Classification de données biographiques : application à des trajectoires migratoires vers Cali (Colombie)

Alexander Estacio-Moreno, Olivier Barbary, Patrick Gallinari, Marie Piron

Rev. Statistique Appliquée, 2004, **LII** (4), 33-54

CLASSIFICATION DE DONNÉES BIOGRAPHIQUES : APPLICATION À DES TRAJECTOIRES MIGRATOIRES VERS CALI (COLOMBIE)

A. ESTACIO-MORENO[1,2], O. BARBARY[3], P. GALLINARI[2] et M. PIRON[1]

[1] *IRD – UR Géodes, 32 ave Henri Varagnat, 93143, Bondy Cedex*
[2] *LIP6, 8 rue du Capitaine Scott, 75015, Paris*
[3] *Cams, EHESS-MARSEILLE, 2 rue de la charité, 13002, Marseille*

RÉSUMÉ

La mobilité des hommes, en tant que mode de caractérisation et de différentiation des individus et des groupes sociaux, est un élément central pour l'analyse et la compréhension des dynamiques et des recompositions urbaines. Cependant, l'analyse des données longitudinales qui décrivent les différentes formes de mobilité (résidentielle, professionnelle, événements familiaux, etc.) pose encore d'importants problèmes méthodologiques.

Cet article présente tout d'abord un rapide état de l'art des principales méthodes d'analyse de données longitudinales. L'une de ces méthodes (la classification par mélange de densités) est ensuite détaillée et appliquée à l'étude de la mobilité résidentielle des habitants de Cali à partir des données d'une enquête rétrospective. Nous mettons en œuvre la méthode (en utilisant des modèles de Markov), et nous montrons comment elle fournit une typologie de trajectoires résidentielles, dont l'interprétation est facilitée par la représentation graphique des matrices de transition résultantes. Enfin la pertinence de la méthode est validée par sa comparaison avec une typologie obtenue par l'analyse harmonique qualitative sur les mêmes données.

Mots-clés : *Données longitudinales, Mobilité résidentielle, Mélange de densités, Modèles de Markov, Analyse harmonique qualitative.*

ABSTRACT

The human mobility, as a mean of characterization and differentiation of individuals and social groups, is an essential issue for analyzing and understanding the urban dynamics and recombinations. However, the analysis of longitudinal datasets describing various forms of mobility (residential, professional, family events, etc.) still poses important methodological problems.

This paper starts with a fast synthesis of scientific studies about the main methods used in longitudinal data analysis. Then, based on a retrospective survey, one of these methods (mixture model clustering using Markov models) is explained and applied to the study of the residential mobility of Cali's inhabitants. We found the method able to provide an interesting typology of residential trajectories. Then, the graphical representation of the resulting transition matrices

facilitates the interpretation. Finally the resulting typology is compared with that obtained by nominal harmonic analysis for the same data.

Keywords : *Longitudinal data, Residential mobility, Mixtures models clustering, Markov models, Nominal harmonic analysis.*

