Usage non classificatoire d'arbres de classification : enseignements d'une analyse de la participation féminine à l'emploi en Suisse

Fabio B. Losa*, Pau Origoni*, Gilbert Ritschard**

*Office statistique du canton du Tessin, Bellinzona, Suisse
fabio.losa@ti.ch ; pau.origoni@ti.ch
**Département d'économétrie, Université de Genève
gilbert.ritschard@themes.unige.ch

Résumé. Cet article présente une application en grandeur réelle des arbres de classification dans un contexte non classificatoire. Les arbres générés visent à mettre en lumière les différences régionales dans la façon dont les femmes décident de leur participation au marché du travail. L'accent est donc mis sur la capacité descriptive plutôt que prédictive des arbres. L'application porte sur des données relatives à la participation féminine au marché du travail issues du Recensement Suisse de la Population de l'an 2000. Ce vaste ensemble de données a été analysé en deux phases. Un premier arbre exploratoire a mis en évidence la nécessité de procéder à des études séparées pour les non mères, les mères mariées ou veuves, et les mères célibataires ou divorcées. Nous nous limitons ici aux résultats de ce dernier groupe, pour lequel nous avons généré un arbre séparé pour chacune des trois régions linguistiques principales. Les arbres obtenus font apparaître des différences culturelles fondamentales entre régions. Du point de vue méthodologique, la principale difficulté de cet usage non classificatoire des arbres concerne leur validation, puisque le taux d'erreur de classification généralement retenu perd tout son sens dans ce contexte. Nous commentons cet aspect et illustrons l'usage d'alternatives plus pertinentes et facilement calculables.

1 Introduction

Les arbres de décision sont, depuis leur popularisation par Breiman et al. (1984), devenus des outils multivariés privilégiés pour prédire la valeur de variables continues ou la classe de variables catégorielles à partir d'un ensemble de prédicteurs. On parle d'*arbre de régression* quand l'attribut à prédire est quantitatif et d'*arbre de classification* lorsqu'il est catégoriel. Bien que leur but premier soit la prédiction et la classification, les arbres présentent bien d'autres intérêts, comme méthode exploratoire pour partitionner et identifier des structures locales dans les bases de données, mais aussi comme alternative aux méthodes statistiques classiques comme la régression linéaire ou logistique par exemple (Wilkinson, 1992).

Cette contribution illustre cet usage *non-classificatoire* des arbres de classification en présentant une application réelle sur des données relatives à la participation féminine

au marché de l'emploi issues du Recensement Suisse de la Population (RSP) de l'an 2000. Le but n'étant pas d'extraire des règles de classification, mais la compréhension des spécificités culturelles des mécanismes qui guident le choix des femmes à participer ou non à l'emploi, le taux d'erreur de classification devient inadéquat comme critère de validation. Nous exploitons pour cela des critères d'ajustement mieux adaptés comme ceux proposés dans Ritschard et Zighed (2003, 2004). Notre expérience fournit des indications nouvelles sur les limites et la praticabilité de ces critères dans le cadre d'applications à grande échelle.

En plus de ces aspects méthodologiques, l'expérience pratique rapportée ici s'avère originale de trois points de vue : 1) le domaine d'application — l'analyse microéconomique — qui est un domaine où les arbres d'induction n'ont jusqu'ici que rarement été utilisés, 2) la génération d'arbres sur un jeu complet de données de recensement de population, 3) un schéma de recherche articulé autour d'un double recours aux arbres, une première fois à des fins exploratoires, puis comme outils d'analyse.

La section 2 rappelle brièvement le principe des arbres de classification et ses principales utilisations. A la section 3, nous présentons les objectifs de notre recherche socio-économique, esquissons le schéma de recherche retenu et commentons les enseignements principaux de l'étude. La section 4 est dévolue au problème de la validation des arbres. Finalement, nous concluons à la section 5 avec une évaluation globale de notre expérience et plus généralement de l'utilisation des arbres de classification à des fins non classificatoires.

2 Arbres de classification, principes et usages

Les arbres de classification sont construits en cherchant, par éclatements successifs de l'ensemble d'apprentissage, des partitions de l'espace des prédicteurs optimales pour prédire la modalité de la variable réponse. Chaque éclatement se fait selon les valeurs d'un prédicteur. Le processus est glouton. A la première étape on teste tous les prédicteurs pour trouver le meilleur, puis le processus est répété à chaque nouveau nœud jusqu'à ce qu'un critère d'arrêt soit satisfait. La détermination du meilleur éclatement à chaque nœud se fait selon un critère local. Le choix du critère est la principale différence entre les diverses méthodes existantes d'induction d'arbres. Parmi celles-ci, CHAID (Kass, 1980), CART (Breiman et al., 1984) et C4.5 (Quinlan, 1993) sont probablement les plus populaires. Pour notre application, nous avons utilisé CART. Cet méthode construits des arbres binaires en choisissant à chaque étape l'éclatement qui maximise le gain en pureté mesuré par l'indice de Gini. CART utilise des règles d'arrêt relativement lâches, mais procède ensuite à un élagage de l'arbre.

