Qualité d'ajustement d'arbres d'induction

Gilbert Ritschard*, Djamel A. Zighed**

*Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch http://mephisto.unige.ch
**Laboratoire ERIC, Université Lumière Lyon 2
zighed@univ-lyon2.fr http://eric.univ-lyon2.fr

Résumé. Cet article discute des possibilités de mesurer la qualité de l'ajustement d'arbres d'induction aux données comme cela se fait classiquement pour les modèles statistiques. Nous montrons comment adapter aux arbres d'induction les statistiques du khi-2, notamment celle du rapport de vraisemblance utilisée dans le cadre de la modélisation de tables de contingence. Cette statistique permet de tester l'ajustement du modèle, mais aussi l'amélioration de l'ajustement qu'apporte la complexification de l'arbre. Nous en déduisons également des formes adaptées des critères d'information AIC et BIC qui permettent de sélectionner le meilleur arbre en termes de compromis entre ajustement et complexité. Nous illustrons la mise en œuvre pratique des statistiques et indicateurs proposés avec un exemple réel.

Mots clés : arbre d'induction, qualité d'ajustement, tests du khi-2, comparaison d'arbres

1 Introduction

Les arbres d'induction (Kass, 1980; Breiman et al., 1984; Quinlan, 1993; Zighed et Rakotomalala, 2000) sont l'un des outils les plus populaires d'apprentissage supervisé. Ils consistent à rechercher par éclatements successifs de sommets, une partition de l'ensemble des combinaisons de valeurs des prédicteurs optimale pour prédire la variable réponse. La prédiction se fait simplement en choisissant, dans chaque classe de la partition obtenue, la modalité la plus fréquente de la variable à prédire. Bien que leur utilisation première soit la génération d'arbres de décisions pour la classification, les arbres d'induction fournissent une description de la façon dont la distribution de la variable à prédire est conditionnée par les valeurs des prédicteurs. Ils nous indiquent par exemple comment la répartition entre clients solvables et insolvables est influencée par les attributs âge, sexe, niveau d'éducation, profession, etc. En ce sens les arbres d'induction sont donc des outils de modélisation de l'influence des prédicteurs sur la variable à prédire au même titre que par exemple la régression linéaire, la régression logistique ou la modélisation log-linéaire de tables de contingence multi-dimensionnelles. C'est essentiellement à cet aspect de modélisation descriptive, et en particulier à l'évaluation de la qualité de la description fournie par un arbre induit que nous nous intéressons dans cet article.

En modélisation statistique, qu'il s'agisse de régression linéaire, d'analyse discri-

minante, de régression logistique ou plus généralement de modèle linéaire généralisé (GLM), il est d'usage d'évaluer la qualité d'ajustement du modèle, c'est-à-dire la qualité de la description fournie par le modèle, avec des mesures descriptives telles que le coefficient de détermination R^2 ou des pseudo R^2, et avec des statistiques de test telles que les khi-2 du score test, de Wald ou du rapport de vraisemblance. Parmi ces dernières, la statistique du rapport de vraisemblance jouit d'une propriété d'additivité qui permet d'évaluer également la pertinence statistique de la simplification d'un modèle de référence par renforcement de contraintes sur ses paramètres, ou, si l'on regarde les choses dans l'autre sens, la significativité statistique de la complexification résultant de l'ajout de paramètres à un modèle donné.

Le cas particulier des tests de significativité globale de l'explication, parfois appelés "tests omnibus", où l'on teste globalement l'apport des facteurs explicatifs par rapport à un modèle de référence naïf — le modèle avec la constante comme seule variable explicative dans le cas de la régression, le modèle d'indépendance dans le cas des modèles log-linéaires — retiendra notre attention. Dans le cas des arbres d'induction, le modèle de référence est naturellement l'arbre de niveau 0 constitué par le seul nœud initial.

Une des difficultés principales à laquelle on se heurte dans la pratique des arbres ou graphes d'induction est le fort degré de complexité des arbres mis en évidence. Il nous parait alors souhaitable de pouvoir disposer aussi de critères tels que le critère d'information AIC d'Akaike (1973) ou le critère d'information bayésien BIC (Schwarz, 1978; Kass et Raftery, 1995). Ces critères sont une combinaison de la qualité d'ajustement (statistique du rapport de vraisemblance) et d'une mesure de la complexité. Ils s'avèrent ainsi une aide précieuse pour arbitrer entre complexité et qualité d'ajustement dans la sélection de modèles.

L'article est organisé comme suit. La section 2 situe la place des mesures de qualité d'ajustement envisagés parmi les mesures classiques de qualité d'un arbre d'induction. La section 3 précise le concept d'ajustement considéré. La section 4 est consacrée aux critères de qualité d'ajustement. On montre comment les statistiques du khi-2 de Pearson et du rapport de vraisemblance utilisées dans le cadre de la modélisation de tables de contingence peuvent s'adapter aux arbres d'induction. Nous discutons ensuite l'amélioration qu'apporte un modèle par rapport à un modèle de référence. On montre comment exploiter la différence des statistiques G^2 du rapport de vraisemblance pour tester la significativité statistique du gain d'information et proposons des indicateurs de type R^2. Nous montrons également comment appliquer les critères d'information d'Akaike (AIC) et bayésien (BIC) aux arbres d'induction et discutons leur intérêt pour guider le choix entre modèles de complexité variable. La section 5 illustre l'utilisation des critères proposés dans un cas concret et la section 6 valide empiriquement le calcul des degrés de liberté et la distribution des statistiques du khi-2 pour arbres. Finalement, la conclusion fait l'objet de la section 7 où nous mentionnons des pistes de développements futurs de la démarche initiée dans cet article.

2 Arbre d'induction et mesures classiques de qualité

Avant de nous concentrer sur la qualité d'ajustement, nous rappelons brièvement le principe des arbres d'induction et leurs critères usuels de qualité. Ceci dans le but

de pouvoir mieux situer le rôle des mesures de qualité d'ajustement proposées dans cet article. Le lecteur intéressé trouvera par une discussion approfondie des arbres et de leurs critères habituels de qualité par exemple dans Zighed et Rakotomalala (2000).

2.1 Rappel du principe des arbres d'inductions

L'objectif est de construire une règle qui permette, à partir de la connaissance d'un vecteur d'attributs $\mathbf{x} = (x_1, \ldots, x_p)$, de prédire une variable réponse y, ou si l'on préfère de classer les cas selon les états de la variable y. La construction de la règle se fait en deux temps. Dans un premier temps, on détermine une partition des valeurs possibles de $\mathbf{x}$ telle que la distribution de la réponse y soit la plus pure possible dans chaque classe de la partition, ou de façon plus ou moins équivalente la plus différente possible d'une classe à l'autre. La règle consiste ensuite à attribuer à chaque cas la valeur de y la plus fréquente dans sa classe.

Les arbres d'induction déterminent la partition par éclatements successifs des sommets. En partant du sommet initial, ils recherchent l'attribut qui permet le meilleur éclatement selon un critère qui peut être par exemple le gain d'entropie (C4.5, Sipina) ou la significativité d'un khi-2 (CHAID). L'opération est répétée à chaque nouveau sommet jusqu'à ce qu'un critère d'arrêt, une taille minimale du sommet par exemple, soit atteint. Le résultat est un arbre tel que celui présenté à la figure 1.

2.2 Taux d'erreur

Le taux d'erreur de classification, c'est-à-dire le pourcentage de cas mal classés est peut-être le critère de qualité le plus utilisé. Il mesure la performance prédictive du modèle. S'agissant de classification, c'est évidemment la performance en généralisation qui importe, c'est-à-dire la performance pour des cas n'ayant pas servi à l'apprentissage. C'est pourquoi il convient de calculer le taux d'erreur sur un échantillon de validation différent de l'échantillon d'apprentissage. Pratiquement, ceci conduit à partitionner l'ensemble de données en deux parties, l'une servant à l'apprentissage, l'autre à la validation.

Le taux d'erreur est souvent évalué par la validation croisée qui donne également des indications sur sa variabilité. Cette méthode consiste à partager les données en

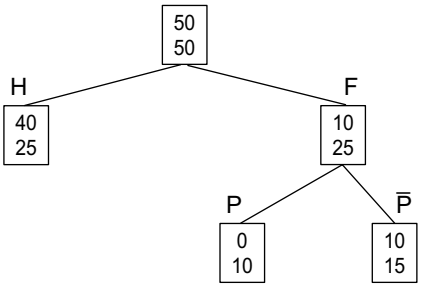


FIG. 1 – Arbre induit

par exemple 10 groupes de tailles approximativement égales et à répéter ensuite l'apprentissage en réservant à chaque fois un groupe différent pour la validation. Le taux d'erreur en validation croisée est la moyenne des taux obtenus. Notons que comme on peut obtenir à chaque fois un arbre différent l'on détermine ainsi le taux d'erreur de la méthode de construction de l'arbre et non l'erreur d'un arbre induit particulier.

La décomposition biais/variance/erreur résiduelle du taux d'erreur (Geurts, 2002) fournit des indications précieuses sur la part de l'erreur due à la variabilité de l'échantillon d'apprentissage. Ces décompositions ne sont cependant quantifiables que numériquement sur la base de simulations ou de méthodes de rééchantillonnage ce qui limite leur utilisation systématique.

