MASTER EN SCIENCES MATHÉMATIQUES

Création et évaluation qualitative d'un modèle de prédiction de défaillance applicable aux starters

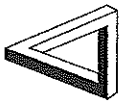
Sancinito Spito, Laurent

Award date:
2000

FUNDP
Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B-5000 Namur Belgique

Création et évaluation qualitative d'un modèle de prédiction de défaillance applicable aux starters



Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

Promoteur : J.-P. Rasson

Laurent SANCINITO SPITO

Année Académique 1999-2000

Au terme de ce mémoire, je tiens à remercier toutes les personnes qui m'ont soutenu dans les moments difficiles.

Tout d'abord, je remercie mon promoteur Mr Jean-Paul Rasson pour la confiance qu'il a placée en moi tout au long de cette année ainsi que pour ses encouragements lorsque j'en avais besoin.

Ensuite, je tiens à remercier Mr Philippe Poulain et Melle Stéphanie Authier, de Fortis Banque, pour leur grande disponibilité, leurs conseils et leurs éclaircissements.

Un tout grand merci à messieurs Jean-Yves Pirçon et Vincent Bertholet pour leur énorme disponibilité, leurs conseils judicieux et leur constante bonne humeur.

Enfin, j'adresse mes remerciements à mes parents pour leur soutien continuel tout au long de ces quatre ans, ainsi qu'à mes frères et mes amis qui ont toujours été là quand il le fallait.

Résumé

Le mémoire est centré sur la notion d'entreprise starter. Les starters sont des jeunes sociétés, qui ont moins de cinq années d'activité. L'objet de ce travail est de créer un modèle de prédiction de défaillance destiné aux starters, dans le but de mieux gérer les risques liés à l'activité commerciale d'un organisme bancaire. Il sera également question de l'apport des variables issues des réponses à un questionnaire d'évaluation qualitative dans ce type de modèle. L'utilisation des techniques statistiques multivariées est devenue fréquente dans le domaine de la gestion du risque de crédit, et pour parvenir à nos fins, nous comparons différentes méthodes statistiques, paramétriques ou non paramétriques, telles que la méthode des noyaux, des k plus proches voisins et des arbres de partitionnement. L'implémentation informatique de ces méthodes représente une part considérable de notre travail. A travers l'utilisation de ces techniques, nous tentons également de mettre en évidence les facteurs de risque qui peuvent mener un starter à la faillite.

Abstract

This thesis is based on the notion of starters. By starters we mean companies that are less than five years old. The main aim of this work is to create a failure prediction model for them which will be useful in the credit risk management sector. We will also talk about the contribution data which come from a qualitative evaluation form can bring to that kind of model. Nowadays, the use of multivariate statistical techniques is common in that particular field, so we will compare different types of methods, parametric or not, such as the kernel method, k-nearest-neighbours or decision trees. The programming of those techniques has been a large part of our labour. Through the use of statistics, we will try to point out the risk factors that can lead a starter to bankruptcy.

Table des matières

Introduction

Le mémoire a été réalisé en étroite collaboration avec Fortis Banque, et plus particulièrement avec le département de la gestion du risque de crédit. Les organismes bancaires sont à la recherche d'outils qui permettent d'évaluer les risques qu'ils encourent lorsqu'ils confient une somme d'argent à une entreprise. Au sein de ce département, une équipe travaille à l'élaboration de ces différents outils et propose à l'ensemble du réseau des solutions adaptées à leurs besoins. A ce jour, Fortis Banque ne dispose pas de modèle de prédiction de défaillance développé spécifiquement pour les starters, c'est-à-dire pour les entreprises débutantes qui ont moins de cinq ans d'existence. Cette absence constitue un manque cruel, puisque les starters sont particulièrement fragiles. L'objet de ce travail est de créer un modèle de prédiction de défaillance destiné à ces entreprises, dans le but de mieux gérer les risques liés à l'activité commerciale d'un organisme bancaire.

Notre démarche sera d'utiliser des méthodes statistiques paramétriques ou non paramétriques d'analyse discriminante dans le but de construire un modèle qui soit capable de retrouver les éléments annonciateurs de la défaillance dans l'analyse de la santé financière d'un starter et dans l'évaluation qualitative de celui-ci. Les données sur lesquelles nous travaillons ont la particularité d'avoir deux classes très peu distinctes. Cela est notamment dû au biais de la base, causé par le fait que tous les individus de celle-ci se sont vus accorder un crédit au terme d'une analyse préalable du risque qu'ils présentaient. Par conséquent, il n'y a pas dans notre base d'individu particulièrement mauvais.

Nous commencerons par présenter le cadre bancaire dans lequel nous avons travaillé, pour ensuite détailler les techniques statistiques que nous avons utilisées. La troisième partie sera dédiée à la présentation et à l'analyse des différents résultats. Enfin, nous présenterons les conclusions générales de notre travail, en essayant de proposer une méthode optimale.

Première partie

Le cadre bancaire

Chapitre 1

Notion d'entreprise et d'entreprise défaillante

1.1 Définition d'entreprise

En théorie économique, la notion d'entreprise correspond à une unité économique qui combine des facteurs de production pour obtenir des biens et des services destinés au marché. En contrepartie de l'utilisation des facteurs de production, elle distribue des revenus.

Nous pouvons illustrer le concept d'entreprise par le schéma suivant :

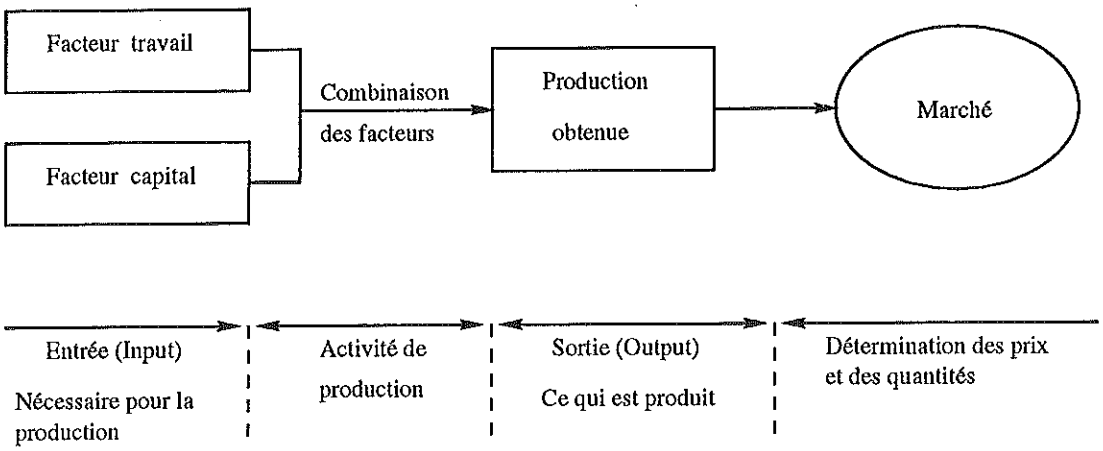


FIG. 1.1: *L'entreprise*

(*Source* : Dictionnaire économique et social, [14])

Remarquons que cette définition ne comporte aucune restriction en terme de taille ou de personnel employé. Elle englobe donc aussi bien les grosses sociétés que les simples indépendants.

1.2 Les starters

Nous appelons **starters** ou **entreprises débutantes** les entreprises ayant moins de cinq années de durée de vie. Nous montrerons dans le deuxième chapitre de cette partie consacrée à l'aspect bancaire de notre travail que les jeunes sociétés sont aussi généralement les plus fragiles, et que leur comportement financier est singulier. En effet, dans le cadre de sa thèse doctorale [7], D. Van Caillie indique que "au cours de ses premières années d'activité, une entreprise affiche un comportement financier non stabilisé, fluctuant d'une année sur l'autre : au cours de cette période, les entreprises sont à la recherche d'un équilibre, tant au niveau de leur taille qu'au niveau de leur structure de financement."

Le comportement atypique des starters a poussé Fortis Banque à créer un modèle de prédiction de faillite qui leur est spécifique, et dont l'évaluation qualitative est l'objet de ce travail.

1.3 Notion d'entreprise défaillante

Dans [2], ch 16, p. 365, Ooghe et Van Wymeersch distinguent une définition économique d'une définition juridique de la notion d'entreprise en difficulté.

1.3.1 Définition économique

"Une entreprise en difficulté peut être définie comme une entreprise qui ne parvient pas à réaliser de manière continue ses objectifs économiques (i.e. maximisation de la valeur de l'entreprise aux actionnaires) compte tenu des contraintes sociales et d'environnement (emploi, fiscalité, bien être des intervenants,...)"

La réalisation continue des objectifs économiques de l'entreprise suppose une *rentabilité* et une *liquidité* suffisantes.

Nous pouvons schématiser la succession des événements financiers qui peuvent mener un entreprise à la défaillance par le "failure path" ou "enchaînement économique menant à la discontinuité financière" présenté à la figure 1.2.

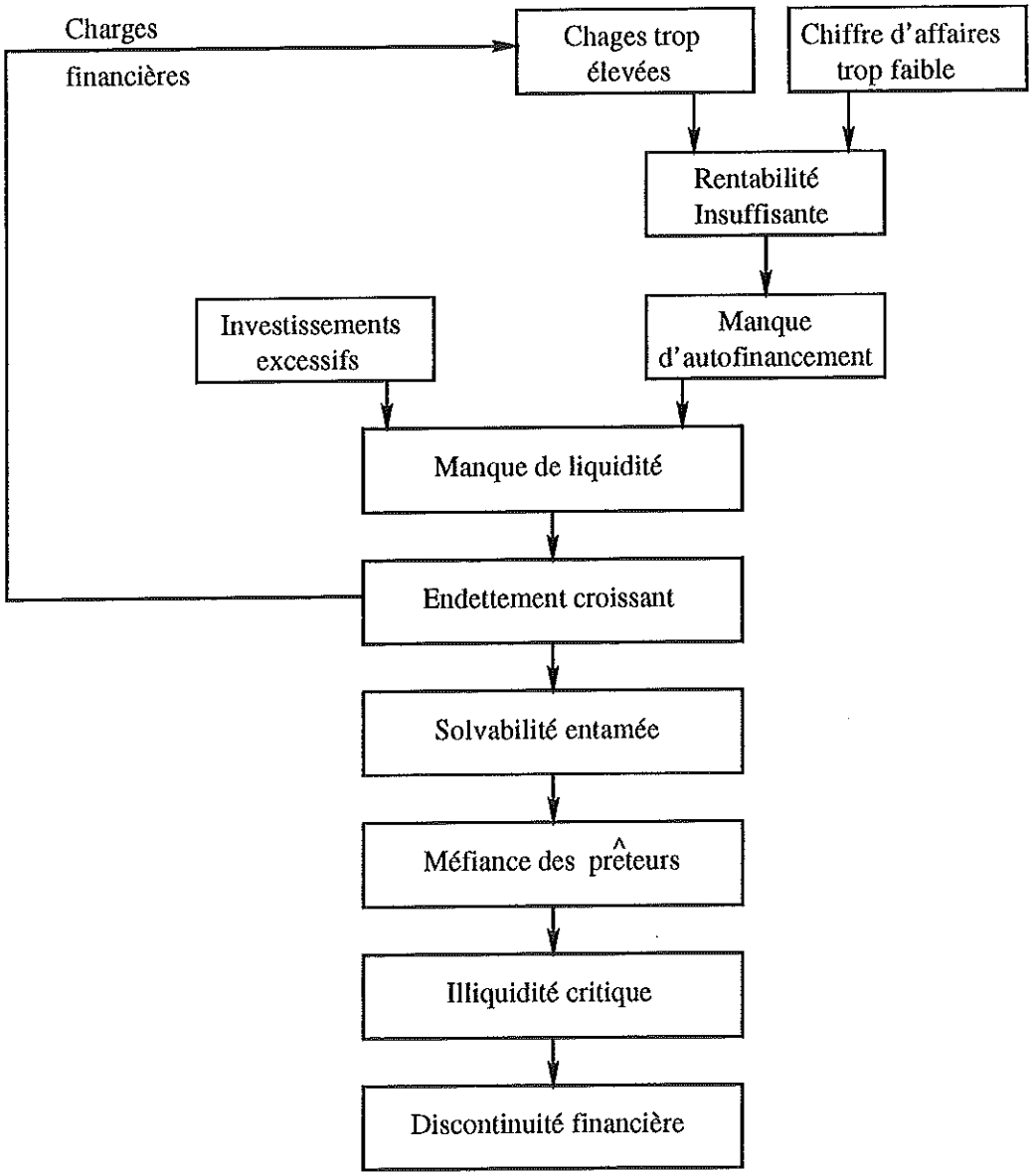


FIG. 1.2: *Enchaînement économique menant à la discontinuité financière*

(*Source* : Traité d'analyse financière, p. 368, [2])

1.3.2 Définition juridique

La définition économique que nous venons de donner est difficilement traduisible en termes statistiques. C'est pourquoi l'analyse statistique de la santé financière d'une entreprise se base sur la définition juridique de la faillite.

Les entreprises en difficulté sont classées juridiquement en deux catégories : les entreprises déclarées en état de faillite et celles bénéficiant d'un concordat.

En Belgique, la loi du 18 avril 1851 précise que "tout commerçant qui cesse ses paiements et dont le crédit se trouve ébranlé est en état de faillite" .

La faillite d'une entreprise est déclarée par le jugement du Tribunal de Commerce dans le ressort duquel se trouve le siège du principal établissement de l'entreprise.

Chapitre 2

Les risques bancaires

2.1 Introduction

L'octroi de crédit aux entreprises constitue l'une des activités principales d'un organisme bancaire. Cependant, la rentabilité de ces transactions n'est nullement assurée : en cas de faillite de l'entreprise, le créditeur risque de perdre une partie du montant exposé. La perte réelle dans un tel cas varie en fonction du pourcentage déjà remboursé et des garanties définies au moment de l'octroi, mais elle est toujours conséquente par rapport au bénéfice qu'aurait généré le prêt s'il avait été remboursé.

L'importance du phénomène de la faillite en Belgique a poussé les banques à s'investir dans le développement d'outils destinés à mieux gérer les risques inhérents à ce type d'activité. Il en va de la survie de la banque elle-même.

L'objectif de ces recherches est de proposer un système d'aide à la décision en attribuant à chaque entreprise une **probabilité de défaillance**, aussi nommée **rating**. Ainsi, une entreprise présentant une probabilité de défaillance élevée pourrait voir sa demande de fonds refusée si le banquier estime le risque trop important. Ce cas de figure n'est pas automatique, car la banque doit également tenir compte de son objectif principal : le profit. Elle peut donc décider de confier du crédit à une entreprise à risque si la rentabilité en cas de remboursement complet le justifie. De plus, la banque privilégie l'erreur de crédit — octroyer du crédit à une entreprise qui fera faillite avant le remboursement complet du montant, aussi appelée *erreur de type I* — au détriment de l'erreur commerciale (*erreur de type II*). En effet, refuser l'octroi de crédit à une entreprise saine aura pour conséquence le départ du client vers une autre banque. C'est ce type de mauvais jugement que Fortis Banque veut éviter avant tout.

Les modèles développés ne tiennent pas compte de ces aspects commerciaux : ils permettent au banquier d'estimer le risque qu'une entreprise tombe en faillite dans l'année afin de prendre une décision en toute connaissance de cause.

2.2 Importance et caractéristiques du phénomène de faillite en Belgique.

Les chiffres présentés dans cette section proviennent de l'Institut National de Statistiques [18], ainsi que de l'ouvrage [2] et des articles [16, 17].

Si nous examinons la situation de la Belgique par rapport aux autres pays de l'Union Européenne, nous constatons que comparativement à sa taille, notre pays présente un taux élevé de faillites. Sur le premier trimestre de 1999, 3.870 entreprises belges ont été déclarées en état de faillite alors qu'aux Pays-Bas, par exemple, ce nombre n'atteignait que 1.694 faillites. Au Portugal et en Espagne, seules 140 et 560 entreprises respectivement ont dû mettre fin à leurs activités. Comparativement à la taille du pays, l'Italie présente un taux de faillite plus faible que la Belgique, puisqu'avec 7.400 faillites début 1999, elle n'a subi que deux fois plus de faillites pour une superficie presque neuf fois plus grande. Les nations qui souffrent le plus des faillites sont la France (21.589), la Grande-Bretagne (20.178), l'Allemagne (13.872) et la Belgique.

La figure 2.2 illustre l'évolution du nombre de faillites en Belgique de 1950 à nos jours. Ces dernières années, le nombre annuel de faillites est reparti à la hausse après s'être stabilisé au début de la décennie à un niveau pourtant déjà préoccupant. L'explosion du nombre de nouvelles entreprises qui s'est marquée à partir de 1987 est une hypothèse avancée pour expliquer partiellement cette croissance. A titre d'exemple, il s'est créé dans notre pays 19.502 nouvelles sociétés en 1998, et un peu plus de 20.000 en 1999.