Un des intérêts majeurs des arbres d'inductions est qu'ils fournissent des résultats sous une forme visuelle aisément interprétables. Cette visualisation, par comparaison par exemple avec les coefficients d'un modèle de régression, a des avantages certains qui facilitent le processus d'extraction de connaissances. De plus, par leur nature, les arbres fournissent une description spécifique de la façon dont les prédicteurs interagissent sur la variable réponse.

Fabbris (1997) distingue huit usages différents des arbres de classification : 1) définir des règles de prédiction et de classification, 2) identifier des groupes déviants, 3) détecter

les cas atypiques, 4) estimer des valeurs manquantes, 5) suggérer des hypothèses de recherche et des modèles explicatifs, 6) étudier les interactions entre prédicteurs, 7) synthétiser l'information de la base de données et 8) tester des relations non linéaires entre prédicteurs et variable à prédire. Les points 2 à 8 sont ce que nous appelons des buts non classificatoires. Ainsi, en plus de leur but prédictif, les arbres s'avèrent utiles pour analyser les liens entre variables prédictives et à prédire, pour structurer les données et pour caractériser des faits empiriques méritant des analyses plus fines.

Comme dans toute modélisation statistique, il est essentiel de s'assurer de la qualité de l'arbre généré avant d'en tirer des conclusions. De ce point de vue, il est important de souligner que les critères de validation utilisés doivent être adaptés à l'usage que l'on veut faire de l'arbre. En particulier, le taux d'erreur de classification, qui est le plus souvent le seul que fournissent les logiciels, n'a guère de sens dans un contexte non classificatoire pour lequel des alternatives plus pertinentes sont discutés à la section 4.

3 L'étude appliquée

3.1 La participation féminine à l'emploi en Suisse

La participation féminine au marché de l'emploi est sujette à des différences significatives entre pays. En Europe, les spécialistes distinguent au moins deux modèles généraux : un modèle méditerranéen (Italie, Grèce, Portugal,...) et un modèle typique de l'Europe centrale et nordique (Reyneri, 1996). Le premier est caractérisé par une courbe en L inversé de l'*activité* ou du *taux de participation* par âge, où après une courte période à taux élevé (à l'entrée sur le marché de l'emploi) la proportion de femmes travaillant ou cherchant un emploi décline fortement jusqu'à la retraite. Le même graphique fait apparaître une courbe en M pour les pays de l'Europe centrale et nordique. La participation est élevée à l'entrée, elle est suivie par un déclin temporaire pendant la période de maternité puis par un retour à l'emploi jusqu'à un certain âge où commence le processus de sortie définitive du marché de l'emploi. Malgré des différences historiques, politiques, administratives, etc., entre pays européens, ces observations empiriques démontrent l'existence de modèles culturels de la participation féminine à l'emploi, découlant des arbitrages entre les valeurs fondamentales que sont la maternité, la famille, les relations hommes-femmes et le travail rémunéré.

De ce point de vue, la Suisse est un cas intéressant. D'une part, il y a trois langues principales, chacune étant parlée dans une région distincte : le français à l'ouest du côté de la frontière française, l'allemand au nord et à l'est du côté des frontières allemande et autrichienne, et l'italien au sud des Alpes du côté de l'Italie. D'autre part, la Suisse se situe à cheval sur les Alpes qui séparent naturellement l'Europe du sud de celle du nord et du centre.

L'existence de trois régions avec des origines historiques, et des particularismes sociaux et culturels très distincts, ainsi que le fait que la partie italophone soit séparée du reste de la Suisse par les Alpes, sont autant d'éléments qui confèrent un intérêt particulier à l'étude des disparités culturelles dans la participation féminine au marché du travail en Suisse. De plus, une étude entre régions d'un même pays garantit, malgré les disparités dues au système fédéraliste suisse, l'homogénéité de l'environnement