2.3 Qualité des partitions

La qualité d'une partition est d'autant meilleure que les sommets terminaux sont purs, c'est-à-dire qu'ils ont des distributions le plus proche possible de la distribution dégénérée qui donne un poids de un à l'une des modalités et zéro aux autres. Cette configuration limite en effet l'incertitude quant à la valeur de la variable réponse à l'intérieur de chaque sommet. Les distributions doivent également être différentes d'un sommet à l'autre.

Il existe plusieurs façons de mesurer la qualité des partitions obtenues (voir Zighed et Rakotomalala, 2000, chap. 9.) On peut mentionner :

- Les mesures issues de la *théorie de l'information* et qui consistent essentiellement à mesurer l'entropie de la réponse pour la partition considérée.
- Celles qui se fondent sur des *distances entre distributions de probabilités*. Le principe consiste ici à vérifier que les distributions diffèrent le plus possible d'une classe à l'autre. La démarche est simple dans le cas de deux classes. Pour le cas plus général d'un nombre quelconque de classes, les solutions proposées restent de portée assez limitée.
- Enfin, les mesures qui s'appuient sur des *indices d'association*. L'idée est ici que la partition est d'autant meilleure que l'association entre les classes de la partition et la variable réponse est forte. Rakotomalala et Zighed (1998) proposent d'utiliser le degré de signification du τ de Goodman et Kruskal ce qui permet de tenir compte également de la taille des échantillons.

2.4 Complexité

Un des intérêts majeurs souvent avancé des arbres et graphes d'induction est la facilité de leur interprétation. Ceci est vrai tant que la complexité de l'arbre reste limitée, d'où l'intérêt de mesurer cette complexité. Les indicateurs couramment utilisés sont :

- *Le nombre de sommets terminaux*. Ce critère correspond au nombre de règles de prédiction ainsi qu'au nombre de classes de la partition finale.
- *La profondeur* ou *le nombre de niveaux de l'arbre*. Ce critère dépend de la procédure de construction. Un arbre n-aire peut définir une même partition avec un nombre plus petit de niveaux qu'un arbre binaire.

– *Le nombre de nœuds.* Ce critère est également lié à la procédure de construction. Il sera en général plus élevé pour un arbre binaire qui tend à multiplier les sommets intermédiaires. Ce critère reflète bien la complexité visuelle de l'arbre induit.
– *La longueur des messages.* Traduit la complexité des règles d'affectation aux différentes classes.

2.5 Place des mesures de qualité d'ajustement envisagées

Les nouveaux critères de la qualité d'ajustement proposés dans cet article concernent la capacité de l'arbre à reproduire la distribution de la réponse y pour les individus ayant un profil $\mathbf{x}$ donné. En d'autres termes, on s'intéresse à la qualité de reproduction de la table de contingence qui croise la variable réponse y avec l'ensemble des prédicteurs. Il s'agit donc de la qualité descriptive de l'arbre par opposition à sa performance en classification. On considère en particulier les deux aspects suivants :

– l'aptitude de l'arbre induit à décrire la distribution de la variable réponse conditionnellement aux valeurs prises par les prédicteurs,
– le gain d'information qu'apporte l'arbre induit par rapport au nœud initial où l'on ne tient pas compte des prédicteurs.

On peut situer les mesures d'ajustement qui nous occupent dans l'optique « qualité des partitions » de la typologie précédente. Il s'agit en effet de voir comment la partition induite par l'arbre ajuste la table cible. Les critères de qualité de partition mentionnés en 2.3 ne mesurent pas explicitement cet ajustement, bien qu'elles s'y réfèrent implicitement. L'entropie de la réponse pour la partition induite aussi bien que l'association entre la partition et la variable réponse constituent en effet des mesures de la proximité entre les distributions prédites par l'arbre et les distributions cibles. Notre objectif est cependant de proposer des mesures d'ajustement qui s'apparentent à celles utilisées en modélisation statistique et qui se prêtent à l'inférence statistique, le test de significativité du défaut d'ajustement ou le test de la différence entre deux arbres en particulier.

La qualité d'ajustement de la table des données d'apprentissage doit être mise en parallèle avec la stabilité de la description fournie, c'est-à-dire la simplicité de l'arbre. En effet, en développant trop loin un arbre, on a tendance à décrire les spécificités de l'échantillon plutôt que la structure du phénomène sous-jacent. Un arbre trop complexe dépend trop étroitement de l'échantillon pour fournir une description généralisable des liens liant la variable réponse aux prédicteurs. On cherchera ainsi par exemple l'arbre le plus simple qui ajuste la table de façon satisfaisante. Alternativement, on peut s'intéresser à l'arbre qui offre le meilleur compromis entre qualité d'ajustement et complexité. Nous proposons à cet effet des adaptations pour les arbres des critères d'information AIC et BIC, critères qui par construction relèvent simultanément de la complexité et de la qualité des partitions. Les adaptations proposées reposent sur le modèle paramétré de reconstruction des données à partir de l'arbre que nous introduisons à la section 4.3 et qui permet de mesurer la complexité de l'arbre en termes de nombre de paramètres. Cette dernière possibilité constitue en soi une contribution à la panoplie des mesures de complexité.

3 Concept de qualité d'ajustement d'un arbre

3.1 Table cible et table prédite

De façon générale, la qualité d'ajustement d'un modèle se réfère à sa capacité à reproduire les données. Dans le cas de la prédiction quantitative d'une variable Y, par exemple dans le cas de la régression linéaire, l'objectif est clair. Il s'agit d'obtenir des valeurs prédites $\hat{y}_\alpha$ qui s'ajustent le mieux possible aux valeurs observées y_α, pour $\alpha = 1, \ldots, n$, n étant le nombre d'observations. De même, dans l'optique de la classification, les états prédits $\hat{y}_\alpha$ doivent correspondre le plus souvent possible aux vraies valeurs y_α. Le taux d'erreur est dans ce cas un indicateur naturel de qualité d'ajustement.

Dans certaines situations, en particulier dans les sciences de comportement (sociologie, sciences politiques, marketing, ...), les arbres ou graphes d'induction sont utilisés plus dans une optique descriptive que prédictive comme outil de mise en évidence des principaux déterminants de la variable à prédire. Ils sont utilisés comme outil d'aide à la compréhension de phénomènes et non pas comme outil de classification.

Ce ne sont plus alors les états particuliers y_α que l'on cherche à reproduire. Pour comprendre comment les prédicteurs interagissent sur la variable réponse Y, il convient en effet d'examiner comment la distribution de Y change avec le profil $\mathbf{x}$. Dans cette optique, la qualité d'ajustement considérée ici se réfère à la qualité de la reproduction de l'ensemble des distributions conditionnelles.

Formellement, dans l'optique classification il s'agit de caractériser une fonction $f(\mathbf{x})$ qui prédit la valeur de y compte tenu de $\mathbf{x}$. Avec les arbres, la construction de cette fonction se fait en deux étapes :

1. Caractériser un modèle descriptif $\mathbf{p}(Y|\mathbf{x}) = \big(p_1(\mathbf{x}), \ldots, p_\ell(\mathbf{x})\big)$, où la notation $p_i(\mathbf{x})$ désigne la probabilité conditionnelle $p(Y = y_i|\mathbf{x})$.
2. Prédire par la règle majoritaire $f(\mathbf{x}) = \arg\max_i p_i(\mathbf{x})$.

Contrairement à la régression logistique par exemple, où $\mathbf{p}(Y|\mathbf{x})$ s'exprime analytiquement en fonction de $\mathbf{x}$, le modèle descriptif prend ici la forme non paramétrique d'un ensemble fini de distributions :

$$\left\{\mathbf{p}_{|j} \in [0,1]^\ell \mid \mathbf{p}_{|j} = \mathbf{p}(Y|\mathbf{x}_j), j = 1, \ldots, c\right\} \ .$$

Les prédicteurs étant supposés prendre un nombre fini de valeurs, l'ensemble des distributions conditionnelles observées est représentable sous forme d'une table de contingence $\mathbf{T}$ croisant y avec la variable composite définie par le croisement de tous les prédicteurs. Le tableau 1 est un exemple d'une telle table dans le cas où la variable à prédire est le statut marital et les prédicteurs le genre et le secteur d'activité. Le nombre de lignes de $\mathbf{T}$ est le nombre ℓ d'états de la variable Y. Si chaque prédicteur x_ν, $\nu = 1, \ldots, p$ a c_v valeurs différentes, le nombre c de colonnes de $\mathbf{T}$ est au plus le produit des c_v, soit : $c \leq \prod_\nu c_\nu$, l'inégalité étant stricte lorsque certaines combinaisons de valeurs des attributs sont structurellement impossibles. On s'intéresse donc à la capacité de l'arbre à reproduire cette table $\mathbf{T}$.[1]

[1] Il n'est pas sans intérêt de relever comme nous l'avons fait dans Ritschard et Zighed (2003), que dans cette optique de reconstitution de T, les arbres peuvent constituer un outil alternatif et complémentaire à la modélisation log-linéaire des tables de contingence multidimensionnelle.

	homme			femme			
marié	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	total
non	11	14	15	0	5	5	50
oui	8	8	9	10	7	8	50
total	19	22	24	10	12	13	100

TAB. 1 – Exemple de table de contingence $\mathbf{T}$

Un élément de la table $\mathbf{T}$ est noté n_{ij} et représente le nombre de cas avec profil $\mathbf{x}_j$ qui dans les données prennent la valeur y_i de la variable réponse. On note $\hat{\mathbf{T}}$ la table prédite à partir d'un arbre et $\hat{n}_{ij}$ désigne un élément générique de cette table. Formellement, la qualité d'ajustement se réfère ainsi à la divergence entre les tables $\mathbf{T}$ et $\hat{\mathbf{T}}$.

