Analyse de dissimilarités par arbre d'induction

Matthias Studer*, Gilbert Ritschard*, Alexis Gabadinho*, Nicolas S. Müller*

*Département d'économétrie et Laboratoire de démographie, Université de Genève
{matthias.studer, gilbert.ritschard, alexis.gabadinho, nicolas.muller}@unige.ch,
http://www.unige.ch/ses/metri/

Résumé. Dans cet article, nous considérons des objets pour lesquels nous disposons d'une matrice des dissimilarités et nous nous intéressons à leurs liens avec des attributs. Nous nous centrons sur l'analyse de séquences d'états pour lesquelles les dissimilarités sont données par la distance d'édition. Toutefois, les méthodes développées peuvent être étendues à tout type d'objets et de mesure de dissimilarités. Nous présentons dans un premier temps une généralisation de l'analyse de variance (ANOVA) pour évaluer le lien entre des objets non mesurables (p. ex. des séquences) avec une variable catégorielle. La clef de l'approche est d'exprimer la variabilité en termes des seules dissimilarités ce qui nous permet d'identifier les facteurs qui réduisent le plus la variabilité. Nous présentons un test statistique général qui peut en être déduit et introduisons une méthode originale de visualisation des résultats pour les séquences d'états. Nous présentons ensuite une généralisation de cette analyse au cas de facteurs multiples et en discutons les apports et les limites, notamment en terme d'interprétation. Finalement, nous introduisons une nouvelle méthode de type arbre d'induction qui utilise le test précédent comme critère d'éclatement. La portée des méthodes présentées est illustrée à l'aide d'une analyse des facteurs discriminant le plus les trajectoires professionnelles.

1 Introduction

L'analyse des dissimilarités concerne un vaste ensemble de domaines. On y retrouve ainsi la biologie avec l'analyse des gènes et des protéines (alignement de séquences), l'écologie avec la comparaison d'écosystèmes, la sociologie, l'analyse de réseau dont la notion de similarité constitue la base ou encore l'analyse de textes pour n'en citer que quelques-uns. Lorsque les objets analysés sont complexes, des séquences ou des écosystèmes par exemple, il est souvent plus simple de réfléchir en termes de dissimilarités entre objets. Il est d'usage, lorsque l'on a su mesurer les dissimilarités, de procéder à une analyse en cluster qui facilite l'interprétation en réduisant la variabilité de ces objets. Une fois les groupes identifiés, on peut mesurer les liens entre ces objets et d'autres variables d'intérêt à l'aide de tests d'association ou de régression logistique sur la clusterisation obtenue.

Cette façon de procéder présente l'avantage de réduire les dimensions des objets considérés et de faciliter l'interprétation. Toutefois, cette réduction peut amener à des conclusions abusives, en particulier pour les objets se trouvant en périphérie des clusters. De même, il est

possible que certaines associations perdent de leur significativité à cause de cette réduction de l'information. En effet, cette dernière n'est pas contrôlée et les choix, généralement opérés sur des critères statistiques, peuvent en masquer d'autres qui auraient pu s'avérer meilleurs pour montrer les associations avec d'autres variables.

Dans cet article, nous présentons un ensemble de méthodes pour analyser directement des dissimilarités sans passer par une clusterisation préalable. Elles nous permettront de mesurer les liens entre, d'une part, une ou plusieurs variables indépendantes et, d'autre part, des objets décrits à l'aide de dissimilarités. Dans un premier temps, nous abordons le lien avec une seule variable à l'aide du test introduit par Anderson (2001). Nous étendons ces analyses en introduisant un nouveau test d'homogénéité des variabilités et une visualisation originale des résultats lorsque les objets sont des séquences d'états. Nous présentons dans un deuxième temps la méthode de McArdle et Anderson (2001) qui permet d'inclure plusieurs variables explicatives. Finalement, nous introduisons une méthode basée sur les arbres d'induction qui permet une meilleure interprétation des résultats. La méthode est similaire à celle de Geurts et al. (2006) mais plus générale car non limitée aux distances exprimées sous forme de noyaux. Tout au long de cet article, nous appliquons ces méthodes à l'étude des trajectoires professionnelles.

2 Présentation des données

Nous proposons une introduction à l'étude des trajectoires professionnelles afin que les exemples et leurs interprétations apparaissent plus clairs au lecteur. Nous cherchons à comprendre la construction des parcours professionnels et les facteurs qui l'influencent. Nous nous centrons sur l'étude des taux d'activités en suivant les travaux de Levy et al. (2006). On sait que si les trajectoires des hommes sont relativement homogènes et composées des phases « études », « travail à plein temps » et « retraite », celles des femmes sont beaucoup plus variées. Ainsi, leur courbe du taux d'activité moyen présentent une forme en dos de chameau avec une baisse du taux d'activité lorsque les enfants sont en bas âge et une reprise par la suite. Cette courbe moyenne rassemble en fait un nombre très varié de trajectoires. Certaines femmes arrêtent complètement de travailler ou diminuent leur taux pour reprendre ou non l'activité par la suite. Par ailleurs, certaines femmes font de fréquents aller et retour entre activités professionnelles et vie au foyer.