1. Introduction

A bien des égards, et plus encore dans le contexte contemporain de la généralisation de l'urbanisation, la ville peut être considérée comme le produit de la mobilité des hommes et des biens. La mobilité, dans ses différentes dimensions (spatiale, sociale, étapes de la carrière éducative et de la constitution de la famille, etc.), caractérise et différencie les individus et les groupes sociaux et devient un élément central pour l'analyse et la compréhension des dynamiques et des recompositions urbaines (Dureau *et al.*, 2000). Les sources statistiques principales qui décrivent ces différents types de mobilité à différentes échelles spatiales et temporelles sont les enquêtes biographiques rétrospectives, où sont recueillies sur un échantillon d'individus les trajectoires définies par les changements d'état des variables résidentielles, professionnelles, concernant les événements familiaux, etc. Ces dernières années ont vu des avancées significatives dans les domaines de la production (Antoine *et al.*, 1999) et de l'analyse des données biographiques (données appelées aussi longitudinales. Deville et Saporta, 1980, Courgeau et Lelièvre, 1989, Barbary, 1996). L'un des débats théoriques importants qui traverse l'évolution récente de ce champ des sciences sociales est celui de la conciliation, plus que de l'opposition, entre « individualisme méthodologique » et « déterminisme »; c'est-à-dire, en termes plus explicites, sur quels concepts et méthodes fonder une analyse qui considère l'individu comme un acteur central dans l'ensemble social, doté par conséquent d'une autonomie d'action, sans pour autant ignorer les déterminations et les contraintes que les structures sociales font peser en retour sur les stratégies et les comportements individuels et collectifs (*cf.* par exemple, Antoine *et al.*, 1999). Ainsi la recherche sur le traitement des données biographiques se trouve face à des problèmes méthodologiques nouveaux que posent l'analyse simultanée des différents types de mobilité et leur articulation avec les dynamiques des structures du milieu urbain (différentiation et recomposition sociodémographique et fonctionnelle des espaces, concentration ou ségrégation résidentielle, etc.).

Dans le champ de la statistique, on peut distinguer deux familles de méthodes pour analyser les données biographiques (Van Der Heijden, 1987, Fénelon, 1997). Une première approche modélisatrice, souvent utilisée par les économistes, les épidémiologues et les démographes, est basée sur la théorie des processus stochastiques. On définit une variable d'état évoluant dans le temps et on cherche à quantifier les effets d'autres variables sur celle-ci en utilisant des modèles log-linéaires, des modèles logit et probit, des modèles de survie de Cox ... (voir Malinvaud, 1981, Cox et Oakes, 1984, Davies et Crouchley, 1984, Courgeau et Lelièvre, 1989). L'autre approche est celle de l'analyse typologique. Elle est basée sur des méthodes désormais classiques en analyse des données : analyses factorielles et classifications automatiques (ACP, AFC, méthodes hiérarchiques d'agrégation ou de partitionnement, nuées dynamiques etc., *cf.* Lebart *et al.*, 2002). L'analyse harmonique qualitative

– AHQ –, développée par Deville et Saporta (Deville et Saporta, 1980, Saporta, 1981, Deville, 1982), a été utilisée pour le traitement de données biographiques sur les trajectoires résidentielles (Barbary et Pinzon, 1998, *cf. infra*). Cette approche permet de différencier des types de trajectoires à partir d'un certain nombre de caractéristiques sans en privilégier une.

Parallèlement à ces travaux, différentes communautés issues de l'intelligence artificielle ont développé ces dernières années des techniques pour l'analyse des trajectoires ou des séquences. Elles ont appliqué ces techniques à différents domaines où l'individu joue un rôle central, comme l'analyse de données client, la modélisation du comportement d'utilisateurs sur le Web ou à d'autres domaines comme l'analyse de séquences biologiques, du langage, etc. Parmi celles-ci, les modèles stochastiques utilisant des algorithmes d'apprentissage qui permettent d'ajuster leurs paramètres, sont les plus employées (par exemple Cadez *et al.*, 2000, Smyth, 1997, Smyth, 1999, Gaffney et Smyth, 1999).

Ces derniers travaux ont pour cadre l'estimation de densité de probabilité semi paramétrique (Bishop, 1995), et utilisent, en tant que modèle, un mélange de densités. Cette approche concilie d'une certaine façon la modélisation et l'analyse typologique car on obtient en sortie une classification des trajectoires individuelles et des modèles de trajectoire moyenne pour chacune des classes. Elle est donc particulièrement intéressante si l'on veut essayer d'enrichir les connaissances obtenues d'une part de l'analyse causale (modèles) et d'autre part de l'analyse descriptive (typologies), en s'attaquant à l'étude des interactions entre les individus et les structures auxquelles ils appartiennent. Les modèles générateurs de chacune des classes constituent alors un résultat intégrable dans un système qui viserait à appréhender ces interrelations. Plus modestement, dans cet article, à partir du cadre théorique dû à Cadez *et al.* (2000), nous abordons la classification de trajectoires comme étant un problème d'estimation de densité de probabilité, et nous proposons d'utiliser un mélange de densités qui permet d'obtenir une classification tout en fournissant des modèles pour chaque classe. En utilisant ensuite des données issues d'une enquête effectuée à Cali (Colombie) en 1998, nous montrons comment cette méthode est applicable à des données biographiques et conduit à des résultats descriptifs pertinents pour l'analyse des mobilité résidentielles.

