Les arbres d'induction comme outils de modélisation de tables de contingence multidimensionnelles

Gilbert Ritschard, Dép. économétrie, Université de Genève
juin 2003

Plan

1 Motivation
2 Arbres d'induction et table cible
3 Critères classiques de validation
4 Ajustement de la table cible
5 Mesure et test de la qualité d'ajustement
6 Illustration: le Titanic et SES98
7 Conclusion et perspectives

Ritschard and Zighed (2002),

http://mephisto.unige.ch

1 Motivation

Etude réussite étudiants SES 98

Variable réponse :

- bilan octobre 1999 (éliminé, redouble, réussi)

prédicteurs :

- âge
- date immatriculation
- tronc commun choisi
- type diplôme secondaire
- lieu obtention diplôme secondaire
- âge obtention diplôme secondaire
- nationalité
- domicile de la mère

Cas de données catégorielles (Tableau croisé multidimensionnel)

Sociologues habitués à :

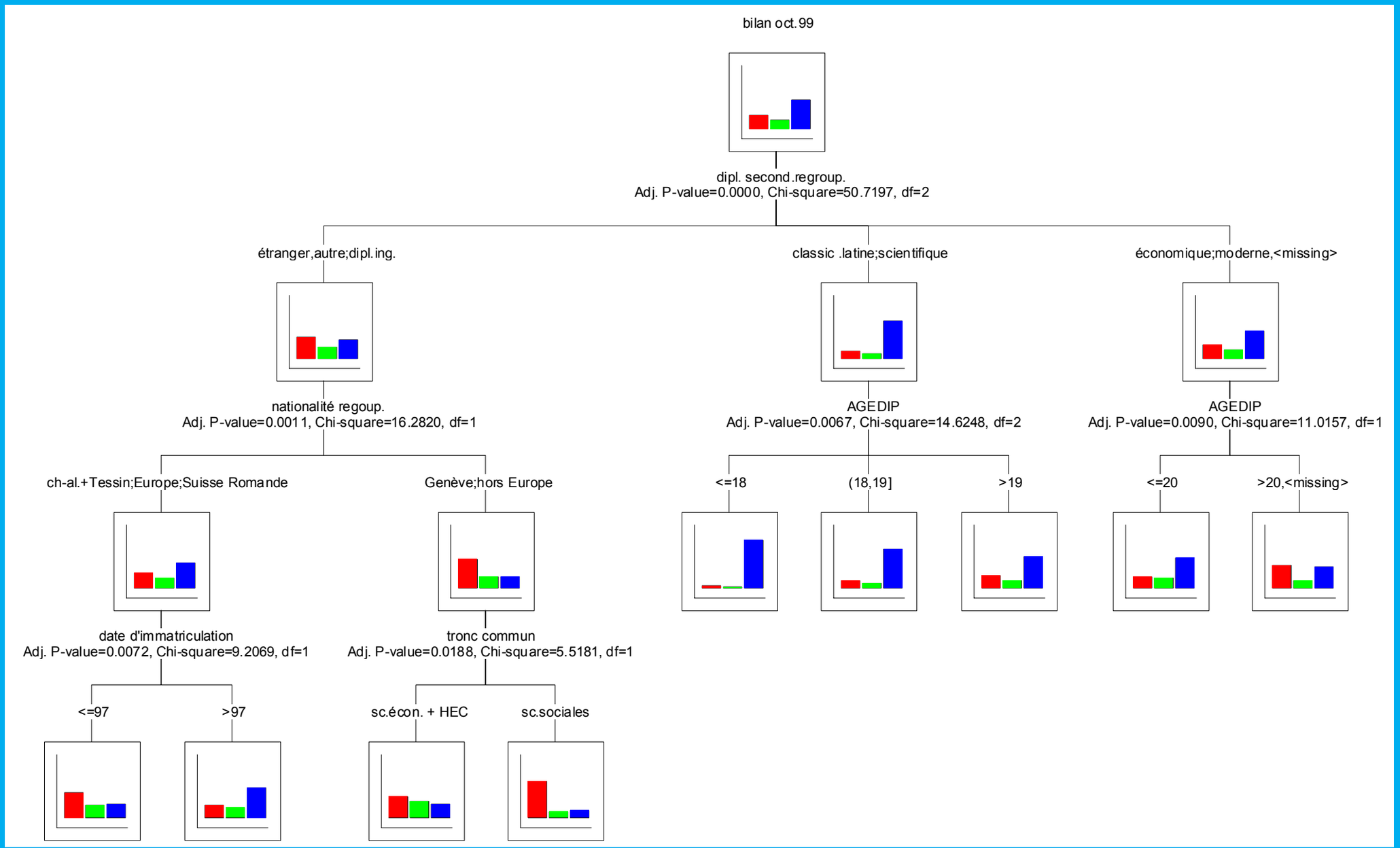
Analyse de la structure d'association $\Rightarrow$ modèles log-linéaires

