée dans le modèle, trois autres variables sont considérées comme significatives dans l'explication des rejets. Ces 4 variables ne sont pas corrélées entre elles. Selon l'ingénieur, deux des nouvelles variables peuvent être liées au problème, mais de manière moins déterminante. Quant à la dernière variable restante, aucun lien évident n'est connu avec le problème.

Dans cet exemple, chaque technique présente des spécificités qui montrent plus leur complémentarité que leurs oppositions pour atteindre l'objectif recherché. Les analyses de variances préalables donnent déjà une indication sur l'équipement mais les arbres ne parviennent pas à identifier l'étape concernée. La caractérisation de la défaillance est finalement obtenue par l'identification du ou des tests de procédé les plus liés. Arbres et régression logistique indiquent bien le bon test mais diffèrent sur les critères et stratégies

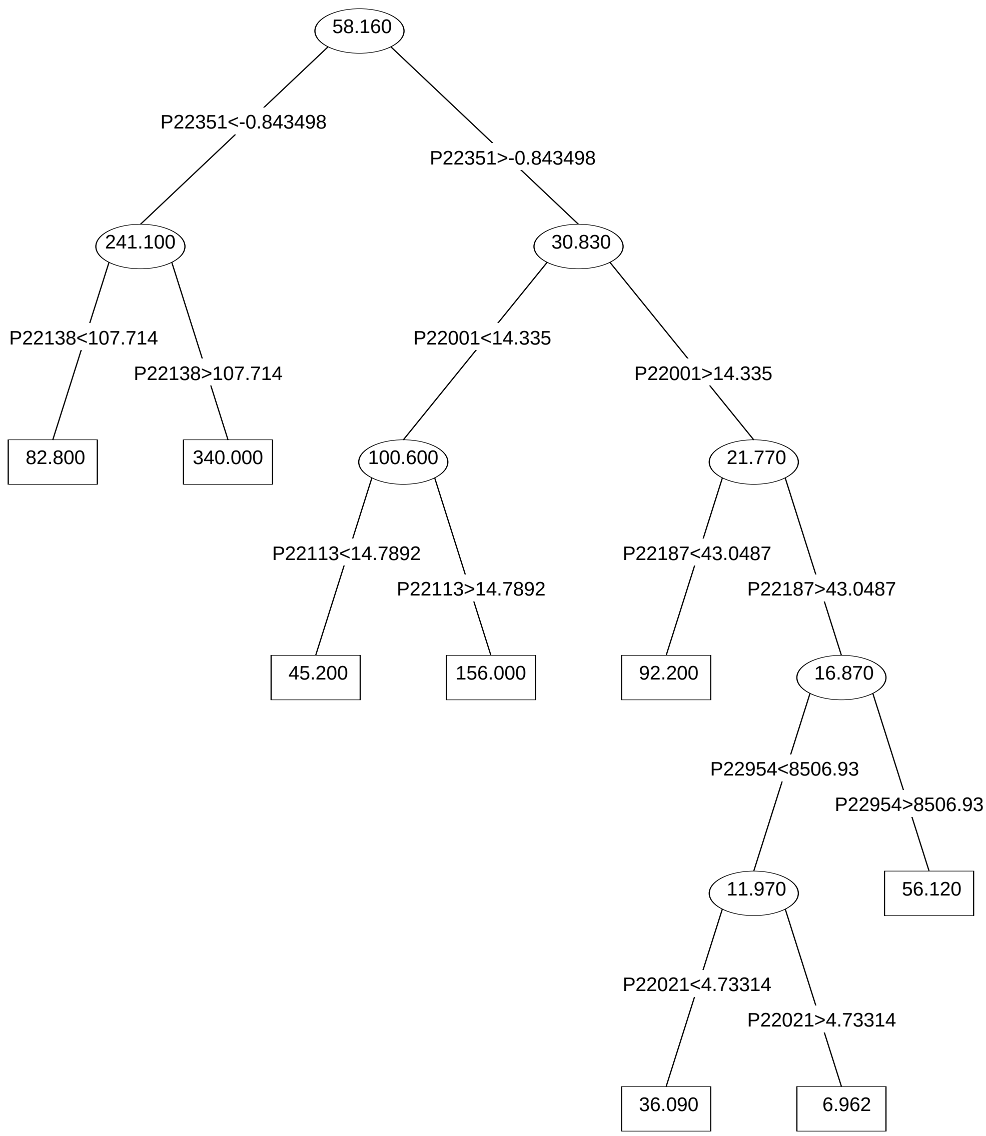

FIG. 7 – *Arbre de régression non élagué pour la caractérisation de rejets à un test électrique. Seule la première division associée à la variable P22351 est conservée lors de l'élagage.*