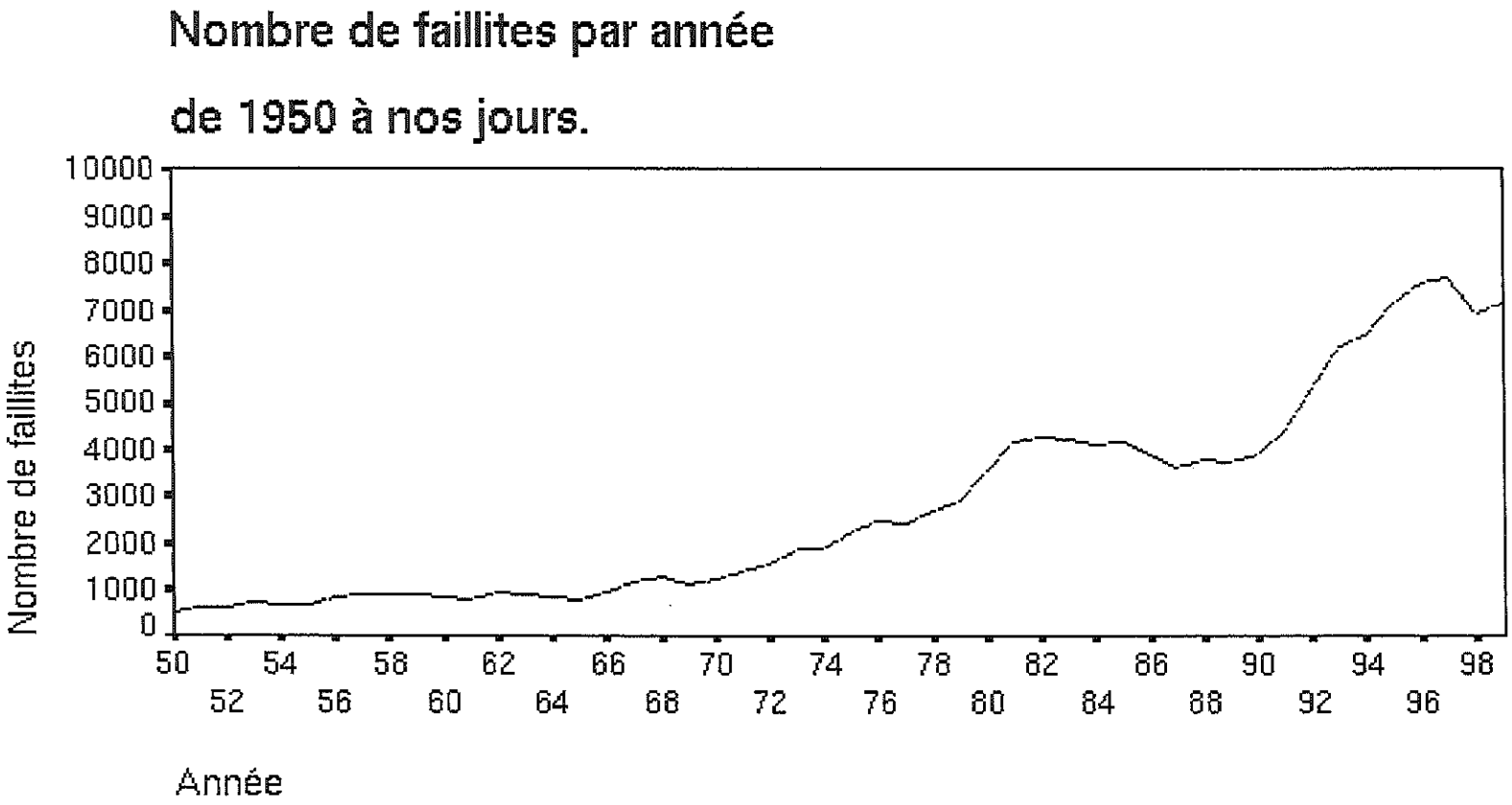


Fig. 2.2 Evolution du nombre de faillites en Belgique

Source: Institut National de Statistiques

Si nous examinons plus en détail les statistiques belges, nous constatons que ce sont les petites et moyennes entreprises qui souffrent du phénomène de faillite. Au mois de novembre 1999, pas moins de 85,7% des faillites déclarées concernaient des entreprises de moins de 5 personnes. En mars 2000, ce nombre s'élevait à 87%.

Parmi les petites entreprises, ce sont en majorité les starters qui se montrent les plus fragiles : sur la période 1978-1988, un tiers des faillites concernaient des entreprises en activité depuis moins de 4 ans. Au mois de décembre 1999, 30% des entreprises faillies étaient des starters, de même que 34,7% des faillites de février 2000.

Nous le constatons, le phénomène de faillite en Belgique est un problème relativement sérieux, qui touche plus particulièrement les très petites entreprises et les starters. Il est donc primordial pour un organisme bancaire tel que Fortis Banque de pouvoir évaluer le risque présenté par une telle entreprise. Par ailleurs, il pourrait être intéressant de distinguer les types d'activité des entreprises fragiles, afin de créer des modèles qui tiennent compte du risque lié au domaine spécifique d'une société.

Chapitre 3

La gestion du risque de crédit chez Fortis Banque

3.1 Le découpage du monde des entreprises

Il est évident que le risque encouru par un organisme créditeur varie en fonction de l'entreprise bénéficiaire. Appliquer un modèle calibré pour les sociétés cotées en bourse à une petite entreprise familiale n'a aucun sens, c'est pourquoi Fortis a découpé le monde des entrepises en 4 segments principaux qui sont :

1. **M.P.R.** (Marché Professionnel & Retail)
   Ce sont les entreprises ou indépendants dont le total actif n'excède pas les 10 millions de francs.

2. **P.M.E.** (Petites et Moyennes Entrprises)
   Les sociétés tombant dans ce segment sont celles dont le total actif varie entre 10 et 800 millions de francs.

3. **Midcap** (Middle Capitalization Enterprises)
   Ce sont les entreprises non cotées en bourse et dont le total actif dépasse les 800 millions de francs.

4. **Les sociétés cotées en bourse**

3.2 Les modèles déjà existants

Pour évaluer les risques présentés par une entreprise, Fortis Banque utilise des outils développés en interne de manière spécifique pour chaque segment décrit ci-dessus, à l'aide de logiciels statistiques tels que SAS, à l'exception du modèle destiné aux sociétés cotées en bourse, Credit Monitor de KMV Corporation, d'origine externe.

Les modèles destinés à chaque type d'entreprise sont très différenciés, ne fut-ce que par le type de données qu'ils exploitent. Si le modèle destiné aux MPR renferme essentiellement des variables qualitatives (réponse à des questionnaires, évaluations faites par les commerciaux de Fortis,...), celui destiné aux PME utilise en majorité des données bilantaires, c'est-à-dire quantitatives.

Le but de ces modèles est d'offrir une classification des entreprises selon le risque qu'elles présentent, classification par couleur allant du bleu au rouge sur une échelle nommée **masterscale**. Cette échelle de risque a été créée pour des raisons de clarté d'explication et d'utilisation au sein du réseau. Comme nous l'avons précisé précédemment, la décision d'accorder ou non du crédit ne se fait pas uniquement en fonction du risque. Le montant mis en jeu rentre fortement en compte, c'est pourquoi la banque essaie d'estimer la perte attendue (*Expected Loss, EL*). Celle-ci dépend du taux de défaillance présumé de l'entreprise (*Expected Default Frequency, EDF*), du montant exposé et des sûretés (*Loss Given Default, LGD*). La formule de la perte attendue est donnée par

$$EL = EDF * EAD * LGD$$

Notons que cette notion ne coïncide pas avec la perte réelle subie lors d'une faillite.

Bien que la panoplie des modèles existants semble couvrir l'ensemble des quatre types différents de société, les travailleurs de l'unité de recherche sur la gestion des risques de crédit de Fortis Banque ont éliminé toutes les données relatives à des entreprises starters lors de l'élaboration des modèles destinés aux deux premières catégories. Ils ont en effet estimé que le comportement singulier des entreprises débutantes pourrait perturber la performance de leurs modèles, et ont décidé de leur dédier un modèle propre.

Par ailleurs, quand il en existe, les données disponibles sur un starter sont presque toujours qualitatives. Il s'agit généralement de réponses à des questionnaires soumis lors de la création de l'entreprise. En ce qui concerne les indépendants, aucune donnée bilantaire ne viendra s'ajouter, puisque ceux-ci ne publient pas de bilan. Quand il s'agit d'une société, certaines données bilantaires sont accessibles après la première année d'existence, évidemment. La difficulté liée à

l'obtention de données fiables et exploitables dans leurs autres modèles a également poussé Fortis à ne pas insérer les starters dans la création de ceux-ci.

Dans ce mémoire, nous avons tenté de construire un modèle destiné à évaluer la probabilité de défaillance dans l'année d'une entreprise débutante au moyen de techniques statistiques non paramétriques, et en particulier à l'aide de la méthode des noyaux.

Chapitre 4

Les données de Fortis Banque concernant les starters

L'étape de collection des données fut certainement une des étapes les plus difficiles lors de la réalisation de ce travail. Les renseignements sur les starters ne sont pas légions, c'est pourquoi nous avons dû travailler avec une base de donnée relativement limitée.

Dans un premier temps, nous avons extrait un certain nombre de ratios financiers du CD-Rom **Bel-first** n^{0}17 d'Août 1999, édité par le bureau Van Dijk. Cet ouvrage disponible à la vente permet de consulter certaines informations publiées par les entreprises belges et luxembourgeoises tels que les bilans et les ratios financiers. Il rassemble des données en provenance de la Banque Nationale de Belgique qui concernent plus de 250 000 entreprises de notre pays ainsi que les 200 plus grandes entreprises du Luxembourg, et offre un accès par plusieurs critères de recherche.

Nous nous sommes intéressés à des entreprises ayant publié un bilan en 1996 et ayant été créées entre 1993 et 1996, de sorte que notre échantillon ne contienne que des starters. Parmi les entreprises sélectionnées, nous avons séparé celles ayant été déclarées en faillite au cours de l'année suivante de celles qui avaient publié un bilan en 1997, c'est-à-dire les entreprises encore vivantes un an plus tard. A ce stade, notre échantillon était composé de 6485 entreprises, dont 5000 étaient encore en activité en 1997 et 1485 avaient fait faillite à cette date.

Les ratios que nous avons choisis sont répartis en différentes catégories :

1. Les ratios principaux.

   - Ratios d'exploitation

- Ratios de rentabilité
- Ratios de structure
- Ratios d'investissement

2. Les ratios dits d'Ooghe et Van Wymeersch.
   - Ratios de liquidité
   - Ratios de solvabilité
   - Ratios de rentabilité
   - Ratios de valeur ajoutée
3. Les ratios ONSS.

En outre, nous avons extrait certains renseignements pratiques tels que l'adresse, la forme juridique, le numéro de TVA, le code NACE-BEL (code représentant le type d'activité de l'entreprise),... qui n'ont pas nécessairement été introduits dans la construction des modèles, mais qui se sont avérés utiles.

Nous ne pouvions pas nous satisfaire de ces données exclusivement bilantaires pour construire un modèle destiné aux entreprises débutantes puisque de l'aveu même des employés de Fortis Banque, les informations disponibles sur les starters sont généralement des réponses à des questionnaires soumis à l'entrepreneur lors de la création de la société, ainsi que des évaluations réalisées par des agents commerciaux. En ce qui concerne les indépendants et les entreprises de moins d'un an, ce sont par ailleurs les seuls renseignements dont nous pouvons disposer du fait de l'absence d'un quelconque bilan. En outre, la qualité des données bilantaires est souvent faible en raison de l'existence en Belgique de deux schémas de présentation des comptes, l'un nommé *schéma complet* et l'autre, privilégié par les petites entreprises, nommé *schéma abrégé*.

Par conséquent, nous avons extrait les numéros de TVA des entreprises de notre échantillon afin de les comparer avec les dossiers de Fortis Banque. Pour des raisons internes à la banque, seules 575 entreprises étaient communes aux deux sources de données, ce qui réduisait considérablement notre base de travail.

Ces fichiers internes contiennent essentiellement les réponses à un questionnaire d'évaluation figurant en **Annexe C**, réactualisées chaque trimestre. Les autres variables renseignent des valeurs prises par certaines procédures de rating effectuées au sein de la banque, et par conséquent inintéressantes pour notre travail, sauf éventuellement à des fins de comparaison.

Par ailleurs, nous avons dû faire face à un nombre élevé de valeurs manquantes. Tout d'abord, nous avons été forcés de nous limiter aux données les plus récentes. En effet, nous disposions de cinq bilans consécutifs allant de 1993 à 1997, mais seul celui de 1996 avait été publié par l'ensemble des entreprises. Nous avons été confronté au même problème au niveau des renseignements internes à la banque, puisque seul le dernier trimestre était exploitable.

Ensuite, à cause des différences entre les deux schémas bilantaires belges, beaucoup de ratios étaient incalculables et donc absents. Nous avons décidé de garder les variables qui présentaient un taux de valeurs manquantes inférieur à 15% des 575 observations de notre échantillon. Sur les 99 variables extraites de Bel-First, une cinquantaine dépassaient ce seuil, et nous avons été contraints de les supprimer de l'étude. Nous avons alors constaté que dans la plupart des cas, les mêmes entreprises étaient à l'origine du déficit des différentes variables qui approchaient ce taux de 15%. Nous avons par conséquent supprimé ces entreprises, de manière à réduire grandement le taux de valeurs déficitaires.

Malgré nos efforts, quelques variables présentaient toujours des valeurs manquantes. Nous avons décidé de les corriger en les remplaçant par la valeur de la moyenne sur les 510 observations subsistantes. Signalons tout de même que ces corrections ne concernaient que peu de variables et peu de valeurs.

En outre, nous avons effectué une analyse univariée sur notre base de données afin de mettre en évidence d'éventuelles corrélations entre variables. De nouveau, nous avons supprimé certaines variables lors de cette étude, parce que plusieurs d'entre elles se répétaient et d'autres étaient extrêmement corrélées. Cette suppression ne s'est pas faite uniquement sur base de l'étude statistique de la corrélation, mais également en tenant compte de la signification des ratios utilisés. En effet, certains ratios mesurent sensiblement la même chose, comme par exemple les *fonds de roulement nets* et les *besoins en fonds de roulement nets*. Mais généralement une impression était confirmée par la statistique, et nous avons toujours pris une décision sur base des deux aspects.

Au final, notre base de données est composée de 510 entreprises (113 faillites et 397 non-faillites) sur lesquelles sont mesurées 31 variables : 19 ratios, 11 réponses au questionnaire et la forme juridique de l'entreprise. La liste des ratios que nous avons pu exploiter figure en **Annexe B**.

Afin de tester le modèle dans une dernière étape de validation, nous avons séparé aléatoirement cette base de donnée en une base d'entraînement de 400 individus (répartis en 309 entreprises saines et 91 faillites) et en une base test de 110 entreprises (22 faillites et 88 en activité).

Deuxième partie

Le cadre statistique

Chapitre 5

La classification et l'analyse discriminante

5.1 La classification

Le principe de la classification est de retrouver dans une population donnée des groupes d'individus homogènes, appelés **classes**, de telle manière que les individus d'un même groupe ont des caractéristiques communes, qui s'opposent à celles des individus d'un autre groupe. Par exemple, lors des études de marchés, la classification est utilisée pour mettre en évidence des groupes de consommateurs aux comportements spécifiques, dans le but de mieux cibler la publicité.

Pour ce faire, les observations sont décrites par un ensemble de caractéristiques, ou autrement dit de variables, grâce auxquelles il est possible d'établir une distance entre individus.

Il existe deux grandes catégories de méthodes de classification. D'une part, il y a celles qui considèrent au départ que l'échantillon total forme une grande classe et qui essaient de diviser cette classe en sous-classes. Ces méthodes sont dites *divisives* ou *hiérarchiques descendantes*. L'autre catégorie est celle des méthodes dites *agglomératives* ou *hiérarchiques ascendantes*, qui considèrent que chaque individu est une classe à lui seul, et qui tentent de regrouper des classes semblables.

A partir des années 70, le développement des techniques de classification a été considérable grâce à l'arrivée des ordinateurs car les problèmes sont souvent de grande taille. Les capacités de stockage et la puissance de calcul des ordinateurs ont permis de traiter de tels problèmes en des temps acceptables.

Une des principales difficultés en classification est de déterminer le nombre de classes naturelles présentes dans l'échantillon.

5.2 L'analyse discriminante

En analyse discriminante, le problème est sensiblement différent puisque nous connaissons le nombre de classes, disons g, ainsi que les individus qui les composent. Il s'agit ici de déterminer la classe d'un point additionnel, non présent lors de la constitution des classes. Notre problème est donc un problème d'analyse discriminante, puisque nous disposons d'un fichier décrivant les starters et leur classe (faillite ou non), et nous devons retrouver la classe de chaque entreprise.