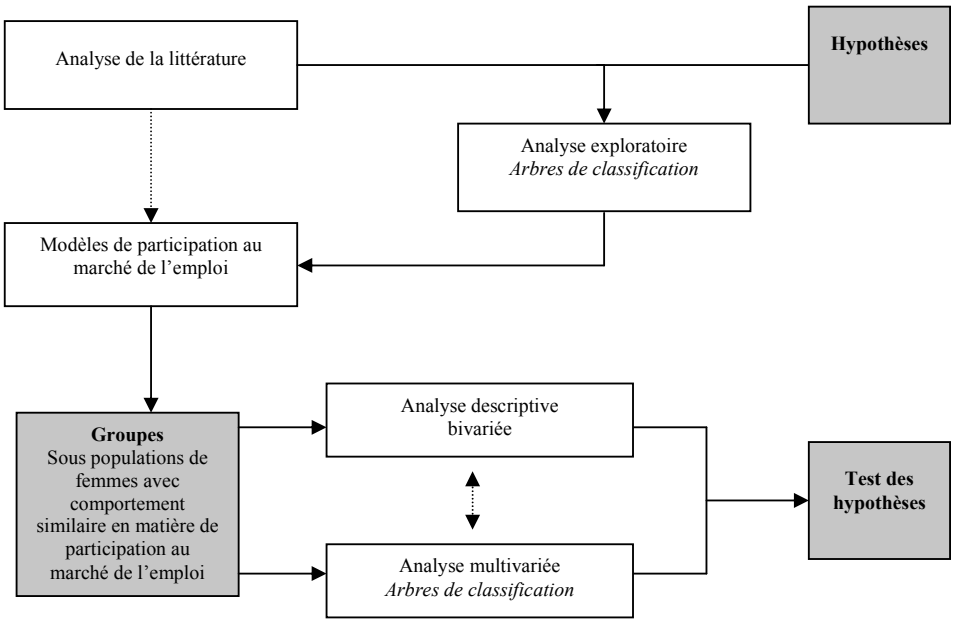

FIG. 1 – *La procédure de recherche appliquée*

institutionnel et politique des régions étudiées qui permet de contrôler, en partie du moins, les effets parasites de cet environnement.

L'objectif du projet de recherche était de vérifier, sur la base des données du RSP 2000, l'existence de modèles culturels pour la participation féminine à l'emploi propres à chacune des trois régions. Pratiquement, pour mesurer la participation, nous avons considéré la proportion de femmes actives et leur répartition entre temps plein et temps partiels.

3.2 Le double recours aux arbres de classification

L'étude que nous présentons est un exemple non seulement de l'utilisation non classificatoire des arbres de classification mais aussi d'un protocole d'étude (figure 1) qui s'articule autour d'un double recours aux arbres avec des buts différents mais complémentaires.

L'algorithme CART a été utilisé tout d'abord dans une phase préliminaire du processus,[1] dans le but de déterminer une partition pertinente de l'ensemble de la population en un nombre limité de groupes aussi homogènes que possible du point de vue des comportements de choix entre participer ou non à l'emploi et entre participer à plein temps ou à temps partiel, l'idée étant de faire ensuite des études approfondies de chaque groupe. En décomposant ainsi l'étude, il s'agissait essentiellement d'éviter de

[1]Le choix de CART a été motivé par le grand pouvoir discriminant et la facilité de lecture des arbres binaires qu'il génère.

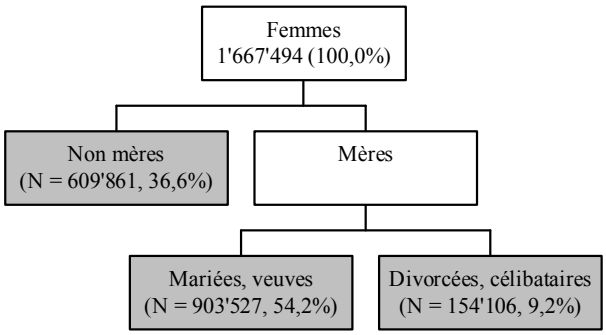

FIG. 2 – *Les trois groupes analysés*

ne capter que des effets moyens ou agrégés, jouant au niveau global et masquant les particularismes recherchés.

L'étude porte sur la population féminine suisse âgée de 20 à 61 ans, pour laquelle on a analysé le *statut sur le marché du travail*[2] (variable réponse) en fonction d'une série de prédicteurs potentiels : âge, profession, niveau d'éducation, mère/non-mère, nombre d'enfants, âge du dernier enfant, type de ménage, etc. Il en est ressorti une partition en trois groupes comme meilleur compromis entre niveau de détail pour les analyses subséquentes et taille des sous populations concernées. Les trois groupes sont les *non-mères*, les *mères mariées ou veuves*, et les *mères célibataires ou divorcée* (figure 2).

La seconde utilisation des arbres de classification s'est faite dans le cadre de l'analyse du comportement en termes de taux d'activité de chaque groupe de femmes. Avant d'induire des arbres, nous avons procédé à des analyses bivariées entre le statut d'emploi et les prédicteurs potentiels. Ceci nous a permis d'identifier les attributs les plus pertinents dans la perspective de mettre à jour les différences culturelles. L'analyse de ces impacts bruts sur le statut d'emploi a fourni déjà de premières indications sur le comportement des femmes et a conduit à l'identification de sous groupes déviants.