Il reste à préciser comment l'on déduit la table estimée $\hat{\mathbf{T}}$ à partir d'un arbre. On utilise pour cela le modèle de reconstruction suivant où l'on note $\mathbf{T}_j$ la j-ème colonne de $\mathbf{T}$:

$$\mathbf{T}_j \quad = \quad n\, a_j \mathbf{p}_{|j}, \quad j = 1, \ldots, c \tag{1}$$

Les paramètres sont le nombre total n de cas, les proportions a_j de cas par colonne $j = 1, \ldots, c$, et les c vecteurs de probabilités $\mathbf{p}_{|j} = \mathbf{p}(Y|\mathbf{x}_j)$ correspondant à la distribution de Y dans chaque colonne j de la table. Nous verrons que chaque arbre donne lieu à des estimations $\hat{\mathbf{p}}_{|j}$ différentes des vecteurs $\mathbf{p}_{|j}$ et par suite à une estimation $\hat{\mathbf{T}}$ différente. Les a_j seront estimés par les proportions $\hat{a}_j = n_{\cdot j}/n$ observées dans l'échantillon, $n_{\cdot j}$ désignant le total de la j-ème colonne de $\mathbf{T}$. Contrairement aux $\hat{\mathbf{p}}_{|j}$, les estimations $\hat{a}_j$ sont indépendantes de l'arbre induit.

3.2 Arbre saturé et arbre étendu

En modélisation statistique, on appelle modèle saturé un modèle avec le nombre maximal de paramètres libres qui peuvent être estimés à partir des données. En modélisation log-linéaire de tables de contingence multidimensionnelles, le modèle saturé permet de reproduire exactement la table modélisée. Par analogie, nous introduisons le concept d'arbre saturé qui permet de reproduire exactement la table $\mathbf{T}$.

Définition 1 (Arbre saturé) *Pour des variables prédictives catégorielles, on appelle* arbre saturé, *un arbre qui résulte de tous les éclatements successifs possibles selon les modalités des variables prédictives.*

Les distributions $\mathbf{p}_{|j}$ conditionnelles aux feuilles de l'arbre saturé sont estimées par les vecteurs de fréquences relatives, soit les vecteurs d'éléments $n_{ij}/n_{\cdot j}$, $i = 1, \ldots, \ell$.

L'arbre saturé n'est pas unique, des variantes étant possibles selon l'ordre dans lequel les variables sont prises en compte. Tous les arbres saturés conduisent cependant aux mêmes feuilles (sommets terminaux). Ces feuilles correspondent aux colonnes de la table de contingence $\mathbf{T}$.

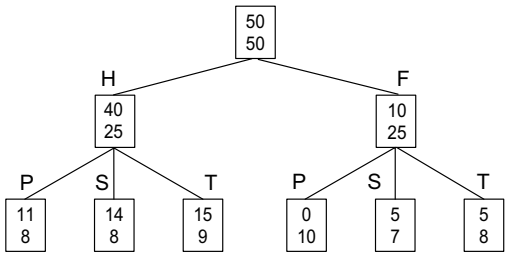


FIG. 2 – Arbre saturé

Par exemple, la figure 2 donne l'arbre saturé correspondant au cas du tableau 1 où l'on cherche à prédire le statut marital connaissant le genre (H = homme, F = femme) et le secteur d'activité (P = primaire, S = secondaire, T = tertiaire). Les distributions conditionnelles sont :

$$\hat{\mathbf{p}}_{|HP} = \begin{pmatrix} 11/19 \\ 8/19 \end{pmatrix}, \quad \hat{\mathbf{p}}_{|HS} = \begin{pmatrix} 14/22 \\ 8/22 \end{pmatrix}, \quad \hat{\mathbf{p}}_{|HT} = \begin{pmatrix} 15/24 \\ 9/24 \end{pmatrix},$$
$$\hat{\mathbf{p}}_{|FP} = \begin{pmatrix} 0/10 \\ 10/10 \end{pmatrix}, \quad \hat{\mathbf{p}}_{|FS} = \begin{pmatrix} 5/12 \\ 7/12 \end{pmatrix}, \quad \hat{\mathbf{p}}_{|FT} = \begin{pmatrix} 5/13 \\ 8/13 \end{pmatrix}$$

Notons qu'un algorithme d'induction d'arbres ne peut en général générer un arbre saturé que si (i) toutes les cellules de la table de contingence des variables prédictives sont non vides et si (ii) la distribution de la variable réponse est différente dans chacune des cellules.

Pour comparer les distributions des feuilles de l'arbre induit à celles de l'arbre saturé, on doit étendre l'arbre induit pour obtenir des feuilles de même définition et en particulier en même nombre que celles de l'arbre saturé.

Définition 2 (Extension maximale d'un arbre induit) *Pour des variables prédictives catégorielles, on appelle* extension maximale de l'arbre induit *ou* arbre induit étendu, *l'arbre obtenu à partir de l'arbre induit en procédant à tous les éclatements successifs possibles de ses sommets terminaux et en appliquant aux feuilles de l'extension la distribution* $\hat{\mathbf{p}}^a_{|k}$ *observée dans le nœud terminal parent de l'arbre initial.*

Par exemple, si l'arbre induit est l'arbre avec les sommets blancs de la figure 3, son extension maximale s'obtient en ajoutant les sommets gris et en répartissant dans ceux-ci l'effectif selon la distribution du sommet dont ils sont issus. Les distributions des six sommets terminaux de l'extension se déduisent de celles des trois feuilles de l'arbre induit, soit pour notre exemple :

$$\begin{aligned} \hat{\mathbf{p}}_{|HP} = \hat{\mathbf{p}}_{|HS} = \hat{\mathbf{p}}_{|HT} &= \hat{\mathbf{p}}^a_{|H} = \begin{pmatrix} 40/65 \\ 25/65 \end{pmatrix} \\ \hat{\mathbf{p}}_{|FP} &= \hat{\mathbf{p}}^a_{|FP} = \begin{pmatrix} 0/10 \\ 10/10 \end{pmatrix} \\ \hat{\mathbf{p}}_{|FS} = \hat{\mathbf{p}}_{|FT} &= \hat{\mathbf{p}}^a_{|F\bar{P}} = \begin{pmatrix} 10/25 \\ 15/25 \end{pmatrix} \end{aligned}$$

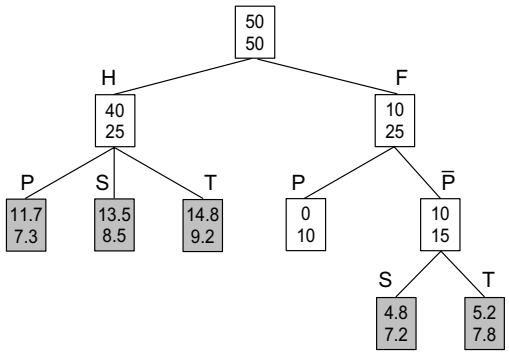


FIG. 3 – Arbre induit (sommets blancs) et son extension maximale

Les feuilles terminales de l'extension de l'arbre induit donnent lieu à la table prédite $\hat{\mathbf{T}}$ représentée au tableau 2.

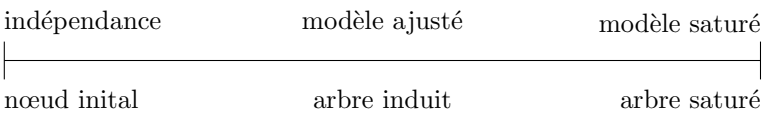
	homme			femme			
marié	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	total
non	11.7	13.5	14.8	0	4.8	5.2	50
oui	7.3	8.5	9.2	10	7.2	7.8	50
total	19	22	24	10	12	13	100

TAB. 2 – Exemple de table de contingence prédite $\hat{\mathbf{T}}$

4 Qualité d'ajustement d'un arbre induit

Ayant précisé le concept d'ajustement d'un arbre qui nous occupe, et en particulier les notions de tables cible et prédite, nous proposons dans cette section des statistiques et indicateurs permettant d'en évaluer la qualité. Il s'agit essentiellement d'adaptations des mesures classiquement utilisées en modélisation statistique pour juger de la qualité globale d'un modèle.

Nous proposons d'utiliser des statistiques de test du khi-2 pour mesurer la divergence entre la table prédite $\hat{\mathbf{T}}$ et la table cible $\mathbf{T}$. Comme en modélisation multinomiale log-linéaire (Agresti, 1990) où cette divergence sert d'indicateur de l'écart entre modèle ajusté et modèle saturé, elle renseigne ici sur l'écart entre l'arbre induit et l'arbre saturé.



Plutôt que de chercher à savoir à quelle distance l'on se trouve du modèle saturé correspondant à la partition la plus fine, il peut être utile de savoir ce que l'on a gagné par rapport à la situation d'indépendance où l'on ne tient pas compte des prédicteurs. Cet aspect correspond à l'écart entre l'arbre induit et le modèle d'indépendance représenté par l'arbre constitué du seul nœud initial.

