Classification de parcours de vie à l'aide de l'optimal matching*

Nicolas S. Müller, Matthias Studer, Gilbert Ritschard

Département d'économétrie, Université de Genève
nicolas.muller@metri.unige.ch

Mots clés : application, classification automatique, séquences.

Ce travail analyse les parcours de vie familiale en les considérant comme des séquences. Le but est de parvenir à observer le caractère temporel des parcours de vie en prenant en compte la durée entre chaque événement constitutif de ce parcours, mais aussi l'ordre dans lequel ils surviennent. Nous proposons d'appliquer aux données de l'enquête biographique rétrospective du Panel suisse de ménages une méthode d'analyse des séquences dans le but d'obtenir une typologie des parcours de vie du 20ème siècle. Celle-ci nous permettra ensuite de mieux approcher les changements qui ont pu intervenir dans leur structure.

1 Méthode

Afin de représenter les parcours de vie familiale sous forme de séquences, quatre événements ont été retenus. Il s'agit de l'âge au départ de chez les parents, l'âge au premier mariage, l'âge au premier enfant et l'âge au premier divorce. Ces événements sont considérés comme les étapes qui jalonnent la vie des individus. Ils sont comme des « bornes » qui délimitent les différents états qu'un individu traverse au cours de sa vie. La vie familiale d'un individu est donc représentée sous la forme d'une « séquence », une suite d'années de sa vie avec comme information l'état dans lequel il se trouve chaque année. A partir des quatre événements retenus, huit états distincts ont été définis. Ils sont représentés dans le tableau 1.

Les individus sont observés de l'âge de quinze ans à celui de trente ans. Ceux âgés de moins de trente ans au moment de l'enquête ont été éliminés afin de ne conserver que des séquences complètes et sans données manquantes. Le nombre total d'observations est de 4318. Dans le but de limiter le nombre d'états, il a été décidé de considérer tout divorce comme un état unique, tout en ayant conscience que le nombre de divorces avant l'âge de trente ans est très faible.

La méthode d'analyse de séquences que nous utilisons dans ce travail est celle dite d' « optimal matching ». L'algorithme retenu, connu aussi sous le nom d'alignement de séquences, a été développé à l'origine pour l'analyse rapide des protéines et des séquences d'ADN. Ce type de méthode a été conçu pour permettre la comparaison rapide de nombreuses séquences afin de trouver des correspondances parmi celles-ci. Les premiers algorithmes d'optimal matching sont apparus au début des années 70 et leur première utilisation dans les sciences sociales remonte à l'article d'Abbott et Forrest sur leur application à des données historiques [1]. On doit à Abbott de nombreux articles méthodologiques sur l'utilisation de ces méthodes dans les sciences sociales, et notamment en sociologie [2], [3], [4].

*Etude soutenue par le Fonds national suisse de la recherche (FNS) FN-100012-113998, et réalisée avec les données collectées dans le cadre du projet « Vivre en Suisse 1999-2020 », piloté par le Panel suisse de ménages et supporté par le FNS, l'Office fédéral de la statistique et l'Université de Neuchâtel.

Tab. 1 – Liste des états

	départ	mariage	enfant	divorce
0	non	non	non	non
1	oui	non	non	non
2	non	oui	oui/non	non
3	oui	oui	non	non
4	non	non	oui	non
5	oui	non	oui	non
6	oui	oui	oui	non
7	oui/non	oui/non	oui/non	oui

Concrètement, cette méthode évalue la « distance » entre une séquence A et une séquence B en calculant le nombre minimum d'opérations nécessaires pour passer de l'une à l'autre. Les opérations disponibles sont de deux types, soit l'insertion ou la suppression d'un état, soit la substitution d'un état par un autre. Un coût est attribué à ces deux types d'opération selon le type de données. Dans ce travail, les opérations de substitution ont été favorisées par rapport aux opérations d'insertion/suppression dans le but d'éviter la déformation du temps (allongement ou rétrécissement) qu'elles provoqueraient. Les coûts de substitution entre états sont définis par une matrice des coûts calculée en fonction des taux de transition observés dans les données. Ainsi, plus un passage d'un état à un autre est observé fréquemment, plus son coût est bas.