Des arbres de classification ont ensuite été induits séparément pour chacune des trois régions de Suisse et pour chacun des trois groupes de femmes. Leur comparaison nous a permis, comme décrit ci-après, de mettre en évidence les similitudes et particularités des schéma de comportement en vigueur dans chacune des régions en matière de participation féminine à l'emploi. On peut relever ici que la complémentarité des enseignements fournis par les arbres et les méthodes classiques de description statistique s'est avérée particulièrement stimulante et productive pour l'extraction de connaissances pertinentes et la compréhension du phénomène étudié.

3.3 Résultats

Nous discutons d'abord les caractéristiques des trois groupes de femmes avant d'analyser celui des mères célibataires ou divorcées.

[2]Le *statut sur le marché du travail* est une variable catégorielle avec quatre modalités : actif à plein temps (au moins 90% de la durée standard de travail hebdomadaire), temps partiel long (50% à 89%), temps partiel court (moins de 50%) et non actif, où actif signifie travailler ou chercher un emploi.

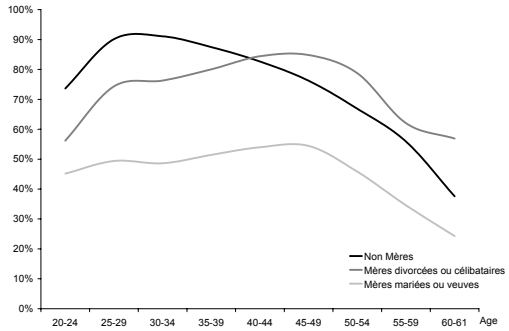

FIG. 3 – *Taux de participation selon l'âge, pour les trois groupes sélectionnés*

3.3.1 Caractéristiques des groupes analysés

Le recours à un arbre de classification dans l'étape préliminaire, qui nous a conduits à caractériser les trois groupes décrits à la figure 2, s'est avéré fort utile. La focalisation sur chacun de ces trois groupes de notre population féminine nous a permis d'éviter une échelle d'analyse trop générique en déterminant des groupes particulièrement pertinents car exhibant une forte diversité inter groupes et une remarquable homogénéité intra groupes. La forte diversité est illustrée notamment par les taux de participation à l'emploi selon l'âge représentés à la figure 3.[3]

Ce contraste est encore renforcé par l'examen des degrés d'occupation qui fait apparaître des choix très différents de la part des femmes en emploi de chacun des groupes. Les non mères pratiquent essentiellement le plein temps durant toute leur carrière professionnelle, les mères célibataires ou divorcées passent du temps partiel pendant la période de grossesse et de maternité au plein temps (ou temps partiel long), et les mères mariées ou veuves optent dans la majorité des cas pour le temps partiel court.

3.3.2 Les déterminants des choix des mères célibataires ou divorcées

L'identification de modèles culturels de la participation des femmes à la force de travail a été réalisée en induisant un arbre par région linguistique pour chacun des trois groupes.[4] Les arbres obtenus — en combinaison avec les analyses uni- et bivariées — ont été méticuleusement examinés et comparés en termes de structure et de distribution dans les noeuds. Nous donnons ci-après un bref aperçu des principaux enseignements pour le groupe des mères célibataires ou divorcées.[5] Le lecteur intéressé par plus de

[3]La figure 3 montre également que les courbes en M ou L que l'on rencontre au niveau d'études nationales résultent en fait de la superposition de courbes de groupes spécifiques.

[4]Les arbres ont été initialement induits pour les trois régions confondues en incluant la région de résidence comme attribut prédictif. Cette analyse préliminaire a confirmé le rôle significatif de cette variable qui est retenue comme meilleur prédicteur pour éclater plusieurs nœuds intermédiaires.

[5]Pour des raisons de place, nous ne présentons ici, figures 4 et 5, que les arbres (légèrement simplifiés) des régions italophones et francophones de la Suisse. Un fonds blanc indique les nœuds où la majorité des femmes est non active, le gris clair les cas où l'on a une majorité de temps partiels, et le gris foncé les cas avec une majorité de temps pleins.