Nous supposons que tout individu de la $i^{\text{ème}}$ classe ($i = 1, 2$) est la réalisation d'un vecteur aléatoire de $\mathbb{R}^d$ de densité $f_i(x)$, $i = 1, 2$. Nous considérerons tour à tour chaque entreprise $\mathbf{x}$ de notre échantillon et tenterons de retrouver sa classe.

La règle d'affectation que nous utiliserons dans notre problème est dérivée de la **règle d'affectation bayesienne**, qui minimise le risque bayesien et qui s'énonce comme suit :

$$\textit{Classer } \mathbf{x} \textit{ dans la classe 1 si}$$
$$p_1 f_1(\mathbf{x}) \geq p_2 f_2(\mathbf{x})$$

où p_1 et p_2 sont respectivement les probabilités a priori qu'une entreprise appartienne à la classe 1 et 2. Nous supposerons l'équiprobabilité des classes, c'est-à-dire que nous utiliserons **la règle d'affectation du maximum de vraisemblance** :

$$\textit{Classer } \mathbf{x} \textit{ dans la classe 1 si}$$
$$f_1(\mathbf{x}) \geq f_2(\mathbf{x})$$

Nous pourrons donc résoudre le problème qui nous est posé si nous connaissons les densités de chacune des deux classes de notre échantillon. Or nous les ignorons. Nous devrons par conséquent estimer ces fonctions. Il existe deux types de méthodes d'estimation de densité : les méthodes *paramétriques* et les méthodes *non paramétriques*.

Chapitre 6

Les méthodes paramétriques

Les méthodes paramétriques supposent généralement que les densités sont normales multivariées, c'est-à-dire

$$f(\mathbf{x}) = \frac{1}{\det \mathrm{S}^{\mathrm{d}/2}} \exp\left(-\frac{(\mathbf{x}-\mu)^T S^{-1} (\mathbf{x}-\mu)}{2}\right)$$

où $\mu = \mathrm{E}[\mathrm{x}]$ est l'espérance du vecteur aléatoire ;
$S = \mathrm{E}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T]$ est sa matrice de dispersion.

Les deux paramètres (μ, S) sont inconnus et donc remplacés par leurs estimateurs du maximum de vraisemblance $(\hat{\mu}, \hat{S})$.

6.1 L'analyse discriminante linéaire de Fisher

Supposons que les matrices de dispersion sont égales. La règle de Fisher vise à déterminer l'hyperplan qui sépare le mieux les deux populations, c'est-à-dire que la surface de décision entre les classes 1 et 2 est linéaire et égale à

$$w^t x = w_0$$

Or, discriminer au moyen d'un hyperplan revient à discriminer les projections des échantillons sur une droite normale à cet hyperplan. Fisher a proposé la droite pour laquelle, projetée sur cette droite, la distance entre les moyennes pondérée par la déviation standard de l'échantillon, est maximale. Cette maximisation fournit l'hyperplan défini par

$$\begin{aligned} w &= \hat{S}^{-1}(\bar{x}_1 - \bar{x}_2) \\ w_0 &= \log(p_2) - \log(p_1) + \frac{\bar{x}_2^T \hat{S}^{-1} \bar{x}_2 - \bar{x}_1^T \hat{S}^{-1} \bar{x}_1}{2} \end{aligned}$$

où p_1 et p_2 sont les probabilités a priori d'appartenance à une classe.

6.2 La régression logistique

Soient

- Y, une variable telle que

$$Y = \begin{cases} 0 & \text{si l'observation appartient à la classe 1} \\ 1 & \text{si l'observation appartient à la classe 2} \end{cases}$$

- $\mathbf{x}$ le vecteur représentant une entreprise
- β une constante à estimer
- α le vecteur des coefficients d'une combinaison linéaire à estimer
- μ une perturbation supposée de moyenne nulle et de variance 1, suivant une loi logistique dont la fonction de répartition est donnée par

$$\mathrm{F}(x) = \frac{1}{1 + e^{-x}}$$

En outre, supposons que

$$Y = \begin{cases} 0 & \text{si } \beta + \alpha\mathbf{x} + \mu \leq 0 \text{ ou encore } \mu \leq -\beta - \alpha\mathbf{x} \\ 1 & \text{si } \beta + \alpha\mathbf{x} + \mu > 0 \text{ ou encore } \mu > -\beta - \alpha\mathbf{x} \end{cases}$$

Le modèle de la régression logistique est

$$\text{logit } p = \ln\left(\frac{p}{1-p}\right)$$

où $p = \mathrm{P}(Y = 1|\ \mathbf{x})$.

Nous avons donc

$$\begin{aligned} p &= \mathrm{P}(Y = 1|\ \mathbf{x}) \\ &= \mathrm{P}(\mu > -\beta - \alpha\mathbf{x}) \\ &= 1 - \mathrm{P}(\mu \leq -\beta - \alpha\mathbf{x}) \\ &= 1 - \mathrm{F}(-\beta - \alpha\mathbf{x}) \\ &= \frac{1}{1 + e^{\beta + \alpha\mathbf{x}}} \end{aligned}$$

Nous obtenons que le modèle s'écrit

$$\ln\left(\frac{p}{1-p}\right) = \beta + \alpha\mathbf{x}$$

Le calcul des paramètres α et β se fait en maximisant la vraisemblance. En général, ces valeurs se trouvent par un processus itératif.

Chapitre 7

Les méthodes non paramétriques

7.1 Introduction

L'utilisation de méthodes non paramétriques permet de mettre en évidence certains aspects dans la structure d'un ensemble de données que les méthodes paramétriques occultent généralement. Or, ces aspects peuvent se révéler d'une importance capitale, notamment dans le domaine de l'économie.

La différence fondamentale entre ces deux types de méthodes réside dans le fait que les méthodes paramétriques forcent l'estimateur à appartenir à une certaine classe de fonctions. Il faut donc choisir a priori une forme spécifique pour la fonction de densité, forme qui peut être parfois trop rigide, voire inadaptée au problème. Les méthodes non paramétriques permettent de se passer d'une expression analytique imposée à la fonction de densité, ce qui a pour effet de mieux "capter" les spécificités ou les petites variations présentes dans les données.

Soit un échantillon aléatoire univarié de taille n, c'est-à-dire $X_1, X_2, \ldots, X_n$, n variables aléatoires de densité commune f. Notre but est d'estimer f au moyen de méthodes non paramétriques.

7.2 Les histogrammes

La méthode non paramétrique la plus simple et la plus utilisée est l'histogramme. Il est construit en divisant la droite réelle en intervalles de tailles égales. L'estimation de la fonction de densité est *une fonction en escalier dont la valeur sur chaque intervalle est égale à la proportion d'observations qui y est contenue divisée par la longueur de l'intervalle* :

$$\hat{f}_H(x;b) = \frac{\sharp \textit{ d'observations dans l'intervalle comprenant } x}{nb}$$

où $\hat{f}_H(x;b)$ désigne la valeur de l'estimateur au point x pour des intervalles de longueur b ; et n désigne la taille de l'échantillon.

Le paramètre b est appelé **paramètre de lissage**. Sa valeur influe sur le résultat obtenu. Le terme de *paramètre de lissage* signifie que si b est trop petit, l'estimateur accordera trop d'importance aux petites variations de densité (on parle de *sous-lissage*), tandis que s'il est trop grand, il y accordera trop peu d'importance (on parle de *sur-lissage* dans ce cas). Il est donc nécessaire de trouver un juste milieu qui rende compte des spécificités de l'ensemble de données.

L'effet du paramètre de lissage est illustré par la figure 7.1. Cette dernière montre également que le résultat obtenu dépend significativement du choix des frontières des intervalles.

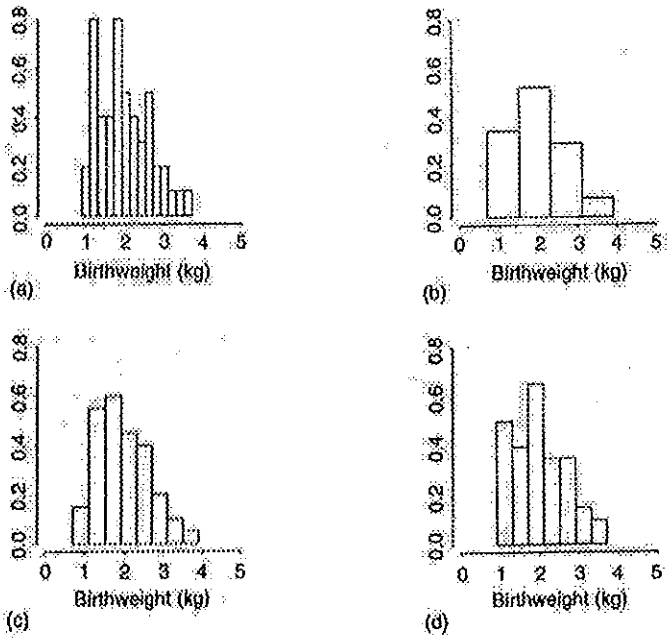


FIG. 7.1: *Les histogrammes (a) et (b) ont été construits à partir des mêmes données, mais leur paramètre de lissage vaut respectivement 0.2 et 0.8. Les graphes (c) et (d) utilisent le même paramètre de lissage mais le premier intervalle débute en 0.7 pour l'un, et en 0.9 pour l'autre.*

(*Source* : Kernel Smoothing, [1])

La diversité des résultats obtenus en fonction des choix réalisés fait de l'histogramme un estimateur trop peu précis.

7.3 La méthode des noyaux

7.3.1 Estimation de densité par des noyaux univariés

L'estimateur noyau univarié

L'estimateur noyau univarié est donné par

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

où K est une fonction satisfaisant $\int_{\mathbb{R}} K(x)\,dx = 1$ appelée **noyau** ;
h est appelé **largeur de fenêtre.**

Généralement, K est choisi comme étant une fonction de densité de probabilités symétrique autour de 0, ce qui assure que $\hat{f}(x;h)$ est aussi une densité. Cependant, nous pourrions utiliser des noyaux qui ne sont pas des densités.

En posant $K_h(u) = h^{-1}K(u/h)$, l'estimateur noyau devient

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{7.1}$$

La formule (7.1) indique que l'estimateur noyau est construit en centrant un noyau sur chaque observation. Sa valeur en un point x est la moyenne des valeurs des n noyaux transformés en ce point. De cette manière, les régions où se trouvent beaucoup d'observations apportent une large contribution à l'estimateur noyau, à l'inverse de celles contenant peu d'observations, exactement comme nous l'attendons d'une fonction de densité.

Exemple

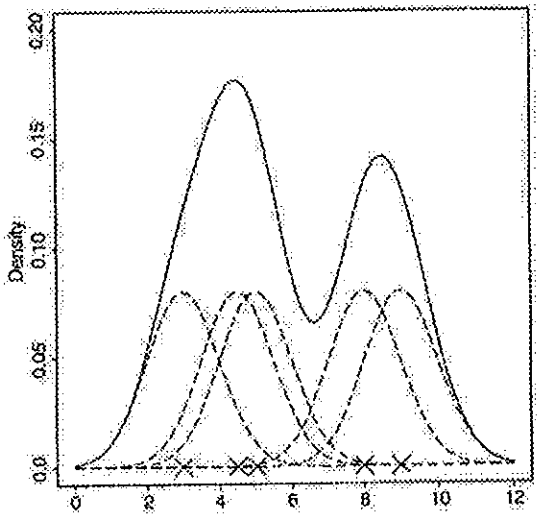


FIG. 7.2: *Estimation de densité par le noyau normal, basée sur 5 observations*

(*Source* : Kernel Smoothing, [1])

Cette figure montre l'estimation de densité obtenue sur base d'un échantillon de 5 individus en utilisant un noyau normal $N(0,1)$. Dans ce cas, K_h est en fait une $N(0,h^2)$.

La largeur de fenêtre h fonctionne comme un paramètre de lissage avec les risques de sur-lissage ou de sous-lissage que cela comporte (fig 7.3), c'est pourquoi nous allons, dans la section suivante, déterminer la façon dont l'erreur dépend de ce paramètre. Nous pourrons ainsi mettre en évidence un choix optimal pour h.

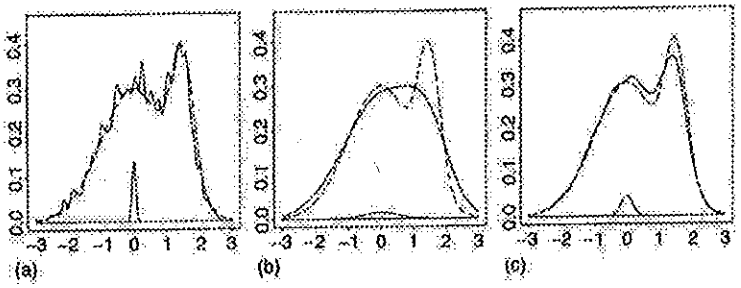


FIG. 7.3: *Effet de sur-lissage et de sous-lissage en fonction de la largeur de fenêtre : (a) h=0.06, (b) h=0.54 et (c) h=0.18*

(*Source* : Kernel Smoothing, [1])

Les critères d'erreur : MSE et MISE

Pour déterminer l'efficacité de l'estimateur noyau, il nous faudrait pouvoir mesurer l'erreur commise lors de l'estimation de la densité $f(x)$. Nous allons pour cela utiliser un critère d'erreur largement répandu en statistiques pour mesurer l'écart entre un estimateur $\hat{\theta}$ et le paramètre θ qu'il estime : *MSE* ou *Mean Square Error*, donné par

$$\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

MSE possède la caractéristique intéressante de pouvoir se décomposer en la somme du biais de l'estimateur et de sa variance :

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

Considérons l'estimateur noyau $\hat{f}(x;h)$. En sommant sa variance et son biais, nous obtiendrons $\mathrm{MSE}\{\hat{f}(x;h)\}$, c'est-à-dire l'erreur commise lors de l'estimation de $f(x)$ en un point x fixé. Wand et Jones [1] ont montré que

$$\begin{aligned} E[\hat{f}(x;h)] - f(x) &= (K_h * f)(x) - f(x) \\ \mathrm{Var}[\hat{f}(x;h)] &= n^{-1}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\} \end{aligned}$$

Nous obtenons donc

$$\mathrm{MSE}\{\hat{f}(x;h)\} = \{(K_h * f)(x) - f(x)\}^2 + n^{-1}\{(K_h^2 * f)(x) - (K_h * f)^2(x)\} \quad (7.2)$$

Nous remarquons que l'expression de l'erreur fait intervenir la fonction de densité $f(x)$, fonction que nous désirons estimer et que nous ne connaissons donc pas.

Pour mesurer l'erreur commise non plus en un seul point mais sur toute la droite réelle, nous allons utiliser *MISE* ou *Mean Integreted Square Error* :

$$\mathrm{MISE}\{\hat{f}(.;h)\} = E\int_{\mathbb{R}} (\hat{f}(x;h) - f(x))^2\, dx$$

En changeant l'ordre d'intégration,

$$\mathrm{MISE}\{\hat{f}(.;h)\} = \int_{\mathbb{R}} E[\hat{f}(x;h) - f(x)]^2\, dx = \int_{\mathbb{R}} \mathrm{MSE}\{\hat{f}(x;h)\}\, dx$$

En intégrant (7.2), nous obtenons

$$\begin{aligned} \mathrm{MISE}\{\hat{f}(.;h)\} \;=\; & (nh)^{-1}\int_{\mathbb{R}} K^2(x)\, dx + (1 - n^{-1})\int_{\mathbb{R}} (K_h * f)^2(x)\, dx \\ & -2\int_{\mathbb{R}} (K_h * f)(x)\, f(x)\, dx + \int_{\mathbb{R}} f(x)^2\, dx \end{aligned}$$

Nous constatons que l'erreur dépend effectivement de h, mais cette dépendance n'est pas claire. Nous voudrions la rendre plus évidente, afin de faciliter l'interprétation de l'influence de h sur la performance de l'estimateur noyau. Nous utiliserons à cette fin les approximations asymptotiques de MSE et MISE. Introduisons d'abord deux notations :

$$R(K) = \int_{\mathbb{R}} K^2(x)\, dx$$

et

$$\mu_2(K) = \int_{\mathbb{R}} z^2 K(z)\, dz$$

Wand et Jones [1] ont montré que

$$\mathrm{MSE}\{\hat{f}(x;h)\} \approx (nh)^{-1} R(K)\, f(x) + \frac{1}{4} h^4 \mu_2(K)^2 f''(x)^2 + o\{(nh)^{-1} + h^4\}$$

et que

$$\mathrm{MISE}\{\hat{f}(.;h)\} \approx (nh)^{-1} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 R(f'') + o\{(nh)^{-1} + h^4\} \qquad (7.3)$$

Dans l'approximation asymptotique de MISE (7.3) le biais est fonction de h^4 tandis que la variance est fonction de $(nh)^{-1}$. Par conséquent, pour faire diminuer le biais, il nous faudrait choisir h petit, mais ce choix ferait augmenter la variance. Il faudra donc trouver le juste compromis, de manière à minimiser chaque composante de MISE.