Nous commençons donc avec les tests d'ajustement du khi-2, puis discutons de la comparaison avec un modèle de référence, et plus généralement de la différence entre deux modèles, en introduisant les statistiques de test de l'amélioration de l'ajustement. Dans la même optique nous proposons plusieurs indicateurs de type R^2. Enfin, dans la perspective de sélectionner l'arbre offrant le meilleur compromis entre ajustement et complexité, nous complétons la discussion en établissant des formes des critères d'information AIC et BIC applicables aux arbres.

4.1 Statistiques du khi-2 pour arbres d'induction

L'objectif est de mesurer la divergence entre $\mathbf{T}$ et $\hat{\mathbf{T}}$ avec une statistique permettant d'évaluer la significativité de l'écart observé. Rappelons que $\mathbf{T}$ est reproduit exactement par l'arbre saturé. La divergence que l'on se propose d'évaluer ici s'interprète donc également comme l'écart entre l'arbre induit et l'arbre saturé.

Les $\mathbf{p}_{|j}$ étant estimés par le maximum de vraisemblance, on peut simplement appliquer les statistiques de divergence de Cressie et Read (1984), dont les cas particuliers les plus connus sont les khi-2 de Pearson que l'on note X^2 et la statistique du rapport de vraisemblance noté G^2 :

$$X^2 = \sum_{i=1}^{\ell}\sum_{j=1}^{c}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{2}$$

$$G^2 = 2\sum_{i=1}^{\ell}\sum_{j=1}^{c} n_{ij}\,\ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right) \tag{3}$$

Sous l'hypothèse que le modèle est correct et sous certaines conditions de régularité, voir par exemple Bishop et al. (1975, chap. 4), ces statistiques suivent une même distribution du χ^2 avec pour degrés de liberté le nombre de cellules de la table de contingence moins le nombre de paramètres linéairement indépendants du modèle de prédiction de $\mathbf{T}$. Dans notre cas $\mathbf{T}$ est prédit par le modèle de reconstruction (1) dont il s'agit alors de préciser le nombre de paramètres linéairement indépendants.

Un arbre induit non saturé définit une partition de l'ensemble $\mathcal{X}$ des profils $\mathbf{x}$ possibles. Chacun de ses q sommets terminaux correspond donc à un sous-ensemble $\mathcal{X}_k \subseteq \mathcal{X}$, $k = 1, \ldots, q$ de profils $\mathbf{x}_j$ pour lequel, on impose la contrainte

$$\mathbf{p}_{|j} = \mathbf{p}^a_{|k} \quad \text{pour tout } \mathbf{x}_j \in \mathcal{X}_k \qquad k = 1, \ldots, q \tag{4}$$

où $\mathbf{p}^a_{|k}$ désigne la distribution dans le sommet terminal k de l'arbre.

Chaque vecteur $\mathbf{p}^a_{|k}$ contient $\ell - 1$ termes indépendants et il y a $q - 1$ distributions conditionnelles $\mathbf{p}^a_{|k}$ indépendantes. On en déduit le décompte du tableau 3 des paramètres indépendants pour un nombre q fixé de sommets terminaux.[2] En retranchant

[2]Bien que q soit induit des données, on raisonne ici conditionnellement à q comme cela se fait en

paramètres	nombre	dont indépendants
$p_{i\|j}$, $i = 1, \ldots, \ell$, $j = 1, \ldots, c$	$c\ell$	$q(\ell - 1)$
a_j, $j = 1, \ldots, c$	c	$c - 1$
n	1	1
Total	$c(\ell + 1) + 1$	$q(\ell - 1) + c$

TAB. 3 – Calcul du nombre de paramètres indépendants

au nombre $c\ell$ de cellules de $\mathbf{T}$ le nombre $q(\ell - 1) + c$ de paramètres indépendants, on obtient les degrés de liberté d_M de l'arbre induit, soit

$$\text{degrés de liberté} = d_M = (c - q)(\ell - 1) \ .$$

Notons que ce nombre correspond au nombre de contraintes (4). Pour le modèle d'indépendance, noté I, on a $q = 1$ et l'on retrouve la valeur usuelle des degrés de liberté du test d'indépendance, soit $d_I = (c-1)(\ell-1)$. De même, pour l'arbre saturé, noté S, on a $q = c$ et donc $d_S = 0$.

Sous réserve des conditions de régularité, les statistiques X^2 et G^2 de tables associées à des arbres suivent donc, lorsque le modèle est correct, une distribution du χ^2 avec $(c - q)(\ell - 1)$ degrés de liberté.

Pour l'arbre de la figure 1, on trouve par exemple, $X^2 = 0.1823$ et $G^2 = 0.1836$. On a $c = 6$, $q = 3$ et $\ell = 2$, et donc $d_M = (6 - 3)(2 - 1) = 3$ degrés de liberté. Les valeurs des statistiques sont très petites et, avec un degré de signification de l'ordre de 98% dans les deux cas, indiquent clairement que $\hat{\mathbf{T}}$ ajuste de façon satisfaisante $\mathbf{T}$. La qualité de l'ajustement de l'arbre induit aux données est donc dans ce cas excellente.

Théoriquement, les statistiques X^2 de Pearson (2) et G^2 du rapport de vraisemblance (3) devraient permettre de tester si l'arbre induit s'ajuste de façon satisfaisante aux données. Il est bien connu cependant que la portée du test reste limitée lorsque l'échantillon est grand, le moindre écart devenant alors statistiquement significatif. Par ailleurs, les conditions de régularité requises pour que les statistiques soient distribuées selon une loi du χ^2, et en particulier les conditions d'intériorité (pas d'effectifs attendus nuls), sont difficilement tenables lorsque l'arbre saturé compte un grand nombre de feuilles.

Plus intéressante nous semble être l'utilisation de la statistique G^2 pour comparer des modèles imbriqués, un modèle M_2 (restreint) étant inclus dans M_1 (non restreint) si l'espace de ses paramètres est un sous-ensemble de M_1, c'est-à-dire, en d'autres termes, si les paramètres de M_2 s'obtiennent en imposant des contraintes sur ceux du modèle M_1. En effet dans ce cas la déviance entre les deux modèles est (voir par exemple Agresti, 1990, p. 211 ou Powers et Xie, 2000, p. 105) :

$$G^2(M_2|M_1) \quad = \quad G^2(M_2) - G^2(M_1) \tag{5}$$

modélisation log-linéaire où l'on suppose données les interactions prises en compte, alors qu'elles sont en fait induites des données par le biais du processus de sélection du modèle.

qui, sous l'hypothèse que M_2 est correct, est approximativement distribuée selon une loi du χ^2 avec pour degrés de liberté la différence $d_2 - d_1$ des degrés de liberté des modèles M_2 et M_1.

Cette dernière propriété permet en particulier de tester la significativité d'un éclatement. La déviance entre le modèle après l'éclatement et celui avant l'éclatement nous renseigne en effet sur la pertinence statistique de cet éclatement. Par exemple, si M_1 est l'arbre induit de la figure 1 et M_2 l'arbre avant l'éclatement du sommet « femme ». On a $G^2(M_1) = 0.18$ avec 3 degrés de liberté et $G^2(M_2) = 8.41$ avec 4 degrés de liberté. La déviance est alors

$$G^2(M_2|M_1) = 8.41 - 0.18 = 8.23 \quad \text{avec} \quad d_2 - d_1 = 4 - 3 = 1$$

Son degré de signification (p-valeur) est 0.4%, donc inférieur au seuil généralement admis de 5%, ce qui indique que l'éclatement est statistiquement significatif.

4.2 Comparaison avec un modèle de référence

Cette section est consacrée aux indicateurs de type R^2 qui mesurent le gain relatif de qualité d'un arbre induit M par rapport à l'arbre trivial I constitué par le seul nœud initial. Nous discutons successivement le pourcentage d'amélioration du taux d'erreur, le pourcentage de réduction de l'entropie, l'amélioration de l'ajustement et les pseudo R^2. Les mesures considérées ici pour la comparaison entre l'arbre induit et le nœud initial se généralisent aisément, bien que nous ne le traitions pas explicitement, au cas général de la comparaison de deux modèles dont l'un est inclus dans l'autre.

4.2.1 Remarque sur le pourcentage de réduction du taux d'erreur

Bien qu'on ne s'intéresse pas ici à la prédiction de valeurs individuelles, nous aimerions souligner que l'idée de comparer la performance du modèle avec le modèle qui ne tient pas compte des prédicteurs est également pertinente en terme de taux d'erreur. On peut noter que, sur l'échantillon d'apprentissage, le pourcentage de réduction de l'erreur de classification correspond à la mesure d'association $\lambda_{y|\text{partition}}$ de Guttman (1941) et Goodman et Kruskal (1954) entre la variable réponse y et la partition définie par le graphe induit. Il existe une forme analytique de la variance asymptotique de cette mesure qui permet d'en tester la significativité sous certaines conditions de régularité (Goodman et Kruskal, 1972; Olszak et Ritschard, 1995). Nous n'approfondissons pas cet aspect ici, notre objectif étant les qualités descriptives du graphe induit plutôt que ses qualités prédictives.