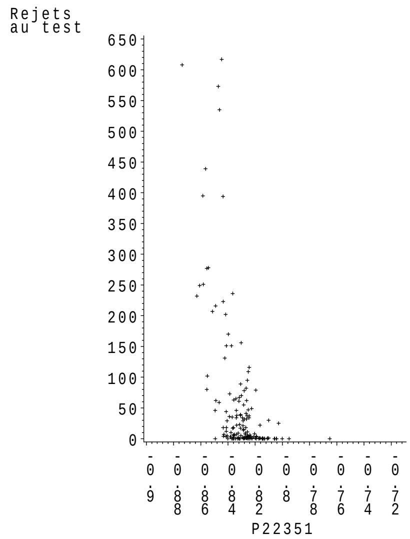

Fig. 8 – *Graphique des nombres de rejets par plaquette au test électrique en fonction de la variable de contrôle P22351. La proportion de rejet est plus importante en dessous de la valeur seuil (-0,84) identifée par l'arbre de régression.*

Tab. 4 – Résultats de la régression logistique.

Summary of Stepwise Procedure

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	P22351		1	14.1106	.	0.0002
2	P22954		2	16.4431	.	0.0001
3	P22011		3	11.4519	.	0.0007
4	P22194		4	11.8078	.	0.0006
5	P22044		5	20.0683	.	0.0001
6		P22044	4	.	1.3621	0.2432

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-82.3508	67.0387	1.5090	0.2193	.	.
P22954	1	0.000542	0.0002	7.3442	0.0067	1.379020	1.001
P22011	1	-0.00062	0.000224	7.6379	0.0057	-1.865285	0.999
P22194	1	-52.5976	20.1700	6.8001	0.0091	-1.136925	0.000
P22351	1	-334.5	119.3	7.8690	0.0050	-2.243490	0.000

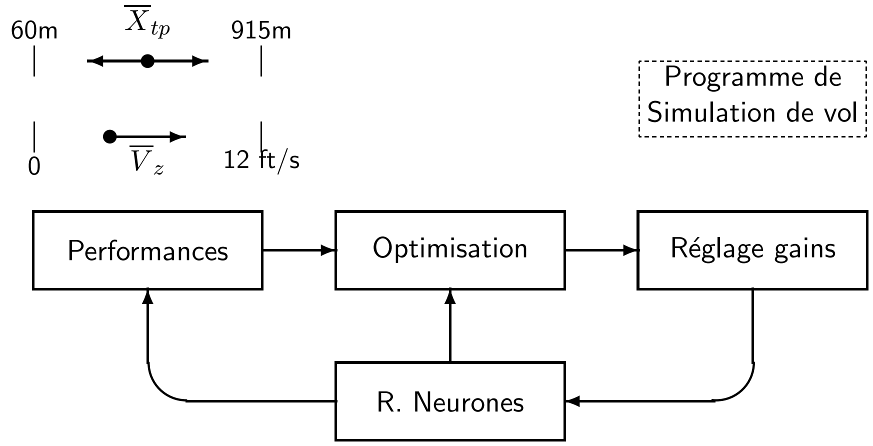

FIG. 9 – *À partir d'un comportement de PA défini, on cherche les variations des gains qui permettent d'obtenir les performances statistiques exigées par la réglementation. On construit d'abord un modèle de prévision des statistiques résumant les variables de l'impact en fonction des gains, puis on inverse le processus.*

pour le choix d'autres variables. Comme dans l'exemple précédent, un arbre pourrait déceler une éventuelle interaction mais il serait intéressant de compléter les résultats en fournissant la liste des variables concurrentes (Ghattas, 1999) lors du premier niveau de segmentation de l'arbre. De son côté la régression logistique, sur la base d'un autre critère, fournit une liste de variables pouvant établir un faisceau de présomptions.