Outre les effets de genre sur les trajectoires, nous nous intéressons aux effets de génération (2 catégories), à ceux engendrés par la vie familiale — nombre d'enfants (4 cat.) et état civil (4 cat.) — ainsi qu'aux différences de milieux sociaux — classe sociale du père (10 cat.), revenu (4 cat.) et niveau d'éducation (3 cat.). Nous sommes également intéressés par savoir si les trajectoires des jeunes générations sont plus diversifiées que les anciennes, montrant ainsi une pluralisation des parcours de vie.

Pour répondre à ces questions, nous utilisons les données de l'enquête biographique rétrospective réalisée par le Panel suisse de ménages[1] en 2002. Nous disposons, pour chaque individu et chaque année de sa situation professionnelle décrite à l'aide des états suivants : plein temps, temps partiel, interruption négative (p. ex., chômage), interruption positive (p. ex., voyage), foyer et formation. Nous centrons notre analyse sur la période de 25 à 40 ans puisque c'est gé-

[1] http ://www.swisspanel.ch

néralement dans cette période que se construit la carrière professionnelle. Nous avons retenu l'ensemble des cas sans données manquantes, ce qui revient à considérer 1560 trajectoires.

Pour mesurer les dissimilarités entre trajectoires, on utilise généralement l'« optimal matching » (OM), plus connu en biologie sous le nom d'alignement de séquence. Cette méthode permet de calculer les distances entre deux trajectoires en les considérant comme des chaînes de caractères, chaque caractère correspondant à un état. L'OM, également appelé distance d'édition ou de Levenshtein, permet donc de construire une matrice des distances deux à deux. C'est cette matrice de dissimilarité que nous utiliserons par la suite.

3 Mesurer le lien sur la base des dissimilarités

Nous présentons une méthode pour évaluer l'association entre, d'une part, des objets caractérisés par une matrice des dissimilarités et, d'autre part, une variable catégorielle à l'aide d'une généralisation des principes de l'ANOVA. Cette méthode a été introduite par Anderson (2001) pour l'analyse des écosystèmes. Notre présentation des fondements théoriques se base toutefois sur les développements et l'approche plus géométrique de Batagelj (1988) exposée dans sa généralisation du critère de Ward. Nous appliquons ensuite ces méthodes sur notre exemple.

L'analyse des dissimilarités a souvent recours au clustering pour réduire les dimensions de l'analyse. On peut alors utiliser le critère de Ward qui est connu pour donner de bons résultats (Batagelj, 1988). Les principes de ce critère ressemblent ainsi fortement à ceux du k-means. On cherche à minimiser la somme des carrés résiduelle (ou inertie intra-classe). Autrement dit, on cherche la partition qui « explique » la plus grande part de la « variance » de la population dans son ensemble. Le critère de Ward se base sur la relation formalisée par l'équation (1) : la somme des carrés (ou inertie, notée SC) peut être exprimée comme une moyenne des distances euclidiennes deux à deux élevées au carré (notée de_{ij}^2).

$$SC = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j=1}^{n}(y_i - y_j)^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=i+1}^{n} de_{ij}^2 \quad (1)$$

La méthode peut être généralisée à d'autres types de distances en remplaçant la distance euclidienne au carré de_{ij}^2 par une autre mesure de dissimilarité d_{ij} au sein de l'équation (1). D'un point de vue formel, la mesure de dissimilarité d_{ij} doit satisfaire les conditions mathématiques de la distance (Batagelj, 1988). Toutefois, Anderson (2001) montre que la dissimilarité peut être semi-métrique, si on utilise d_{ij}^2. Nous développons plus loin les implications du non respect de l'inégalité triangulaire.

La somme des carrés nous permet de définir une mesure de la variance de notre échantillon. En effet, la variance s^2 peut être obtenue directement puisque $s^2 = \frac{1}{n}SC$. A notre sens, il est toutefois difficile de continuer de parler de variance lorsque l'on considère d'autres mesures de dissimilarité, et l'on préférera parler de « pseudo-variance ». Cette mesure de la variabilité des objets est assez intuitive. Si l'on s'en réfère à l'équation (1), on remarque qu'elle correspond à la dissimilarité moyenne divisée par deux puisque $s^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}$.

A titre d'exemple, la variabilité de nos trajectoires professionnelles vaut 0.501 ce qui correspond à la moitié de la distance d'édition moyenne qui est égale à 1.02. Il est intéressant de noter que cette variabilité est égale à 0.118 pour les hommes alors qu'elle vaut 0.614 pour

les femmes indiquant par là même qu'elles ont une diversité de trajectoires professionnelles beaucoup plus élevée que les hommes.