Avec une variable réponse (catégorielle)

$\Rightarrow$ Régression logistique (binaire, multinomiale)

On peut aussi modéliser ce type de données par des arbres

ou par d'autres procédures relevant de l'apprentissage


bilan oct.99
dipl. second.regroup.
Adj. P-value=0.0000, Chi-square=50.7197, df=2
étranger,autre;dipl.ing.
classic .latine;scientifique
économique;moderne,<missing>
nationalité regoup.
Adj. P-value=0.0011, Chi-square=16.2820, df=1
AGEDIP
Adj. P-value=0.0067, Chi-square=14.6248, df=2
AGEDIP
Adj. P-value=0.0090, Chi-square=11.0157, df=1
ch-al.+Tessin;Europe;Suisse Romande
Genève;hors Europe
<=18
(18,19]
>19
<=20
>20,<missing>
date d'immatriculation
Adj. P-value=0.0072, Chi-square=9.2069, df=1
tronc commun
Adj. P-value=0.0188, Chi-square=5.5181, df=1
<=97
>97
sc.écon. + HEC
sc.sociales

Points communs et différences entre modélisation et apprentissage

Modélisation est apprentissage ont en commun :

- Le contexte : une (des) variable(s) réponse(s) y, des prédicteurs $\mathbf{x}$.
- Recherche d'une fonction $f(\mathbf{x})$ pour prédire ou expliquer les valeurs prises par y.
- Induction de f à partir de données d'apprentissages (estimation)

Les différences portent sur

Modélisation :

- Utilise souvent une approche paramétrique : on postule une forme de distribution de y, sa dépendance par rapport à $\mathbf{x}$ et on estime les paramètres.
- Objectif prioritaire : décrire les mécanismes liant y à $\mathbf{x}$
- Validation par mesure de la qualité d'ajustement (des données d'apprentissage), test d'hypothèses.

Apprentissage supervisé :

- Utilise en général une approche non-paramétrique : pas d'hypothèses sur la forme des distributions (k-ppv, arbres, réseau de neurones). f est le plus souvent considéré comme une boite noire.
- Objectif prioritaire : prédire y (classer) à l'aide de $\mathbf{x}$
- Validation par taux d'erreur (prédiction ou classement) en généralisation

2 Arbres d'induction et table cible

Arbres d'induction : apprentissage supervisé

(Kass (1980), Breiman et al. (1984), Quinlan (1993), Zighed and Rakotomalala (2000), Hastie et al. (2001))

$\Rightarrow$ 1 variable réponse catégorielle y (statut marital)
prédicteurs, attributs catégoriels ou métriques $\mathbf{x} = (x_1, \ldots, x_p)$
(âge, secteur d'activité)

(variable réponse métrique $\Rightarrow$ arbre de régression)

Apprentissage supervisé

A partir d'un échantillon $\{(\mathbf{x}_\alpha, y_\alpha)\}_{\alpha=1,\ldots,n}$,
construire une fonction prédictive (ou de classification) $f(\mathbf{x})$ qui permette de prédire la valeur ou classe y de cas dont on ne connaît que $\mathbf{x}$.
(prédire le statut marital à partir de la classe d'âge et du secteur d'activité)