2. Mélange de densités

Dans l'estimation de densités on essaie de modéliser une densité de probabilité $p(x)$ à partir des données observées $X = (x_1, x_2, \ldots x_N)$, que l'on suppose issues de cette densité. Avec le mélange de densités, la densité de probabilité est modélisée par une combinaison linéaire de densités composantes. Ainsi :

$$p(x) = \sum_{k=1}^{K} p(x/k)P(k)$$

où K est le nombre de composantes du mélange, les $P(k)$ sont les paramètres du mélange (la probabilité *a priori* pour que la donnée ait été générée par la composante k du mélange, avec $\sum_{k=1}^{K} P(k) = 1$), et les $p(x/k)$ sont les densités composantes.

Une propriété importante de ce modèle est que, pour différents choix de composantes, on peut approcher une densité quelconque avec une précision qui ne dépend que du nombre de composantes et du choix des paramètres.

2.1. Classification par Mélange de Densités (CMD)

Soit $D = \{D_1, D_2, \ldots D_N\}$ un ensemble de données.

D_i sont les données observées pour l'individu i : $D_i = \{d_{i1}; d_{i2}, \ldots d_{in_i}\}$, où d_{ij} est la $j^{\text{ème}}$ trajectoire de l'individu i, et $1 \leqslant j \leqslant n_i$. Rappelons qu'une trajectoire est définie par les changements d'état d'une variable résidentielle, professionnelle, etc. Dans le processus considéré ici l'individu i est décrit par n_i trajectoires. Cadez *et al.* (2000) proposent un cadre de classification basée sur un mélange de densités, que l'on peut résumer comme suit :

1. Un individu i est tiré aléatoirement de la population.
2. L'individu est affecté à l'une des K classes, notée $k(1 \leqslant k \leqslant K)$, de poids $P(k)$ telle que $\sum_k P(k) = 1$.
3. A chaque classe k, on associe un modèle de génération de données [1] $p(D_i/\Theta_k)$, où D_i est la donnée de l'individu i, et Θ_k sont les paramètres de cette distribution de probabilité.

La distribution de probabilité des individus est donc vue comme une combinaison linéaire des modèles composants :

$$p(D_i/\Theta) = \sum_{k=1}^{K} p(D_i/\Theta_k)P(k) \tag{1}$$

Si on suppose que les observations sont conditionnellement indépendantes sachant les classes, alors la probabilité d'un individu de la classe k est donnée par :

$$p(D_i/\Theta_k) = \prod_{j=1}^{n_i} p(d_{ij}/\Theta_k) \tag{2}$$

Sous la même hypothèse d'indépendance conditionnelle, mais cette fois ci entre individus, on obtient la vraisemblance totale de l'ensemble des observations D :

$$p(D/\Theta) = \prod_{i=1}^{N} p(D_i/\Theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} p(D_i/\Theta_k)P(k) \tag{3}$$

Pour spécifier complètement le modèle, il faut estimer les paramètres Θ à partir des données D, c'est-à-dire les paramètres de chaque classe, $\Theta_k = \{\Theta_1, \Theta_2, \ldots \Theta_K\}$, et les K poids $P(k)$. Cette estimation peut être faite par l'algorithme Espérance –

[1] Un modèle de génération de données est un modèle permettant de calculer la probabilité qu'un individu ait un vecteur D_i de données, sachant qu'il appartient à la classe k.