En généralisant la notion de somme des carrés à d'autres mesures de dissimilarité, le théorème de Huyghens (équation 2) est toujours vérifié (Batagelj, 1988). Autrement dit, la somme des carrés totale (SC_T) est égale à la somme des carrés expliqués (SC_{EXPL}) par une partition plus une somme des carrés résiduels (SC_{RES}). On retrouve ainsi les bases de l'analyse de variance (ANOVA).

$$SC_T = SC_{EXPL} + SC_{RES} \tag{2}$$

L'ensemble des termes de cette équation est calculable, SC_T et SC_{RES} à l'aide de l'équation (1) puisque SC_{RES} est simplement égale à l'addition des sommes des carrés de chaque sous-groupe. On peut ensuite en déduire SC_{EXPL}. L'équation (2) nous permet d'évaluer la part de la variabilité, au sens défini par la mesure de dissimilarité, expliquée par une partition ou, autrement dit, par une variable catégorielle ou une variable continue discrétisée. Dans l'esprit de l'ANOVA, cette diminution de la variabilité est due à une différence de positionnement des centres de gravité (ou centroïde) de chaque classe. Cette vision peut être étendue à tout type de distance même si la notion de centre de classe n'est pas clairement définie pour l'objet considéré (Batagelj, 1988). Il est fort probable que ce centre n'appartienne pas à l'espace des objets, au même titre que la moyenne peut ne pas appartenir à l'ensemble des valeurs (par exemple, si les mesures sont discrètes). Conceptuellement, nous cherchons donc à expliquer une part de la variation en cherchant des différences de positionnement. On se propose de mesurer la part de la variabilité expliquée à l'aide de la formule du R^2. Soit n le nombre d'individu et m le nombre de paramètres :

$$R^2 = \frac{SC_{EXPL}}{SC_T} \tag{3}$$

$$F = \frac{SC_{EXPL}/(m-1)}{SC_{RES}/(n-m)} \tag{4}$$

Il n'est pas possible de mesurer la significativité de l'association à l'aide de la statistique F (équation 4) comme dans l'analyse de variance classique. En effet, la statistique F ne suit pas une loi de Fisher, l'hypothèse de normalité n'étant pas vérifiée. Dès lors, on se propose de mesurer la significativité de l'association à l'aide de test de permutation (Moore et al., 2003). Ces tests fonctionnent de la manière suivante. On réalise une permutation des lignes et des colonnes de la matrice des distances (sans permuter l'appartenance aux catégories) et on calcule les F_{perm} obtenu. On répète cette permutation p fois en enregistrant la distribution de la statistique F_{perm}. Implicitement, ces permutations supposent une absence de relation entre les catégories et les distances. Dès lors, la distribution observée de F_{perm} correspond à la distribution non paramétrique de F sous l'hypothèse nulle, c'est-à-dire que la partition n'explique aucune part de la variabilité. A partir de cette distribution, on peut évaluer la significativité du F par la proportion de F_{perm} supérieur au F_{obs}. On considère généralement qu'il est nécessaire de réaliser 5000 permutations pour un seuil de significativité de l'ordre de 1% et 1000 pour un seuil de 5%.

L'interprétation des résultats doit être faite avec précautions si la dissimilarité est semi-métrique. En effet, la généralisation que nous avons présentée implique que la contribution d'un objet x à la pseudo-variance, qui peut également être interprétée comme la dissimilarité

entre x et son centre de gravité $\tilde{g}$, est donné par $d_{x\tilde{g}} = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\big(2 \cdot d_{ix} - d_{ij}\big)$. Cette contribution peut être négative si l'inégalité triangulaire n'est pas respectée. Pour s'en convaincre, il suffit d'observer que $d_{x\tilde{g}}$ est minimal lorsque d_{ij} est maximal, soit égal à $d_{xi} + d_{xj}$ selon l'inégalité triangulaire, ce qui amène au résultat $d_{x\tilde{g}} = 0$. La «dissimilarité» $d_{x\tilde{g}}$ peut être négative lorsque x diminue la variance entre les autres individus, ce qui pourrait se produire lorsque la distance entre deux autres observations y et z est réduite quand on passe par x, autrement dit si $d_{y,z} > d_{y,x} + d_{x,z}$. Ce cas est relativement courant en analyse de réseau. Par exemple dans un réseau entre x, y et z où la dissimilarité est égale à 1 lorsque deux individus se fréquentent souvent et est égale à 10 s'ils ne se fréquentent jamais, la contribution de x à la variance sera négative si x entretient des relations fréquentes avec y et z mais que ces derniers ne se voient jamais.

Ces considérations ne concernent pas notre exemple puisque l'«optimal matching» respecte l'inégalité triangulaire. En effet, celle-ci garantit que la dissimilarité correspond au coût minimal de la transformation de la séquence y en z. Ainsi $d_{y,z} \leq d_{y,x} + d_{x,z}$, sinon l'algorithme transformerait d'abord la séquence y en x puis en z.