2.1 Table cible

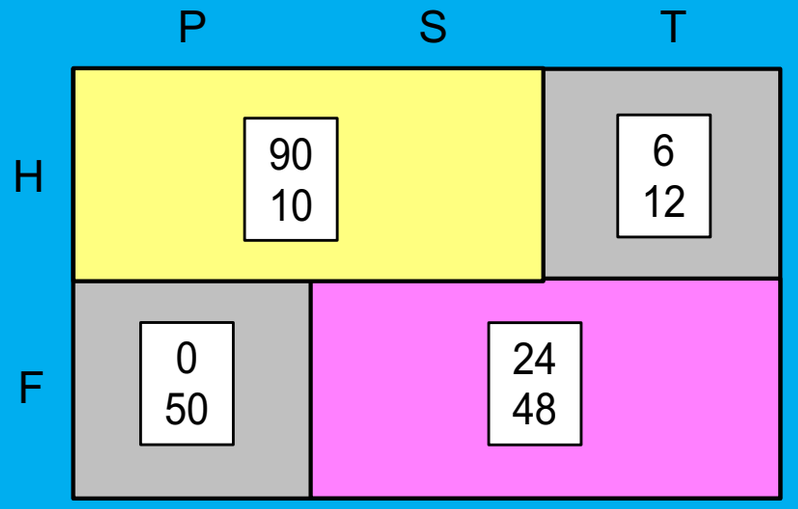
Si toutes les variables sont catégorielles, on peut représenter les données sous forme d'une table de contingence croisant la variable réponse avec une variable composite définie par le croisement de tous les prédicteurs.

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

	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

Arbres d'induction construit la règle $f(\mathbf{x})$ en deux temps :

1. Déterminer une partition des profils possibles $\mathbf{x}$ telle que la distribution p_y de la réponse Y soit la plus différente possible d'une classe à l'autre.



2. La règle consiste ensuite à attribuer à chaque cas la valeur de y la plus fréquente dans sa classe.

$$\hat{y} = f(\mathbf{x}) = \arg\max_i \hat{p}_i(\mathbf{x})$$

2.2 Principe des arbres d'induction

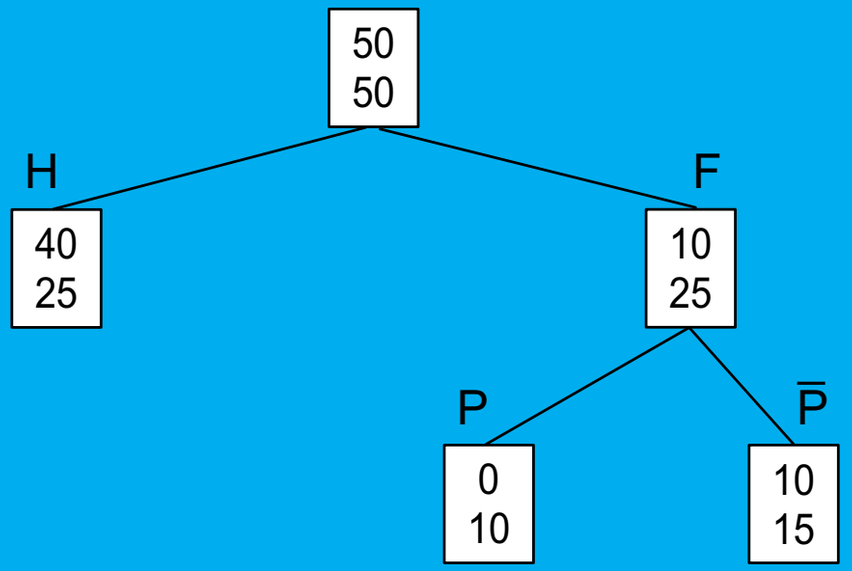


Fig. 1 – Arbre induit

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 donné. 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.

