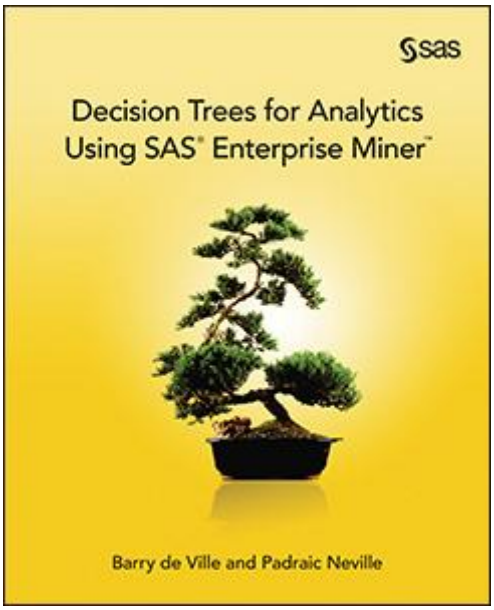
Decision Trees for Analytics Using SAS Enterprise Miner

Barry de Ville and **Padraic Neville**

SAS Institute, June 2013.



L'ouvrage effectue un large tour d'horizon des arbres décision et de leur implémentation dans SAS Enterprise Miner (SAS EM). Il s'intéresse à la fois aux aspects historiques, théoriques et pratiques.

Les auteurs sont des références incontournables, tant concernant les aspects méthodologiques liées aux arbres que concernant leur programmation dans des logiciels. **Barry de Ville** a dirigé le développement de Knowledge Seeker, le premier logiciel d'induction d'arbres interactifs qu'il m'ait été donné de voir lorsque je me suis intéressé à ce thème dans le milieu des années 90. Je participais à l'époque au développement de la version 2 de Sipina (que j'ai vite abandonné pour repartir de zéro avec la version 3). J'ai vite mesuré le chemin qu'il me restait à faire – immense – pour produire un outil avec un tel niveau d'efficacité (rapidité de calcul, fonctionnalités interactives, etc.). Barry de Ville est aujourd'hui « Solution Architect » chez SAS. J'imagine qu'il s'y connaît un peu en matière de développement de logiciels. **Padraic Neville** est statisticien. Il a produit la première version commerciale de la méthode CART (Breiman et al., 1984). Là également, on imagine aisément

le sacré background qu'il y a derrière. Il a développé – entres autres – les procédures hautes performances de Random Forest (HPFOREST) de SAS EM.

Avec un tel pédigrée, vous imaginez très bien que j'étais particulièrement excité à l'idée de lire l'ouvrage, avec deux arrières pensées avouées : décortiquer les algorithmes sous-jacents aux arbres implantées dans SAS EM, avec pourquoi pas des détails sur les procédures statistiques utilisées ; obtenir des informations détaillées sur les fonctionnalités interactives proposées par l'outil, si certaines qui sont bonnes à reprendre dans l'actuelle version de Sipina, je ne m'en priverai pas.

Après coup, je me rends compte que l'ouvrage se situe clairement sur un autre registre. Il s'adresse clairement aux non-spécialistes susceptibles de s'intéresser à ce que l'on peut obtenir avec les arbres. Ni le statisticien, ni l'informaticien qui sont en moi n'y ont trouvé leur compte à vrai dire. Il n'en reste pas moins que j'ai lu l'ouvrage en détail et, certains éléments glanés au détour des pages m'ont paru particulièrement intéressants. Je vais essayer de rendre compte de cela dans les paragraphes qui suivent.

Le **chapitre 1** est introductif. Il décrit succinctement ce que recouvrent les arbres de décision, à quel type de problématique ils répondent, pourquoi cette approche est intéressante dans un processus d'analyse prédictive. A partir de copies d'écran d'arbres, la lecture des règles est détaillée. Un arbre peut servir à la fois pour la régression (la variable cible est quantitative) et pour le classement (la cible est qualitative).