5 Applications aéronautiques

Deux exemples sont présentés dans cette section, chacun posant des problèmes particuliers. Dans le premier, les variables à modéliser sont quantitatives et issues de simulations. Elles peuvent donc être générées en fonction des besoins de l'étude contrairement à la situation plus habituelle en *data mining* de données provenant d'un entrepôt souvent construit à d'autres fins. Le deuxième exemple est encore un cas de discrimination d'une variable binaire mais, destiné à être embarqué comme aide au pilotage, il doit valider les procédures légales de certification.

5.1 Réglage d'un Pilote Automatique d'Atterrissage

Objectifs

Le contexte de l'étude (Raimbault et coll., 2001) est le réglage de la loi d'atterrissage automatique d'un avion. Schématiquement, un tel système doit guider l'avion selon une trajectoire spécifique jusqu'à l'impact sur la piste. Le guidage s'effectue à l'aide de gouvernes dont les mouvements sont calculés par le pilote automatique (PA) d'atterrissage.

Les ordres donnés aux gouvernes sont le résultat d'équations différentielles (fonction de l'altitude, de la vitesse, etc...) dont les paramètres doivent être optimisés. En effet, le comportement de l'avion doit être satisfaisant quelles que soient les conditions de chargement (masse, centrage) et les conditions météorologiques. Les performances du pilote automatique sont évaluées par la position de l'impact sur la piste (X_{TP}) et la vitesse verticale (V_z). La réglementation aéronautique impose un périmètre de sécurité au milieu de la piste où l'avion peut atterrir ainsi qu'une vitesse verticale raisonnable préservant le train d'atterrissage. On doit garantir que la probabilité de ne pas respecter ces contraintes de sécurité est inférieure à 10^{-9}.

Il s'agit donc de trouver les meilleurs paramètres (ou gains) de la loi en fonction de la robustesse recherchée. Cette robustesse s'exprime en termes de statistiques (moyenne et écart-type) des performances du pilote automatique (X_{TP} et V_z). On utilise ces quantités ainsi que les bonnes caractéristiques gaussiennes des résidus pour les “démonstrations de risque”.

Un logiciel de simulation d'atterrissage d'un avion permet de faire ces calculs de robustesse et de bon comportement du PA. Mais cette simulation procède par intégration d'équations de la mécanique du vol, ce qui rend le calcul trop lent pour son intégration dans une procédure d'optimisation. La simulation doit donc être approchée par un modèle suffisamment précis pour rendre possible la recherche de gains optimaux.

La méthode globale est décrite dans la Figure 9. On dispose d'une loi d'atterrissage bien définie dont la robustesse en termes statistiques n'est pas encore optimale. Dans un premier temps, il faut construire un modèle de prévision des performances statistiques en fonction des variations que l'on peut appliquer aux gains (paramètres). Le processus est ensuite inversé pour trouver les paramètres optimaux de la loi en fonction des performances recherchées. Une nouvelle loi d'atterrissage est ainsi obtenue suffisamment robuste vis à vis des contraintes de la réglementation si la stabilité n'a pas été perturbée.

Données

Chaque donnée simulée est une statistique (moyenne et écart-type de V_Z et X_{TP}) sur 200 approches pour lesquelles un paramètre de vol (masse, centrage, vent,..) est bloqué à une valeur critique (minimum ou maximum)

TAB. 5 – *Comparaisons des qualités d'ajustement et de prédiction (R^2) des différents modèles par resubstitution (sur l'échantillon d'apprentissge) et sur l'échantillon test.*

Modèle	Apprentissage	Test
régression quad.	0,49	0,40
régression cub.	0,57	0,51
Rés. Neurones	0,92	0,89

alors que les autres conditions de vol sont choisies aléatoires uniformément dans leur domaine de variation. On dispose de 1200 données soit 1200×200 atterrissages simulés que l'on répartit en données d'apprentissage (600), de validation (400) et de test (200).

Les variables d'entrée sont les paramètres de la loi (gain) plus un code indiquant le paramètre de vol fixé à son extremum, soit 18 entrées. En sortie, on calcule la moyenne et la variance des deux variables X_{TP} et V_z, soit 4 sorties.

Modélisation

L'objectif est constitué de modèles de prévision de 4 variables expliquées par 18 autres reliées par un phénomène connu pour être non-linéaire. On s'intéresse ici à la comparaison des performances des modèles polynomiaux vis à vis des réseaux de neurones pour approcher ce modèle. En effet, compte tenu du nombre réduit de données, des modèles polynomiaux pourraient s'avérer efficaces.