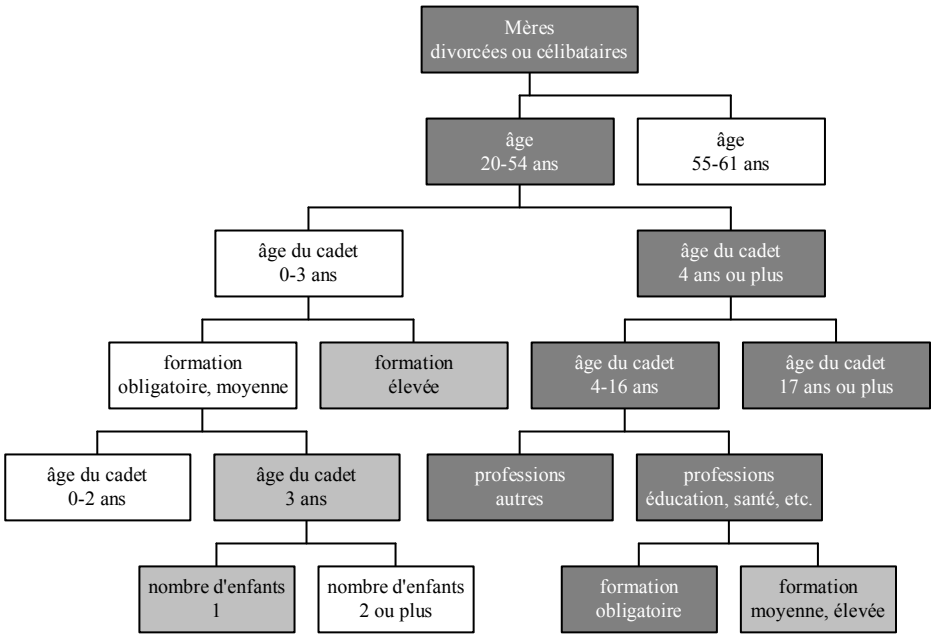

FIG. 4 – *Arbre pour la participation des mères célibataires ou divorcées, Suisse italienne*

détails peut consulter le rapport complet de l'étude (Losa et Origoni, 2004).

Parmi tous les attributs considérés, les arbres induits pour les mères célibataires ou divorcées n'en ont exploités effectivement que quelques uns, soit l'âge du cadet des enfants, le niveau d'éducation, le type de ménage, le nombre d'enfants, la profession, l'âge et le statut d'emploi du partenaire.

Les attributs *profession* et *âge de la mère* caractérisent des groupes à comportement très distincts. Le premier met en lumière le groupe des professions dans des domaines tels que la santé, l'éducation et les sciences, où le temps partiel est très répandu. L'âge joue un rôle central dans l'arbre de la Suisse italienne et dans celui (non présenté) de la Suisse alémanique, en faisant ressortir une période active (jusqu'à 54-55 ans) et une période de retrait définitif du marché de l'emploi.

L'*âge du cadet* ressort clairement comme le facteur le plus discriminant dans les trois régions, soulignant l'importance du fait d'être mère dans le partage du temps entre famille et travail pour ces femmes seules à assumer la responsabilité du ménage. La différence la plus significative entre régions linguistiques porte clairement sur cette variable, qu'il s'agisse de sa position dans l'arbre, des seuils de discrétisation retenus ou des distributions dans les classes des partitions engendrées.

Le *niveau d'éducation* a dans les trois régions une forte influence sur la participation des femmes à la force de travail. Plus le niveau est élevé, plus la proportion d'actives est grande et plus le taux de plein temps est faible. Ce double effet est particulièrement marqué en Suisse italienne et Suisse alémanique lorsque le dernier enfant est très jeune (respectivement moins de 4 et 6 ans). Les mères avec un niveau d'éducation élémentaire

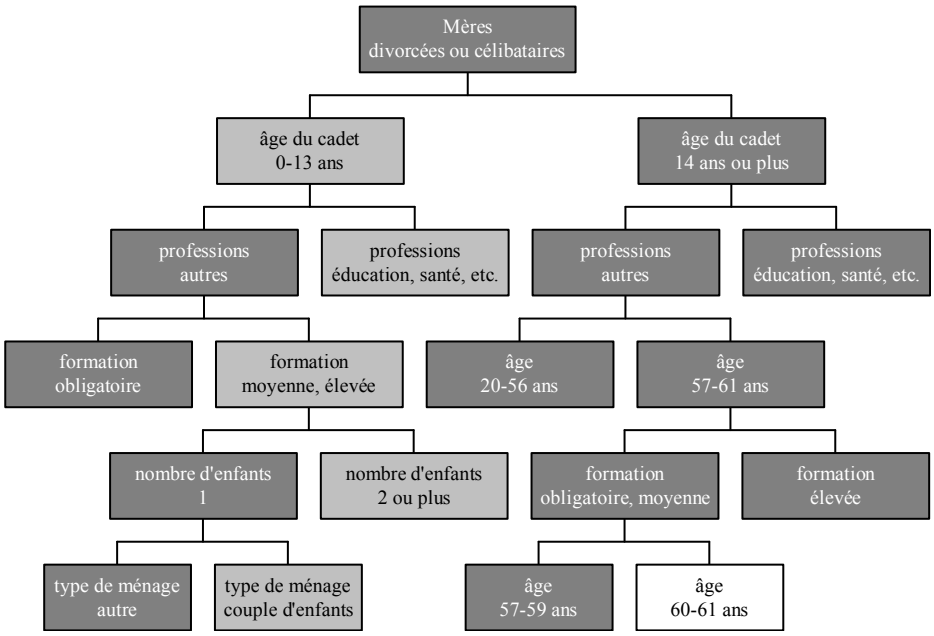

FIG. 5 – *Arbre pour la participation des mères célibataires ou divorcées, Suisse romande*

ou intermédiaire décident dans la majorité des cas de quitter leur emploi pour rester à la maison pendant cette période, tandis que les mères plus instruites optent pour le temps partiel.