4.2.2 Gain d'information

Le gain d'information peut être mesuré par la réduction de l'entropie de la distribution de la variable réponse que permet la connaissance des classes de la partition définie par l'arbre induit. Précisons que nous nous intéressons à la réduction globale d'entropie que permet l'arbre par rapport à la distribution marginale, c'est-à-dire celle dans le nœud initial. Le gain discuté ici se distingue donc du gain partiel et conditionnel à un

nœud que certains algorithmes cherchent à maximiser à chaque étape de construction de l'arbre.

Le gain d'information relativement au nœud initial est mesuré par exemple par les deux indicateurs suivants :

$$\hat{\tau}_{M|I} = \frac{n \sum_i \sum_j \frac{\hat{n}_{ij}^2}{n_{\cdot j}} - \sum_i n_{i\cdot}^2}{n^2 - \sum_i n_{i\cdot}^2} \quad (6)$$

$$\hat{u}_{M|I} = \frac{\sum_i \frac{n_{i\cdot}}{n} \log_2 \frac{n_{i\cdot}}{n} - \sum_j \frac{n_{\cdot j}}{n} \sum_i \frac{\hat{n}_{ij}}{n_{\cdot j}} \log_2 \frac{\hat{n}_{ij}}{n_{\cdot j}}}{\sum_i \frac{n_{i\cdot}}{n} \log_2 \frac{n_{i\cdot}}{n}} \quad (7)$$

Le $\hat{\tau}_{M|I}$ est la deuxième mesure d'association nominale proposée par Goodman et Kruskal (1954). Il mesure la proportion de réduction de l'entropie quadratique $H_Q(\mathbf{p}) = \sum_i p_i(1-p_i)$, connue aussi comme l'indice de variation de Gini. Le $\hat{u}_{M|I}$ est connu en statistique sous le nom de coefficient d'incertitude de Theil (1967, 1970). Il mesure la proportion de réduction de l'entropie de Shannon $H_S(\mathbf{p}) = -\sum_i p_i \log_2 p_i$.

Pour l'arbre M de la figure 1, on trouve par exemple $\hat{\tau}_{M|I} = 0.145$ et $\hat{u}_{M|I} = 0.132$, valeurs qui indiquent une réduction d'entropie d'environ 14%. Si l'on compare ces valeurs à celles du modèle saturé, soit respectivement $\hat{\tau}_{S|I} = 0.146$ et $\hat{u}_{S|I} = 0.134$, il apparaît que l'arbre capte à peu près toute l'information que l'on peut tirer des attributs prédictifs retenus.

Pour tester la significativité statistique du gain d'information, soit les hypothèses $H_0 : \tau_{M|I} = 0$ et $H_0 : u_{M|I} = 0$, une possibilité est d'utiliser les variances asymptotiques que l'on trouve dans la littérature (voir par exemple Olszak et Ritschard, 1995) et une approximation par la loi normale. Il est préférable cependant d'utiliser les transformations suivantes des indicateurs :

$$C(I|M) = (n-1)(\ell-1)\,\hat{\tau}_{M|I} \quad (8)$$

$$G^2(I|M) = \Big(-2\sum_i n_{i\cdot} \log(n_{i\cdot}/n)\Big)\,\hat{u}_{M|I} \quad (9)$$

La première C est due à Light et Margolin (1971) qui montrent que dans le cas où le tableau prédit est le modèle saturé ($M = S$), $C(I|S)$ est, sous l'hypothèse d'indépendance ($\tau_{S|I} = 0$), asymptotiquement distribuée comme un χ^2 à d_I degrés de liberté. Dans le cas d'un modèle M plus restrictif que S, il suffit d'adapter les degrés de liberté. Ainsi, de façon générale $C(I|M)$ suit un χ^2 avec $d_I - d_M$ degrés de liberté.

La transformation du coefficient $\hat{u}_{M|I}$ montre que tester la significativité de $u_{M|I}$ est équivalent à tester la significativité de la différence d'ajustement entre I et M avec $G^2(I|M) = G^2(I) - G^2(M)$. Les deux transformations (8) et (9) suivent donc, sous H_0, asymptotiquement la même loi du χ^2 à $d_I - d_M$ degrés de liberté. Le test avec ces statistiques est plus puissant qu'avec une approximation normale de $\hat{\tau}_{I|M}$ ou $\hat{u}_{I|M}$.

Pour notre exemple d'arbre induit, on trouve $C(I|M) = 14.32$ et $G^2(I|M) = 18.36$. Ces valeurs sont très grandes compte tenu des degrés de liberté $d_I - d_M = 5 - 3 = 2$. Elles confirment donc la significativité statistique du gain d'information de l'arbre par rapport au modèle d'indépendance.

4.2.3 Pseudo R^2

Dans une optique purement descriptive, on peut envisager, comme on le fait par exemple dans la modélisation log-linéaire, des pseudo R^2 qui mesurent la proportion de la déviance entre indépendance I et arbre saturé que l'arbre d'induction reproduit. On peut par exemple utiliser le pseudo R^2

$$R^2 \quad = \quad 1 - \frac{G^2(M)}{G^2(I)}$$

ou sa version corrigée des degrés de liberté

$$R^2_{\text{ajust}} \quad = \quad 1 - \frac{G^2(M)/d_M}{G^2(I)/d_I}$$

Pour notre exemple, on a $G^2(I) = 18.55$, $d_I = 5$, $G^2(M) = .18$ et $d_M = 3$, d'où $R^2 = .99$ et $R^2_{\text{ajust}} = .984$. Ces valeurs confirment que l'arbre capte presque le 100% de l'écart entre l'indépendance et la table cible représentée par l'arbre saturé.

Dans une optique de réduction d'entropie, nous proposons comme alternative au pseudo R^2 ci-dessus, de calculer la part de la proportion maximale de réduction d'entropie possible que l'on atteint avec l'arbre induit. La proportion maximale de réduction d'entropie est obtenue avec la partition la plus fine, c'est-à-dire le modèle saturé. Elle correspond à $\hat{\tau}_{S|I}$ pour l'entropie quadratique et $\hat{u}_{S|I}$ pour l'entropie de Shannon. Les parts de ces valeurs atteintes avec l'arbre induit sont donc :

$$R^2_\tau = \frac{\hat{\tau}_{M|I}}{\hat{\tau}_{S|I}} \quad \text{et} \quad R^2_u = \frac{\hat{u}_{M|I}}{\hat{u}_{S|I}}$$

Pour notre exemple, on obtient respectivement $R^2_\tau = .993$ et $R^2_u = .985$.

4.3 Ajustement et complexité

Dans le but de pouvoir arbitrer entre ajustement et complexité, on peut recourir aux critères d'information AIC d'Akaike (1973) ou au critère bayésien BIC (Schwarz, 1978; Kass et Raftery, 1995).

Dans notre cas, ces critères peuvent par exemple s'écrire :

$$\begin{aligned} \text{AIC}(M) &= G^2(M) + 2(q\ell - q + c) \\ \text{BIC}(M) &= G^2(M) + (q\ell - q + c)\log(n) \end{aligned}$$

Ici, la complexité est représentée par le nombre $q\ell - q + c$ de paramètres indépendants. Le critère BIC pénalise plus fortement la complexité que le critère AIC, la pénalisation augmentant avec le nombre de données n. Notons qu'il existe des formes alternatives du coefficient BIC. Raftery (1995), par exemple, propose BIC $= G^2 - d\log(n)$, où d est le nombre de degrés de liberté, soit dans notre cas $d = (c-q)(\ell - 1)$. Comme ce nombre d diminue d'une unité chaque fois que l'on ajoute un paramètre indépendant, la pénalisation reste évidemment la même. Les deux formulations sont équivalentes à une translation $c\ell$ près.

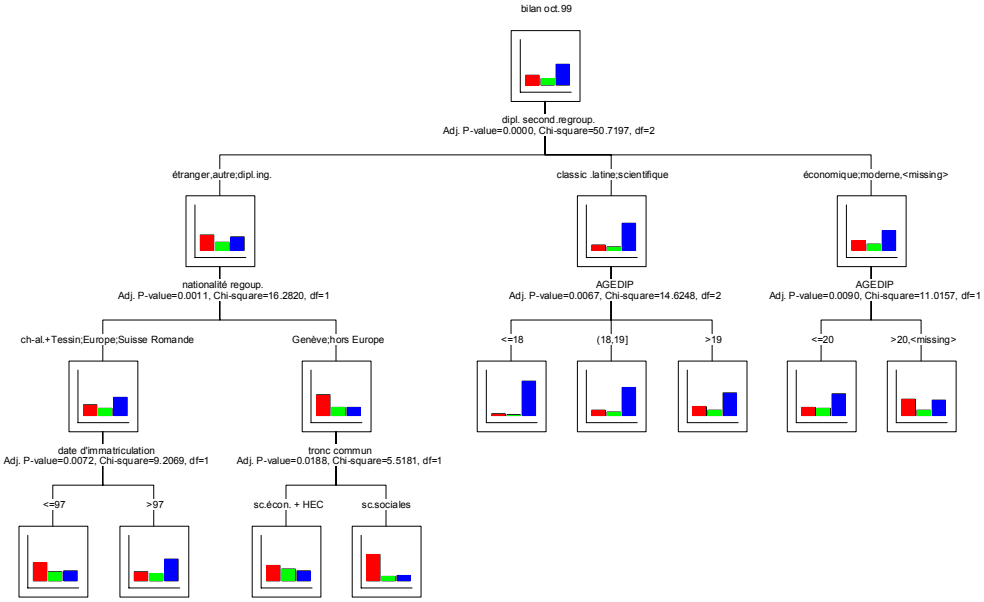
Ces critères d'information offrent une alternative aux tests statistiques pour la sélection de modèles. Parmi plusieurs modèles, celui qui minimise le critère réalise le meilleur compromis entre ajustement et complexité. Le modèle qui minimise BIC en particulier, est, dans une approche bayésienne, optimal compte tenu de l'incertitude des modèles.