2.3 Les critères

Critères issus de

la théorie de l'information : entropies (incertitude) de la distribution

Entropie de Shannon : $h_S(p) = -\sum_{i=1}^{c} p_i \log_2 p_i$

Entropie quadratique (Gini) : $h_Q(p) = \sum_{i=1}^{c} p_i(1-p_i) = 1 - \sum_{i=1}^{c} p_i^2$

$\Rightarrow$ maximiser la réduction d'entropie (ou entropie standardisée)

Par exemple, C4.5 maximise le Gain Ratio $\left(\frac{h_S(p_y) - h_S(p_y|x)}{h_S(p_x)}\right)$

association statistique Khi-2 de Pearson, mesures d'association

$\Rightarrow$ maximiser l'association, minimiser la p-valeur du test de l'association nulle.

3 Critères classiques de validation

La qualité d'un arbre (graphe) se mesure par

- Performance en classification
- Complexité
- Qualité des partitions

3.1 Performance en classification

Chaque cas est classé dans la catégorie la plus fréquente du sommet final où il se trouve.

Taux d'erreur (pourcents de cas mal classés)

- Sur échantillon d'apprentissage
- Sur échantillon de validation indépendant
- Par validation croisée
- Par bootstrap

On peut comparer avec le taux d'erreur du classement naïf (tous dans catégorie la plus fréquente du sommet initial).

$\Rightarrow$ mesures de type R^2 (proportion de réduction de l'erreur)

- sur l'échantillon d'apprentissage : $\lambda_{Y|P}$ de Goodman-Kruskal, où P est la partition.

3.2 Complexité

Complexité : nombre de sommets, nombre de niveaux, longueur des messages (règles)

On peut réduire la complexité

- a priori en renforçant les critères d'arrêt
  (par exemple nombre maximum de niveaux, taille minimale des sommets)
- a posteriori par des procédures d'élagages
  (procédure automatique par exemple dans CART)

3.3 Qualité des partitions

On peut calculer l'amélioration totale du critère

- Gain d'information entre sommet initial et ensemble des sommets finaux.
- Degré d'association entre partition finale et variable dépendante ($\tau_{Y|P}$ de GK, $u_{Y|P}$ de Theil, v de Cramer, ...).
  remarque : $u_{Y|P}$ et $\tau_{Y|P}$ mesurent respectivement la proportion de réduction de l'entropie de Shannon et de l'entropie quadratique.
- Degré de signification de l'association.

Les logiciels (Answer Tree, Sipina, ...) ne donnent pas ces valeurs et ne permettent pas de récupérer l'information nécessaire (no du sommet final).

Question : Peut-on mesurer la qualité de l'ajustement fourni par un arbre, comme on mesure la qualité d'ajustement d'une régression linéaire ou d'un modèle log-linéaire par exemple ?

Mesures de type R^2 : $\lambda_{Y|P}, \tau_{Y|P}$ et $u_{Y|P}$

$\Rightarrow$ gain par rapport au modèle naïf

Quid de la qualité de reproduction des données (distance prédictions - observations) ?

Peut-on tester la significativité des effets pris en compte par l'arbre ?

4 Ajustement de la table cible

Qualité d'ajustement : capacité du modèle à reproduire les données.

Deux types d'ajustement

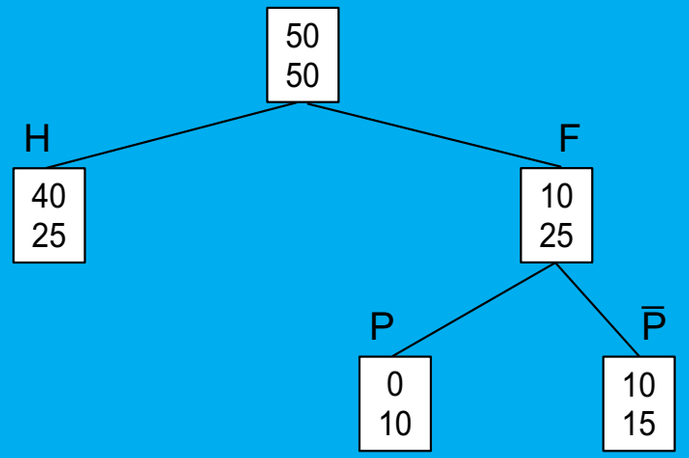
1. ajustement des données individuelles y_α
2. ajustement de la représentation synthétique (table cible $\mathbf{T}$)

En apprentissage supervisé, l'objectif est en général la classification
$\Rightarrow$ ajustement des cas individuels (qualité de la règle $f(\mathbf{x})$).