2 Résultats

L'application de l'optimal matching a permis le calcul d'une « distance » entre chaque individu, en fonction du nombre de transformations nécessaires pour passer d'une séquence à une autre. Le résultat se présente sous la forme d'une matrice symétrique de distances qui est ensuite utilisée dans une analyse de classification hiérarchique ascendante. La méthode de regroupement des cas est celle de Ward. Nous avons décidé de retenir une solution de classification à cinq classes. Une visualisation des résultats sous la forme d'histogrammes avec pour chaque année de vie la contribution de chaque groupe au nombre total d'individus (barres empilées à 100%) nous permet de distinguer les cinq groupes selon plusieurs caractéristiques. Les graphiques pour les groupes 1 et 2 sont reproduits plus loin (voir figure 1 et 2).

Le premier groupe (n = 952 ; 22% du total) contient une proportion remarquable d'individus ayant eu des enfants sans mariage et d'individus ayant rencontré un divorce. Le deuxième groupe (n = 1051 ; 24,3%) se caractérise par une période entre le départ de chez les parents et le premier mariage qui est courte. Les mariages sont relativement précoces : 50% des individus qui composent ce groupe sont mariés à l'âge de 22 ans, et 100% à 27 ans (l'âge médian au premier mariage de l'échantillon = 26 ans ; 19,2% mariés à 22 ans). Le troisième groupe (n = 1228 ; 28,4%), contient des individus au départ plutôt tardif ; les premiers départs ne commencent qu'à l'âge de 23 ans (l'âge médian au premier départ dans l'échantillon est de 22 ans). Le quatrième groupe (n = 872 ; 20,2%) se caractérise par des départs précoces suivis d'une période relativement longue avant les premiers mariages. A l'âge de 30 ans, seulement 30% des individus sont mariés, dont moins de la moitié avec des enfants. Le dernier groupe, (n = 215 ; 5%) est constitué d'individus qui ne quittent

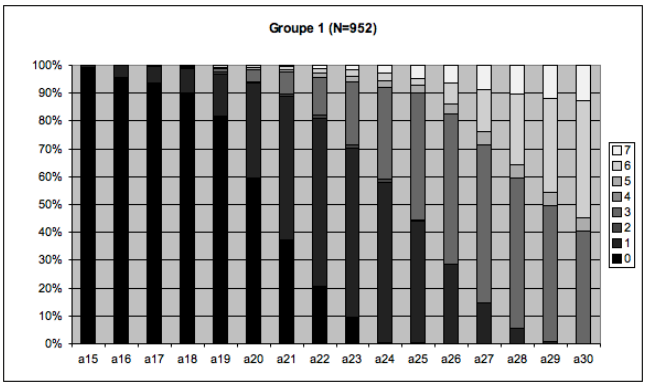

FIG. 1 – Groupe 1

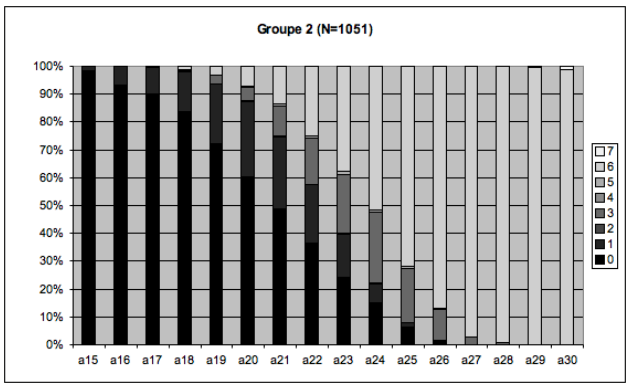

FIG. 2 – Groupe 2

pas le foyer parental. Il est intéressant de noter que ce groupe est réuni avec le groupe trois (départs tardifs) dans la solution de classification à quatre groupes. Des tests du χ^2 nous permettent de rejeter d'emblée l'hypothèse d'indépendance entre l'appartenance à l'un des groupes et les variables cohorte de naissance et genre. Les cohortes de naissance ont été définies selon la discrétisation optimale opérée par la procédure CHAID d'induction d'arbre dans le cas où la variable prédite est l'appartenance à un groupe et la variable prédictrice l'année de naissance. En fait on obtient ainsi 5 cohortes dont on a réuni la première et la dernière avec leur cohorte adjacente respective pour faciliter l'interprétation et éviter d'avoir des cohortes avec trop peu d'individus. Les trois cohortes résultantes sont définies selon les dates de naissance suivantes : de 1909 à 1947, de 1948 à 1956 et de 1957 à 1972.