Calcul du h optimal

Pour déterminer le meilleur choix pour h, il suffit de dériver l'expression (7.3) par rapport à h et d'annuler. Nous obtenons alors

$$h_{opt} = \left[\frac{R(K)}{\mu_2(K)\, R(f'')\, n}\right]^{1/5}$$

Une fois encore, il nous faut souligner que l'expression du choix optimal de h fait intervenir la fonction de densité f que nous cherchons à estimer, ce qui rend cette expression inutilisable. Cependant, nous pourrions utiliser le résultat d'une estimation préalable de f que nous saurions moins performante mais plus directe que l'estimation par les noyaux.

Noyaux optimaux et théorie du noyau optimal

Nous avons abondamment discuté le choix du paramètre h. Mais qu'en est-il de celui de la forme du noyau K ?

La plupart du temps, le choix se porte sur des fonctions de densité symétriques et unimodales telles que la loi normale, et ce pour des raisons de simplicité d'interprétation. Cette sélection repose aussi sur le fait que certains estimateurs de densité basés sur des noyaux ne satisfaisant pas ces deux conditions se sont avérés inadmissibles (Cline, 1988). Comme il existe beaucoup de noyaux qui rencontrent ces critères, il nous est nécessaire de limiter les possibilités de choix.

Pour ce faire, considérons l'approximation asymptotique de MISE, dans laquelle la valeur optimale pour h a été substituée, comme une fonction de K. Si nous minimisons cette fonction, nous obtenons un noyau optimal appelé **noyau d'Epanechnikov**. Il est de forme quadratique et vaut

$$K_e(t) = \begin{cases} \frac{3}{4\sqrt{5}}(1 - \frac{1}{5}t^2) & \text{si } -\sqrt{5} \leq t \leq \sqrt{5} \\ 0 & \text{sinon} \end{cases}$$

Définissons **l'efficacité** d'un noyau par

$$\text{eff}(K) = \{C(K_e)/C(K)\}^{5/4}$$

où

$$C(K) = \left\{\int_{\mathbb{R}} t^2 K(t)dt\right\}^{2/5} \left\{\int_{\mathbb{R}} K(t)^2 dt\right\}^{4/5}$$

Les noyaux les plus répandus tels que le noyau normal par exemple ont une efficacité proche du maximum de 1 atteint par le noyau d'Epanechnikov. Même le noyau uniforme a une efficacité de plus de 0.9.

7.3.2 Estimation de densité par des noyaux multivariés

L'estimateur noyau multivarié est une généralisation directe du cas univarié.

Soient $X_1, \ldots, X_n$ un échantillon aléatoire "d-varié" de taille n. Nous avons donc que $X_i = (X_{i1}, \ldots, X_{id})^T$.

Nous noterons $\int \cdots \int_{\mathbb{R}^d}$ par $\int$ et $dx_1 \ldots dx_d$ par $d\mathbf{x}$.

L'estimateur noyau multivarié

Dans sa forme la plus générale, l'estimateur noyau multivarié de dimension d est

$$\hat{f}(\mathbf{x}, H) = n^{-1} \sum_{i=1}^{n} K_H(\mathbf{x} - \mathbf{x}_i)$$

où H est une matrice symétrique définie positive de dimension $d \times d$ appelée **matrice de largeur de fenêtre**,

K est une fonction **noyau** de dimension d telle que $\int K(\mathbf{x})d\mathbf{x} = 1$,

$K_H(\mathbf{x}) = |H|^{-1/2}K(H^{-1/2}\mathbf{x})$.

La fonction noyau est souvent choisie parmi les fonctions de densité de probabilités à d dimensions.

Il y a plusieurs manières de générer un noyau multivarié à partir d'un noyau symétrique univarié κ. Nous nous limiterons à la technique que nous avons effectivement implémentée dans les programmes informatiques que nous avons utilisés, à savoir le *produit de noyaux*. La fonction noyau de dimension d construite à partir des noyaux univariés κ_i par le produit de noyau est définie comme suit :

$$K(\mathbf{x}) = \prod_{i=1}^{d} \kappa_i(\mathbf{x}_i)$$

Par exemple, le noyau produit construit à partir du noyau normal univarié (i.e. $\kappa_i = N(\mu, \sigma^2)\ \forall i$) n'est autre que la densité normale d-variée $K(\mathbf{x}) = (2\pi)^{-d/2}\exp(-\frac{1}{2}\mathbf{x}^T\mathbf{x})$. Ce noyau est souvent utilisé dans la pratique, et dans ce cas, $K_H(\mathbf{x} - \mathbf{x}_i)$ est le vecteur gaussien $N(\mathbf{x}_i, H)$.

En général, H contient $\frac{1}{2}d(d+1)$ variables indépendantes ce qui, même pour un d raisonnable, est un nombre élevé de paramètres de lissage à choisir. On peut y préférer la simplification suivante : choisir H dans l'ensemble $\mathcal{S}$ défini par

$$\mathcal{S} = \{A \mid A = h^2 I\,, h > 0\}$$

$\mathcal{S}$ est un sous ensemble de l'ensemble des matrices symétriques définies positives. Dans ce cas, l'estimateur noyau multivarié devient

$$\begin{aligned} \hat{f}(\mathbf{x}, h) &= \frac{1}{nh^d}\sum_{i=1}^{n} K\left\{\frac{(\mathbf{x} - X_i)}{h}\right\} \\ &= \frac{1}{nh^d}\sum_{i=1}^{n}\prod_{j=1}^{d} \kappa_j\left\{\frac{(\mathbf{x}_j - X_{ij})}{h}\right\} \end{aligned}$$

7.4 La méthode des k plus proches voisins.

Le principe des plus proches voisins est extrêmement simple : si $P(b|\mathbf{x})$ désigne la probabilité qu'une entreprise caractérisée par le vecteur $\mathbf{x}$ a de faire faillite dans l'année (c'est-à-dire que nous notons b pour "bad", et g pour "good"), alors l'estimation par les k plus proches voisins de $P(b|\mathbf{x})$ est k_b/k, où k_b est le nombre d'entreprises ayant fait faillite parmi les k entreprises les plus semblables à $\mathbf{x}$, c'est-à-dire les plus proches de $\mathbf{x}$ au sens d'une distance que nous aurons définie. De manière analogue, l'estimateur de $P(g|\mathbf{x})$ est k_g/k. Il convient alors de classer $\mathbf{x}$ dans la classe des entreprises faillites si

$$\frac{k_b}{k} > \frac{k_g}{k}$$

Dans le cas contraire, nous reclasserons $\mathbf{x}$ dans la classe des entreprises saines.

Le cas pathologique où $k_b = k_g$ peut être évité en privilégiant les valeurs impaires du paramètre de lissage k, mais nous avons préféré tester également les valeurs paires afin d'être complets. Notre décision a été de ne pas classer un tel point, bien que dans la pratique ce cas ne se soit jamais présenté.

7.5 Les arbres de partitionnement

7.5.1 Introduction

Dans cette section, nous changeons radicalement d'optique, puisque les arbres de partitionnement ne sont pas une méthode d'estimation de densité. Le problème est toujours de déterminer une règle pour discriminer entre deux classes.

Pour y parvenir, la méthode construit des coupures binaires qui définissent des partitions emboîtées. Les sous-ensembles qui ne sont pas subdivisés sont appelés noeuds terminaux ou feuilles. Le but est de recueillir dans chaque noeud terminal la population la plus homogène possible quant à son appartenance à un des groupes. C'est-à-dire que :

- au mieux nous souhaitons que le noeud terminal contienne des éléments d'un seul des groupes,
- à défaut qu'il contienne un mélange des 2 classes, mais où une des classes domine nettement.

Le noeud terminal est alors affecté à ce groupe. L'ensemble de ces affectations constitue la **règle de décision**. Plusieurs noeuds terminaux peuvent être affectés au même groupe.

La construction de l'arbre s'effectue par

- la détermination des coupures successives,
- à chaque noeud la décision de déclarer le noeud terminal ou de continuer les coupures,
- l'affectation de chaque noeud terminal à une classe.

Le problème est donc de déterminer à chaque noeud t

- un **critère de coupure** ou **une question** à laquelle nous répondons par "oui" ou "non" :
  - la réponse "oui" détermine la branche gauche de l'arbre aboutissant au noeud t_g
  - la réponse "non" détermine la branche droite de l'arbre aboutissant au noeud t_d
- une **mesure de qualité de la coupure** afin de subdiviser au mieux à chaque pas
- un **critère d'arrêt** de la construction de l'arbre
- une **règle d'affectation** de chaque noeud terminal.

Exemple

L'exemple suivant suppose que notre population est divisée en trois groupes de 100 observations chacun. Nous ne montrons que la première itération de l'algorithme.

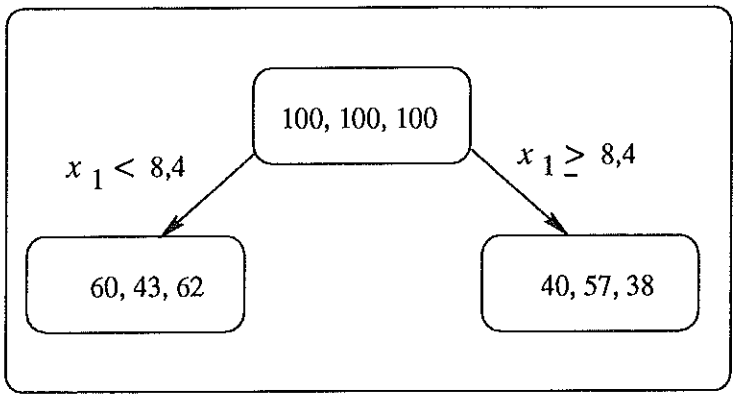


FIG. 7.4: *Dans une première étape, la méthode détermine le meilleur seuil pour chaque variable.*

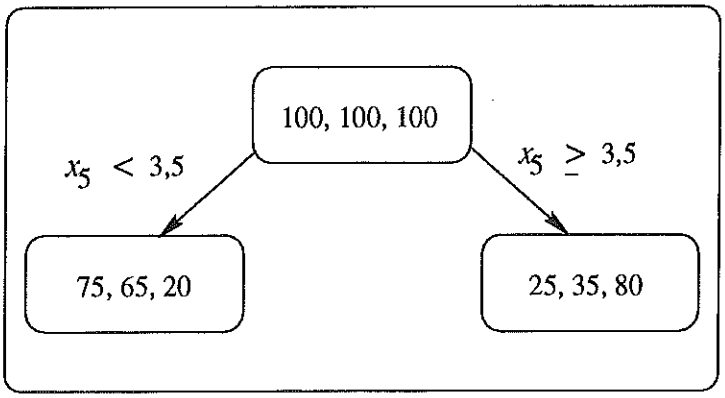


FIG. 7.5: *Ensuite, parmi toutes les coupures possibles, elle retient la meilleure en terme de séparation entre groupes.*

(*Source* : Les arbres de partitionnement, [12])

Il suffit alors de répéter l'itération sur les deux noeuds fils jusqu'à ce qu'un critère d'arrêt soit vérifié.

Nous allons à présent détailler les différentes étapes de l'algorithme.

7.5.2 Définition de la coupure

La coupure correspond à différents types de questions, selon le type des variables de notre base de données :

- La division d'un noeud à l'aide d'une variable **continue** x_j est définie par le choix d'un seuil c. La question est de la forme : "Avons-nous $x_j \leq c$? ". Si la réponse est oui, l'observation est mise dans le noeud à gauche t_g, sinon elle est mise dans le noeud à droite t_d.
- Si la variable est **qualitative ordonnée** à m modalités, il y a $m-1$ seuils possibles.
- Si la variable est **qualitative non ordonnée** à m modalités, il y a 2^{m-1} seuils possibles.

7.5.3 L'impureté

Pour mesurer la qualité d'une coupure au noeud t, puis le pouvoir discriminant de l'arbre, nous allons définir l'impureté qui caractérisera le degré de mélange dans un noeud, ou dans les noeuds terminaux d'un arbre.

Notons $n_k(t)$ le nombre d'observations provenant de la classe k et se trouvant dans le noeud t, $n(t)$ le nombre total d'individus du noeud t, et N le nombre total d'observations.

L'**impureté** d'un noeud t est une fonction $i(t)$ des probabilités d'appartenir au groupe k sachant que l'observation se trouve dans le noeud t :

$$i(t) = \phi(p(1|t), p(2|t), \ldots, p(J|t))$$

où J est le nombre de classes ;

$p(k|t) = \dfrac{n_k(t)}{n(t)}$.

Dans notre cas, $J = 2$.

L'impureté doit satisfaire aux propriétés suivantes :

- être une fonction symétrique des $p(k|t)$ pour $k = 1, \ldots, J$
- être minimum si le noeud est pur, i.e. si
  $(p(1|t), p(2|t), \ldots, p(J|t)) = (1, 0, \ldots, 0)$ ou $(0, 1, 0, \ldots, 0) \ldots$ ou $(0, \ldots, 0, 1)$
- être maximum si le mélange est parfait, i.e. si
  $$(p(1|t), p(2|t), \ldots, p(J|t)) = (1/J, 1/J, \ldots, 1/J)$$
- être une fonction concave afin que quelque soit le noeud t et la question Q, la diminution d'impureté soit toujours positive ou nulle. Nous aurons la nullité si $\forall k \quad p(k|t) = p(k|t_g)$ ou $p(k|t_d)$.

La division d'un noeud est opérée de manière à maximiser la diminution d'impureté, notée Δi :

$$\Delta i = i(t) - \left[\frac{n(t_g)}{n(t)} i(t_g) + \frac{n(t_d)}{n(t)} i(t_d)\right]$$

L'impureté totale de l'arbre T est définie par

$$I(t) = \sum_{t \in T} i(t) p(t)$$

où $p(t)$ est la probabilité qu'une observation se trouve dans le noeud t,
i.e. $p(t) = \dfrac{n(t)}{N}$.

Nous donnons en exemple deux fonctions d'impureté souvent rencontrées. La première est nommée **indice de diversité de Gini** et vaut

$$i(t) = \sum_{r=1}^{J} \sum_{\substack{s=1 \\ s \neq r}}^{J} p(r|t)\, p(s|t).$$

La seconde est l'**entropie de Shannon** :

$$i(t) = -\sum_{r=1}^{J} \sum_{\substack{s=1 \\ s \neq r}}^{J} p(r|t)\, \log p(s|t).$$

7.5.4 Règle de décision

Le noeud t est affecté à la classe j $\iff$ $p(j|t) \geq p(k|t)$, $\forall k \neq j$

Une façon de mesurer la qualité de la règle de décision est d'évaluer le taux apparent de mauvais classements du noeud t affecté à la classe j :

$$r(t) = \sum_{k \neq j} p(k|t) = 1 - \max_k p(k|t).$$

Le taux apparent de mauvais classements total est $R(t) = \sum_{t \in T} r(t) p(t)$

7.5.5 Arrêt de l'arbre

Au noeud t, nous ne poursuivons pas les coupures

- si le noeud est pur,

- si l'impureté est descendue en dessous d'un seuil : $i(t) < i_0$,

- si la variation d'impureté est trop faible : $\max_{s \in \mathcal{S}} \Delta I(s,t) < \beta$ où s désigne une coupure et $\mathcal{S}$ l'ensemble de toutes les coupures possibles,

- si le nombre d'individus arrivés au noeud t est inférieur à un seuil : $n(t) < n_0$.

Lorsque plus aucune coupure n'est possible, nous obtenons l'arbre maximal.

7.5.6 Elaguage

Généralement, le taux de bons classements diminue significativement si nous appliquons à un ensemble test l'arbre construit à partir d'un échantillon d'apprentissage. Cela signifie que les ramifications sont trop nombreuses et trop dépendantes de l'échantillon pour que l'arbre de partitionnement soit robuste. Pour éviter ce désagrément, nous aurons recours à l'élaguage de l'arbre.

Pour élaguer l'arbre, plusieurs méthodes sont envisageables :

1. **Le taux de bons classements.**
   Le sous-arbre élagué est celui qui maximise le taux de bons classements sur l'échantillon test parmi tous les sous-arbres possibles.
2. **Critère sur la qualité des noeuds descendants du noeud t.**
   Nous élaguons en dessous de t si $\Pi(t)$ est trop faible :

$$\Pi(t) = \frac{n_{mc}(t) - n_{mc}(\text{descendants } t)}{n(t)\,[n(\text{descendants } t) - 1]}, \quad \text{où}$$

   $n_{mc}(t)$ est le nombre de mauvais classements au noeud t
   $n_{mc}(\text{descendants } t)$ est le nombre de mauvais classements parmi les noeuds descendants du noeud t.
3. **Critère de réduction de complexité.**
   Pour α donné, nous choisissons l'arbre élagué T_α qui minimise sur un échantillon test le risque moyen $R_\alpha(T)$.

$$R_\alpha(T) = R(T) + \alpha \,.\, (\text{nombre de noeuds terminaux})$$

   où α est un coefficient de pénalité de la complexité.
   Ensuite, parmi les arbres choisis pour différentes valeurs de α, nous sélectionnons le meilleur.