En sciences sociales, on s'intéresse plutôt aux mécanismes (influences des prédicteurs sur la variable à prédire)
$\Rightarrow$ examiner effets de $\mathbf{x}$ sur distribution de Y
$\Rightarrow$ ajustement de la table de contingence (qualité du modèle $\mathbf{p}(\mathbf{x})$).

4.1 Table générée par l'arbre induit

$\mathbf{T}^a$ table croisant la variable à prédire avec la partition générée par l'arbre.

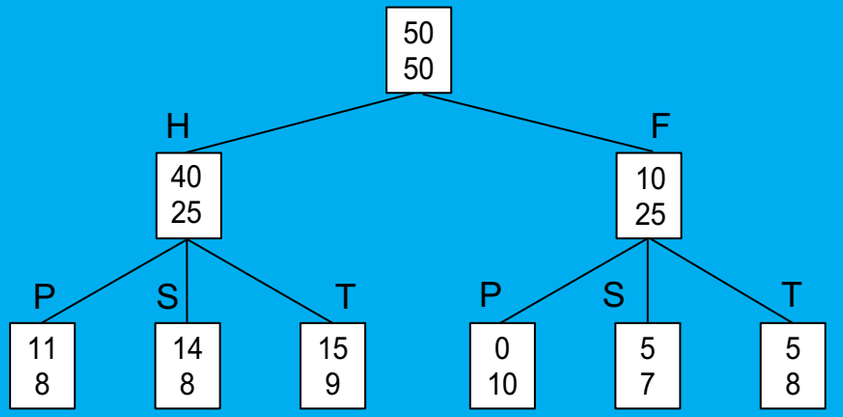


Tab. 2 – Table de contingence générée par l'arbre $\hat{\mathbf{T}}^a$

marié	homme	femme secteur primaire	femme autre secteur	total
non	40	0	10	50
oui	25	10	15	50
total	65	10	25	100

Arbre saturé et table cible

Arbre saturé : arbre qui génère exactement la table cible $\mathbf{T}$

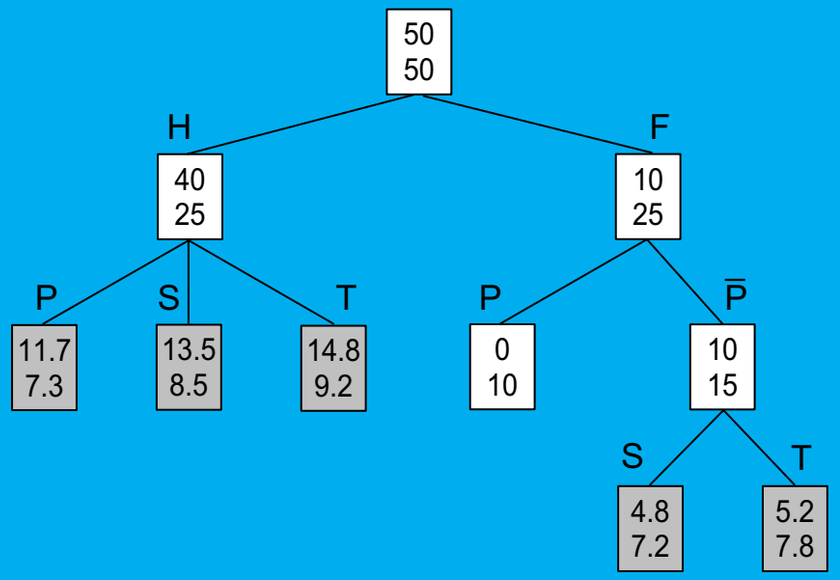


Tab. 3 – Table de contingence cible $\mathbf{T}$

	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

Arbre étendu et table prédite

Arbre induit (sommets blancs)
et son extension maximale



Tab. 4 – Table de contingence prédite $\hat{\mathbf{T}}$

marié	homme			femme			total
	primaire	secondaire	tertiaire	primaire	secondaire	tertiaire	
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