La *présence d'un partenaire* dans le ménage exerce un effet négatif tant sur la décision de participer que sur le taux d'activité. L'effet est clair pour les mères mariées. Pour les mères célibataires ou divorcées, l'effet intervient dans les arbres de la Suisse alémanique et de la Suisse romande (francophone) lorsqu'il y a un seul enfant très jeune et que la mère a dépassé le niveau élémentaire d'éducation. Il introduit une distinction entre les familles traditionnelles et les autres types de ménages (avec un seul parent).

On relève un effet similaire du *nombre d'enfants*, mais avec une nuance culturelle : tant que le cadet est jeune, les mères suisses italiennes restent plus volontiers à la maison. Les femmes des autres régions, par contre, ne quittent pas leur emploi et se contentent de réduire leur volume hebdomadaire de travail. Celles de Suisse romande optent plus souvent pour le temps partiel long, et les suisses alémaniques pour le temps partiel court.

4 Validation d'arbres en contexte non classificatoire

Comme nous l'avons déjà relevé, le taux d'erreur de classification n'est pas satisfaisant pour évaluer des arbres dans un contexte non classificatoire. Par exemple, si la modalité majoritaire de la variable réponse est la même dans toutes les feuilles de l'arbre, la réduction du taux d'erreur de classification que permet l'arbre par rapport

au nœud initial sera évidemment nulle. Malgré ce gain nul en termes de classification, l'arbre peut faire apparaître des différences significatives entre les distributions au sein des feuilles. Ces différences sont des informations utiles du point de vue descriptif et il convient donc d'en tenir compte.

Ritschard et Zighed (2004) proposent diverses alternatives qui reposent sur une mesure de déviance. Cette dernière évalue en fait à quelle distance l'arbre ajusté se trouve du tableau cible associé à la partition la plus fine qui puisse être générée à partir des données. Bien que l'idée soit intéressante, le calcul de la déviance requiert de pouvoir construire le tableau cible, ce qui est relativement simple tant que l'on utilise qu'un nombre limité d'attributs avec chacun un petit nombre de valeurs. Dans notre application en grandeur réelle, la construction du tableau cible s'avéra cependant être une tâche pratiquement irréalisable, la combinaison des modalités des attributs considérés donnant en effet lieu à plus d'un million de profils différents, c'est-à-dire de colonnes de la table cible.

Pour pallier à cette limite, nous avons considéré une déviance partielle $D(m|m_{T^*})$ qui mesure l'écart par rapport à la partition m_{T^*} définie par les seules valeurs de partage utilisées dans l'arbre induit. En d'autres termes, on compare la partition définie par l'arbre avec la partition la plus fine que l'on peut atteindre en combinant les classes de valeurs engendrées par l'arbre. La table cible ainsi obtenue est évidemment quelque peu arbitraire, avec pour conséquence que la déviance partielle n'a pas de véritable sens en soi. Cependant, en notant m_T la vraie table cible, nous avons grâce à une propriété d'additivité de la déviance, $D(m|m_T) = D(m|m_{T^*}) + D(m_{T^*}|m_T)$. Ainsi, la différence entre les déviances partielles de deux arbres emboîtés m_1 et m_2 reste la même, quelle que soit la table cible m_{T^*} utilisée. En comparant par exemple la déviance de l'arbre à celle du nœud initial on peut alors tester la significativité du gain de l'arbre par rapport au cas où l'on ne tient compte d'aucun attribut prédictif. Il s'agit là en fait d'un test du khi-deux du rapport de vraisemblance similaire à ceux utilisés par exemple en régression logistique.