Chapitre 8

L'implémentation informatique des méthodes

En ce qui concerne les méthodes paramétriques, nous avons utilisé les procédures DISCRIM et LOGISTIC insérées dans le logiciel SAS.

Les autres méthodes ont été implémentées par nos soins, souvent sur base de programmes existants que nous avons modifié pour les adapter à notre problème. Ce fut le cas pour les algorithmes des k plus proches voisins (avec et sans la procédure de sélection de variables), ainsi que pour certaines implémentations de la méthode des noyaux (noyaux simples et produit de noyaux uniformes).

Nous avons par ailleurs construit de nouveaux programmes pour implémenter le produit de noyaux autres que le noyau uniforme. En effet, le programme existant était inutilisable sur les autres noyaux à cause de la particularité du noyau uniforme d'être construit en comptant le nombre de points tombant dans un intervalle. Nous sommes donc repartis de la formule du produit de noyaux

$$\hat{f}(\mathbf{x}, h) = \frac{1}{nh^d} \sum_{i=1}^{n} \prod_{j=1}^{d} \kappa_j \left\{ \frac{(\mathbf{x}_j - X_{ij})}{h} \right\}$$

pour combiner un noyau uniforme sur les composantes qualitatives et un autre noyau — normal, d'Epanechnikov, de Cauchy ou de Hilbert — sur les composantes quantitatives.

L'obstacle principal que nous devions surmonter pour traiter simultanément les variables continues et les variables qualitatives était l'énorme différence de variation entre ces deux types de données. Les variables qualitatives de notre base sont de deux types : à quatre ou à cinq modalités. Les variables continues s'étendent dans des intervalles beaucoup plus larges. Nous avons par conséquent procédé à la normalisation des variables continues. Nous avions le choix entre deux

types de normalisations. La première était de normaliser dans l'intervalle [0,255] (le choix de cet intervalle particulier s'explique par le fait que les programmes originaux ont été conçus pour la reconnaissance d'images satellites). La seconde était de normaliser de sorte que toutes les variables aient la même moyenne et la même variance. Nous avons testé ces deux normalisations et il s'est avéré que les résultats étaient insensibles à la normalisation utilisée. Par contre, nous avons observé qu'une normalisation des données était bel et bien nécessaire, puisque les performances diminuaient si nous la supprimions. Nous avons choisi de garder la normalisation dans l'intervalle [0,255] par défaut, même si nous aurions pu faire l'inverse.

Nous avons, dans tous nos programmes, réalisé les estimations en **leaving-one-out**, c'est-à-dire que le point que nous reclassons ne participe pas à l'élaboration des estimateurs de densité des classes, au contraire de la **resubstitution**. Cette dernière présente généralement des taux de bon reclassement meilleurs que le leaving-one-out, il est d'ailleurs possible d'obtenir 100% de bon reclassement avec la méthode des noyaux, mais elle ne représente pas la situation réelle d'application du modèle lorsque celui-ci est calibré, et l'utiliser sur une entreprise extérieure à notre base de donnée pourrait fortement décevoir.

Enfin, signalons que la forte disproportion entre les classes de notre échantillon — la classe 1 représente 77% de l'ensemble des entreprises — nous a forcés à utiliser des coûts inversément proportionnels pour calculer le nombre de points mal classés. En effet, dans la méthode des noyaux, nous tentons de trouver une valeur pour le paramètre de lissage qui maximise le taux global de bon reclassement. Supposons qu'une certaine valeur de h favorise le reclassement des points dans la classe 1, alors le taux global de bon reclassement pourrait avoisiner les 77%, ce qui est un bon taux, mais la méthode ne retrouverait quasi aucune entreprise défaillante. Pour éviter ce désagrément, nous donnons plus de poids aux points mal classés provenant de la classe 2, et nous retenons les valeurs de h qui présentent un bon taux de reclassement global tout en reclassant suffisamment de points dans chacune des classes.

Troisième partie

Présentation des résultats

Chapitre 9

Choix des variables

Afin de déterminer les variables discriminantes pour notre problème, nous avons utilisé différents critères qui ont mené à différents choix de variables.

Dans ce qui suit, les variables dont le nom est du type "Axyz "(où xyz est un nombre à trois chiffres) sont des ratios financiers, et donc des variables continues ou encore quantitatives.

Les variables contenant les réponses au questionnaire de Fortis Banque, qui commencent par SQEDQ, sont des variables qualitatives à 5 modalités, en raison de la nature du questionnaire qui se trouve en annexe. Donc, pour ces variables, les valeurs peuvent être résumées comme suit

1 = Très bon

2 = Moyen

3 = Faible

4 = Non significatif

5 = Je ne sais pas

Enfin, la variable HH12 est une variable qualitative à 4 modalités qui renseigne sur la forme juridique de l'entreprise. Les différentes possibilités sont

1 = S.A.

2 = S.P.R.L.

3 = S.C.

4 = S.C.S./S.N.C.

Tout d'abord, nous avons utilisé les procédures d'analyse discriminante linéaire de Fisher et de régression logistique programmées dans le logiciel SAS. La sélection de variables a été effectuée en stepwise, forward et backward en ayant fixé le niveau d'entrée ou de sortie à 0.05. La plupart des variables choisies par les

six procédures que nous avons exécutées au total étaient identiques. Nous avons dès lors décidé de former une première base de données regroupant toutes ces variables. Le tableau 9.1 montre la composition de cette première base, que nous appellerons base SAS.

Base SAS	
A506	Valeur ajoutée / Immobilisation corporelles brutes
A513	Rentabilité brute de l'actif total avant impôts et ch. fin.
A523	Capitaux propres /Total du passif
A601	Fonds de roulement nets
A650	Rotation des imm. d'exploit. dans la valeur de la prod.
SQEDQ1	Réponse question 1 trimestre actuel -1
SQEDQ4	Réponse question 4 trimestre actuel -1
SQEDQ10	Réponse question 10 trimestre actuel -1

FIG. 9.1: *Base de données retenue par les deux procédures de SAS*

Le second critère que nous avons mis en oeuvre est celui des k plus proches voisins. La première étape consiste à déterminer la valeur de k qui offre des performances optimales, c'est-à-dire un taux de bon reclassement élevé et bien réparti au niveau du classement dans chacune des classes. Nous avons testé des valeurs allant de 1 à 30 et avons retenu deux possibilités : $k = 9$ et $k = 22$. Nous avons alors lancé un programme de sélection de variables qui choisit à chaque étape la variable qui maximise le taux de bon classement en minimisant la disparité entre les classes. Dans les résultats obtenus, nous avons recherché le taux de reclassement maximum et avons extrait les variables choisies. Nous obtenions deux nouvelles bases de données, une pour chaque valeur de k, dont une ne contenait que 4 variables : 2 ratios et 2 variables qualitatives (tableau 9.3). Lors d'une discussion avec Melle Authier dans les locaux de Fortis, nous avons décidé d'abandonner cette dernière en raison du trop faible nombre de variables. En effet, si plusieurs d'entre elles sont manquantes, il est impossible d'appliquer le modèle. La base de donnée obtenue avec $k = 9$ est décrite dans le tableau 9.2.

Nous devons ici faire une remarque qui aura son importance lors de l'analyse des résultats. En réalité, la variable A523 (Capitaux propres / Total du passif) n'a pas été choisie par le critère des k plus proches voisins, mais nous avons estimé bon de l'ajouter aux variables sélectionnées en raison de sa signification économique. En effet, lors d'une entrevue chez Fortis Banque, nous avons appris que ce ratio est souvent révélateur de la santé financière de l'entreprise, et avons

décidé de l'inclure dans l'analyse.

Base K9	
HH12	Forme juridique
A508	Amort., reduction valeur, prov. risque ch. / Valeur ajoutée
A523	Capitaux propres / Total du passif
A603	Trésorerie nette
A631	Rentabilité des actifs d'exploitation après amortissement
A650	Rotation des imm. d'exploit. dans la valeur de la prod.
SQEDQ1	Réponse question 1 trimestre actuel -1
SQEDQ2	Réponse question 2 trimestre actuel -1
SQEDQ3	Réponse question 3 trimestre actuel -1
SQEDQ4	Réponse question 4 trimestre actuel -1
SQEDQ7	Réponse question 7 trimestre actuel -1

FIG. 9.2: *Base de donnée retenue par le critère des k plus proches voisins,* $k = 9$

Base K22	
A603	Trésorerie nette
A623	Couverture fonds tiers par cash flow avant distribution
SQEDQ4	Réponse question 4 trimestre actuel -1
SQEDQ8	Réponse question 8 trimestre actuel -1

FIG. 9.3: *Base de donnée retenue par le critère des k plus proches voisins,* $k = 22$

Chapitre 10

Présentation des résultats

Nous ne présenterons pas dans ce chapitre l'ensemble des résultats que nous avons obtenus. En effet, à cause du grand nombre de méthodes que nous avons testées sur les différentes bases, nous n'avons pas moins d'une trentaine de résultats à analyser, et nous pensons que la clarté de la rédaction serait diminuée si nous les présentions tous ici. Nous allons donc nous limiter aux résultats qui nous permettent de tirer un certain nombre de conclusions intéressantes, et citerons les autres si nécessaire.

Pour rappel, les chiffres présentés concernent les résultats obtenus sur les bases d'entraînement SAS et K9 de 400 entreprises, réparties en 309 starters sains et 91 en faillite.

Dans un premier temps, nous tenterons de mettre en évidence le choix d'une méthode et d'une base de données de manière à proposer un modèle de prédiction de défaillance applicable aux starters. Ensuite, nous étudierons l'influence des variables qualitatives sur l'efficience de la méthode proposée.

10.1 Résultats des méthodes non paramétriques

10.1.1 Les k plus proches voisins

Nous disposions de deux versions du programme des k plus proches voisins. Une de ces versions incluait une procédure de sélection de variables en stepwise que nous avons utilisé pour construire une base de données basée sur ce critère, comme nous l'avons déjà signalé. La deuxième version permettait de prendre en compte toutes les variables d'une base de données pour construire l'estimateur. Il nous a fallu procéder à quelques modifications pour adapter ces programmes à nos données. Lors de l'étape de constitution des bases sur lesquelles nous avons travaillé, la valeur de k qui offrait les meilleures performances était $k = 9$. Nous

avons de nouveau testé plusieurs valeurs de k (de 1 à 30), et les meilleurs résultats sont apparus avec la même valeur. Nous les présentons ci-dessous :

$k = 9$	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	220	89	71,2%	70,5%	211	98	68,2%	66%
classe 2	29	62	68,1%		38	53	58,2%	

Nous pourrions être très surpris à la lecture de ces résultats, parce que le résultat de la procédure semble être moins bon sur la base qui a été construite à partir du critère des k plus proches voisins. Il nous faut alors nous rappeler que la variable A523 n'a pas été selectionnée par le critère, et qu'elle est sûrement la cause de cette baisse de performance. Pour nous en assurer, nous avons supprimé cette variable de la base K9 et avons relancé le programme. Nous obtenons alors :

$k = 9$	classe		pourcentage	pourcentage
	1	2	par classe	total
classe 1	232	77	75 %	71,5 %
classe 2	37	54	59,3%	

Comme nous pouvions nous y attendre, le résultat est meilleur sur la véritable base de variables extraites avec le critère des neuf plus proches voisins. Cependant, avec cette première méthode, le résultat obtenu sur l'ensemble des variables choisies par les procédures de SAS est relativement proche, avec en outre une meilleure répartition des points bien reclassés dans chacune des classes.

10.1.2 La méthode des noyaux

Nous avons deux possiblités d'utilisation de la méthode des noyaux selon que nous utilisons l'expression univariée de l'estimateur noyau ou son expression multivariée.

Dans le premier cas, l'estimateur s'écrit

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Nous l'appelerons **noyau simple**, puisque cette expression n'utilise qu'un noyau. Il implique la recherche d'une distance appropriée qui combine les deux types de variables, quantitatives et qualitatives. La distance que nous utilisons est celle mise en place dans le mémoire [4], à savoir

$$\forall x \neq y \quad d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{k=1}^{d} \phi(\mathbf{x}_k,\mathbf{y}_k)^2}$$

où

$$\phi(\mathbf{x}_k,\mathbf{y}_k) = \begin{cases} \frac{|\mathbf{x}_k - \mathbf{y}_k|}{\text{étendue de la } k^{\text{ème}} \text{ variable}} & \text{si } k \text{ est une variable quantitative} \\ \frac{1}{\sharp \text{modalité}} & \text{si } k \text{ est une variable qualitative.} \end{cases}$$

Cette distance permet de ne pas privilégier les variables quantitatives ayant une étendue considérable par rapport aux autres variables.

Par ailleurs, ce procédé ne requiert qu'un paramètre de lissage par fonction de densité, puisque nous centrons un noyau sur chaque point de l'espace $I\!R^d$ sans le décomposer. L'estimateur donne alors la même importance à chaque direction.

Dans le second cas, que nous appelerons **produit de noyaux**, l'estimateur noyau est

$$\hat{f}(\mathbf{x},h) = \frac{1}{nh^d}\sum_{i=1}^{n}\prod_{j=1}^{d} \kappa_j\left\{\frac{(\mathbf{x}_j - X_{ij})}{h}\right\}.$$

Le produit de noyau permet de différencier les deux types de variables en leur appliquant un noyau et une distance propres. Sur les variables quantitatives, nous avons utilisé la distance euclidienne classique, tandis que la distance sur la partie qualitative du vecteur observation compte le nombre de variables à j modalités ($j = 4$ ou 5 dans notre cas) qui diffèrent dans ce vecteur. L'inconvénient de cette méthode est le plus grand nombre de paramètres de lissage à déterminer, puisque nous traitons de manière spécifique les variables quantitatives, qualitatives à 4 modalités, et qualitatives à 5 modalités.

Les noyaux simples

Nous avons implémenté trois noyaux différents en utilisant la formule du noyau simple. Ce sont les noyaux d'Epanechnikov, de Cauchy et le noyau normal. Nous présentons ci-dessous les résultats que nous avons obtenus sur chacune des deux bases, ainsi que les valeurs des paramètres de lissage qui ont été retenues dans l'estimation de la densité de chaque classe.

Noyau d'Epanechnikov

Le noyau d'Epanechnikov s'écrit $K(\mathbf{x}) = \begin{cases} 1 - \left(\frac{d(\mathbf{x},X_j)}{h}\right)^2 & \text{si } d(\mathbf{x},X_j) \leq h \\ 0 & \text{sinon} \end{cases}$

	Base SAS				Base K9			
	classe 1	classe 2	% par classe	% total	classe 1	classe 2	% par classe	% total
classe 1	192	117	62,1%	66,5%	218	91	70,5%	68,75%
classe 2	17	74	81,3%		34	57	62,6%	

Les valeurs retenues pour le paramètre de lissage sont

- pour la base SAS : 0,40 et 0.38
- pour la base K9 : 0.66 et 0.66

Noyau normal

Le noyau normal centré en chaque observation est $K(\mathbf{x}) = \exp\left\{-\left(\frac{d(\mathbf{x},X_j)}{h}\right)^2\right\}$

	Base SAS				Base K9			
	classe 1	classe 2	% par classe	% total	classe 1	classe 2	% par classe	% total
classe 1	148	161	47,8%	57,25%	205	104	66,3%	66,5%
classe 2	10	81	89%		30	61	67%	

Les valeurs retenues pour le paramètre de lissage sont

- pour la base SAS : 0,2 et 0,18
- pour la base K9 : 0,24 et 0,24

Noyau de Cauchy

Le noyau de Cauchy est $K(\mathbf{x}) = \left[1 + \left(\frac{dist(\mathbf{x},X_j)}{h}\right)^{d+1}\right]^{-1}$

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	262	47	84,8%	75,2%	214	95	69,3%	66,75%
classe 2	52	39	42,9%		38	53	58,2%	

Les valeurs retenues pour le paramètre de lissage sont

- pour la base SAS : 0,24 et 0,24
- pour la base K9 : 0,02 et 0,02

Résumé des résultats :

Nous voyons que les résultats sont meilleurs sur la base K9 quelque soit le noyau utilisé. En effet, même si le taux global de bon reclassement obtenu sur la base SAS par le noyau simple de Cauchy semble largement meilleur, nous ne pouvons pas nous en satisfaire car notre but est de prédire au mieux la défaillance des starters, et cette méthode ne retrouve pas 50% des faillites. En fait, l'application du noyau simple sur la base SAS privilégie souvent une classe par rapport à l'autre, comme nous pouvons le voir dans les autres résultats. Nous verrons que cette tendance se réduit lors de l'utilisation du produit de noyaux.