Pour chacune des 4 variables de sortie, un polynôme est ajusté sur les données. Les modèles quadratiques donnant de très mauvais résultats, des polynomes d'ordre s'avèrent nécessaires. On obtient malgré tout un modèle de qualité médiocre car la recherche d'un bon modèle parmi toutes les interactions ne peut pas être prise en compte par les algorithmes automatiques classiques de choix de modèle : 1957 paramètres à estimer pour 600 données.

Dans un deuxième temps, un réseau de neurones a été estimé pour chaque variable. Il s'agit d'un perceptron à une seule couche cachée composée de 12 ou 14 neurones suivant la variable à expliquer. L'apprentissage est conduit par rétro-propagation sur les 600 données. L'architecture est validée sur les données de validation. Les données de test servent de référence pour l'évaluation de la capacité de généralisation et les comparaisons. Les résultats des ajustements et qualité de prédiction des modèles sont comparés dans le tableau 5 qui montre un net avantage à l'approche neuronale.

Les réseaux de neurones présentent dans ce cas de très bonnes performances de modélisation malgré leur complexité relativement au faible nombre de données. Les modèles polynomiaux sont difficiles à mettre en œuvre à cause du grand nombre de variables et de la nécessité de faire intervenir de nombreuses interactions. Ils perdent même alors leur caractère explicatif et se trouvent surpassés par la flexibilité du perceptron. Les résultats des ajustements des différents modèles sont résumés dans le tableau 5.

5.2 Détection de pompage piloté

Objectif

Le pompage piloté (Pilot Induced Oscillations ou PIO) est un phénomène critique pour la sécurité des avions. Il s'agit d'interactions en boucle fermée entre le pilote, l'avion et les lois de pilotage qui conduisent à des mouvements oscillatoires de l'avion (cf. figure 10). Lorsque ces oscillations sont entretenues ou divergentes, leurs grandes amplitudes peuvent rendre l'avion instable, mettant en cause la sécurité. De nombreux facteurs de déclenchement du pompage piloté ont été mis en évidence : non-linéarités dans les lois de pilotage, retards (calculateurs, visualisation). Ces constats ont donné lieu à l'élaboration de critères spécifiques de conception des lois de pilotage dans le but d'éradiquer la tendance au pompage. Néanmoins, le facteur humain (comportement du pilote) reste une composante incontournable du déclenchement de ce phénomène. Des cas de pompage piloté peuvent ainsi apparaître de manière inattendue pendant les phases de développement d'un nouvel avion. Un modèle de détection d'amorce de pompage piloté en temps réel associé à un dispositif de compensation sur les commandes de vol constitue une solution de secours.

L'objectif de l'étude (Raimbault & Fabre 2001) est donc de développer un algorithme évaluant en temps réel la tendance au PIO. Le détecteur doit être conçu pour détecter le PIO avec la même acuité qu'un expert. Il vise en effet à reproduire de manière automatique et fiable l'expertise d'un ingénieur.

L'analyse du phénomène de pompage piloté permet de caractériser le phénomène PIO comme un couplage entre les signaux pilote (manche) et avion (assiette ou roulis) de type sinusoïde de grande amplitude, de fréquence comprise entre 0.3 et 0.6 Hertz, associé à une saturation des gouvernes. Ce constat oriente donc l'étude vers l'utilisation d'une décomposition fréquentielle des signaux pilote et avion et l'évaluation de la saturation des gouvernes. Il s'agit de concevoir un modèle permettant de synthétiser toutes ces données pour évaluer, en temps réel, la tendance au pompage.

Modélisation

Un travail préalable de collecte de vols comportant ou non des phases de pompage fournit une base de données pour l'estimation des paramètres. Ces données sont analysées par un expert qui détermine précisément quelles

