Usage non classificatoire d'arbres de classification

Enseignements d'une analyse de la participation féminine à l'emploi en Suisse

Fabio Losa Pau Origoni
Office statistique du Canton du Tessin

Gilbert Ritschard
Dept Econométrie, Université de Genève

EGC, janvier 2005

Plan

http://mephisto.unige.ch

1 Motivation

1.1 Application des arbres en socio-économie

Etude de la participation féminine à l'emploi

- Contexte socio-économique
- Objectif : mieux comprendre (et non pas classifier)
- Données du recensement suisse de la population 2000 (RSP 2000)

1.2 Arbres comme outil descriptif

- Avantages
  - Résultat visuel facile à interpréter.
  - Eclairage sur interactions entre prédicteurs.
  - Facile à mettre en œuvre
    - non-paramétrique
    - prédicteurs quantitatifs et catégoriels
- Limites
  - Comment valider la qualité descriptive de l'arbre ?
  - Instabilité

2 L'étude appliquée

2.1 Le contexte

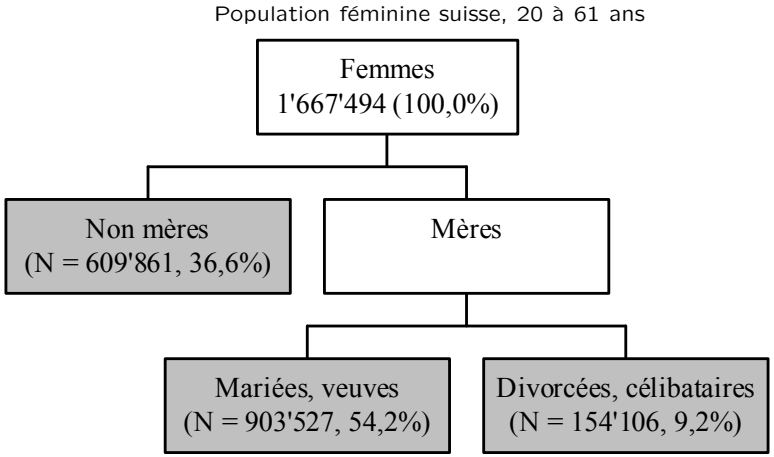
- Etude participation féminine à l'emploi en Suisse.

  $\Rightarrow$ variable réponse à 4 états :
  - plein temps ($> 90\%$ temps hebdomadaire standard)
  - temps partiel long (entre 50% et 90%)
  - temps partiel court ($< 50\%$)
  - non actif

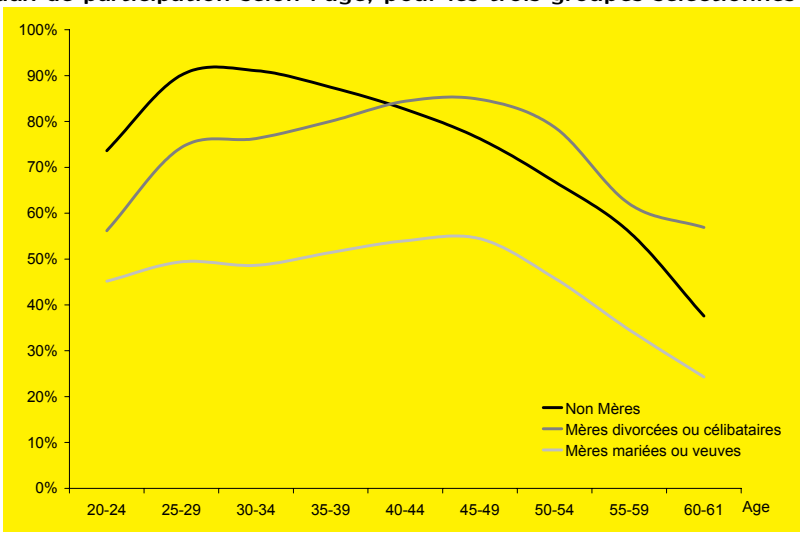
- Données du recensement suisse de la population.
  Population féminine, 20 à 61 ans (1'667'494 cas)

Voir Losa and Origoni (2004)