Maximisation (EM, *cf.* Dempster *et al.*, 1977, McLachlan et Krishnan, 1997), en la considérant comme un problème de données manquantes. Ici les données manquantes sont les classes des individus.

2.2. Cas de modèles markoviens

Considérons maintenant que chaque individu n'a qu'une trajectoire et que les modèles pour chaque classe $p(D_i/\Theta_k)$ sont des modèles de Markov de premier ordre. Ce choix des modèles est justifié dans la section 3.1. Les paramètres du modèle pour chaque classe (Θ_k) sont donc : un vecteur de probabilité d'état initial $\pi_k(e)$, et une matrice $m \times m$ de probabilités de transition $a_k(e_t/e_{t-1})$, où les e sont des états discrets de l'espace d'états E $(1 \leqslant e \leqslant m)$. De la définition d'une chaîne de Markov on déduit :

$$p(D_i/\Theta_k) = \pi_k(e_{i,1}) \prod_{t=2}^{T_i} a_k(e_{i,t}/e_{i,t-1}), \quad \text{pour } 1 \leqslant k \leqslant K \tag{4}$$

L'équation (4) est la vraisemblance d'une trajectoire individuelle conditionnée par l'appartenance de l'individu à une classe particulière de paramètres Θ_k, où :

$e_{i,t}$, est l'état à l'instant t de la trajectoire de l'individu i.

T_i, est la longueur de la trajectoire de l'individu i.

Les équations (1), (3) et (4) spécifient complètement le modèle à partir des données observées D. On modélise d'abord les trajectoires individuelles (équations 1 et 4) et ensuite les données pour l'ensemble des individus (équation 3).

Pour l'estimation de paramètres l'algorithme EM s'applique dans ce contexte de la manière suivante :

Etape *E* : on calcule les probabilités conditionnelles d'appartenance aux classes de chaque individu, $p(i \in k/D_i, \Theta)$, pour chacun des K modèles de classe avec les paramètres courants Θ.

$$p(i \in k/D_i, \Theta) = \frac{p(D_i/\Theta_k)P(k)}{\sum_{u=1}^{K} p(D_i/\Theta_u)P(u)} \qquad 1 \leqslant k \leqslant K \tag{5}$$

Etape *M* : on met à jour les paramètres courants Θ, en pondérant chaque individu par sa probabilité conditionnelle d'appartenance aux classes, $p(i \in k/D_i, \Theta)$:

$$P(k)^{Nouveau} = \frac{1}{N} \sum_{i=1}^{N} p(i \in k/D_i, \Theta)$$

$$\pi_k^{Nouveau}(e) = \frac{\sum_{i=1}^{N} p(i \in k/D_i, \Theta)\delta(e, e_{i,1})}{\sum_{i=1}^{N} p(i \in k/D_i, \Theta)}$$

$$a_k^{Nouveau}(e_t/e_{t-1}) = \frac{\sum_{i=1}^{N} p(i \in k/D_i, \Theta) r_i^{e_{t-1} \to e_t}}{\sum_{i=1}^{N} p(i \in k/D_i, \Theta) r_i^{e_{t-1} \to}} \quad (6)$$

où :

$$\delta(e, e_{i,1}) = \begin{cases} 1 & Si \quad e = e_{i,1} \\ 0 & Sinon \end{cases}, \quad \text{et} \quad 1 \leqslant e \leqslant m.$$

$r_i^{e_{t-1} \to e_t}$ est le compte des transitions depuis l'état e_{t-1} à l'état e_t dans la trajectoire de l'individu i.

$r_i^{e_{t-1} \to}$ est le compte des transitions depuis l'état e_{t-1} à n'importe quel état dans la trajectoire de l'individu i.

Et $1 \leqslant e_t, e_{t-1} \leqslant m$.

A partir de l'équation (4) on calcule les nouveaux $p(D_i/\Theta_k)$ qui vont servir dans l'étape E à recalculer, par l'équation (5), les nouveaux $p(i \in k/D_i, \Theta)$.