Pour illustrer l'utilisation de ces critères, on se propose de comparer notre arbre induit M avec la variante M^* où l'on éclate le sommet «femme» selon les trois secteurs P, S, I au lieu du partage binaire entre primaire P et non primaire $\bar{P}$. Dans les deux cas on a $n = 100$, $c = 6$ et $\ell = 2$. Pour M, on a $q = 3$ et donc $(q\ell - q + c) = 9$ et pour M^*, $q = 4$ et donc $(q\ell - q + c) = 10$. Comme $G^2(M) = 0.18$ et $G^2(M^*) = .16$, on obtient $\text{AIC}(M) = 18.18$ et $\text{AIC}(M^*) = 20.16$. De même, on trouve $\text{BIC}(M) = 41.63$ et $\text{BIC}(M^*) = 46.21$. Les deux critères indiquent que l'arbre M plus simple est préférable à l'arbre M^*. Le gain en qualité d'ajustement de M^* n'est pas assez important pour justifier l'accroissement de la complexité. Remarquons que du point de vue de ces critères d'information, l'arbre induit est supérieur tant au modèle d'indépendance ($\text{AIC}(I) = 32.55$, $\text{BIC}(I) = 50.78$) qu'au modèle saturé ($\text{AIC}(S) = 24$, $\text{BIC}(S) = 55.26$).

5 Illustration

Nous illustrons ici les enseignements apportés par les critères de qualité d'ajustement proposés sur un exemple concret. On considère pour cela les données relatives aux 762 étudiants qui ont commencé leur première année d'études à la Faculté des sciences économiques et sociales de Genève en 1998. Il s'agit de données administratives réunies par Petroff et al. (2001). On rapporte quelques résultats d'une analyse visant à évaluer les chances de respectivement réussir, redoubler ou être éliminé à la fin de la première année d'études selon les caractéristiques personnelles portant notamment sur l'origine et le cursus scolaire. La figure 4 montre l'arbre obtenu avec la procédure CHAID (Kass, 1980) implémentée dans Answer Tree (SPSS, 2001). Parmi une trentaine de prédicteurs potentiels, CHAID en a sélectionné 5 dont deux quantitatifs, l'année d'immatriculation à l'université et l'âge à l'obtention du diplôme de l'école secondaire. Les cinq variables avec les discrétisations et regroupements de modalités proposés par CHAID sont le type de diplôme secondaire (3 modalités), l'âge de son obtention (4), la date d'immatriculation (*datimma*, 2), le tronc commun choisi (*troncom*, 2) et la nationalité (*nationa*, 2). La table cible $\mathbf{T}$ définie par ces variables contient 88 colonnes. Elle a 3 lignes correspondant aux 3 situations possibles de l'étudiant après sa première année d'étude.

Le tableau 4 donne la taille de la partition, la déviance G^2 avec ses degrés de liberté et son degré de signification et les critères AIC et BIC. Ces valeurs peuvent être comparées à celles de plusieurs variantes. CHAID2 est CHAID sans l'éclatement du sommet 4 ($nationa \notin$ {GE, hors Europe}) et CHAID3 sans l'éclatement des sommets 4 et 5 ($nationa \in$ {GE, hors Europe}). Le modèle Sipina correspond au graphe de la figure 5 obtenu avec la procédure Sipina (Sipina for Windows V2.5, 2000; Zighed et Rakotomalala, 2000) qui, comme on peut le voir, autorise également des fusions de sommets. On donne également les valeurs trouvées pour les partitions qui donnent respectivement



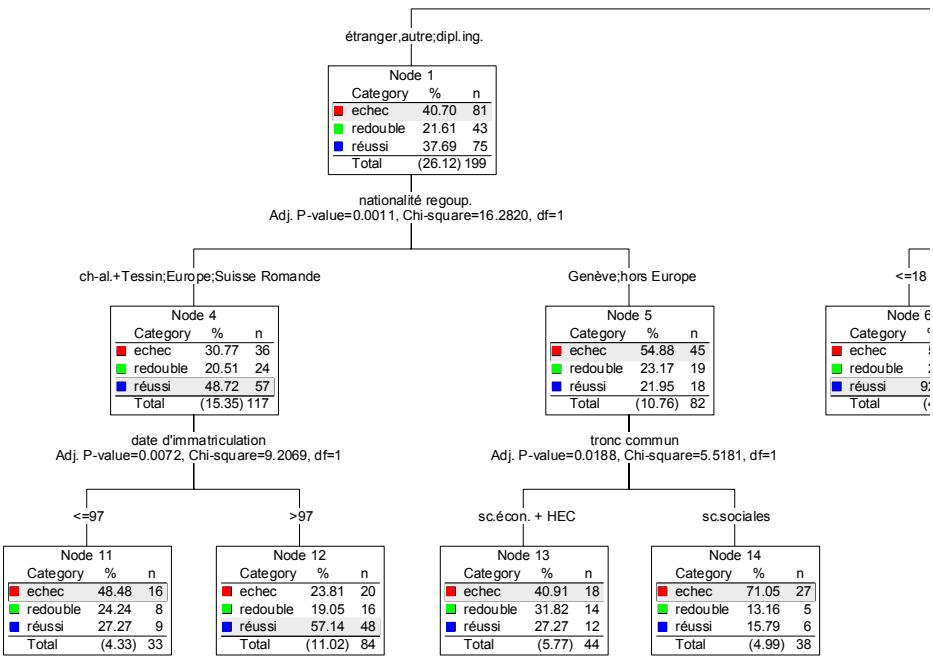


FIG. 4 – Bilan après une année en SES : arbre CHAID et détail de la branche gauche

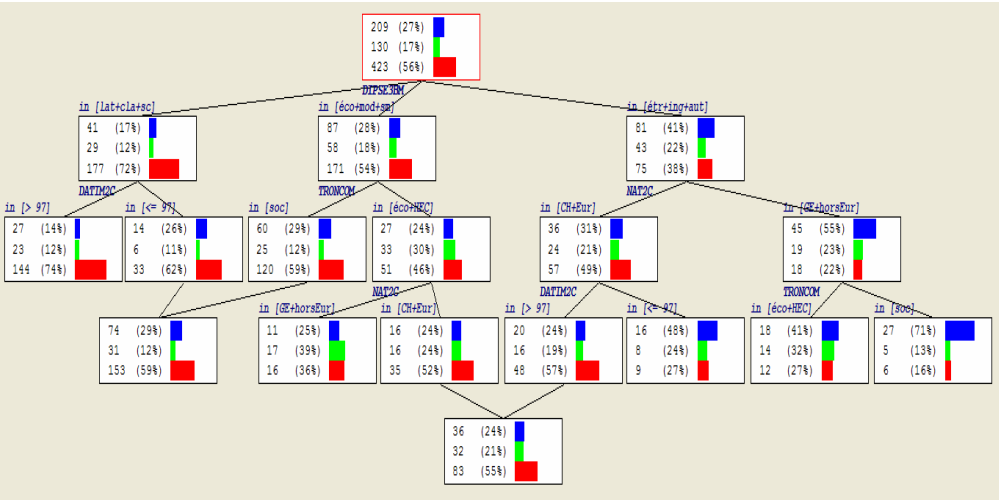


FIG. 5 – Graphe induit avec Sipina

les plus petits AIC et BIC. Enfin, le modèle saturé correspond à la partition la plus fine et le modèle d'indépendance au cas où tous les profils sont regroupés en seul groupe.