Les trois types de mères analysés



Taux de participation selon l'âge, pour les trois groupes sélectionnés



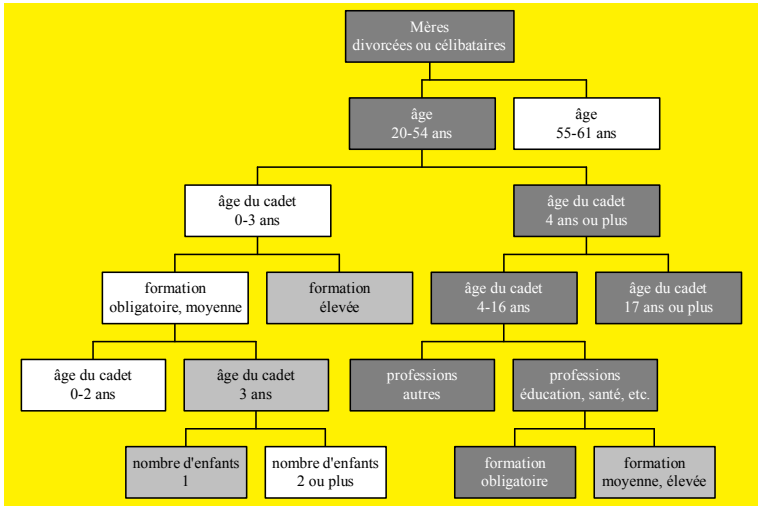
2.2 Exemples d'arbres obtenus

Nous avons générés 9 arbres

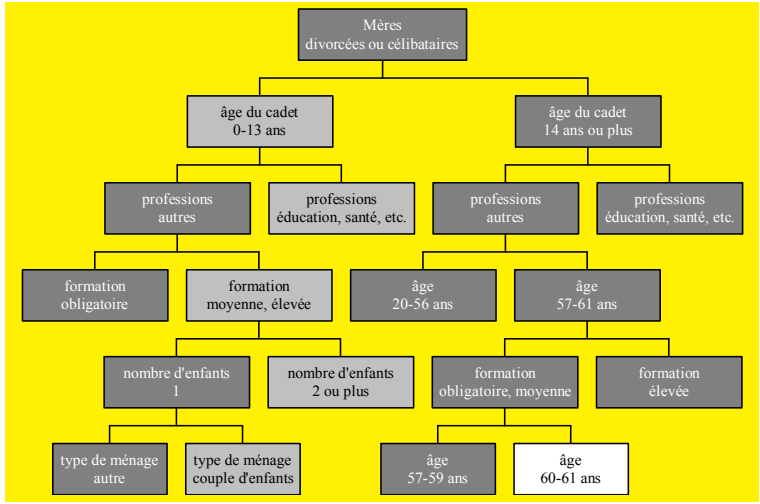
– 3 régions linguistiques

– 3 groupes de mères
(non, mariées ou veuves, divorcées ou célibataires)

Statut emploi, mères célibataires ou divorcées, Suisse italienne



Statut emploi, mères célibataires ou divorcées, Suisse romande

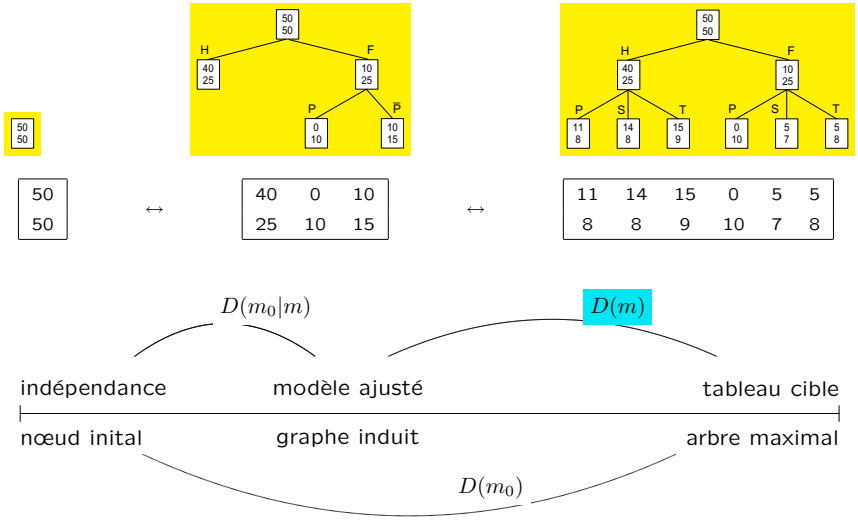


3 Validation des arbres

Erreur de classification non pertinente pour juger de la qualité descriptive

⇒ besoin de critères mieux adaptés (Ritschard and Zighed, 2004)