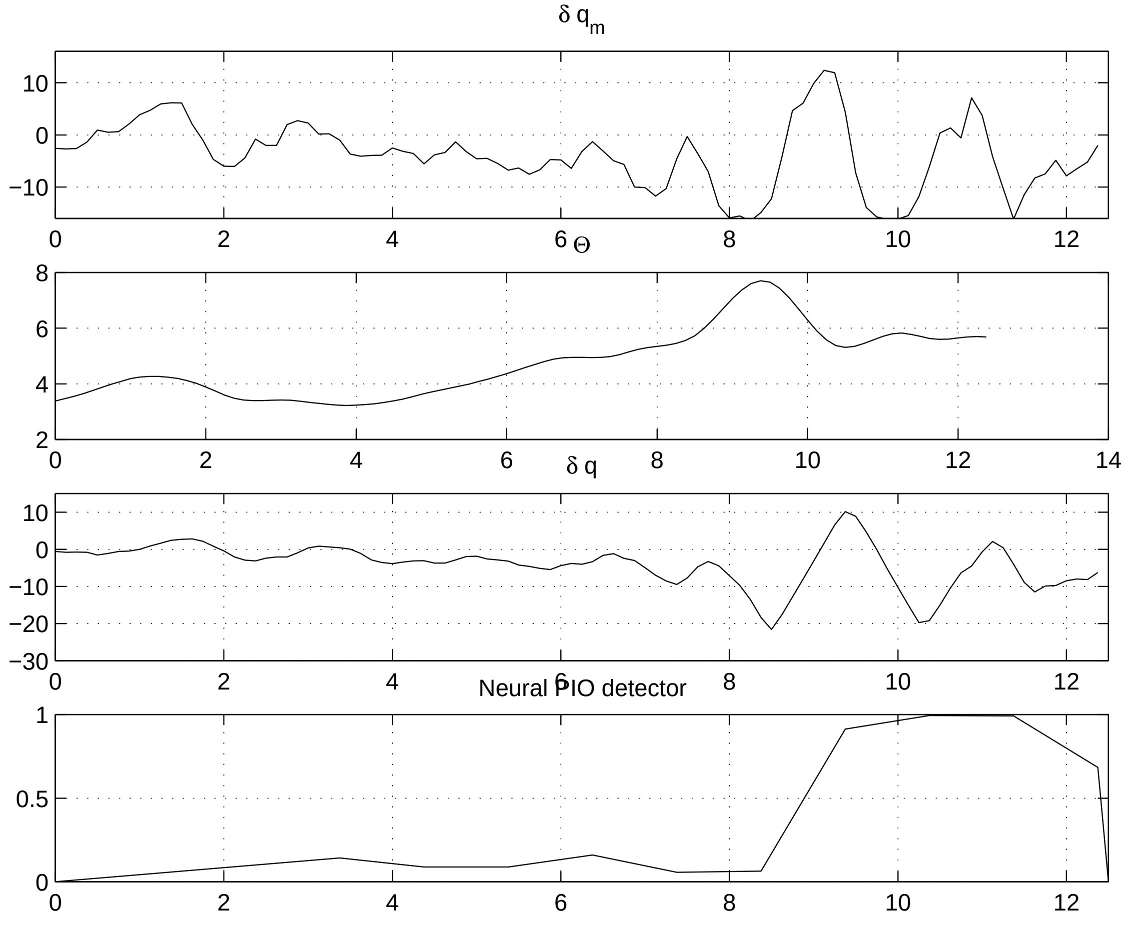

FIG. 10 – *Les trois premiers graphiques (de haut en bas) représentent les paramètres de vol en fonction du temps : le manche, l'avion (assiette), la gouverne (profondeur). Le quatrième est la réponse du réseau de neurones modélisant la propension au PIO et donc susceptible de déclencher une réaction du sytème. Ces graphiques analysent le comportement longitudinal de l'appareil mais les mêmes résultats sont disponibles pour le comportement latéral en remplaçant assiette par roulis, et gouverne par aileron.*

sont les zones de temps où un pompage apparaît et quelle est sa sévérité. Ces informations permettent de régler et de valider la capacité de détection des modèles.

À partir de l'analyse du phénomène de pompage et des données disponibles, un réseau de neurones est estimé. La méthode est basée sur un traitement préalable des signaux (décomposition en série de Fourier, approximation polynomiale). Les résultats de ces traitements constituent les entrées du réseau tandis que la sortie est associée à l'indice de pompage piloté. Il suffit d'un perceptron à une seule couche cachée pour obtenir une très bonne qualité de détection. L'apprentissage se fait à partir des données précédemment décrites par rétro-propagation. Un détecteur de PIO fiable autant dans les phases calmes qu'agitées a ainsi été estimé avec succès (cf. figure 10).