On constate tout d'abord qu'à l'exception du modèle d'indépendance et de CHAID3 avec un degré de signification légèrement inférieur à 5%, tous les modèles reproduisent de façon satisfaisante la table **T**. La simplification de l'arbre CHAID en CHAID2 ou CHAID3 se traduit comme attendu par une détérioration du G^2. Les écarts sont $G^2(\text{CHAID2}|\text{CHAID}) = 9.5$ et $G^2(\text{CHAID3}|\text{CHAID}) = 17.3$ qui pour un gain de respectivement 2 et 4 degrés de liberté sont clairement significatifs, ce qui valide statistiquement l'éclatement des nœuds 4 et 5. Les différences de G^2 avec les autres modèles qui ne sont pas des sous graphes de l'arbre CHAID ne peuvent être testés. On peut par contre comparer avec les AIC ou BIC de ces modèles. On remarque tout d'abord que CHAID

Modèle	q	d	G^2	sig(G^2)	AIC	BIC
Saturé	88	0	0	1	528	1751.9
Meilleur AIC	14	148	17.4	1	249.4	787.2
CHAID	9	158	177.9	0.133	390.0	881.3
CHAID2	8	160	187.4	0.068	395.4	877.5
CHAID3	7	162	195.2	0.038	399.2	872.1
Sipina	7	162	185.8	0.097	389.8	862.6
Meilleur BIC	6	164	75.2	1	275.2	738.8
Indépendance	1	174	295.1	0.000	475.8	892.3

CHAID2 : CHAID sans éclatement *datimma* du sommet 4 (*nationa*≠ GE, hors Europe)
CHAID3 : CHAID2 sans éclatement *troncom* du sommet 5 (*nationa*= GE, hors Europe)

TAB. 4 – SES 98 : qualités d'ajustement d'un choix de modèles

Modèle	proportion de réduction d'entropie			relativement au modèle saturé			pseudo
	$\hat{\tau}_{M\|I}$	$\hat{u}_{M\|I}$	$\hat{\lambda}_{M\|I}$	$\hat{\tau}_{M\|I}$	$\hat{u}_{M\|I}$	$\hat{\lambda}_{M\|I}$	R^2_{ajust}
Saturé	0.193	0.197	0.183	1	1	1	1
Meilleur AIC	0.159	0.185	0.179	0.824	0.939	0.978	.941
CHAID	0.094	0.078	0.109	0.487	0.396	0.596	.336
CHAID2	0.086	0.072	0.088	0.446	0.365	0.481	.309
CHAID3	0.08	0.067	0.088	0.415	0.340	0.481	.289
Sipina	0.087	0.073	0.103	0.451	0.371	0.563	.324
Meilleur BIC	0.149	0.147	0.142	0.772	0.746	0.776	.745
Indépendance	0	0	0	0	0	0	0

TAB. 5 – SES 98 : mesures de type R^2 pour un choix de modèles

et ses deux variantes ont des AIC et BIC très voisins, sensiblement meilleurs que ceux du modèle saturé et dans une moindre mesure que ceux du modèle d'indépendance. La partition générée par Sipina obtient un AIC équivalent au modèle CHAID, mais son BIC est inférieur à celui des 3 modèles CHAID. L'écart supérieur à 10 traduit, selon l'échelle postulée par Raftery (1995), une supériorité très forte de cette partition. On peut noter toutefois, que les valeurs des AIC et BIC obtenues pour le graphe Sipina restent très nettement supérieures aux valeurs optimales possibles avec les attributs retenus. Notons cependant que les partitions AIC et BIC optimales ne peuvent être décrites par des arbres, les règles (non données ici) qui caractérisent les classes de ces partitions consistant en des mélanges de conjonctions ('et') et d'alternatives ('ou') de conditions. Les partitions AIC et BIC optimales ont été obtenues avec la procédure décrite dans Ritschard (2003).

Le tableau 5 récapitule les mesures de type R^2. Le $\hat{\lambda}_{M|I}$ qui indique la proportion de réduction du taux d'erreur sur données d'apprentissage est donné pour comparaison avec les proportions de réduction d'entropie. Les proportions maximales de réduction d'entropie par rapport au modèle d'indépendance sont évidemment obtenues avec la partition la plus fine, c'est-à-dire le modèle saturé. On note que ces maxima sont inférieurs à 1. La part de cette réduction maximale réalisée par chaque modèle est également donnée. On voit que ces dernières valeurs sont très similaires au pseudo R^2 ajusté. Elles nous indiquent par exemple, que les modèles CHAID et Sipina, malgré un ajustement satisfaisant, ne captent qu'environ 1/3 du potentiel de réduction d'entropie possible avec les prédicteurs retenus. Les meilleures partitions du point de vue tant du BIC que de l'AIC font nettement mieux de ce point de vue.

6 Étude par simulations des statistiques X^2 et G^2

Nous rapportons ici les principaux enseignements de simulations menées pour étudier empiriquement la distribution des statistiques du khi-2 considérées à la section précédente. L'objectif principal de ces analyses est de conforter empiriquement le nombre de degrés de liberté que nous avons calculé, à savoir $(c-q)(\ell-1)$. Lebart et al. (2000, p. 362) mentionnent par exemple que des calculs similaires dans le cadre de l'analyse des

correspondance a donné des résultats faux. Rappelons que les degrés de liberté d'une variable du χ^2 représentent son espérance mathématique et la moitié de sa variance.

Plusieurs séries de simulations ont été réalisées. Nous présentons ici les résultats pour deux tailles de tables $\mathbf{T}$, soit 2×6 et 3×88, cette dernière étant la taille de la table cible de l'illustration de la section 5. Pour chacun des cas, nous avons considéré plusieurs partitions des profils (colonnes des tables). Pour la petite table, le regroupement en une seule classe (indépendance) et une partition en 3 générée par un arbre. Pour la grande table, l'indépendance et les partitions correspondant à l'arbre CHAID et au meilleur BIC de l'illustration SES 98. A chaque fois, nous avons imposé au niveau de la population l'égalité des distributions de la réponse pour les profils d'une même classe. Pour la grande table, ceci a été fait en modifiant la structure de la population des étudiants SES 98. Nous avons d'autre part aussi étudié le cas d'une population répartie uniformément entre les 264 cases de la table (qui vérifie nécessairement toutes les contraintes.) Dans chacune des populations ainsi définies, nous avons tirés aléatoirement 200 échantillons pour lesquels nous avons calculé les statistiques X^2 et G^2 mesurant l'écart entre la table $\mathbf{T}$ échantillonnée et la table $\hat{\mathbf{T}}$ prédite en imposant la partition.

Le tableau 6 donne la valeur moyenne et la demi-variance des 200 X^2 et G^2 obtenus dans chaque cas ainsi que l'erreur standard des moyennes. Pour chaque série de simulations sont indiqués la taille q de la partition, le nombre zs de zéros structurels et le nombre théorique de degrés de liberté $d = (c-q)(\ell-1) - zs$. Notons que si l'on peut ici déterminer le nombre de zéros structurels, ceci n'est pas le cas avec des données réelles issues de populations qui restent inconnues.

				moyenne				variance/2	
Modèle	q	zs	d	X^2	(err std)	G^2	(err std)	X^2	G^2
table cible 2×6									
indépendance	1	0	5	4.98	(0.22)	4.99	(0.22)	4.69	4.75
arbre	3	0	3	3.05	(0.17)	3.06	(0.17)	2.70	2.70
table cible 3×88, population SES98, taille échantillon 762									
indépendance	1	0	174	173.4	(1.26)	197.5	(1.46)	157.6	123.1
CHAID	9	0	158	142.9	(1.20)	156.7	(1.13)	142.0	126.0
BIC opt.	6	39	125	123.2	(1.24)	135.5	(1.22)	153.4	146.7
table cible 3×88, population SES98, taille échantillon 100000									
indépendance	1	0	174	173.4	(1.45)	173.8	(1.45)	207.9	209.7
CHAID	9	0	158	159.0	(1.20)	159.4	(1.19)	144.2	142.1
BIC opt.	6	39	125	127.2	(1.17)	127.7	(1.18)	135.2	138.7
table cible 3×88, population uniforme, taille échantillon 100000									
indépendance	1	0	174	173.0	(1.32)	173.1	(1.32)	173.7	173.1
CHAID	9	0	158	159.1	(1.24)	159.2	(1.24)	153.2	153.6
BIC opt.	6	0	164	165.2	(1.28)	165.2	(1.28)	163.0	163.2

TAB. 6 – Moyennes et demi-variances observées des statistiques X^2 et G^2. Les valeurs encadrées indiquent les moyennes empiriques de X^2 et G^2 qui s'écartent significativement de la valeur d théorique.

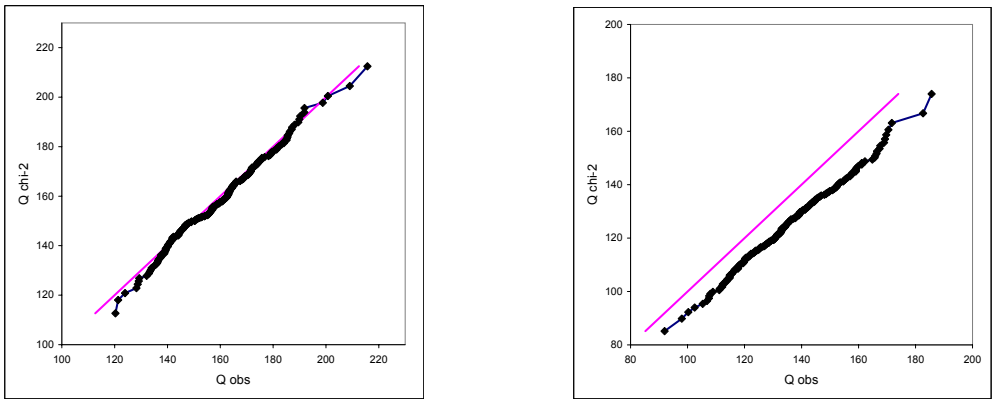


FIG. 6 – QQ-plot : à gauche partition CHAID en 9, échantillons de taille 100000, $d = 158$ degrés de liberté théoriques et à droite partition BIC optimale en 6, échantillons de taille 762, $d = 125$

Les simulations réalisées confirment de façon générale le bien fondé des résultats théoriques, en particulier lorsque l'échantillon est suffisamment grand pour assurer des effectifs attendus conséquents dans chaque cellule. Dans le cas d'échantillons de taille 762 pour la table de 3×88, on a en moyenne moins de 3 cas par case, ce qui est insuffisant pour justifier la distribution du χ^2. Ceci se traduit par des écarts significatifs entre la valeur moyenne et les degrés de liberté théoriques. On remarque également que dans cette situation le G^2 excède la valeur de X^2 de façon importante, de l'ordre de 10%. Les demi-variances, qui devraient théoriquement aussi être égales aux degrés de liberté, sont moins convaincantes excepté pour la population uniforme et, dans une moindre mesure, les petites tables.