Par ailleurs, la méthode du noyau simple déçoit légèrement au niveau de ses performances moins bonnes que celles des k plus proches voisins. Ces résultats décevants trouvent peut-être leur cause dans le fait que le noyau simple traite les variables qualitatives de la même manière que les variables quantitatives au travers d'une distance qui englobe les deux types de variables. Pour nous en assurer, nous avons appliqué le produit de noyau, qui sépare ces deux aspects des données.

Les produits de noyaux

Dans un premier temps, nous avons appliqué le programme du produit de noyaux uniformes réalisé dans le cadre du mémoire [4]. Puis, nous avons construit de nouveaux programmes qui permettent de centrer un noyau uniforme sur les variables qualitatives et un autre noyau sur les variables quantitatives. Les noyaux que nous avons testés sont les mêmes noyaux que nous avions implémentés en tant que noyaux simples, c'est-à-dire le noyau d'Epanechnikov, le noyau de Cauchy et le noyau normal. Nous avons en outre testé un autre noyau, celui de Hilbert, qui a la particularité de ne pas se servir d'un paramètre de lissage. Comme dans la section précédente, nous donnons les résultats et les vecteurs de paramètres de lissage.

Produit de noyaux uniformes

Sur les données quantitatives, le noyau prend la forme du noyau uniforme classique, c'est-à-dire

$$K(\mathbf{x}) = \begin{cases} \frac{1}{2^{d''}} & \text{si } dist(\mathbf{x}, X_j) \leq h \\ 0 & \text{sinon} \end{cases}$$

où d'' est le nombre de variables continues. Sur les données qualitatives, le noyau uniforme compte le nombre de points qui se trouvent dans une fenêtre de largeur h.

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	191	118	61,8%	65,5 %	218	91	70,5%	69,25%
classe 2	20	71	78%		32	59	65%	

Les valeurs retenues pour les paramètres de lissage sont

- pour la base SAS : 81, 78, 1, 1.
- pour la base K9 : 21, 17, 1, 0, 1, 2.

Les deux premières valeurs sont les valeurs des paramètres pour les variables continues et pour chacune des classes, les deux suivantes sont celles qui concernent les variables qualitatives à quatre modalités (elles sont absentes pour la base SAS puisque cette base ne contient pas de telle variable), et les deux dernières concernent les variables qualitatives à 5 modalités.

Produit de noyaux d'Epanechnikov et uniforme

Pour rappel, nous utilisons le noyau d'Epanechnikov sur les variables continues et le noyau uniforme sur les variables qualitatives.

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	231	78	74,8 %	71,75%	224	85	72,5%	68,75%
classe 2	35	56	61,5%		40	51	56%	

Les valeurs retenues pour les paramètres de lissage sont

- pour la base SAS : 72, 71, 3, 3.
- pour la base K9 : 44, 44, 1, 1, 5, 5.

Produit de noyaux normal et uniforme

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	236	73	76,4%	73,25%	218	91	70,5%	69,5%
classe 2	34	57	62,6%		31	60	66 %	

Les valeurs retenues pour les paramètres de lissage sont

- pour la base SAS : 26, 26, 3, 3.
- pour la base K9 : 15, 15, 1, 1, 5, 5.

Produit de noyaux de Cauchy et uniforme

	Base SAS				Base K9			
	classe 1	classe 2	% par classe	% total	classe 1	classe 2	% par classe	% total
classe 1	239	70	77,3%	73,75%	219	90	70,8%	69%
classe 2	35	56	61,5%		34	57	62,6%	

Les valeurs retenues pour les paramètres de lissage sont

- pour la base SAS : 37, 37, 3, 3.
- pour la base K9 : 19, 19, 1, 1, 5, 5.

Produit de noyaux de Hilbert et uniforme

	Base SAS				Base K9			
	classe 1	classe 2	% par classe	% total	classe 1	classe 2	% par classe	% total
classe 1	236	73	76,3 %	69,5%	221	88	71,5%	65,5%
classe 2	49	42	46,1%		50	41	45%	

Les valeurs retenues pour les paramètres de lissage sont

- pour la base SAS : 1, 3.
- pour la base K9 : 0, 1, 1, 5.

Le noyau de Hilbert est un peu particulier en ce sens qu'il tente de se passer du paramètre de lissage. L'estimateur de densité basé sur le produit de noyaux de Hilbert a la forme suivante

$$\hat{f}(\mathbf{x}) = \frac{d!}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{2 \log n |\mathbf{x}_j - X_{ij}|}.$$

où n est le nombre d'observations et $d = d' + d''$ est la dimension du vecteur d'observation. Bien sûr, comme nous ne l'utilisons que sur les variables quantitatives, l'expression de l'estimateur que nous utilisons est légèrement différente. Par ailleurs, ceci explique la présence dans nos résultats de certaines valeurs du

paramètre h, qui sont les valeurs pour le noyau uniforme appliqué aux variables qualitatives. Nous avons testé ce noyau parce que l'absence du paramètre de lissage pourrait simplifier considérablement l'utilisation de la méthode des noyaux, dont la principale difficulté est justement de déterminer les paramètres adéquats.

Résumé des résultats :

Nous pouvons affirmer que les résultats sont globalement meilleurs par rapport aux noyaux simples, surtout si nous appliquons un autre noyau que le noyau uniforme sur les variables continues. Contrairement à ce que nous attendions, c'est la base SAS qui se montre la plus performante, puisque le taux global de bon reclassement est élevé et que ce taux est assez bien réparti sur les deux classes, notamment pour le noyau normal et le noyau de Cauchy. Notons que les performances de ces deux noyaux sont extrêmement proches, qu'ils soient appliqués à la base SAS ou à la base K9. Nous remarquons aussi que dans ces deux résultats, la base K9 permet de retrouver plus de points de la classe des starters faillis que la base SAS, ce qui explique aussi ses taux globaux de reclassement moins élevés. Nous avons également testé ces différents programmes en éliminant la variable A523 de la base K9 pour nous assurer qu'elle n'est pas la cause des performances moins bonnes de cette base. Ces tests nous ont permis d'infirmer cette hypothèse, puisque les résultats ne sont pas meilleurs que ceux qui sont présentés ci-dessus.

En ce qui concerne le noyau de Hilbert, il semble moins performant que les autres noyaux. Il semblerait donc que l'absence du paramètre de lissage nuise à la qualité de la méthode. Il serait peut-être nécessaire de réaliser davantage de comparaisons pour s'en assurer, mais ce n'était notre but.

Si nous faisons abstraction de ce noyau particulier, nous remarquons également que, contrairement aux enseignements de la théorie, le noyau d'Epanechnikov ne donne pas de meilleurs résultats que les autres noyaux, à l'exception du noyau uniforme. Cette différence entre théorie et pratique s'explique par le fait que ces deux noyaux sont à support compact. Il semble donc qu'il faille privilégier les noyaux à support non compact.

10.1.3 Les arbres de partitionnement

L'algorithme des arbres de partitionnement, que nous devons à V. Bertholet, contient de manière intrinsèque une étape de sélection de variables. Nous avons donc dû laisser de côté les bases SAS et K9 pour reprendre la base d'entraînement qui contenait l'ensemble des variables exploitables (voir **Annexe B**). Lorsque nous aurons les résultats, nous dresserons la liste des variables qui ont été utilisées lors de la constitution de l'arbre (tableau 10.1).

La fonction d'impureté que nous employons n'est pas habituelle. En effet, nous supposons à chaque étape qu'une coupure est faite pour affecter un des deux noeuds fils à une classe, et l'autre noeud à la seconde classe, et ce même s'il s'avère que ces noeuds seront subdivisés à leur tour. Nous évaluons pour chaque coupure la somme des proportions de points mal classés si le fils droit est affecté à la classe 1 (et donc si le fils gauche est affecté à la classe 2), puis l'inverse. Nous sélectionnons la coupure qui minimise cette valeur sur toutes les sommes possibles.

En ce qui concerne l'élaguage, le programme utilise le critère de réduction de complexité que nous avons présenté dans la section dédiée à la théorie des arbres.

Le résultat de cette procédure est présenté dans le tableau suivant :

	classe 1	classe 2	pourcentage par classe	pourcentage total
classe 1	290	19	93%	89%
classe 2	24	67	74%	

Le résultat est très bon mais nous devons considérer le taux de bon classement comme un taux de resubstitution, puisque l'arbre est constitué à partir de toutes les observations de la base d'entraînement. Nous tenterons de tester la solidité de cette méthode en appliquant à l'arbre construit notre échantillon test.

Nous donnons dans le tableau qui suit les différentes variables qui ont été sélectionnées lors de la construction de l'arbre.

Base ARBRE	
A506	Valeur ajoutée / Immobilisation corporelles brutes
A508	Amort., reduction valeur, prov. risque ch. / Valeur ajoutée
A509	Charges financières / Valeur ajoutée
A515	Liquidité au sens large
A524	Acquisitions d'immobilisations corporelles / Valeur ajoutée
A612	Ratio de trésorerie nette
A630	Rentabilité des actifs d'exploitation avant amortissement
A639	Degré de levier financier
A647	proportion de valeur ajoutée brute affectée au résultat ajouté
A649	taux de valeur ajoutée
A650	Rotation des imm. d'exploit. dans la valeur de la prod.
SQEDQ2	Réponse question 2 trimestre actuel -1
SQEDQ3	Réponse question 3 trimestre actuel -1
SQEDQ4	Réponse question 4 trimestre actuel -1
SQEDQ5	Réponse question 5 trimestre actuel -1
SQEDQ7	Réponse question 7 trimestre actuel -1
SQEDQ8	Réponse question 8 trimestre actuel -1
SQEDQ9	Réponse question 9 trimestre actuel -1
SQEDQ10	Réponse question 10 trimestre actuel -1

FIG. 10.1: *Variables retenues lors de la constitution de l'arbre*

10.1.4 Conclusions des résultats obtenus en non paramétrique

La méthode non paramétrique la plus performante au vu de nos résultats est la méthode des arbres de partitionnement. Cependant, nous attendrons les chiffres de l'échantillon test avant d'émettre un avis définitif.

Nous noterons également le redressement de la méthode des noyaux, au travers du produit de noyaux. La meilleure performance a été obtenue en appliquant à la base SAS le noyau de Cauchy sur les variables continues et le noyau uniforme sur les variables qualitatives. Le fait que beaucoup de méthodes non paramétriques fonctionnent mieux sur cette base est quelque peu surprenant, puisque la base K9 a été constituée à l'aide du critère des k plus proches voisins, et qu'appliquer une méthode sur une base de donnée extraite à partir du même critère devrait donner de meilleurs résultats.

Enfin, nous soulignons également les résultats presque similaires du produit de noyaux Cauchy-uniforme et du produit normal-uniforme.

10.2 Résultats des méthodes paramétriques

Nous commencerons par présenter les résultats obtenus en utilisant la procédure de régression logistique incluse dans le logiciel SAS. Il ne nous a pas été possible de l'appliquer en leaving-one-out. Les résultats présentés ci-dessous sont par conséquent issus d'une procédure de resubstitution et nous devrons en tenir compte dans notre analyse.

Par contre, la routine d'analyse discriminante linéaire de Fisher qui se trouve dans ce même logiciel permettait de calculer les taux de reclassement en leaving-one-out, et nous les présentons également afin de pouvoir comparer avec les résultats des méthodes non paramétriques. Signalons que le leaving-one-out de cette procédure est moins pénalisant que dans les méthodes non paramétriques, puisqu'ici la méthode estime l'espérance de l'échantillon et sa matrice de variance-covariance. L'absence d'une observation ne modifie pas énormément ces estimations.

En resubstitution

Régression logistique

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	289	20	93%	81,8%	281	28	91%	78%
classe 2	53	38	41,7%		60	31	34%	

Analyse discriminante linéaire de Fisher

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	247	62	80%	77,75%	216	93	70%	70,25%
classe 2	27	64	70%		26	65	71,4%	

En leaving-one-out

Comme nous l'avons déjà signalé, nous ne présentons ici que les résultats de l'analyse discriminante linéaire de Fisher.

	Base SAS				Base K9			
	classe		%	%	classe		%	%
	1	2	par classe	total	1	2	par classe	total
classe 1	242	67	78,3%	75,5 %	212	97	68,6%	67,75%
classe 2	31	60	65,9 %		32	59	64,8%	

Résumé des résultats :

Les taux de bon reclassement obtenus par la régression logistique semblent très bons, mais nous émetterons deux réserves. D'une part, ces taux sont des taux obtenus en resubstitution, et nous savons que celle-ci est sur-optimiste et peu réaliste. D'autre part, si nous examinons les pourcentages de points bien classés dans chacune des classes, nous remarquons que cette méthode retrouve énormément de points de la première classe et peu de points de la deuxième. De nouveau, de tels résultats ne peuvent pas nous satisfaire, puisque notre but est de retrouver le maximum d'entreprises dans les deux classes.

Les résultats obtenus en appliquant une analyse linéaire discriminante de Fisher sont meilleurs. En effet, le taux global de bon reclassement est très bon (75,5% en leaving-one-out sur la base SAS) et la répartition de ce taux entre les deux classes est acceptable.

Les fonctions logistiques et discriminantes construites par ces deux méthodes sur la base SAS se trouvent respectivement en **Annexe D** et **Annexe E**.

10.3 Etude de la stabilité des différentes méthodes

Afin d'évaluer la stabilité des méthodes que nous avons testées, nous les avons lancées sur un échantillon test composé de 110 starters parmi lesquels 88 étaient encore en activité en 1997 et 22 avaient été déclarées en faillite.

Nous avons décidé de tester la stabilité des méthodes les plus performantes, puisque ce sont uniquement ces dernières qui pourraient être utilisées par un organisme bancaire pour gérer les risques de crédit. Par conséquent, nous avons soumis à ce test la méthode du produit de noyaux (Cauchy-uniforme), la méthode des arbres de partitionnement et l'analyse discriminante linéaire de Fisher. Pour chacune de ces méthodes, nous avons appliqué l'ensemble test en prenant les valeurs des différents paramètres telles qu'elles avaient été fixées lors de la construction des modèles au moyen de la base d'entraînement.

Les tableaux qui suivent décrivent les résultats qui résultent de ces tests :

L'analyse discriminante linéaire de Fisher

	classe 1	classe 2	pourcentage par classe	pourcentage total
classe 1	71	17	80,7%	73,6%
classe 2	12	10	45,5%	

Le produit des noyaux de Cauchy et uniforme

	classe 1	classe 2	pourcentage par classe	pourcentage total
classe 1	77	8	87,5%	82,7%
classe 2	8	14	63,7%	

Les arbres de partitionnement

	classe 1	classe 2	pourcentage par classe	pourcentage total
classe 1	70	18	79,5%	74,5%
classe 2	10	12	54,5%	

Nous devons, dans l'analyse de ces résultats, être particulièrement attentifs à la capacité de chaque méthode à retrouver des entreprises de la classe des faillites. En effet, lors de l'étape de constitution des modèles, nous avons tenté de mettre en évidence les méthodes qui parvenaient à retrouver suffisamment d'entreprises en faillite malgré la très grande proportion d'entreprises saines dans notre échantillon d'entraînement. Comme notre échantillon test est réparti sensiblement dans les mêmes proportions, nous verrons si l'aptitude à retrouver les entreprises défaillantes a été conservée.

Ce test nous montre que les bons résultats que nous avions obtenus en appliquant l'analyse discriminante linéaire de Fisher sont un peu surestimés. En effet, cette dernière est la seule méthode à ne pas retrouver 50% des starters faillis de notre échantillon test. Il semblerait donc que les méthodes non paramétriques soient plus robustes que les méthodes paramétriques.

La méthode des noyaux se montre la plus stable, puisque c'est elle qui retrouve le plus d'observations dans la classe des faillites (14/22) tout en proposant un taux très élevé de bon classements dans l'autre classe. Le résultat du produit de noyau sur l'échantillon d'entraînement était peut-être moins spectaculaire que celui des arbres de partitionnement, mais ceux-ci ont le désavantage de dépendre beaucoup trop de l'échantillon sur lequel ils sont constitués, même si l'étape de l'élaguage permet d'atténuer cette dépendance.

10.4 Evaluation qualitative des modèles

Dans cette ultime étape, nous analysons l'influence des variables qualitatives issues du questionnaire sur nos résultats. Pour savoir si elles ont été exploitées par les différentes méthodes, nous avons repris quelques uns des meilleurs résultats et nous avons supprimé ces variables des différentes bases de données. Nous avons relancé les programmes suivants : le produit de noyaux de Cauchy et uniforme, les k plus proches voisins, les arbres de partitionnement et l'analyse discriminante linéaire de Fisher.