Le tableau 1 récapitule les résultats pour l'ensemble de la population ainsi que pour les hommes et les femmes séparément. Nous avons réalisé mille permutations pour évaluer la significativité des tests. Sans surprise, le sexe est de loin la dimension la plus explicative de la

	Total			Hommes			Femmes		
Variable	F	R^2	Sig	F	R^2	Sig	F	R^2	Sig
Sexe	477.995	0.235	**0.000**						
Classe soc. père	1.578	0.009	**0.029**	2.085	0.026	**0.005**	1.205	0.013	0.163
Revenu	1.349	0.003	0.182	3.086	0.013	**0.006**	3.553	0.013	**0.000**
Formation	18.486	0.023	**0.000**	20.632	0.054	**0.000**	6.287	0.015	**0.000**
Cohorte	17.037	0.011	**0.000**	6.330	0.009	**0.001**	14.911	0.018	**0.000**
Enfants	13.704	0.026	**0.000**	1.006	0.004	0.391	25.740	0.085	**0.000**
Etat civil	9.744	0.018	**0.000**	1.783	0.007	**0.047**	18.078	0.061	**0.000**

TAB. 1 – *Test d'association avec les trajectoires professionnelles.*

variabilité des trajectoires. Ainsi, le R^2 atteint 0.235. Autrement dit, la variable sexe permet «d'expliquer» 23.5% de la pseudo-variance des trajectoires. La relation est statistiquement significative puisque le $F_{obs} = 477.995$ n'est jamais observé à l'aide de mille permutations. On remarquera également que les variables économiques semblent avoir un plus grand impact sur les trajectoires masculines que féminines. Au contraire, les trajectoires féminines semblent plus influencées par les facteurs familiaux (enfant et état civil).

Nous avons montré que plusieurs facteurs sont significativement liés aux trajectoires professionnelles. Nous avons également vu que théoriquement ces tests nous montrent une différence de positionnement des centres de gravité des groupes de séquences. Malheureusement, à ce stade, il est difficile de comprendre ces différences.

La figure 1 présente une nouvelle méthode pour visualiser des différences entre groupes de séquences. Les deux premiers graphiques montrent les séquences des hommes et des femmes sous la forme d'index-plot (Scherer, 2001). Dans ces graphiques, chaque ligne correspond à

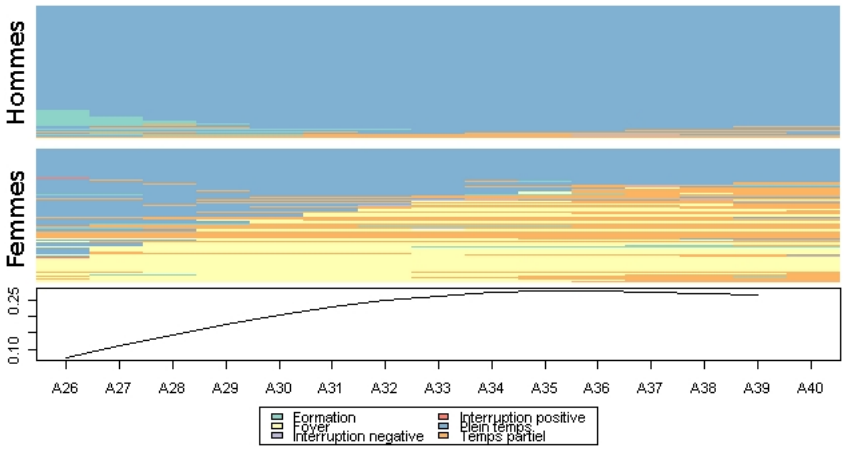

FIG. 1 – *Différences de trajectoires selon le sexe.*

une séquence. Chaque séquence est représentée sous la forme d'une succession de carré de couleur, chaque carré représentant une unité de temps et sa couleur l'état occupé.

Contrairement aux méthodes habituellement utilisées, nous avons ordonné les séquences selon la première dimension d'une PCA («Principal Coordinates Analysis») (Gower, 1966). Nous verrons dans la section 5 que cette méthode est fortement reliée aux analyses que nous avons déjà présentées. Elle a l'avantage d'ordonner les séquences selon une dimension sous-jacente, facilitant ainsi la lecture du graphique. Notons que le graphique permet, par la même occasion, de donner une interprétation aux axes de la PCA. On remarque ainsi que les séquences sont organisées dans un continuum allant des trajectoires à plein temps aux trajectoires entièrement au foyer en passant par les trajectoires à temps partiel.

Le dernier graphique montre l'évolution de l'association entre la variable catégorielle et une sous-séquence de la trajectoire calculée sur deux périodes. Pour chaque unité de temps, on extrait une sous-séquence de deux unités. On calcule la matrice des distances et la part de la variabilité expliquée par la variable indépendante. Cette représentation permet d'identifier les périodes qui sont le plus différenciées selon le genre. On observe ainsi que les différences de genre atteignent leur paroxysme aux alentours de 35 ans.