La figure 6 montre les qq-plots qui comparent les quantiles observés (Q obs) des G^2 aux quantiles (Q chi-2) de la distribution du χ^2 théorique pour deux cas : à gauche pour la partition CHAID en 9 classes et des échantillons de taille 100000, à droite le cas défavorable de la partition BIC optimale avec des échantillons de taille 762. Le fait que sur le graphique de gauche les points soient pratiquement alignés sur la droite théorique montre que, non seulement la moyenne des G^2 correspond aux degrés de liberté, mais que la distribution empirique est très proche de celle du khi-2. Sur le graphique de droite, on observe que les quantiles observés excèdent systématiquement les quantiles théoriques. Les points étant cependant alignés, la forme de χ^2 de la distribution ne semble pas remise en cause. A titre indicatif, un test d'ajustement de Kolmogorov-Smirnov sur ces deux exemples, donne respectivement un degré de signification de 98% et de 0.0001%. Ce même test de Kolmogorov-Smirnov pour le cas BIC optimal avec échantillons de taille 100000 donne un degré de signification de 65%, bien que pour ce même cas l'écart entre la moyenne et les degrés de liberté soit légèrement supérieur à deux erreurs standards.

Ces analyses par simulations nous invitent à une certaine prudence en ce qui concerne les degrés de signification donnés dans le cadre de l'illustration de la section 5 au tableau 4. Les valeurs indiquées donnent cependant clairement des ordres

de grandeur raisonnables. Pour le G^2, les degrés de liberté observés dans nos simulations sont toujours supérieurs ou égaux aux degrés de liberté théoriques. Les p-valeurs calculées semblent donc être des bornes inférieures ce qui indiquerait que les tests de significativité sont conservateurs. La présence d'éventuels zéros structurels non décelés agit cependant en sens inverse. Une borne supérieure de leur nombre zs est donnée par le nombre zs_{sup} de zéros dans la table prédite $\hat{\mathbf{T}}$. En retenant les degrés de liberté corrigés $(c-q)(\ell-1) - zs_{sup}$, il est néanmoins possible de s'assurer des degrés de signification conservateurs.

7 Conclusion

Cet article aborde la question de la qualité de l'ajustement des arbres d'induction. Il s'agit d'un aspect peu discuté dans la littérature sur l'extraction de connaissances alors même que la qualité d'ajustement fait partie des outils classiques d'évaluation de modèles en statistique. La qualité d'ajustement fournit des indications complémentaires aux indicateurs de qualité traditionnellement utilisés pour les arbres d'induction en permettant, en particulier, d'évaluer la pertinence statistique d'un arbre induit.

Concrètement, nous avons montré, en introduisant les notions d'arbre saturé et d'arbre étendu, comment adapter aux arbres d'induction les statistiques du khi-2 de Pearson et du rapport de vraisemblance utilisés dans le cadre de la modélisation de tables de contingence. Nous avons également considéré la question de la comparaison de modèles pour laquelle la différence des statistiques G^2 du rapport de vraisemblance permet de tester la significativité statistique du gain d'information d'un modèle par rapport à un modèle de référence. Pour la comparaison avec le modèle d'indépendance ne tenant pas compte des prédicteurs, nous avons examinés divers indicateurs de type R^2. Enfin, nous avons vu que l'on pouvait exploiter les critères d'information AIC et BIC pour guider le choix entre arbres de complexité variable.

Ce travail avait pour objectif de montrer comment appliquer des critères statistiques bien établis aux arbres d'induction. Il reste évidemment encore beaucoup à faire. D'une part, il convient de généraliser la mise en œuvre des statistiques et indicateurs discutés dans des cas concrets, et en particulier de les implémenter dans une procédure de construction d'arbres d'induction.

L'approche retenue dans cet article qui s'appuie notamment sur les concepts d'arbre saturé et d'extension maximale de l'arbre, s'applique lorsque le croisement de toutes les modalités des attributs donne lieu à un nombre raisonnable de catégories. Dans le cas particulier de variables quantitatives continues, il y a lieu de discrétiser les valeurs. La difficulté tient ici au fait que, dans les arbres d'induction, la discrétisation ne se fait en règle générale pas a priori mais est déterminée en cours de processus de façon à optimiser la discrimination entre classes.

Une approche possible est de retenir la discrétisation des variables continues définie par l'ensemble des seuils utilisés par le graphe induit. Les variables continues étant ainsi rendues catégorielles, la construction de l'arbre saturé et de l'arbre étendu devient possible et la démarche précédente s'applique. C'est la démarche que nous avons adoptée dans l'illustration de la section 5. Les résultats sont alors conditionnels à la discrétisation retenue, ce qui en limite évidemment la portée. On peut songer à d'autres

approches prenant en particulier en compte le fait que les seuils de discrétisation sont également des paramètres du modèle. Ceci mérite cependant une réflexion approfondie qui dépasse le cadre de cet article

Enfin, de façon plus générale, il reste encore beaucoup de questions ouvertes sur la pertinence statistique des arbres d'induction. Par exemple, la mesure de la fiabilité des estimations des paramètres du modèle de reconstruction (1) issu de l'arbre et celle de la stabilité de l'arbre induit sont à nos yeux essentielles pour apprécier la confiance à accorder à un arbre.

Références

Agresti, A. (1990). *Categorical Data Analysis*. New York : Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox et F. Caski (Eds.), *Second International Symposium on Information Theory*, pp. 267. Budapest : Akademiai Kiado.

Bishop, Y. M. M., S. E. Fienberg, et P. W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge MA : MIT Press.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. New York : Chapman and Hall.

Cressie, N. et T. R. Read (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society 46*, 440–464.

Geurts, P. (2002). *Contributions to Decision Tree Induction : Bias/Variance Tradeoff and Time Series Classification*. Liège : Faculté des sciences appliquées. PhD Thesis.

Goodman, L. A. et W. H. Kruskal (1954). Measures of association for cross classifications. *Journal of the American Statistical Association 49*, 732–764.

Goodman, L. A. et W. H. Kruskal (1972). Measures of association for cross classifications IV : simplification of asymptotic variances. *Journal of the American Statistical Association 67*, 415–421.

Guttman, L. (1941). An outline of the statistical theory of prediction. In P. Horst et others (Eds.), *The Prediction of Personal Adjustment*, Volume 8, New York, pp. 253–318. Social Science Research Council.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Kass, R. E. et A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

Lebart, L., A. Morineau, et M. Piron (2000). *Statistique exploratoire multivariée* (Troisième ed.). Paris : Dunod.

Light, R. J. et B. H. Margolin (1971). An analysis of variance for categorical data. *Journal of the American Statistical Association 66*(335), 534–544.

Olszak, M. et G. Ritschard (1995). The behaviour of nominal and ordinal partial association measures. *The Statistician 44*(2), 195–212.

Petroff, C., A.-M. Bettex, et A. Korffy (2001, Juin). Itinéraires d'étudiants à la faculté des sciences économiques et sociales : le premier cycle. Technical report, Université de Genève, Faculté SES.

Powers, D. A. et Y. Xie (2000). *Statistical Methods for Categorical Data Analysis*. San Diego, CA : Academic Press.

Quinlan, J. R. (1993). *C4.5 : Programs for Machine Learning*. San Mateo : Morgan Kaufmann.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC : The American Sociological Association.

Rakotomalala, R. et D. A. Zighed (1998). Mesures PRE dans les graphes d'induction : une approche statistique de l'arbitrage généralité-précision. In G. Ritschard, A. Berchtold, F. Duc, et D. A. Zighed (Eds.), *Apprentissage : des principes naturels aux méthodes artificielles*, pp. 37–60. Paris : Hermes Science Publications.

Ritschard, G. (2003). Partition BIC optimale de l'espace des prédicteurs. *Revue des nouvelles technologies de l'information 1*, 99–110.

Ritschard, G. et D. A. Zighed (2003). Modélisation de tables de contingence par arbres d'induction. *Revue des sciences et technologies de l'information — ECA 17*(1–3), 381–392.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Sipina for Windows V2.5 (2000). http ://eric.univ-lyon2.fr. Logiciel.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago : SPSS Inc.

Theil, H. (1967). *Economics and Information Theory*. Amsterdam : North-Holland.

Theil, H. (1970). On the estimation of relationships involving qualitative variables. *American Journal of Sociology 76*, 103–154.

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

Summary

This paper is concerned with the fit of induction trees. Namely, we explore the possibility to measure the goodness-of-fit as it is classically done in statistical modeling. We show how Chi-square statistics and especially the Log-likelihood Ratio statistic that is abundantly used in the modeling of contingency tables can be adapted for induction trees. Not only is the Log-likelihood Ratio statistic suited for testing the fit. It allows us also to test the significance of the fit improvement provided by the complexification of a tree. In addition, we derive from it adapted forms of the Akaike (AIC) and Bayesian (BIC) information criteria that prove useful in selecting the best compromise tree between fit and complexity. The practical use of the statistics and indicators proposed is illustrated on an real example.