Le **chapitre 2** nous décrit les principales caractéristiques d'un arbre. Les auteurs distinguent 3 finalités : la description sert à rendre compte des relations existantes dans les données ; la prédiction vise à produire un modèle purement prédictif qui peut s'appliquer sur de nouvelles observations non étiquetées, s'assurer des conditions de reproductibilité des relations entres les variables prédictives et la cible est primordial dans ce cas ; l'explication cherche plutôt à établir les relations de cause à effet c.-à-d. comprendre le sens de l'influence des prédictives sur les valeurs prises par la cible. Bien sûr, ces 3 aspects ne sont pas antinomiques. Une prédiction soit s'appuyer sur des relations que l'on constate et que l'on comprend. L'interprétation du modèle est un élément essentiel pour s'assurer de la reproductibilité des résultats sur les observations futures. Les arbres permettent d'associer ces différents prismes.

Les auteurs proposent plusieurs exemples pour comprendre l'intérêt des arbres lors de l'interprétation des résultats. En effet, les segmentations successives permettent d'identifier les relations entre variables dans des contextes différents, introduits par les partitionnements en amont. Ils donnent pour exemple (page 23, figure 2.9) le contraste opéré sur la liaison entre le type de promotion et le comportement d'achat selon le segment de marché.

Puis, de manière un peu abrupte quand même, les auteurs enchaînent par l'historique de l'induction par arbres (à partir de la page 24). A l'origine était AID (Automatic Interaction Detection) de Morgan et Sonquist (1963). Ils se plaçaient dans le cadre de régression. Ils présentaient l'arbre comme une alternative non linéaire à la régression linéaire multiple. Les arbres permettaient également de dépasser les problèmes de multi colinéarité. L'approche a connu une évolution importante en se généralisant aux problèmes de classement (variable cible qualitative) avec la méthode CHAID (chi-square AID ; Kass, 1980) et CART (Classification and regression trees ; Breiman, Friedman, Olshen et Stone, 1984). Les deux ont une approche différente du problème de surapprentissage que l'on reprochait à AID (lorsque l'arbre est trop grand, il « colle » trop aux données d'apprentissage au point d'en ingérer les spécificités). La première s'appuie sur une correction de la p-value lors du test de significativité des partitionnements lors de l'expansion de l'arbre. Basé sur le principe des comparaisons multiples, CHAID utilise la correction de Bonferroni pour éviter des segmentations intempestives qui ne reflètent pas des relations effectivement présentes dans la population. La seconde, CART, s'appuie sur une technique novatrice (à l'époque) : le post-élagage. Après avoir construit l'arbre le plus grand possible (phase d'expansion – growing phase), il est réduit en optimisant un critère en liaison directe avec la finalité de la modélisation (phase de post-élagage – pruning phase). Dans un problème de classement, on cherche à minimiser le taux d'erreur.

Le chapitre va au-delà de ces thèmes. Il aborde de nombreux domaines : la validation statistique ; l'induction de règles, corollaire à la construction des arbres de décision ; les contributions de Ross Quinlan, père de la méthode C4.5 ; les différences entres les communautés statistiques et informatiques. La distinction, attribuée à Breiman (2002), reposerait sur l'opposition : les statisticiens cherchent à répondre à la question « why things works », alors que les informaticiens seraient surtout intéressés par « whether things works ». Le texte est intéressant. Il se lit agréablement. Mais on ne voit pas la trame, le fil directeur.

Notons tout de même une revue des principales caractéristiques des arbres en fin de chapitre (page 50).

Le **chapitre 3** est intitulé « The mechanics of decision trees construction ». Nous voici dans le vif du sujet. Les auteurs identifient 6 grandes étapes dans le processus d'induction à partir de données :

1. La préparation des données pour l'algorithme d'induction d'arbre ;
2. La spécification de la variable cible et des variables prédictives ;
3. La définition des paramètres de l'algorithme d'apprentissage ;
4. Le calcul des segmentations candidates, cela inclut les éventuels regroupements des modalités de la variable de partitionnement ;
5. Sélection et application de la meilleure segmentation candidate ;
6. La poursuite du processus récursivement jusqu'à obtention de l'arbre de décision définitif, le modèle final.