4 Mesurer l'homogénéité de la variabilité

Dans certaines situations, il peut être intéressant de tester si les pseudo-variances de chaque catégorie diffèrent significativement. D'un point de vue géométrique, on ne s'intéresse plus ici à des différences de positionnement des centres de gravité dans l'espace des objets, mais à des différences de diamètre ou d'ampleur. Dans l'analyse de variance classique, on utilise généralement le test de Bartlett (Snedecor et Cochran, 1989) qui suppose, sous H_0, l'égalité des variances ou, en d'autres termes, l'homogénéité des variances. Ce test se base sur la distribution de la statistique T de l'équation (5), où s_i^2 désigne la pseudo-variance du groupe i. L'ensemble

des termes de cette équation peut être calculé à partir de ce qui a été présenté. Ici encore, il n'est pas possible de faire l'hypothèse que cette statistique suit une loi prédéfinie. Nous nous baserons donc sur des tests de permutations pour évaluer la significativité des différences de variabilité.

$$T = \frac{(n-m)\ln\left\{\sum_{i=1}^{m}\frac{(n_i-1)s_i^2}{(n-m)}\right\} - \sum_{i=1}^{m}(n_i-1)\ln(s_i^2)}{1+\frac{1}{3(m-1)}\left[\sum_{i=1}^{m}\frac{1}{n_i-1}-\frac{1}{n-m}\right]} \tag{5}$$

Dans la section 3, nous avions relevé que la pseudo-variance des hommes vaut 0.118 contre 0.614 pour les femmes. Nous observons dans ce cas un T_{obs} de 460.017, une valeur jamais observée à l'aide de mille permutations. Dès lors, le test d'homogénéité des pseudo-variances nous permet d'attester que la différence des pseudo-variances selon le sexe est significative.

5 Modèles à plusieurs facteurs explicatifs

Les tests que nous avons présentés permettent d'évaluer l'association avec une variable explicative. Dans cette partie, nous présentons la généralisation de ceux-ci réalisée par McArdle et Anderson (2001) pour analyser des écosystèmes à l'aide de la distance semi-métrique de Bray-Curtis. Cette généralisation part de la même propriété de décomposition de la variance dans le cas multivarié et peut contenir plusieurs variables catégorielles ou continues. Ici encore, la méthode permet d'évaluer la significativité et le pouvoir explicatif du modèle. Plusieurs variantes existent, nous présentons ici la méthode et le formalisme de McArdle et Anderson (2001) où l'on trouvera également l'ensemble des développements mathématiques. On peut également citer ici les articles suivants : Excoffier et al. (1992); Gower et Krzanowski (1999); Anderson (2001); Zapala et Schork (2006).

Dans notre notation, les matrices et les vecteurs sont spécifiés en gras. Nous nous intéressons au modèle de régression multivarié suivant : $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, où $\mathbf{Y}$ est une matrice $n \times t$ contenant les valeurs de t variables expliquées pour n individus et $\mathbf{X}$ est une matrice $n \times m$, contenant les valeurs des m prédicteurs pour chaque individu et dont la première colonne est un vecteur dont toutes les valeurs sont égales à 1.

La somme de la somme des carrés (pour chaque variable) peut être dérivée à l'aide de la seule matrice de Gower utilisée dans la PCA. Cette matrice se base sur celle des distances euclidiennes (Gower, 1966). McArdle et Anderson (2001) procèdent dès lors à une généralisation à tout type de dissimilarités. Soit $\mathbf{1}$ un vecteur de longueur n où toutes les valeurs sont égales à 1 et $\mathbf{A}$ une matrice telle que $a_{ij} = -\frac{1}{2}d_{ij}$, avec d_{ij} la dissimilarité entre l'individu i et j.[2]

$$\mathbf{G} = \left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{A}\left(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'\right) \tag{6}$$

La somme totale des carrés SC_T est alors égale à la trace de $\mathbf{G}$. McArdle et Anderson (2001) montrent que la somme SC_{EXPL} des carrés expliquée par le modèle et la somme SC_{RES} des

[2]Dans notre généralisation, nous avons considéré que la dissimilarité utilisée pour des variables continues est la distance euclidienne au carré. McArdle et Anderson (2001) considèrent qu'il s'agit de la distance euclidienne. Ils utilisent donc la dissimilarité au carré, soit $a_{ij} = -\frac{1}{2}d_{ij}^2$.

carrés résiduelle s'écrivent :

$$SC_{EXPL} = tr(\mathbf{HGH}) \quad (7)$$

$$SC_{RES} = tr\big[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})\big] \quad (8)$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est connue sous le nom de matrice chapeau dans la régression linéaire. Ces équations nous permettent de calculer un R^2 et une statistique F à l'aide de (3) et (4). Cette formulation permet d'inclure tout type de variable, les variables continues et les catégorielles à l'aide d'une codification indicatrice ou par contrastes.