3.1 Déviance



Principe de construction de la table prédite

Table prédite $\hat{T}$

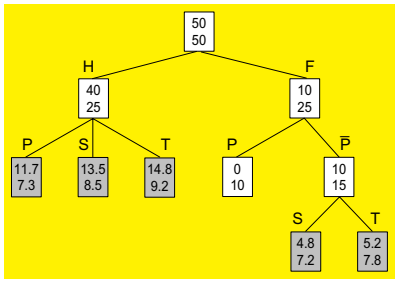
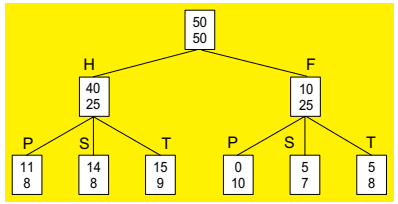


Table cible T



3.2 Calcul de la déviance

$T = (n_{ij})$ tableau $\ell \times c$ cible :

ℓ lignes = catégories de la variable à prédire

c colonnes = profils différents en termes des prédicteurs

$\hat{T} = (\hat{n}_{ij})$ tableau $\ell \times c$ prédit par l'arbre

Total de chaque colonne réparti selon distribution de la feuille contenant le profil correspondant.

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

Difficulté : construction des tableaux T et $\hat{T}$ car c peut être très grand

Déviance partielle $D(m|m_{T^*})$

T^* tableau $\ell \times c^*$ cible

défini avec les c^* profils différents en termes de prédicteurs et groupements de valeurs retenus par l'arbre induit

Perte d'intérêt de l'interprétation de la déviance en tant que distance par rapport à la cible.

Différences de déviance entre arbres emboîtés restent les mêmes, par exemple :

$$D(m_0|m) \; = D(m_0) - D(m) \quad = \quad D(m_0|m_{T^*}) \; - \; D(m|m_{T^*})$$

mesure gain par rapport au nœud initial.

Déviance et rapport de vraisemblance

$D(m_0|m)$ = statistique du khi-2 du rapport de vraisemblance
pour test indépendance sur tableau associé à l'arbre induit.

$D(m_0)$ = statistique du khi-2 du rapport de vraisemblance
pour test indépendance sur tableau cible.

Ces deux valeurs s'obtiennent avec les logiciels statistiques (SPSS, SAS, ...)

On obtient la déviance partielle par différence

$$D(m|m_{T^*}) \quad = \quad D(m_0|m_{T^*}) \; - \; D(m_0|m)$$

3.3 Indicateurs dérivés de la déviance

Indicateurs dérivés de la déviance :

- BIC = déviance pénalisée pour la complexité (nbre de paramètres)
  défini à une constante additive près ⇒ seules variations sont pertinentes
- pseudo $R^2 = 1 - D(m)/D(m_0)$,
  pas pertinent avec déviance partielle
- u Theil, taux de réduction de l'entropie de Shannon

$$u = \frac{D(m_0|m)}{-2\sum_i n_{i.} \ln(n_{i.}/n)}$$

  évolution quadratique ⇒ $\sqrt{u}$ plus pertinent

3.4 Applications aux arbres obtenus

	q	c^*	p	n	$D(m_0\|m)$	d	sig.
CHI	12	263	299	5770	822.2	33	.00
CHF	10	644	674	35239	4293.3	27	.00
CHG	11	684	717	99641	16258.6	30	.00

	ΔBIC(m_0, m)	ΔBIC(m, m_{T^*})	u Theil	$\sqrt{u}$
CHI	536.4	3235.7	.056	.237
CHF	4010.7	4160.0	.052	.227
CHG	15913.3	-17504.3	.064	.253

4 Conclusion

Notre étude a mis en évidence

- l'intérêt des arbres pour l'analyse de problématique socio-économique (où les arbres doivent faire sens)
- la faisabilité de telles analyses à grande échelle (recensement)
- l'absence de mesures de validation appropriées dans les logiciels

Nous avons montré comment calculer des déviances (et les indicateurs dérivés) en recourant à des logiciels statistiques standard

Il serait très utile d'inclure les critères proposés dans les logiciels d'induction d'arbres parmi les critères de validation fournis

⇒ Appel aux développeurs

Références

Losa, F. B. and 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.

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