Certification

Une telle application souligne l'intérêt des réseaux de neurones dans leurs applications en traitement du signal, reconnaissance de forme ou de caractéristiques particulières. Cette application à l'aéronautique, domaine à exigence sécuritaire s'il en est, soulève le problème de la démonstration des performances d'un tel modèle. En effet, le détecteur neuronal de PIO est prévu pour déclencher, soit une alarme, soit une compensation pour arrêter le phénomène. Ce type de dispositif doit vérifier des contraintes strictes de fiabilité ; par exemple une probabilité plus petite que 10^{-5} de déclenchement de l'alarme alors qu'il n'y a pas de pompage piloté. Pour démontrer de telles performances, il faut pouvoir estimer un niveau de confiance dans la capacité de généralisation du réseau de neurones, i.e. en son estimation du pompage piloté sur des données non apprises. Il faut pouvoir mesurer l'erreur de généralisation et garantir avec un niveau α donné que l'erreur ne dépasse pas l'erreur maximum tolérable.

Des utilisations de plans d'expériences peuvent permettre de bien choisir les points d'apprentissage et de validation pour le réseau tandis que la validation croisée (Efron, 1983) permet d'optimiser à la fois le choix des points d'apprentissage parmi les échantillons disponibles et la structure du réseau. Il existe par ailleurs pour les perceptrons à une seule couche des règles de choix (Baum & Haussler, 1989) du nombre de données d'apprentissage en fonction de la précision souhaitée et de la complexité du réseau (par exemple 10 fois plus de données que de poids pour 10% des cas de validation mal modélisés). Tous ces outils permettent d'obtenir le meilleur modèle possible. Le bootstrap (Tibshirani, 1996) peut permettre d'estimer l'erreur de généralisation. Il existe enfin des méthodes de construction d'intervalles de confiance (Hwang & Ding, 1997 ; De Veaux et coll., 1998) et des tests basés sur la normalité des erreurs permettant de contrôler le risque que l'erreur dépasse le maximum toléré.

Cette revue bibliographique permet de poser les bases d'une application au problème concret de l'industrialisation du détecteur neuronal de pompage piloté. Cette problématique est loin d'être propre à l'aéronautique et constitue actuellement un thème de réflexion intéressant et concret autour de l'utilisation des réseaux de neurones. Elle est de toute façon incontournable pour espérer voir en vol, dans un avenir proche, un dispositif neuronal.

6 Conclusion

Cet article propose un tour d'horizon partiel donc nécessairement partial des techniques rencontrées en *data mining*. Il présente principalement le point de vue de statisticiens et néglige donc certains des apports de l'Intelligence Artificielle (machine learning). De plus les exemples présentés ne visent pas à l'exhaustivité des types de problèmes rencontrés mais nous allons tâcher d'en tirer quelques enseignements relativement aux questions soulevées en introduction.

Choix de méthodes. Beaucoup de méthodes d'origine et de conception très différentes poursuivent les mêmes objectifs de modélisation en vue d'une prévision. Dans les bons cas, données bien structurées, elles fournissent des résultats très similaires, dans d'autres une méthode peut se révéler plus efficace compte tenu de la taille de l'échantillon ou géométriquement mieux adaptée à la topologie des groupes à discriminer. Enfin, ces méthodes ne présentent pas toutes les mêmes capacités d'interprétation. Il n'y a donc pas de choix *a priori* meilleur, seuls l'expérience et un protocole de *test* soigné permettent de se déterminer, à moins d'opter pour une combinaison (*bagging*) de modèles. Les exemples présentés abondent en ce sens. C'est la raison pour laquelle des logiciels généralistes comme SAS (SEM, 2001) ne font pas de choix et offrent ces méthodes en parallèle pour mieux s'adapter aux données, aux habitudes de chaque utilisateur ou client potentiel et à la mode.