Dans l'arbre étendu, on applique aux feuilles (grises) de l'extension la distribution des sommets (blancs) de l'arbre induit dont ils sont issus

$$\hat{\mathbf{p}}_{|HP} = \hat{\mathbf{p}}_{|HS} = \hat{\mathbf{p}}_{|HT} = \mathbf{p}^a_{|H} = \begin{pmatrix} 40/65 \\ 25/65 \end{pmatrix}$$

$$\hat{\mathbf{p}}_{|FP} = \mathbf{p}^a_{|FP} = \begin{pmatrix} 0/10 \\ 10/10 \end{pmatrix}$$

$$\hat{\mathbf{p}}_{|FS} = \hat{\mathbf{p}}_{|FT} = \mathbf{p}^a_{|F\bar{P}} = \begin{pmatrix} 10/25 \\ 15/25 \end{pmatrix}$$

5 Mesure et test de la qualité d'ajustement

Qualité d'ajustement : distance entre $\hat{\mathbf{T}}$ et $\mathbf{T}$

Mesures de divergence du khi-2 : X^2 de Pearson et G^2 du rapport de vraisemblance (déviance)

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

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

Lorsque le modèle est correct, et sous réserve des conditions de régularité, X^2 et G^2 sont distribuées selon loi du χ^2.

Quels sont les degrés de liberté ?

Modèle de reconstruction et degrés de liberté

L'arbre induit donne lieu au modèle de reconstruction suivant où l'on note $\mathbf{T}_j$ la j-ème colonne de $\mathbf{T}$:

$$\hat{\mathbf{T}}_j = n\, a_j \hat{\mathbf{p}}_{|j}, \quad j = 1, \ldots, c \tag{3}$$

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

où $\mathcal{X}_k$ est la classe de profils $\mathbf{x}$ défini par la $k$ème feuille finale de l'arbre.

Les paramètres sont

- n le nombre total de cas,
- a_j les proportions de cas par colonne $j = 1, \ldots, c$, et
- $\mathbf{p}_{|j}$, les c vecteurs $\mathbf{p}(Y|j)$ de ℓ probabilités définissant la distribution de Y dans chaque colonne j de la table.

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	$c\ell+\ell+c+1$	$q\ell-q+c$

Degrés de liberté $= c\ell$ cellules $-\ (q(\ell-1)+c)$ paramètres indépendants, soit

$$\boxed{d_M=(c-q)(\ell-1)}$$

Ce nombre correspond au nombre de contraintes (4).

Pour modèle d'indépendance : $q=1$ et donc $d_I=(c-1)(\ell-1)$.
Pour l'arbre saturé : $q=c$ et donc $d_S=0$.

5.1 Comparaison de modèles

La statistique du G^2 permet de tester la différence de modèles imbriqués.

Si modèle restreint M_2 correct lorsque M_1 l'est,

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

Permet de tester la significativité d'une expansion (branche).

Exemple : M_1 notre arbre induit et M_2 arbre avant éclatement de «femme»

$G^2(M_1) = 0.18$ avec $d_{M_1} = 3$ et $G^2(M_2) = 8.41$ avec $d_{M_2} = 4$,

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

Degré de signification : $p(\chi^2_1 > 8.23) = .004 \Rightarrow$ effet significatif

Pseudo R^2

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

ou sa version corrigée des degrés de liberté

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

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 $R^2_{\text{ajust}} = .984$.