Afin de capter de manière plus précise l'effet de cohorte et l'effet de genre sur l'appartenance à l'un des groupes découverts, une régression logistique a été faite pour chacun des cinq groupes. Plusieurs variables ont été introduites dans le modèle, comme la catégorie socio-professionnelle des parents, le niveau d'éducation des parents ou encore la langue dans laquelle a été rempli le questionnaire (et qui permet d'obtenir une approximation de la région linguistique dans laquelle habite l'individu). Les variables retenues pour chaque groupe sont la cohorte de naissance ainsi que le sexe, excepté pour le groupe 4. Dans celui-ci, le genre ne semble pas être un facteur favorisant l'appartenance ou non. Les coefficients des régressions logistiques sont présentés dans le tableau 2.

On observe qu'un individu a environ 2,5 fois plus de chances d'être dans le groupe 5 ou 2 fois plus de chances d'être dans le groupe 2 s'il est une femme plutôt qu'un homme. Un homme a quant à lui 2,3 fois plus de chances qu'une femme d'être dans le groupe 3

Tab. 2 – Régressions logistiques

	groupe 1	groupe 2	groupe 3	groupe 4	groupe 5
cohorte (1909-1947)	1	1	1	1	1
cohorte (1948-1956)	1.334 ***	0.839*	0.650 ***	2.309 ***	0.347 ***
cohorte (1957-1972)	1.256 **	0.506 ***	0.861*	3.127 ***	0.202 ***
homme	0.787 ***	0.517 ***	2.404 ***	1.082	0.485 ***
problèmes argent	0.814 **	-	-	-	-
constante	0.284 ***	0.564 ***	0.293 ***	0.120 ***	0.130 ***

*** Significatif au seuil de 1% ** Significatif au seuil de 5% * Significatif au seuil de 10%

plutôt qu'un autre. La variable de cohorte nous permet de voir l'évolution des chances d'appartenir à un groupe plutôt qu'à un autre ; ainsi, le groupe 2 et le groupe 3 montrent que les chances d'appartenir à ce groupe baissent dans les deux dernières cohortes par rapport à la première. Cet effet est encore plus marqué pour le groupe 5 (celui des individus ne quittant pas le foyer parental). Une variable supplémentaire a été introduite dans le modèle pour le groupe 1 ; ainsi, un individu n'ayant pas eu de problèmes d'argent dans sa jeunesse a moins de chances de se trouver dans le groupe 1.

3 Conclusion

Nous pouvons conclure que l'utilisation de l'optimal matching présente un intérêt certain pour l'analyse des parcours de vie. Couplée à une classification hiérarchique, elle a permis de mettre en évidence des groupes aux caractéristiques bien définies. Elle permet surtout d'aborder le parcours de vie dans sa totalité, en prenant en compte plusieurs événements, leur durée et leur chronologie et offre donc une perspective exploratoire intéressante pour l'analyse des séquences de manière générale.

Références

[1] A. Abbott, J. Forrest, « Optimal Matching Methods for Historical Sequences » *in Journal of Interdisciplinary History*, 26, 1986, pp. 471-494.

[2] A. Abbott, « A Primer on Sequence Methods » *in Organization Science*, Vol.1, No. 4, 1990, pp. 375-392.

[3] A. Abbott, A. Hrycak, « Measuring Resemblance in Sequence Data : An Optimal Matching Analysis of Musicians' Careers » *in The American Journal of Sociology*, Vol. 96, No. 1, Jul. 1990, pp. 144-185.

[4] A. Abbott, A. Tsay, « Sequence Analysis and Optimal Matching Methods in Sociology » *in Sociological Methods & Research*, Vol. 29, No. 1, Aug. 2000, pp. 3-33.

[5] J. B. Kruskal, «An overview of sequence comparison» *in David Sankof and Joseph B. Kruskal (eds), Time warps, string edits, and macromolecules. The theory and practice of sequence comparison*, Don Mills, Ontario : Adison-Wesley, 1983, pp. 1-44.

[6] S.B. Needlman, C. D. Wunsch, «A general method applicable to the search for similarities in the amino acid sequence of two proteins» *in Journal of Molecular Biology*, 48, 1970, pp. 443-453.