Le produit de noyau Cauchy-uniforme a donné exactement les même résultats que sur les bases complètes (voir tableau p. 53). Ce résultat ne nous étonne pas vraiment : nous savons, par les expériences précédentes au département de mathématiques, que la méthode des noyaux exploite mal les données qualitatives. Dans nos résultats, les valeurs des paramètres de lissage relatifs à ces variables sont pratiquement toujours les valeurs maximales, ce qui indique que même les observations pour lesquelles les données qualitatives sont très éloignées d'un point donné jouent un rôle dans l'évaluation des fonctions de densité en ce point. L'idéal serait que seules les entreprises réellement proches s'influencent mutuellement. Le problème provient du fait qu'il est très difficile de trouver une distance sur les variables qualitatives qui convienne à la méthode, c'est-à-dire telle que la distance entre deux points au niveau qualitatif soit du même ordre de grandeur que la distance de ces deux points au niveau quantitatif. Il serait alors possible d'utiliser le même noyau sur les deux types de variables.
Pour vérifier qu'il ne s'agissait pas là d'un effet de hasard, nous avons également testé le produit de noyaux normal-uniforme, qui lui aussi donne les mêmes résultats.

En ce qui concerne la méthode des k plus proches voisins, les résultats sont sensiblement les mêmes avec ou sans les variables du questionnaire. Sur la base SAS, nous avons observé une très petite amélioration (+1,75%) mais ce gain est terni par une disproportion légèrement plus grande entre les classes. De nouveau, nous avons l'impression que les variables qualitatives n'ont que très peu d'influence sur les performances de la méthode, ce qui est tout de même étonnant vu que nous avons constitué la base K9, qui comprend des variables issues du questionnaire, à l'aide du critère des k plus proches voisins. Le principe de la méthode est centré sur la notion de distance, qui est donc très importante ici aussi. Pour ne pas donner trop d'importance aux grandes variations qui peuvent survenir dans les données quantitatives, nous utilisons une normalisation des données et une distance qui tient compte de l'étendue de la variable. Par conséquent la mesure de distance ne semble pas être en cause dans la faible exploitation des données qualitatives.

Parmi les différentes méthodes que nous avons testées, celle des arbres de partitionnement est celle qui devrait profiter au mieux de la présence des variables du questionnaire dans nos observations. Cette intuition s'est effectivement confirmée, puisque les résultats de notre test sont moins bons. Rappelons que la base de donnée utilisée sur le programme des arbres de partitionnement est celle constituée de l'ensemble des variables exploitables qui se trouve en **Annexe B**, puisque la méthode contient en elle-même une procédure de sélection des variables. Il nous a suffit de relancer le programme en ignorant les variables du questionnaire. Nous perdons 4% de taux global de bon classement sur la base d'entraînement, et les classes sont très disproportionnées. Par ailleurs, la méthode est devenue moins stable. Cet exemple nous montre que l'intérêt des variables qualitatives n'est réel que si la méthode permet de les exploiter. C'est un enseignement important dans le cadre des starters, puisque les données qui les concernent sont parfois exclusivement des données qualitatives.

Enfin, nous avons effectué ce même test sur l'analyse discriminante linéaire de Fisher. Nous savons que cette méthode considère que toutes les variables sont continues. Par conséquent, nous ne nous attendons pas à une baisse conséquente des résultats. Le test a montré que les résultats sont légèrement meilleurs sur les deux bases, mais ce gain est accompagné d'une perte de points de la deuxième classe dans le cas de la base SAS.

L'ensemble des résultats montre que les variables qualitatives issues du questionnaire sont utilisées à bon escient par les techniques construites pour tenir compte de telles variables. C'est le cas des arbres de partitionnement. Pour étayer un peu plus cette hypothèse, nous avons testé également la régression logistique, qui ne fait pas partie de nos meilleurs résultats pour des raisons exposées précédemment, mais qui a l'avantage de traiter de manière spécifique les variables qualitatives. Le test confirme nos pensées, puisque nous avons observé une baisse des résultats sur les deux bases de variables. En ce qui concerne les autres méthodes, si nous éliminons les variables qualitatives issues du questionnaire, soit les performances sont équivalentes, soit elles sont légèrement meilleures.

Conclusions générales

Rappelons que le but du mémoire était de proposer un modèle de prédiction de défaillance applicable aux starters, c'est-à-dire aux entreprises débutantes qui ont moins de cinq années d'activité. Pour ce faire, nous avons procédé à une comparaison d'un bon nombre de méthodes statistiques, paramétriques et non paramétriques, et avons évalué la manière dont elles exploitent les données que nous possédions ainsi que leur performance et leur stabilité.

Il est difficile de trancher catégoriquement en faveur de l'une ou l'autre des méthodes que nous avons testées. Elles ont toutes leurs avantages et leurs inconvénients. Le choix dépendra donc de l'utilisation réelle que nous désirons faire du modèle.

Nous savons que les données qu'une banque peut posséder concernant les starters sont presque exclusivement des données qualitatives, du moins dans les premières années de vie de l'entreprise. Nous avons montré que la méthode qui exploite au mieux ces variables, généralement issues de questionnaires et d'évaluations qualitatives, est la méthode des arbres de partitionnement. C'est d'ailleurs celle qui s'est montré la plus performante sur notre échantillon d'entraînement. Cependant, cette méthode a le désavantage de dépendre très fortement de l'échantillon sur lequel l'arbre est construit, et nous avons vu que lorsque nous la soumettons a un échantillon test, ces résultats sont affaiblis, surtout au niveau de la capacité à retrouver les starters en faillite. Pour résoudre ce problème, il serait intéressant de constituer une base de donnée équilibrée (50% de faillites et 50% de non faillites) car même si cette situation n'est pas la situation réelle d'application — dans la réalité, les entreprises menacées de faillites représentent un faible pourcentage du nombre de sociétés — elle pourrait permettre à la méthode de mieux retrouver les variables discriminantes et leurs seuils justes. Il y aurait alors dans l'arbre construit un plus grand nombre de noeuds terminaux assignés à la classe des faillites, et les performances pourraient se maintenir lors de l'application à un échantillon test.

Les bons résultats obtenus par l'analyse discriminante linéaire de Fisher nous ont fait un peu douter des avantages de travailler en non paramétrique. Mais lors de l'étape du test, il s'est avéré que cette performance est trop optimiste. Il semble donc que les méthodes non paramétriques sont mieux à même de retrouver les faibles variations dans un ensemble de données, ce qui convient à notre problème puisque nos bases de données sont caractérisées par de petites variations, la distinction entre les deux classes n'étant pas énorme. Cette faible séparation des classes trouve son origine dans le biais de nos données, constituées d'entreprises suivies et donc aidées par Fortis Banque. Il n'y a donc pas dans nos bases d'entreprises particulièrement mauvaises, puisqu'elles ont été éliminées avant même la collecte des données.

Si malgré l'amélioration que nous proposons la méthode des arbres de partitionnement se montre encore relativement instable, il est toujours possible d'utiliser la méthode des noyaux ou celle des k plus proches voisins. En effet, ces méthodes ont prouvé leur intérêt par leurs performances très satisfaisantes et par leur robustesse. Ces méthodes sont les plus stables, surtout la méthode du produit de noyaux où nous avons utilisé le noyau de Cauchy sur les variables quantitatives et le noyau uniforme sur les variables qualitatives. Mais la méthode des noyaux exploite mal les variables qualitatives. Pour que la méthode profite de la présence de telles variables, nous pensons qu'une solution serait de rendre toutes les variables continues. Il existe des techniques qui permettent cela, mais le désavantage est l'introduction d'information supplémentaire pas toujours cohérente dans les variables.

Si nous examinons les ratios financiers que les différentes méthodes ont choisi, nous aperçevons que les trois grandes catégories de ratios (liquidité, solvabilité et rentabilité) sont présentes quasi systématiquement dans nos bases de travail. Soulignons que toutes les méthodes ont sélectionné des variables provenant des réponses au questionnaire de Fortis Banque, ce qui montre que ce dernier est particulièrement judicieux.

Nous terminerons par une remarque au niveau des temps d'exécution des programmes informatiques. Le choix de la meilleure estimation de densité peut prendre plus de 24 heures en fonction du nombre de paramètres à tester. Cependant, cette étape n'est à faire qu'une fois et lorsque les paramètres sont fixés, il suffit ensuite d'évaluer les densités en un point pour connaître le verdict, ce qui prend seulement quelques secondes. Le programme des arbres de partitionnment que nous avons utilisé fournit les résultats en quelques minutes, ce qui constitue également un avantage.

Nous le constatons, il n'y a pas de solution unique et optimale au problème qui nous était posé. Cela dépend des objectifs que nous avons, et il faudra toujours se résoudre à un compromis entre les performances pures de la méthode, sa stabilité et sa capacité à utiliser de manière profitable les données qualitatives qui concernent les starters.

Annexes

Annexe A

Description des variables

Renseignements généraux	
STATUT	Statut de l'entreprise (faillite ou non)
HH01	Nom de l'entreprise
HH03	Adresse
HH32	Code postal
HH04	Ville
HH06	Code pays
HH09	Numero de TVA (caractère)
HH09b	Numero de TVA (numérique)
HH10	Date de création
HH12	Forme juridique
HH13	Situation juridique
HH34	Date de la situation juridique
HH11	Date de clôture des comptes
HH15	Comptes consolidés
HH16	Code Nace BEL
HH17	Nace BEL Description

Ratios financiers	
A501	Résultat net / Chiffre d'affaire
A502	Marge brute sur ventes
A503	Marge nette sur ventes
A504	Taux de valeur ajoutée
A505	Valeur ajoutée par personne occupée
A506	Valeur ajoutée / immobilisations corporelles brutes
A507	Part des frais de personnel dans la valeur ajoutée
A508	Amort., réduction valeur, provision risque ch / valeur ajoutée
A509	Charges financières / valeur ajoutée
A510	Rentabilité nette des capitaux propres après impôts
A511	Rendement des ressources durables
A512	Cash flow / Capitaux propres
A513	Rentabilité brute de l'actif total avant impôts et ch. fin
A514	Rentabilité nette de l'actif total avant impôts et ch. fin
A515	Liquidité au sens large
A516	Current ratio
A517	Liquidité au sens strict
A518	Rotation des stocks d'approv. et des marchandises
A519	Rotation stocks en cours de fabrication et des produits finis
A520	Nombre de jours de crédit clients
A521	Nombre de jours de crédit fournisseurs
A522	Dettes à plus de 1 an / Fonds propres
A523	Capitaux propres / Total du passif
A524	Acquisitions d'immobilisations corporelles / Valeur ajoutée
A525	Acquisitions d'immobilisations corporelles / Immob. Corpo ex. précédent
A601	Fonds de roulement net
A602	Besoin en fonds de roulement net
A603	Trésorerie nette
A604	Liquidité au sens large (Current ration)
A605	Liquidité au sens strict (Acid test)
A607	Rotation globale des stocks et commandes en cours d'exécution
A608	Rotation des stocks de biens acquis
A609	Rotation des stocks de fabrication et commandes en cours d'exécution
A610	Délai moyen de paiement client
A611	Délai moyen de paiement fournisseurs
A612	Ratio de trésorerie nette
A613	Degré global d'endettement
A614	Degré global d'endettement
A615	Degré global d'indépendance financière

A616	Degré global d'indépendance financière
A617	Degré d'endettement à long terme
A618	Degré d'endettement à long terme
A619	Degré d'indépendance financière à long terme
A620	Degré d'indépendance financière à long terme
A621	Degré d'autofinancement
A622	Couverture charges financières des fonds tiers, par le résultat après impôts
A623	Couverture fonds tiers par cash flow avant distribution
A624	Couverture fonds tiers long terme par cash flow avant distribution
A625	Couvertures dettes a plus d'1 an échéant dans l'an. par cash flow av. distrib.
A626	Marge brute sur ventes
A627	Marge nette sur ventes
A628	Rentabilité de l'actif total avant amortissement
A629	Rentabilité de l'actif total après amortissement
A630	Rentabilité des actifs d'exploitation avant amortissement
A631	Rentabilité des actifs d'exploitation après amortissement
A632	Rotation des actifs d'exploitation dans les ventes
A633	Rotation des immobilisations d'exploitation dans les ventes
A634	Rotation des actifs circul. d'exploitation dans les ventes
A635	Rentabilité des fonds propres avant impôts
A636	Rentabilité des fonds propres après impôts
A637	Cash flow / Fonds propres
A638	Multiplicateur avant impôts
A639	Degré de levier financier
A640	Résultat net par action
A641	Cash flow par action
A642	Dividende par action
A643	Proportion de valeur ajoutée brute affectée au personnel
A644	Proportion de v.a. brute affectée aux amortis., réduction valeur et provisions
A645	Proportion de v.a. brute affectée aux charges financières des fonds tiers
A646	Proportion de valeur ajoutée brute affectée aux charges fiscales
A647	Proportion de valeur ajoutée brute affectée au résultat ajouté
A648	Valeur ajoutée brute par personne occupée (milliers)
A649	Taux de valeur ajoutée

A650	Rotation des immobilisations d'exploitation dans la valeur de la production
A651	Immobilisation d'exploitation par personne occupée (milliers)
A652	Taux d'investissement
A653	Taux de subsidiation par les pouvoirs publics
A654	Résultat report + réserves / Total passif
A655	Dettes échues envers fisc et ONSS / Fonds tiers à court terme
A656	Valeurs disponibles / Actifs circulant restreints
A657	En cours fabr., stocks prod. finis, comm. / actifs circulants d'exploitation
A658	Dettes < 1 an envers des établissements de crédit / Fonds tiers à court terme
A659	Modéle d'analyse discriminante global

Fichier interne de Fortis Banque	
NCB	Identification intervenant
TVA	Numéro de tva de l'intervenant
STATUTIN	Statut de l'intervenant
STATUTCGH	Date changement statut intervenant
DATERATQ	Date rating qualitatif
DATEIN	Date intervention introduction - dernier changement
SQEDINPE	Intervention personne step 1
SQEDDIEN	Service de la personne step 1
TYPEINTR	Type d'intervention step 2
SQEDINP2	Intervention personne step 2 : Validation rating
SQEDDIE2	Service de la personne step 2
ANSWERNU	Valeur réponse trimestre actuel -1
RATQUAL	Valeur rating qualitatif trimestre actuel -1
CODERAT	Code rating trimestre actuel -11
SQEDOLI0	Limite inférieure algorithme scoring trimestre actuel -1
SQEDBLI0	Limite supérieure algorithme scoring trimestre actuel -1
QUESTNUM	Nombre de questionnaire trimestre actuel -1
SQEDWAN1	Valeur réponse trimestre actuel -2
SQEDWKW1	Valeur rating qualitatif trimestre actuel -2
SQEDRAT1	Code rating trimestre actuel -11
SQEDOLI1	Limite inférieure algorithme scoring trimestre actuel -2
SQEDBLI1	Limite supérieure algorithme scoring trimestre actuel -2
SQEDAAN1	Nombre de questionnaire trimestre actuel -2
SQEDWAN2	Valeur réponse trimestre actuel -3
SQEDWKW2	Valeur rating qualitatif trimestre actuel -3
SQEDRAT2	Code rating trimestre actuel -11
SQEDOLI2	Limite inférieure algorithme scoring trimestre actuel -3
SQEDBLI2	Limite supérieure algorithme scoring trimestre actuel -3
SQEDAAN2	Nombre de questionnaire trimestre actuel -3
SQEDWAN3	Valeur réponse trimestre actuel -4
SQEDWKW3	Valeur rating qualitatif trimestre actuel -4
SQEDRAT3	Code rating trimestre actuel -11
SQEDOLI3	Limite inférieure algorithme scoring trimestre actuel -4
SQEDBLI3	Limite supérieure algorithme scoring trimestre actuel -4
SQEDAAN3	Nombre de questionnaire trimestre actuel -4