Ici encore, il n'est pas possible de considérer que la statistique F suit une loi de Fisher. Mais on peut utiliser les tests de permutations, comme précédemment, pour évaluer la significativité globale du modèle. Plusieurs méthodes sont envisageables pour tester l'apport de chaque variable au modèle global. On retiendra essentiellement les types I et II mis en avant par Shaw et Mitchell-Olds (1993). Dans le type I, on considère que chaque variable apporte une information égale à l'augmentation de SC_{EXPL} lorsqu'on l'introduit dans le modèle. Ainsi, la valeur exacte dépend de l'ordre d'introduction des variables. On considère l'information apportée par une variable, une fois l'effet des variables précédemment introduites pris en compte. Dans le type II, connu pour être robuste en l'absence d'effet d'interaction, on considère la réduction d'information due à la suppression de la variable considérée du modèle complet. Nous retenons ici la deuxième solution. Dans ce cas, on peut calculer la significativité à l'aide du F suivant, généralement utilisé pour comparer deux modèles imbriqués :

$$F_v = \frac{(SC_{EXPL_c} - SC_{EXPL_v})/p}{SC_{RES_c}/(n-m-1)} \quad (9)$$

où l'indice c fait référence au modèle complet, l'indice v à celui obtenu en retirant la variable considérée et p est le nombre de paramètres utilisés pour coder la variable considérée.

Le tableau 2 montre les résultats de l'analyse pour notre exemple. Nous présentons deux modèles. Le premier contient l'ensemble des variables. Le deuxième a été construit à l'aide d'une procédure de type « Backward ». A chaque pas, on retire du modèle la variable la moins significative jusqu'à ce que l'ensemble des variables le soit.

	Complet			« Backward »		
Variable	F_v	ΔR^2_v	Sig	F_v	ΔR^2_v	Sig
Sexe	477.196	0.218	0.000	488.627	0.224	0.000
Formation	8.230	0.008	0.000	10.986	0.010	0.000
Revenu	0.868	0.001	0.542			
Classe sociale du père	1.167	0.005	0.241			
Cohorte	11.586	0.005	0.000	13.670	0.006	0.000
Enfants	9.887	0.014	0.000	10.313	0.014	0.000
Etat civil	4.621	0.006	0.000	5.073	0.007	0.000
	F_{tot}	R^2	Sig	F_{tot}	R^2	Sig
Total	29.557	0.297	0.000	63.602	0.291	0.000

TAB. 2 – *Modèle à plusieurs facteurs explicatifs.*

Pour le modèle complet, on remarque que le sexe est toujours la variable la plus significative. Si on retire la variable « Sexe » le R^2 de l'ensemble du modèle ($= 0.297$) diminue de

0.218. Cette différence est significative puisque le $F_{sexe} = 477.196$, une valeur jamais observée à l'aide de mille permutations. Au contraire, le revenu n'est pas significatif puisque, si on le retire du modèle, le R^2 ne perd que 0.001 et que le $F_{revenu} = 0.868$, une valeur que l'on a atteint $0.542 * 1000 = 542$ fois.

L'évaluation du ΔR_v^2 propre à chaque variable doit être faite avec prudence. En effet, nous avons vu que celui-ci est correct en l'absence d'effets d'interaction. Ce qui n'est pas le cas ici.

La prise en compte de plusieurs facteurs explicatifs nous permet d'observer que l'association avec la classe sociale du père n'est plus significative si l'on insère le niveau de formation. En effet, ces variables sont fortement liées et le niveau d'éducation est plus significatif.

Cette méthode permet d'analyser les trajectoires en incluant plusieurs variables. Elle a l'avantage de prendre en considération, l'information spécifique apportée par chacune et est, dans ce sens, complémentaire à ce que nous avions déjà présenté. Malheureusement, les résultats sont difficilement interprétables. Il n'existe pas de méthode pour visualiser les résultats, ni pour détailler l'influence des facteurs explicatifs.

6 Analyse par arbre d'induction

Nous présentons à présent une nouvelle méthode d'analyse d'objets décrits à l'aide d'une matrice des dissimilarités en prenant appui sur les arbres d'induction. Ces derniers fonctionnent ainsi (Breiman et al., 1984; Kass, 1980). On regroupe l'ensemble des objets dans un nœud initial que l'on cherche à segmenter selon les valeurs d'un prédicteur choisi de manière à ce que les nœuds enfants diffèrent le plus possible l'un de l'autre. On répète ensuite de manière récursive cette opération sur les nœuds enfants. En partitionnant successivement la population, les arbres nous permettront de visualiser les effets des facteurs explicatifs, pour autant que l'on dispose d'une méthode pour visualiser ou présenter un ensemble des objets analysés. Dans le cas de séquences d'état, on pourra ainsi se baser sur les index-plots (Fig. 1).

Dans notre implémentation, nous calculons les différences entre nœuds enfants sur la base du test introduit au début de cet article, puisque ce dernier détecte des différences de positionnement des centres de gravité des nœuds enfants. La qualité de la partition peut être jugée à l'aide du R^2. Plus celui-ci est élevé, meilleure est la partition. Malheureusement, le choix du R^2 nous condamne à ne construire que des arbres binaires, car il ne pénalise pas pour le nombre de catégorie[3]. La procédure utilisée ressemble ainsi fortement à celle de Geurts et al. (2006). Toutefois, la formulation présentée permet d'utiliser des dissimilarités semi-métriques, pour autant que l'interprétation d'une contribution négative à la variance ne soit pas problématique.