Critères d'information

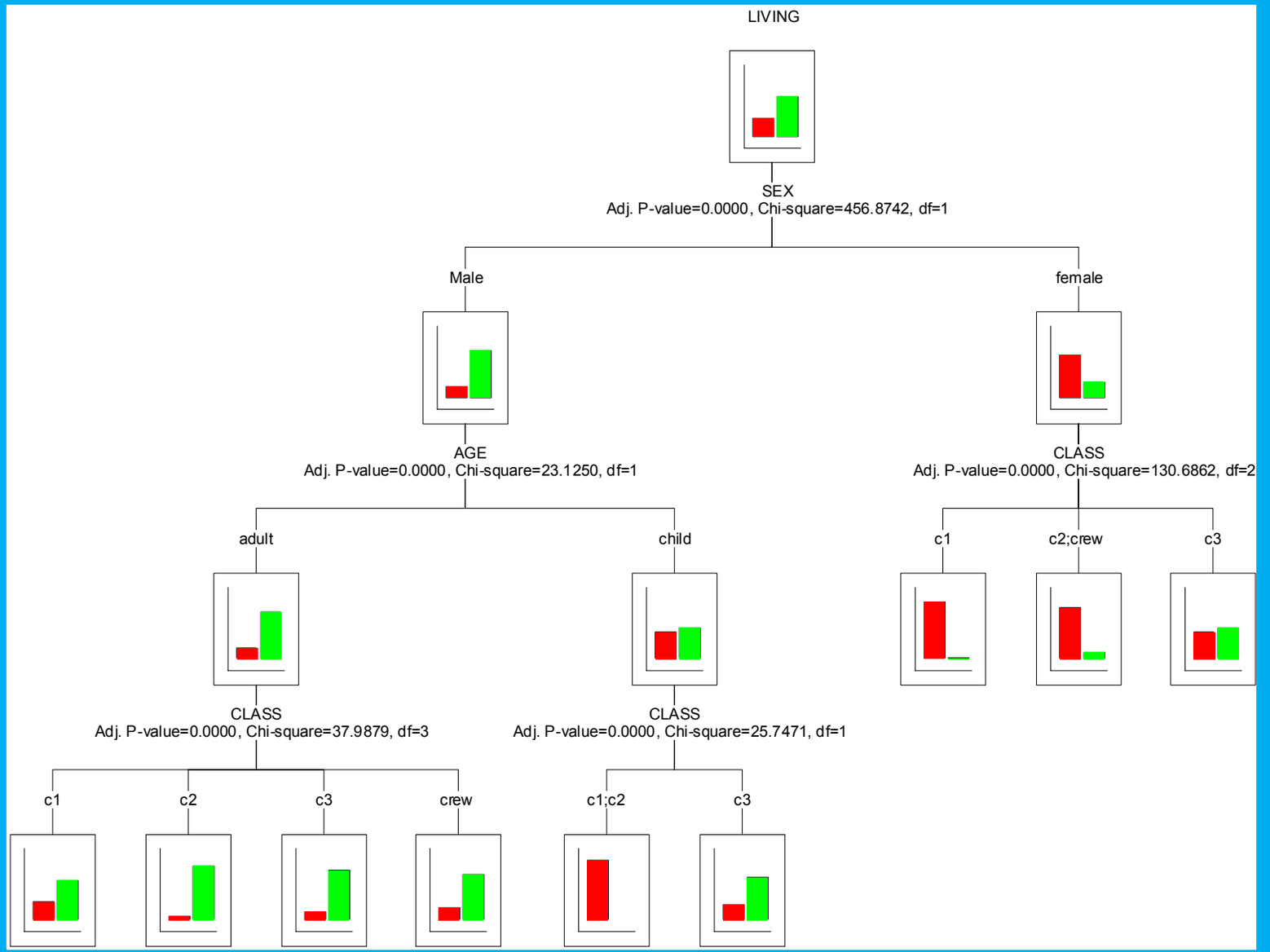
Compromis entre qualité d'ajustement (G^2) et complexité (nbre paramètres indépendants)

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

Permet de comparer des modèles non imbriqués.
$\Rightarrow$ meilleur modèle : celui qui a le plus petit AIC ou BIC.