Automatisation et expertise. Au cours des années 80 nous avons pu assister à l'expansion puis au déclin et à la disparition des logiciels dits *systèmes experts* chargés de simuler le travail d'un expert humain statisticien par un *moteur d'inférences* opérant sur une *base de connaissances*. La plus grande prudence est encore requise face à des procédures visant à remplacer ou automatiser une expertise d'autant que l'exploration manuelle est finalement la meilleure façon de se familiariser avec des données, de s'assurer de leur cohérence ou de leur intégrité. Des automatisations sont possibles et même souhaitables lorsque le volume des données ou le temps réel l'impose, mais cela doit se faire dans un cadre très strict suivant la problématique ainsi que les données disponibles. Rien n'est plus simple que de prendre un artefact trivial pour une pépite de connaissance. L'exemple de marketing bancaire ci-dessus, comme celui de la détection de défaillance, illustrent bien cet aspect. Ainsi, une expertise statistique reste importante car la méconnaissance des limites et pièges des méthodes employées peut en effet conduire à des aberrations discréditant la démarche et rendant caducs les investissements consentis. En cumulant les problèmes de définition, gestion des bases de données, les problèmes de réseau,. . . , les méthodes statistiques ou algorithmiques de modélisation, le champ des compétences requises pour prétendre à une gestion efficace de l'information est d'une étendue redoutable. L'utilisateur peut maintenant disposer d'outils très conviviaux avec lesquels il est facile et rapide d'obtenir des résultats. De façon paradoxale, un petit quart d'heure suffit pour se familiariser avec une interface graphique qui exécute des méthodes dont une compréhension fine nécessite plusieurs heures de cours ou réflexion à Bac+5.

Fiabilité des résultats. L'estimation de la variabilité ou d'un taux d'erreur, que ce soit pour optimiser des modèles, comparer des méthodes ou encore contrôler les capacités de généralisation d'un modèle, pose de réels problèmes. D'une part, une approche imprudente fournit des estimations biaisées (optimistes) à force de vouloir optimiser et réestimer sur le même jeu de données ; d'autre part, l'estimation sur un échantillon test est soumise à une forte variance. Lorsque la preuve mathématique fait défaut du fait de la complexité des méthodes ou des algorithmes mis en jeu, l'expérience du statisticien peut s'avérer utile dans le cadre de cette problématique et devient incontournable pour répondre aux besoins d'une législation comme dans le cas des applications aéronautiques.

Les exemples traités montrent bien qu'il serait illusoire de croire qu'une méthode ou un logiciel est applicable à tout problème de fouilles des données. Comme cela a déjà été mentionné, la caractéristique essentielle du prospecteur est d'initier une démarche qui va dépendre des caractéristiques des données qui lui sont soumises. La recherche de fraudes parmi des millions de transactions journalières par carte bancaire ne fait pas appel aux mêmes outils que l'identification de configurations particulières d'une image en astrophysique. Les mêmes outils de discrimination peuvent éventuellement être utilisés mais le traitement préalable des données (déconvolution, Fourier, ondelettes. . .), souvent essentiel, sera, lui, très spécifique.

Le *Data Mining* ne peut être considéré comme une discipline. Au fil de cet article, il apparaît plus comme le confluent d'approches statistiques et informatiques au service de la discipline à l'origine des données et de leur problématique. De toute évidence, le champ des compétences concernées ne peut être couvert que par une équipe pluridisciplinaire ayant dépassé ses querelles de chapelles. Mais, comme l'association entre *Information* et *Pouvoir* se fait de plus en plus présente, on peut prévoir, avec une bonne certitude, que les blocages et retards de mise place des systèmes d'information publics ou privés sont ou seront d'ordre hiérarchique, institutionnel, politique, financier plus que technique.

Cela nous interroge sur la place et le rôle du statisticien dans une démarche en pleine expansion. Le marché de l'emploi actuel montre qu'il y a de la place pour tous pour prospecter dans les entrepôts de données à la fois eldorado des informaticiens et pays de cocagne des statisticiens. Comme le soulignent Friedman (1997) et Hand (1999), les statisticiens ont tout intérêt à investir ce domaine ou tout du moins à ne pas le négliger. Sur le plan académique, il serait vain et stérile de vouloir opposer les deux disciplines qui apparaissent clairement comme complémentaires. Certes la Statistique a ses us et coutumes mais ils peuvent rapidement s'adapter à condition de ne pas enfermer les statisticiens dans un certain conservatisme, celui des revues de Statistique Mathématique (qui ont leur rôle mais ne sont pas tout) ou, plus grave, celui de quelques institutions. Il serait préjudiciable de freiner ainsi la participation de jeunes chercheurs au développement de nouveaux thèmes de recherche.

Remerciements Nous sommes reconnaissants à Henri Caussinus d'avoir, dans son rôle d'Éditeur, suscité et critiqué cet article. Merci également à Antoine de Falguerolles pour les ouvertures apportées lors de nos discussions informelles.

Références

Académie des Sciences. (2000). La statistique. *Rapport sur la Science et la Technique*. Technique & Documentation.