Les auteurs décrivent les opérations avec un certain recul. Les idées qui sous-tendent les calculs sont parfaitement identifiées. Mais aucune formule n'est vraiment détaillée. Il reste que j'ai pu glaner ici ou là des informations importantes.

Sur les grandes bases, SAS EM utilise des échantillons pour réaliser les calculs – lors de la recherche de la meilleure variable de segmentation - sur un nœud. Le chiffre de 30.000 observations est avancé (page 63). J'étais d'autant plus sensible à l'idée que je l'ai moi-même expérimentée dans Sipina. SAS préconise un échantillonnage. A la sortie, nous n'avons pas la garantie d'obtenir exactement le même arbre que si nous avions utilisé la totalité de l'échantillon. Les performances en prédiction sont en revanche équivalentes. Et le gain en temps de calcul est particulièrement conséquent.

SAS EM propose plusieurs options pour le traitement des données manquantes sur les variables prédictives. Il autorise notamment la propagation des valeurs fractionnaires dans les sous-branches de l'arbre lorsqu'une valeur manque sur un nœud (page 78).

Les auteurs insistent beaucoup sur la correction introduite par Kass pour CHAID pour la détermination du caractère significatif ou non d'une segmentation. SAS EM introduit une sophistication supplémentaire (page 84) tenant compte de la profondeur du sommet (nombre

de niveaux à partir de la racine) pour ajuster la p-value. Aucun détail n'est fourni mais on comprend l'idée. Les parties basses de l'arbre sont tributaires des découpages précédents. Le risque de trouver des segmentations faussement intéressantes est accru. De même, SAS EM tient compte également du nombre de variables explicatives candidates pour ajuster toujours la probabilité critique.

SAS EM sait traiter les coûts de mauvaise affectation dans la construction de l'arbre (page 97). Il sait également corriger les fréquences apparentes des classes si on lui fournit les vraies prévalences dans la population (page 103).

La section « Switching targets » m'a particulièrement intrigué. Il semble possible de changer en cours de route de variable cible durant la construction de l'arbre. Voilà une fonctionnalité interactive que je ne retrouve dans aucun logiciel, du moins à ma connaissance. Quel intérêt me direz-vous ? Les auteurs justifient cela par la nécessité parfois de changer de point de vue ou d'objectif lorsque l'on est sur une sous-population particulière. Par exemple, on cherche à identifier les déterminants du poids des personnes (page 106, figure 3.28). Lorsque nous avons circonscrit dans un nœud de l'arbre les hommes de grande taille, nous « switchons » l'analyse en nous intéressant cette fois-ci aux déterminants de l'aptitude physique dans cette sous-population particulière. Ainsi, tous les calculs sont redéfinis à partir de la nouvelle variable cible dans la deuxième partie l'arbre. Bien évidemment, ce changement est possible avec n'importe quel logiciel. Il suffit d'exporter les individus associés à un nœud et à recommencer une nouvelle analyse avec le sous-échantillon. L'intérêt ici est que le changement est effectué à la volée dans le même arbre.

Je suis très dubitatif par rapport à ce chapitre. On perçoit la trame et les idées qui sous-tendent les traitements. Il n'y a aucun doute là-dessus. Le scientifique est en revanche un peu déçu de ne pas obtenir plus détails concernant les calculs sous-jacents. Manifestement ce n'était pas l'objectif des auteurs.

Le **chapitre 4** fait le parallèle entre la BI (Business Intelligence) et les arbres de décision. Dans un cube, nous explorons les valeurs prises par une variable d'intérêt à partir de croisement de catégories (les dimensions). L'arbre de décision finalement répond à ces spécifications. C'est ce qu'essaient de nous expliquer les auteurs, en arguant notamment que les arbres sont plus souples et disposent de mécanismes statistiques pour détecter les relations les plus

intéressantes. Un tableau comparatif des mérites respectifs des cubes et des arbres est fourni en page 127 (table 4.1).