Akaike (1973), Schwarz (1978), Raftery (1995), Kass and Raftery (1995)

6 Illustration : le Titanic



Tab. 5 – Titanic : effectifs observés et déduits de l'arbre CHAID

feuille					observé		selon arbre		
					living				
j	k	*sex*	*age*	*class*	yes	no	yes	no	Total
1	1	male	adult	c1	57	118	57	118	175
2	2			c2	14	154	14	154	168
3	3			c3	75	387	75	387	462
4	4			crew	192	670	192	670	862
5	5		child	c1	5	0	5	0	5
6	5			c2	11	0	11	0	11
7	6			c3	13	35	13	35	48
8	7	female	adult	c1	140	4	140.03	3.97	144
9	8			c2	80	13	81.47	11.53	93
10	9			c3	76	89	75.77	89.23	165
11	8			crew	20	3	20.15	2.85	23
12	7		child	c1	1	0	0.97	0.03	1
13	8			c2	13	0	11.39	1.61	13
14	9			c3	14	17	14.23	16.77	31
Total					711	1490	711	1490	2201

Tab. 6 – Titanic : qualités d'ajustement d'un choix de modèles

Modèle	d	G^2	sig(G^2)	X^2	sig(X^2)	pseudo R^2_{ajust}	AIC	BIC
CHAID	5	3.72	0.590	2.10	0.835	.986	49.7	180.7
Indépendance	13	671.96	0.000	650.09	0.000	0	702.0	787.4
Saturé	0	0	1	0	1	1	56	215.5
CHAID2	6	35.81	0.000	27.85	0.000	.885	79.8	205.1
CHAID3	6	10.68	0.098	8.44	0.208	.966	54.7	180.0
CART	4	0.08	0.999	0.05	0.999	.999	48.1	184.8
C4.5	6	43.32	0.000	40.10	0.000	.860	87.3	212.6
Sipina	7	5.15	0.642	3.16	0.870	.986	47.2	166.8
Meilleur BIC	8	9.08	0.335	7.82	0.452	.978	49.1	163.0

CHAID2 : regroupe tous les enfants mâles en un seul groupe ($k = 5, 6$).

CHAID3 : regroupe les hommes adultes de 2ème et 3ème classe ($k = 2, 3$).

Exemple étudiants SES 98

Regroupements utilisés par l'arbre ⇒ tableau cible avec 88 colonnes

Tab. 7 – SES 98 : qualités d'ajustement d'un choix de modèles

Modèle	q	d	G^2	sig(G^2)	pseudo R^2_{ajust}	AIC	BIC
Saturé	88	0	0	1	1	528	1751.9
Meilleur AIC	14	148	17.4	1	.941	249.4	787.2
CHAID	9	158	177.9	0.133	.336	390.0	881.3
CHAID2	8	160	187.4	0.068	.309	395.4	877.5
CHAID3	7	162	195.2	0.038	.289	399.2	872.1
Meilleur BIC	6	164	75.2	1	.745	275.2	738.8
Indépendance	1	174	295.1	0.000	0	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)

7 Conclusion et perspectives

- « Arbres » méthode souple pour décrire une table de contingence croisant une variable réponse avec les prédicteurs.
- Il est possible d'utiliser les outils classiques de la statistique pour juger de la qualité de la description fournie de la table.

Développements futurs

- Le cas des prédicteurs numériques (prise en compte de la discrétisation endogène).
- Utilisation des critères d'ajustement lors de l'induction de l'arbre (p.ex : algorithme produisant l'arbre BIC-optimal.)
- Comparaison BIC avec le principe du "Minimum Description Length" (MDL) Rissanen (1983)

Références

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

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

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

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

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

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.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics 11*(2), 416–431.

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

Ritschard, G. and D. A. Zighed (2002). Qualité d'ajustement d'arbres d'induction. Technical report, Groupe Gafo Qualité, CNRS, Paris. 16p.

Ritschard, G. and D. A. Zighed (2003a). Goodness-of-fit measures for induction trees. In Z. W. Ras and N. Zhong (Eds.), *Proceedings of ISMIS 2003*, pp. 1–8. Berlin: Springer-Verlag. forthcoming.

Ritschard, G. and D. A. Zighed (2003b). Modélisation de tables de contingences 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.

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