L'algorithme proposé fonctionne de la manière suivante. A chaque nœud, on teste l'ensemble des regroupements en deux catégories pour chaque prédicteur et on conserve le meilleur. Finalement, nous effectuons un test de permutation sur le meilleur regroupement pour attester sa significativité. Plusieurs optimisations sont possibles. Il n'est pas nécessaire de recalculer SC_{res} pour chaque regroupement. Notre algorithme se base sur une matrice symétrique $\mathbf{E}$, $m \times m$, où m désigne le nombre de catégories et dont les éléments $e_{k\ell} = \sum_{i \in k} \sum_{j \in \ell} d_{ij}$ sont égaux à la somme des dissimilarités entre, d'une part, les individus appartenant à la catégorie k et, d'autre part, ceux appartenant à ℓ. Cette matrice nous permet de calculer la valeur de SC_{res}

[3] Sans connaître la distribution du R^2 ou de SC_{res} sous H_0, il n'est pas possible de construire un indice pénalisant pour le nombre de catégories utilisées comme le BIC par exemple.

pour une partition en deux catégories. La somme des carrés résiduels pour le groupe G de catégories est alors égale à $SC_{G,res} = \frac{1}{n_G}(\sum_{k\in G}\sum_{\ell\geq k,\ell\in G} e_{k\ell})$. On évite ainsi de recalculer les sommes de dissimilarités entre tous les individus. Le R^2 ne peut que se détériorer lors du regroupement en deux catégories. La matrice $\mathbf{E}$ nous permet de calculer le R^2 pour l'ensemble des catégories. Si ce dernier est plus faible que le meilleur R^2 trouvé pour les variables testées précédemment, il est inutile de tester les regroupements de cette variable.

La qualité de l'arbre peut être évaluée en testant l'association entre les dissimilarités et une variable d'appartenance aux nœuds terminaux. On dispose ainsi d'une mesure de la significativité de la partition générée par l'arbre ainsi que d'un R^2 de la pseudo-variance expliquée.

La figure 2 présente un arbre de dissimilarité construit sur la bases des distances entre trajectoires professionnelles. L'arbre a un meilleur pouvoir explicatif que le modèle «Backward» du Tableau 2. En effet, nous observons un R^2 de 0.3 alors même que nous utilisons moins de

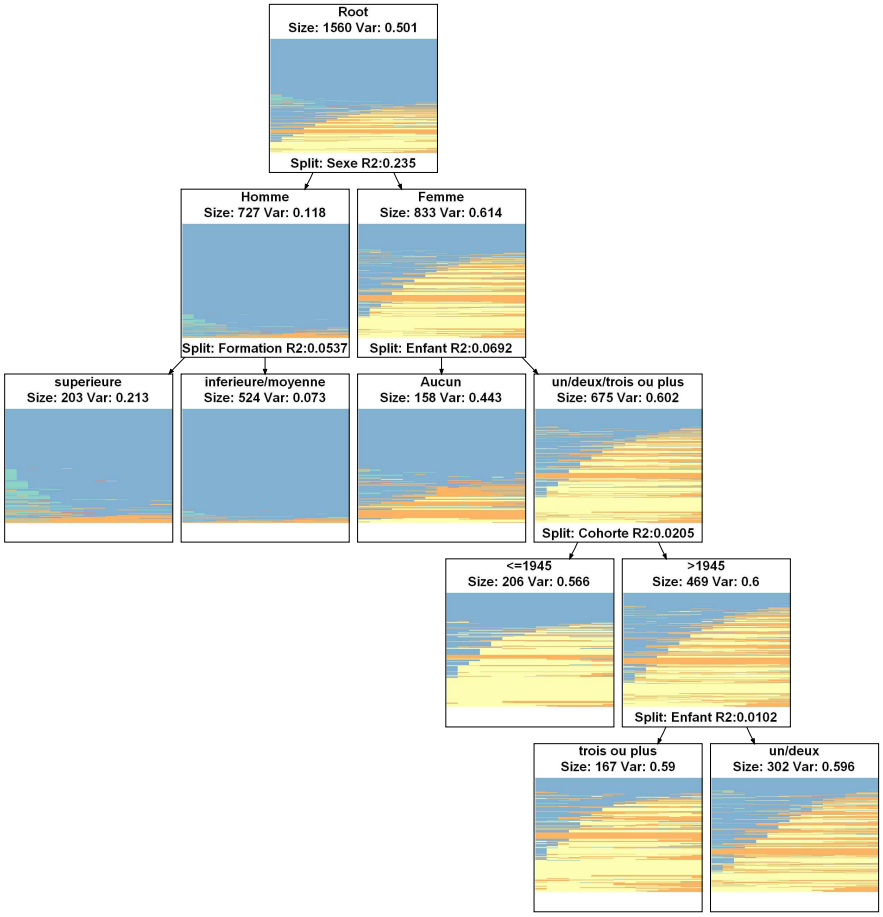

FIG. 2 – *Arbre de séquences.*

paramètres. Cette différence est due à la détection des effets d'interaction entre trajectoire, sexe et nombre d'enfants par exemple. Cette détection est un autre avantage des arbres d'induction.