Pour nos arbres, nous avons déterminé les déviances partielles à l'aide de SPSS. Deux déviances ont été calculées, soit $D(m_0|m_{T^*})$ et $D(m_0|m)$, où m_0 est le nœud initial et m l'arbre induit. Nous avons tout d'abord recodé les attributs de façon à regrouper les modalités qui restent ensemble dans tout l'arbre. Il fut facile ensuite de construire une variable profil qui prend une valeur différente pour chaque combinaison observée des variables recodées. La table cible (m_{T^*}) est précisément la table de contingence qui croise cette variable profil avec la variable réponse, à savoir ici le type de participation au marché du travail. La déviance $D(m_0|m_{T^*})$ est donnée par la statistique du khi-deux du rapport de vraisemblance (RV) du test d'indépendance sur cette table cible. De même, la déviance $D(m_0|m)$ est la statistique RV pour la table qui croise la variable feuille (qui donne le numéro de la feuille d'appartenance) avec la variable réponse. Comme les arbres ont été construits avec Answer Tree (SPSS, 2001), on a déterminé cette variable feuille avec le code SPSS généré par ce logiciel. La déviance $D(m|m_{T^*})$ s'obtient, elle, par différence des deux déviances calculées, $D(m|m_{T^*}) = D(m_0|m_{T^*}) - D(m_0|m)$, une relation similaire permettant d'en calculer les degrés de liberté.

La déviance partielle peut également être exploitée pour définir des critères AIC et BIC, puisque ces derniers sont définis à une constante additive près. Nous avons ainsi

calculé le BIC pour un arbre m comme $\text{BIC}(m) = D(m|m_{T^*}) - \ln(n)(c^* - q)(\ell - 1)$, où n est le nombre de cas, c^* le nombre de profils différents dans la table cible m_{T^*}, q le nombre de feuilles de l'arbre, ℓ le nombre de modalités de la variable réponse, soit dans notre cas les 4 types de participation au marché du travail. Le produit $(c^* - q)(\ell - 1)$ définit les degrés de liberté associés à la déviance partielle. Rappelons que selon Raftery (1995), une différence entre valeurs du BIC supérieure à 10 indique que le modèle avec le plus petit BIC est manifestement un meilleur compromis entre qualité d'ajustement et complexité.

Il est également éclairant de mesurer le gain d'information en termes relatifs. Des pseudo R^2 calculés à partir de déviances partielles ne sont pas très utiles, en raison du caractère arbitraire de la table cible. Il est préférable de considérer le pourcentage de réduction de l'incertitude quant à la distribution de la variable réponse que permet l'arbre par rapport au nœud initial. La mesure d'association τ de Goodman et Kruskal (1954) et le coefficient d'incertitude u de Theil (1970), tous deux fournis par SPSS, en sont deux exemples. Le premier donne la proportion de réduction de l'entropie quadratique et le second de l'entropie de Shannon. Ces deux indices donnent toujours des valeurs proches l'une de l'autre. Ils évoluent de façon à peu près quadratique entre l'indépendance et l'association parfaite. Leur racine carrée s'avère ainsi plus représentative de la position entre ces deux cas extrêmes.

	q	c^*	p	n	$D(m_0\|m)$	d	sig.	BIC	u	$\sqrt{u}$
CHI	12	263	299	5770	822.2	33	.00	536.4	.056	.237
CHF	10	644	674	35239	4293.3	27	.00	4010.7	.052	.227
CHG	11	684	717	99641	16258.6	30	.00	15913.3	.064	.253

TAB. 1 – *Mesures de la qualité des arbres*

Le tableau 1 rapporte quelques indicateurs de qualité d'ajustement que nous avons calculés pour chacun des trois arbres régionaux, CHI pour la partie italophone, CHF pour la partie francophone et CHG pour la partie germanophone, dont le nombre de paramètres libres est donné sous p. Les déviances entre nœud initial et modèle $D(m_0|m)$ sont toutes très grandes comparées à leur degrés de liberté. Ceci établit que les arbres induits améliorent significativement la description par rapport au nœud initial. Les déviances $D(m|m_{T^*})$, non reportées ici, sont également très grandes indiquant qu'il reste des potentialités importante d'améliorer l'ajustement. Les différences des valeurs du BIC entre nœud initial et arbre induit conduisent à des conclusions similaires. Elles sont largement supérieures à 10, établissant ainsi clairement la supériorité des arbres construits sur les nœuds initiaux. Pour CHI et CHF, les BIC sont aussi largement supérieurs à ceux des arbres saturés correspondants. Ceci n'est pas le cas, cependant, pour CHG, indiquant qu'il y a, au moins dans ce cas, des possibilités d'amélioration. Souvenons-nous, cependant, que notre but est de mettre en évidence les mécanismes principaux qui guident le choix des femmes à participer au marché du travail. Nous avons donc un objectif de compréhension pour lequel une plus grande complexité des arbres serait de toute évidence contre productif. Ceci est caractéristique de la modélisation socio-économique où les modèles ne sauraient être déterminés par les seuls critères statistiques. Les arbres doivent faire sens.