SQEDWAN4	Valeur réponse trimestre actuel -5
SQEDWKW4	Valeur rating qualitatif trimestre actuel -5
SQEDRAT4	Code rating trimestre actuel -11
SQEDOLI4	Limite inférieure algorithme scoring trimestre actuel -5
SQEDBLI4	Limite supérieure algorithme scoring trimestre actuel -5
SQEDAAN4	Nombre de questionnaire trimestre actuel -5
SQEDWAN5	Valeur réponse trimestre actuel -6
SQEDWKW5	Valeur rating qualitatif trimestre actuel -6
SQEDRAT5	Code rating trimestre actuel -11
SQEDOLI5	Limite inférieure algorithme scoring trimestre actuel -6
SQEDBLI5	Limite supérieure algorithme scoring trimestre actuel -6
SQEDAAN5	Nombre de questionnaire trimestre actuel -6
SQEDWAN6	Valeur réponse trimestre actuel -7
SQEDWKW6	Valeur rating qualitatif trimestre actuel -7
SQEDRAT6	Code rating trimestre actuel -11
SQEDOLI6	Limite inférieure algorithme scoring trimestre actuel -7
SQEDBLI6	Limite supérieure algorithme scoring trimestre actuel -7
SQEDAAN6	Nombre de questionnaire trimestre actuel -7
SQEDWAN7	Valeur réponse trimestre actuel -8
SQEDWKW7	Valeur rating qualitatif trimestre actuel -8
SQEDRAT7	Code rating trimestre actuel -11
SQEDOLI7	Limite inférieure algorithme scoring trimestre actuel -8
SQEDBLI7	Limite supérieure algorithme scoring trimestre actuel -8
SQEDAAN7	Nombre de questionnaire trimestre actuel -8
SQEDWAN8	Valeur réponse trimestre actuel -9
SQEDWKW8	Valeur rating qualitatif trimestre actuel -9
SQEDRAT8	Code rating trimestre actuel -11
SQEDOLI8	Limite inférieure algorithme scoring trimestre actuel -9
SQEDBLI8	Limite supérieure algorithme scoring trimestre actuel -9
SQEDAAN8	Nombre de questionnaire trimestre actuel -9
SQEDWAN9	Valeur réponse trimestre actuel -11
SQEDWKW9	Valeur rating qualitatif trimestre actuel -10
SQEDRAT9	Code rating trimestre actuel -11
SQEDOLI9	Limite inférieure algorithme scoring trimestre actuel -10
SQEDBLI9	Limite supérieure algorithme scoring trimestre actuel -10
SQEDAAN9	Nombre de questionnaire trimestre actuel -10
SQEDWANA	Valeur réponse trimestre actuel -11
SQEDWKWA	Valeur rating qualitatif trimestre actuel -11
SQEDRATA	Code rating trimestre actuel -11
SQEDOLIA	Limite inférieure algorithme scoring trimestre actuel -11
SQEDBLIA	Limite supérieure algorithme scoring trimestre actuel -11
SQEDAANA	Nombre de questionnaire trimestre actuel -11

SQEDWANB	Valeur réponse trimestre actuel -12
SQEDWKWB	Valeur rating qualitatif trimestre actuel -12
SQEDRATB	Code rating trimestre actuel -11
SQEDOLIB	Limite inférieure algorithme scoring trimestre actuel -12
SQEDBLIB	Limite supérieure algorithme scoring trimestre actuel -12
SQEDAANB	Nombre de questionnaire trimestre actuel -12
SQEDWANC	Valeur réponse trimestre actuel -13
SQEDWKWC	Valeur rating qualitatif trimestre actuel -13
SQEDRATC	Code rating trimestre actuel -11
SQEDOLIC	Limite inférieure algorithme scoring trimestre actuel -13
SQEDBLIC	Limite supérieure algorithme scoring trimestre actuel -13
SQEDAANC	Nombre de questionnaire trimestre actuel -13
RATANNUE	Valeur rating qualitatif année actuelle -1
CODEANNU	Code rating année actuelle -1
SQED0ONL	Limite inférieure algorithme scoring année actuelle -1
SQED0BOL	Limite supérieure algorithme scoring année actuelle -1
SQED1KWR	Valeur rating qualitatif année actuelle -2
SQED1RAC	Code rating année actuelle -2
SQED1ONL	Limite inférieure algorithme scoring année actuelle -2
SQED1BOL	Limite supérieure algorithme scoring année actuelle -2
SQED2KWR	Valeur rating qualitatif année actuelle -3
SQED2RAC	Code rating année actuelle -3
SQED2ONL	Limite inférieure algorithme scoring année actuelle -3
SQED2BOL	Limite supérieure algorithme scoring année actuelle -3
SQED3KWR	Valeur rating qualitatif année actuelle -4
SQED3RAC	Code rating année actuelle -4
SQED3ONL	Limite inférieure algorithme scoring année actuelle -4
SQED3BOL	Limite supérieure algorithme scoring année actuelle -4
SQED4KWR	Valeur rating qualitatif année actuelle -5
SQED4RAC	Code rating année actuelle -5
SQED4ONL	Limite inférieure algorithme scoring année actuelle -5
SQED4BOL	Limite supérieure algorithme scoring année actuelle -5
SQED5KWR	Valeur rating qualitatif année actuelle -6
SQED5RAC	Code rating année actuelle -6
SQED5ONL	Limite inférieure algorithme scoring année actuelle -6
SQED5BOL	Limite supérieure algorithme scoring année actuelle -6
SQED6KWR	Valeur rating qualitatif année actuelle -7
SQED6RAC	Code rating année actuelle -7
SQED6ONL	Limite inférieure algorithme scoring année actuelle -7
SQED6BOL	Limite supérieure algorithme scoring année actuelle -7

SQEDQ1	Réponse question 1 trimestre actuel -1
SQEDQ2	Réponse question 2 trimestre actuel -1
SQEDQ3	Réponse question 3 trimestre actuel -1
SQEDQ4	Réponse question 4 trimestre actuel -1
SQEDQ5	Réponse question 5 trimestre actuel -1
SQEDQ6	Réponse question 6 trimestre actuel -1
SQEDQ7	Réponse question 7 trimestre actuel -1
SQEDQ8	Réponse question 8 trimestre actuel -1
SQEDQ9	Réponse question 9 trimestre actuel -1
SQEDQ10	Réponse question 10 trimestre actuel -1
SQEDQ11	Réponse question 11 trimestre actuel -1

Annexe B

Les variables exploitables

Renseignements généraux	
HH12	Forme juridique
Ratios financiers	
A506	Valeur ajoutée / immobilisations corporelles brutes
A508	Amort., réduction valeur, provision risque ch / valeur ajoutée
A509	Charges financières / valeur ajoutée
A513	Rentabilité brute de l'actif total avant impôts et ch. fin
A515	Liquidité au sens large
A523	Capitaux propres / Total du passif
A524	Acquisitions d'immobilisations corporelles / Valeur ajoutée
A601	Fonds de roulement net
A603	Trésorerie nette
A612	Ratio de trésorerie nette
A621	Degré d'autofinancement
A623	Couverture fonds tiers par cash flow avant distribution
A628	Rentabilité de l'actif total avant amortissement
A630	Rentabilité des actifs d'exploitation avant amortissement
A639	Degré de levier financier
A646	Proportion de valeur ajoutée brute affectée aux charges fiscales
A647	Proportion de valeur ajoutée brute affectée au résultat ajouté
A649	Taux de valeur ajoutée
A650	Rotation des immobilisations d'exploitation dans la valeur de la production

Fichier interne de Fortis Banque	
SQEDQ1	Réponse question 1 trimestre actuel -1
SQEDQ2	Réponse question 2 trimestre actuel -1
SQEDQ3	Réponse question 3 trimestre actuel -1
SQEDQ4	Réponse question 4 trimestre actuel -1
SQEDQ5	Réponse question 5 trimestre actuel -1
SQEDQ6	Réponse question 6 trimestre actuel -1
SQEDQ7	Réponse question 7 trimestre actuel -1
SQEDQ8	Réponse question 8 trimestre actuel -1
SQEDQ9	Réponse question 9 trimestre actuel -1
SQEDQ10	Réponse question 10 trimestre actuel -1
SQEDQ11	Réponse question 11 trimestre actuel -1

Annexe C

Le questionnaire qualitatif

1. **Le fonctionnement de l'entreprise repose-t-il sur la personnalité du dirigeant ?**
    1. Peu, relève assurée
    2. Partiellement
    3. Totalement, relève non assurée
    4. Non significatif
    5. Je ne sais pas

2. **Qualité professionnelle du dirigeant ou de l'équipe dirigeante :**
    1. Grande expérience, formation spécialisée et/ou de haut niveau
    2. Normale, possède la qualification nécessaire
    3. Faible, mauvaise, pas ou peu d'expérience ou de formation spécialisée
    4. Non significatif
    5. Je ne sais pas

3. **Nature des relations sociales au sein de l'entreprise :**
    1. Excellent climat social, personnel stable et motivé, peu ou pas de conflits
    2. Climat normal
    3. Mauvais climat, personnel peu motivé à rotation élevée, conflits sociaux fréquents, arrêts de travail, licenciements
    4. Non significatif
    5. Je ne sais pas

4. **Existence d'instruments réalistes assurant le suivi rapproché des performances de l'entreprise. Existence d'un plan réaliste dépassant l'horizon d'un an :**
    1. Instrument de suivi rapproché des performances et plan à plus d'un an
    2. Existence soit d'un instrument de suivi rapproché des performances, soit d'un plan à plus d'un an
    3. Instrument de suivi périodique de certaines des performances, absence de plan à plus d'un an
    4. Non significatif
    5. Je ne sais pas

5. **Existe-t-il une fonction distincte de contrôle ou d'audit ?**
    1. Permanente interne et externe avec pouvoir de déclencher des actions correctrices à tous les niveaux
    2. Permanente interne
    3. Non ou occasionnelle
    4. Non significatif
    5. Je ne sais pas

6. **Diversification de la clientèle :**
    1. Clientèle diversifiée appartenant à différents secteurs
    2. Clientèle diversifiée appartenant au même secteur
    3. Trois clients représentent 50% ou plus du chiffre d'affaires
    4. Non significatif
    5. Je ne sais pas

7. **Diversification des produits ou services**
    1. Grand nombre de produits à faible risque d'obsolescence ou de substitution
    2. Plusieurs produits à risque d'obsolescence moyen
    3. Un produit représente 50% ou plus du chiffre d'affaires ou plusieurs produits à risque d'obsolescence ou de substitution élevés
    4. Non significatif
    5. Je ne sais pas

8A. Entreprise Industrielle : Qualité de l'outillage, de l'appareil de production, de l'équipement matériel, de l'infrastructure :

1. Neuf, moderne, très performant, renouvelé régulièrement
2. Normalement performant, entretenu
3. Obsolescent, vétuste, dépassé, non renouvelé
4. Non significatif
5. Je ne sais pas

8B. Entreprise Commerciale ou de Service : Gestion de la clientèle et des fournisseurs :

1. Informatique de gestion très performante (facturation, paiements, gestion des stocks)
2. Gestion normale de la clientèle et des fournisseurs avec une informatique de gestion raisonnablement performante
3. Gestion de la clientèle et des fournisseurs faible avec une informatique insuffisante
4. Non significatif
5. Je ne sais pas

9. L'entreprise a-t-elle la volonté ou la capacité de réorienter son activité ou de diversifier ses produits ?

1. L'entreprise est continuellement à la recherche de nouvelles "niches", elle est capable de trouver des débouchés nouveaux et les technologies nécessaires, ainsi que de modifier fondamentalement l'outil
2. L'entreprise est capable de faire évoluer sa gamme de produits dans une mesure raisonnable, sans pour autant se remettre en cause fondamentalement
3. L'entreprise n'a ni les capacités techniques, ni les débouchés, ni l'accès aux technologies et aux fournisseurs nécessaires
4. Non significatif
5. Je ne sais pas

10. L'entreprise pratique-t-elle une politique active de couverture des risques commerciaux, financiers et physiques ?

1. Outre les risques fondamentaux, d'autres risques (sociaux, malfaçon) sont couverts. En outre, les actifs de la société sont protégés par des systèmes de sécurité et/ou gardiennage. Couverture systématique du risque de change et application systématique de clauses de révision des prix
2. Couverture suffisante des risques fondamentaux, couverture occasionnelle du risque de change. L'entreprise est attentive aux clauses de révision des prix (fournisseurs et clients)
3. Couverture faible, limitée à quelques risques fondamentaux (incendie, responsabilité civile) et/ou d'un niveau insuffisant
4. Non significatif
5. Je ne sais pas

11. Comment l'entreprise rencontre-t-elle ses obligations découlant de la législation relative à l'environnement ?

1. Entreprise sensible aux problèmes de l'environnement et menant une gestion active des risques liés à l'environnement
2. Entreprise peu sensible aux problèmes de l'environnement et qui exerce une activité dont l'impact sur l'environnment ne peut être considéré comme une menace immédiate
3. Entreprise peu sensible aux problèmes de l'environnement et dont l'activité présente une menace potentielle importante
4. Non significatif
5. Je ne sais pas

Annexe D

La fonction logistique

Cette annexe présente la fonction logistique construite par le logiciel SAS sur la base SAS.

The SAS System

The LOGISTIC Procedure

Response Profile

Ordered Value	STATUT	Count
1	1	309
2	2	91

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Error
INTERCPT	1	3.2271	0.6406
A506	1	-0.00355	0.000973
A513	1	0.0358	0.0105
A523	1	0.0291	0.00638
A601	1	-9.15E-6	6.141E-6
A650	1	0.00599	0.00443
SQEDQ1	1	-0.3102	0.2075
SQEDQ4	1	-0.3564	0.1577
SQEDQ10	1	-0.2496	0.1672

Concordant = 81.8%	Somers' D = 0.639
Discordant = 17.9%	Gamma = 0.641
Tied = 0.2%	Tau-a = 0.225
(28119 pairs)	c = 0.820

Annexe E

La fonction discriminante de Fisher

Cette annexe présente la fonction discriminante linéaire de Fisher obtenue par la méthode du même nom sur la base SAS.

The SAS System
Discriminant Analysis

400 Observations	399 DF Total
8 Variables	398 DF Within Classes
2 Classes	1 DF Between Classes

Class Level Information

STATUT	Frequency	Prior Probability
1	309	0.500000
2	91	0.500000

Discriminant Analysis Linear Discriminant Function

$$\text{Constant} = -.5\ \bar{X}_j'\ COV^{-1}\ \bar{X}_j \qquad \text{Coefficient Vector} = COV^{-1}\ \bar{X}_j$$

STATUT

	1	2
CONSTANT	-9.32734	-11.76828
A506	0.00590	0.01005
A513	0.00345	-0.02612
A523	0.02231	-0.01004
A601	-0.0000133	-4.3797E-7
A650	0.00973	0.00347
SQEDQ1	3.42453	3.73416
SQEDQ4	1.64273	2.05286
SQEDQ10	2.53123	2.81577

Bibliographie

[1] M.P. WAND & M.C. JONES. *Kernel smoothing.* Chapman and Hall, 1995.

[2] C. VAN WYMEERSCH & H. OOGHE. *Traité d'analyse financière.* Presses Universitaires de Namur, quatrième édition, 1990.

[3] D.J. HAND. *Construction and assesments of classification rules.* Wiley, 1997.

[4] J-Y. PIRÇON. *L'accord de crédit aux particuliers à l'aide de méthodes non paramétriques.* Mémoire de licence, Facultés Universitaires Notre-Dame de la Paix, Namur, 1998-1999.

[5] B. SHMELER. *Prédiction de défaillance d'entreprises par discrimination non paramétrique.* Mémoire de licence, Facultés Universitaires Notre-Dame de la Paix, Namur, 1998-1999.

[6] D.J. HAND & W.E. HENLEY. Construction of a k-nearest-neighbour credit scoring system. *IMA Journal of Mathematics Applied in Business and Industry*, (8) : 305-321, 1997.

[7] D. VAN CAILLIE. *Apports de l'analyse factorielle des correspondances multiples à l'étude de la santé financière des petites ou moyennes entreprises.* PhD thesis, Université de Liège, 1991-1992.

[8] S. AUTHIER, D BAESTAENS, R. OLIESLAGERS & P. POULAIN. Commercial credit rating tools. Generale bank, Central Credit Dept. R&D.

[9] P. POULAIN. Scoring Functions (Installment Loans with Object). Generale bank, Central Credit Dept. R&D, 1998.

[10] B.S. EVERITT. *Cluster analysis.* Arnold, troisième édition. Londres, 1993.

[11] B.W. SILVERMAN. *Density estimation for Statistics and Data Analysis.* Chapman and Hall, 1986.

[12] M. BARDOS & J.-P. RASSON. *Analyse statistique du risque scoring financier.* A paraître chez Dunod, 2001.

[13] A. SILEM. *Encyclopédie de l'économie et de la gestion.* Hachette Education.

[14] J. BRÉMOND & A. GÉLÉDAN. *Dictionnaire économique et social.* Hatiers Paris, 1990.

[15] A. BRIGNONE, J. LAMBERT, A. MARTINET & H. SAVALL. *Encyclopédie de l'économie.* Larousse, 1978.

[16] S GATELIER. La vie des entreprises en Belgique. Cedre n^0 222, Gerling Namur, 1999.

[17] H.C. PÜTZ & B. COUPÉ. Baromètre Européen. Gerling Namur, 1999.

[18] *Annuaire statistique de la Belgique.* Institut National de Statistiques, Ministère des Affaires Economiques. Tomes 70 à 119.
`http://stabel.fgov.be/`

[19] *Bureau Van Dijk Electronic Publishing.* Site internet du Bureau Van Dijk,
`http://www.bvdep.com/`