Baum, E. and D. Haussler (1989). What size net gives valid generalization ? *Neural Computation 1*, 151–160.

Bergeret, F. and Y. Chandon (1999). Improving yield in ic manufacturing by statistical analysis of a large data base. *Micro Magazine*. www.micromagazine.com/archive/99/03/bergeret.html.

Besse, P. (2000). Statistique & data mining. www.ups-tlse.fr/Besse/enseignement.html.

Breiman, L. (1996). Bagging predictors. *Machine learning 26*(2), 123–140.

Breiman, L. (2001). Random forests random features. *Machine learning à paraître.*

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and regression trees.* Wadsworth & Brooks.

De Veaux, R., J. Schumi, J. Schweinsberg, and L. Ungar (1998). Prediction intervals for neural networks via nonlinear regression. *Technometrics 40*(4), 273–282.

Efron, B. (1983). Estimating the error rate of a prediction rule : improvement on cross-validation. *Journal of the American Statistical Association 78*, 316–331.

Elder, J. and D. Pregibon (1996). A statistical perspective on knowledge discovery in databases. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 83–113. AAAI Press/MIT Press.

Fayyad, U. M. (1997). Editorial. *Data mining and Knowledge discovery 1*, 5–10.

Friedman, J. H. (1997). Data mining and statistics. what's the connection ? In *Proc. of the 29th Symposium on the Interface : Computing Science and Statistics*.

Gardner, R., J. Bieker, S. Elwell, R. Thalman, and E. Rivera (2000). Solving tough semiconductor manufacturing problems using data mining. In *IEEE/SEMI Advanced semiconductor manufacturing conference*.

Ghattas, B. (1999). Importance des variables dans les méthodes CART. *La Revue de Modulad 24*, 17–28.

Ghattas, B. (2000). Agrégation d'arbres de classification. *Revue de Statistique Appliquée 48*(2), 85–98.

Goebel, M. and L. Gruenwald (1999). A survey of data mining and knowledge discovery software tools. In *SIGKDD Explorations*, pp. 20–33. ACM SIGKDD.

Hand, D., H. Mannila, and P. Smyth (2001). *Principles of data mining*. MIT Press.

Hand, D. J. (1998). Data mining : Statistics and more ? *The American Statistician 52*(2), 112–118.

Hand, D. J. (1999). Statistics and data mining : intersecting disciplines. In *SIGKDD Explorations*, Volume 1, pp. 16–19. ACM SIGKDD.

Hébrail, G. and Y. Lechevallier (2002). Data mining et analyse de données symboliques. In *Analyse de Données*. Hermes.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*(8), 832–844. citeseer.nj.nec.com/ho98random.html.

Hwang, J. and A. Ding (1997). Prediction intervals for artificial neural networks. *Journal of the American Statistical Association 92*, 748–757.

Jambu, M. (2000). *Introduction au data mining*. Eyrolles.

Michie, D., D. Spiegelhalter, and C. Taylor (1994). *Machine learning, neural and statistical classification*. Harwood.

Mieno, F., T. Sato, Y. Shibuya, K. Odagiri, H. Tsuda, and R. Take (1999). Yield improvement using data mining system. In *Semiconductor Manufacturing Conference Proceedings*, pp. 391 –394. IEEE.

Quinlan, J. (1993). *C4.5 – Programs for machine learning*. Morgan Kaufmann.

Raimbault, N., C. Bes, and P. Fabre (2001). Neural aircraft autopilot gain adjuster. In *15th IFAC Symposium on Automatic Control in Aerospace*.

Raimbault, N. and P. Fabre (2001). Probabilistic neural detector of pilot-induced oscillations (pios). In *AIAA Guidance, Navigation and Control conference*.

S-plus (1997). *S-plus 4 Guide to statistics*. MathSoft.

SAS (1989). *SAS/STAT User's Guide* (fourth ed.), Volume 2. Sas Institute Inc. version 6.

SEM (2001). *SAS/ Enterprise Miner User's Guide*. Sas Institute Inc. version 8.

Shlien, S. (1990). Multiple binary decision tree classifiers. *Pattern Recognition 23*, 757–763.

Tibshirani, R. (1996). A comparison of some error estimates for neural network models. *Neural Computation 8*, 152–163.

Zighed, D. A. and R. Rakotomalala (2000). *Graphes d'induction, apprentissage et data mining*. Hermes.