Le coefficient d'incertitude u de Theil semble refléter un faible pourcentage de gain en incertitude. Si l'on regarde toutefois sa racine carrée, on constate qu'avec l'arbre on a déjà couvert environ 25% de la distance jusqu'à l'association parfaite. De plus, les valeurs obtenues devraient être comparées avec le maximum que l'on peut espérer avec les attributs retenus. Pour la table cible qui retient une partition en c^* classes, le u est respectivement .28, .24 and .23. La racine de ces valeurs est environ .5, soit seulement approximativement le double des valeurs obtenues pour les arbres. Ainsi, avec les arbres qui définissent une partition en q classes (entre 10 et 12) contre c^* (entre 263 à 684) pour la table cible, on se trouve déjà à mi-chemin de la cible.

5 Evaluation et conclusion

L'expérience rapportée démontre le grand potentiel des arbres de classification en tant qu'outil d'analyse pour l'étude de problèmes socio-économiques. La présentation visuelle du résultat lui confère un intérêt particulier. Pour notre étude, cette vue synthétique des mécanismes relativement complexes qui influent sur la façon dont les femmes décident de leur participation au marché du travail nous a apporté des enseignements de grande valeur. Elle nous a notamment permis de mettre en lumière des différences culturelles majeures dans les effets de l'interaction entre attributs comme l'âge du dernier enfant, le nombre d'enfants, la profession et le niveau d'éducation, qu'il aurait été difficile d'identifier avec d'autres techniques, et qui n'ont en particulier pas pu l'être avec les simples analyses de tables de contingence bivariées.

Il est important de souligner que, contrairement aux cas où l'objectif est la classification, il est essentiel lorsque le but est la description et la compréhension du phénomène sous-jacent, de se limiter à des arbres de tailles raisonnables. En effet, des arbres trop complexes avec un grand nombre de niveaux et des centaines de feuilles seraient trop confus pour être utiles. De plus, dans un cadre socio-économique tel que celui considéré dans cet article, les arbres doivent avoir un sens du point de vue social et économique. C'est pourquoi il importe de confronter les enseignements tirés des arbres induits avec ceux d'autres analyses ou tentatives de modélisation. Notre étude a largement profité de cette interaction.

Enfin, notons qu'en tant que simples utilisateurs de logiciels d'induction d'arbres, nous avons été confrontés au manque de mesures de validation adaptées à nos besoins. Bien que nous ayons montrés comment obtenir des statistiques et indicateurs pertinents après coup par le biais d'analyses classiques de tableaux croisés, nous aimerions inciter les développeurs à inclure de tels indicateurs dans les outputs de leur logiciel. Mieux même, nous sommes convaincus qu'une induction d'arbre qui viserait, par exemple, à maximiser la réduction du BIC, devrait générer des arbres de complexité réduite plus pertinents du point de vue descriptif.

Références

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

Fabbris, L. (1997). *Statistica multivariata : analisi esplorativa dei dati.* Milano : McGraw Hill.

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

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

Losa, F. B. et P. Origoni (2004). Partecipazione e non partecipazione femminile al mercato del lavoro. Modelli socioculturali a confronto. Il caso della svizzera italiana nel contesto nazionale. Aspetti statistici, Ufficio cantonale di statistica, Bellinzona.

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.

Reyneri, E. (1996). *Sociologia del mercato del lavoro.* Bologna : Il Mulino.

Ritschard, G. et D. A. Zighed (2003). Goodness-of-fit measures for induction trees. In N. Zhong, Z. Ras, S. Tsumo, et E. Suzuki (Eds.), *Foundations of Intelligent Systems, ISMIS03*, Volume LNAI 2871, pp. 57–64. Berlin : Springer.

Ritschard, G. et D. A. Zighed (2004). Qualité d'ajustement d'arbres d'induction. *Revue des nouvelles technologies de l'information E-1*, 45–67.

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

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

Wilkinson, L. (1992). Tree structured data analysis : AID, CHAID and CART. Paper presented at the Sawtooth/SYSTAT Joint Software Conference, Sun Valley, ID.

Summary

This paper presents a full scaled application of classification trees for non-classificatory purposes. The grown trees are used for highlighting regional differences in the mechanisms that drive the women participation in the female labour market. Hence, the focus is on the descriptive rather than predictive power of the trees. The application is run on female labour data from the Swiss Population Census of year 2000. This large data set was analysed in two stages. A first tree provides evidence for three separate analyses for non-mothers, married or widowed mothers, and divorced or single mothers. We consider here only the latter group for which we have grown a separate tree for each of the three main linguistic regions in Switzerland. The resulting trees exhibit fundamental cultural differences between regions. From the methodological point of view, the main difficulties with such a non-classificatory use of trees concern their validation, since the classical classification error rate does not make sense in this setting. We comment this aspect and consider alternatives that are consistent with our non-classificatory usage and easy to compute.