L'arbre nous permet de voir que, parmi les hommes, le choix du temps partiel semble surtout être le fait de ceux qui ont bénéficié d'une formation supérieure. Chez les femmes, les trajectoires sont bien plus diversifiées. On notera que si elles ont eu au moins un enfant, elles ont beaucoup plus recours au temps partiel quand elles sont nées après 1945. Cet effet cohorte est cependant moins marqué chez les femmes qui ont au moins trois enfants.

7 Conclusion

En nous basant sur les principes de l'analyse de variance, nous avons présenté plusieurs méthodes pour analyser les liens entre, d'une part, des attributs et, d'autre part, des objets décrits à l'aide d'une matrice des dissimilarités. La première généralisation que nous avons exposée permet de détecter des différences de centre de gravité selon une variable catégorielle. Nous avons étendu cette analyse en présentant un test d'homogénéité des pseudo-variances des différentes catégories.

Il est possible d'évaluer le lien entre ces objets et plusieurs facteurs. Malheureusement, cette méthode souffre d'un manque d'interprétabilité des résultats. Il est difficile de comprendre les liens entre les variables et les objets décrits par les dissimilarités. Nous avons ainsi proposé de nous baser sur la méthode des arbres d'induction qui nous permet de visualiser les résultats et de les interpréter. Elle nous a également permis de découvrir des effets d'interaction.

L'ensemble de ces méthodes est disponible dans TraMineR (Gabadinho et al., 2008), un module R librement accessible, et permet ainsi à tout un chacun de les utiliser.

Références

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology 26*, 32–46.

Batagelj, V. (1988). Generalized ward and related clustering problems. In H. Bock (Ed.), *Classification and related methods of data analysis*, pp. 67–74. North-Holland, Amsterdam.

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

Excoffier, L., P. E. Smouse, et J. M. Quattro (1992). Analysis of molecular variance inferred from metric distances among dna haplotypes: Application to human mitochondrial dna restriction data. *Genetics 131*, 479–491.

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2008). Mining sequence data in R with TraMineR. User's guide, Department of Econometrics and Laboratory of Demography, University of Geneva, Geneva. (Released on CRAN the Comprehensive R Archive Network).

Geurts, P., L. Wehenkel, et F. d'Alché Buc (2006). Kernelizing the output of tree-based methods. In W. W. Cohen et A. Moore (Eds.), *ICML*, Volume 148 of *ACM International Conference Proceeding Series*, pp. 345–352. ACM.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika 53*(3/4), 325–338.

Gower, J. C. et W. J. Krzanowski (1999). Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 48*(4), 505–519.

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

Levy, R., J.-A. Gauthier, et E. Widmer (2006). Entre contraintes institutionnelle et domestique : les parcours de vie masculins et féminins en Suisse. *Cahiers canadiens de sociologie 31*(4), 461–489.

McArdle, B. H. et M. J. Anderson (2001). Fitting multivariate models to community data: A comment on distance-based redundancy analysis. *Ecology 82*(1), 290–297.

Moore, D. S., G. McCabe, W. Duckworth, et S. Sclove (2003). *The Practice of Business Statistics: Using Data for Decisions*, Chapter Bootstrap Methods and Permutation Tests. W. H. Freeman.

Piccarreta, R. et F. C. Billari (2007). Clustering work and family trajectories by using a divisive algorithm. *Journal of the Royal Statistical Society A 170*(4), 1061–1078.

Scherer, S. (2001). Early career patterns: A comparison of Great Britain and West Germany. *European Sociological Review 17*(2), 119–144.

Shaw, R. G. et T. Mitchell-Olds (1993). Anova for unbalanced data: An overview. *Ecology 74*(6), 1638–1645.

Snedecor, G. W. et W. G. Cochran (1989). *Statistical methods* (8th ed.). Iowa State University Press.

Zapala, M. A. et N. J. Schork (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proceedings of the National Academy of Sciences of the United States of America 103*(51), 19430–19435.

Summary

In this article we consider objects for which we have a matrix of dissimilarities and we are interested in their links with attributes. We focus on state sequences for which dissimilarities are given for instance by edit distances. The methods discussed apply however to any kind of objects and measures of dissimilarities. We start with a generalization of the analysis of variance (ANOVA) to assess the link of non measurable objects (e.g. sequences) with a given categorical variable. The trick is to show that variability among objects can be derived from the sole dissimilarities, which permits then to identify factors that most reduce this variability. We infer a general statistical test and introduce an original way of rendering the results for state sequences. We then generalize the method to the case with more than one factor and discuss its benefits and limitations especially regarding interpretation. Finally, we introduce a new tree method for general objects that exploits the former test based on dissimilarity measures as splitting criterion. We demonstrate the scope of the various methods presented through a study of the factors that most discriminate professional trajectories.