La section « Multidimensional Analysis with Trees » a attiré mon attention dans ce chapitre. Il aborde le cas des arbres multi-objectifs c.-à-d. avec plusieurs variables cibles à prendre en compte simultanément. Là également, il s'agit d'un thème auquel je suis sensibilisé puisque j'avais travaillé sur les arbres de classification (qui peuvent se muer en système prédictifs lorsque les variables cibles sont distincts des variables de segmentation). A la lecture, on constate qu'il n'existe pas de dispositif dédié dans SAS EM en réalité. Les auteurs décrivent des stratégies de codage permettant de construire une variable synthétique qui retraduit les variations de plusieurs variables cibles.

Dans le **chapitre 5**, les auteurs abordent les questions théoriques relatives à la construction des arbres de décision (Theoretical Issues in the Decision Tree Growing Process). Le premier point concerne la construction interactive des arbres. Les auteurs nous expliquent que la prédiction et la découverte de connaissances ne sont pas antagonistes. La construction d'un arbre « optimal » doit marier les connaissances expertes du domaine et les aspects purement numériques. Un guide pour une présentation convaincante des résultats issus d'une modélisation par arbre est proposé (page 147).

La section « Perspectives on Selection Bias » aborde le problème du biais dans la sélection de la variable de segmentation sur un nœud. C'est une question largement débattue dans la littérature. Les auteurs s'attardent sur 3 solutions statistiques. Là pour le coup, nous sommes vraiment dans la technique. Une approche faiblement biaisée, inspirée de QUEST (Loh, 1997) a semble-t-il été implantée dans la procédure HPFOREST de SAS.

Le dernier point saillant du chapitre concerne les méthodes ensemblistes (Multiple Decision Trees, page 171). Depuis la méthode BAGGING de Breiman (1996), ces méthodes ont pris une place importante dans le data mining, notamment parce qu'elles sont très performantes. En théorie, le principe est applicable à toute méthode d'apprentissage automatique. En pratique, on s'est rendu compte que les arbres constituent un terreau particulièrement fertile pour ces méthodes, notamment parce qu'ils (les arbres) ont une forte variance. Les classifieurs individuels sont très différents les uns des autres (elles sont « décorrélées » dirait-on) et, de fait, elles se complètent. Une discussion est menée, comparant le gradient boosting avec les

random forest. Les auteurs ne se mouillent pas en affirmant que chaque méthode a ses avantages et ses inconvénients (page 186). On s'en doutait un peu.

Le **chapitre 6** traite de l'intégration des arbres avec les autres méthodes de data mining. Divers sujets sont abordés. On remarquera notamment le traitement des séries temporelles. La stratégie préconisée par les auteurs repose sur une transformation des données de manière à être conforme avec une présentation transversale (cross-sectional form). A partir de là, nous retombons sur un schéma que l'on sait classiquement traiter avec les arbres.

Les auteurs nous parlent également de l'induction de règles. J'ai appris que SAS propose une version de RIPPER (Cohen, 1995). L'outil semble particulièrement sophistiqué mais reste rattaché au principe « separate-and-conquer » (page 206, figure 6.13).

Les autres sujets abordés concernent la sélection de variables basée sur l'importance de variables mesurées à l'aide des arbres, la prise en compte des interactions entre les variables à partir des sous-branches mis en évidence par les arbres. Un tableau des contributions croisées entre les arbres et les autres méthodes est proposé en page 199 (tableau 6.1).

En **conclusion**, s'il fallait rebaptiser l'ouvrage, je dirais « Quelques considérations autour des arbres de décision et de régression ». Les auteurs, qui ont manifestement un background conséquent, nous proposent une sorte de flânerie. Ils nous parlent de différents thèmes relatifs à la construction, la mise en œuvre et l'exploitation des arbres, s'attardant plus ou moins selon l'intérêt qu'ils leur portent. Il ne s'agit pas d'un ouvrage d'initiation à l'induction des arbres de décision et régression, ni un ouvrage à destination des scientifiques, encore moins un tutoriel pour SAS EM. Il faut en être conscient simplement lorsqu'on en entame la lecture.