On itère ces deux pas jusqu'à trouver un maximum local de la fonction cible de vraisemblance totale et les valeurs finales des paramètres. Le problème lié au choix initial des probabilités $p(i \in k/D_i, \Theta)$ sera discuté dans le cadre de l'application à la section 3.3.

3. Mise en œuvre

3.1. Présentation des données

Les données que nous utilisons proviennent d'une enquête biographique effectuée à Cali (Colombie) en 1998. Elle a été réalisée dans le cadre d'un programme de recherche en coopération entre le CIDSE (Centre de Recherche et de Documentation Socio-économique de l'Université du Valle) et l'IRD (Institut de Recherche pour le Développement), portant sur les différentes modalités de la mobilité spatiale et sociale, de l'insertion résidentielle, économique et culturelle des populations afrocolombiennes et non afrocolombiennes à Cali (Barbary, 1999, 2001). Parmi l'ensemble d'information que fournit l'enquête, nous ne nous intéressons ici qu'à celles concernant la mobilité résidentielle (les changements de résidence des individus au cours de leur vie). Chaque observation individuelle est donc constituée d'une seule trajectoire.

Nous considérons la trajectoire résidentielle résultant du codage des lieux de résidence successivement habités par un individu. La variable d'état contenant

l'information (Lieux de résidence) est constituée du code du pays, du code de la localité (département, municipe), et de l'appartenance à la zone urbaine ou rurale du municipe. Etant donnée la taille relativement restreinte de l'échantillon (1179 individus migrants), conserver tout le détail de cette information conduit à un ensemble de trajectoires très dispersées, au sens des états possibles, et donc à une variabilité extrême des objets à classer. Pour «rapprocher» les trajectoires et diminuer la taille de l'espace des états, les lieux de résidence ont été recodés en 36 agrégats géographiques («régions») qui définissent les 36 états possibles de la variable (voir annexe A).

Nous choisissons donc des chaînes de Markov de premier ordre pour modéliser les séquences de changement d'état dans chaque classe. Les états de la chaîne de Markov représentent, selon un codage qu'on précisera plus loin, les étapes des trajectoires individuelles. Nous cherchons ainsi une approximation simple à la dynamique des trajectoires comme base pour mettre au point la méthode. De fait, une chaîne de Markov d'ordre 1 ne rend compte que de l'ordre de la séquence d'états d'une trajectoire, en prenant en compte la probabilité de l'état initial et la dépendance entre deux états consécutifs de la trajectoire.

3.2. Prise en compte du temps, gestion des données censurées et codage

Dans l'analyse des trajectoires biographiques le temps est un paramètre fondamental, qui doit être géré de manière pertinente en fonction du thème et des objectifs de l'étude. Il faut définir donc *le temps de l'analyse, c'est-à-dire l'origine et l'étendue de la période temporelle d'observation et d'analyse des trajectoires* (voir Pinzon, 1998). Ici, le temps d'analyse choisi est le temps biographique par rapport à un événement, en considérant les années antérieures à la dernière arrivée à Cali (événement de référence). Nous nous fixons une période d'analyse de 30 ans (durée totale de suivi des individus). De même nous fixons à 34 le nombre de classes à obtenir. Dans le cadre de ce travail, ces choix nous sont imposés par le souci d'évaluation des résultats et de validation du modèle proposé par rapport à une typologie qui a déjà été obtenue à partir de l'application de l'AHQ (voir Barbary *et al.*, 2002). En effet l'analyse par AHQ qui a été faite sur ces données, et qui a une pertinence au sens des résultats obtenus et de leur interprétation, est fondée sur ces mêmes choix.

Dans l'observation longitudinale, quel que soit le type de temps et la longueur de la période d'analyse, il y a en général des individus qui ne sont pas observés durant la période d'analyse complète. C'est le cas si l'on considère les mouvements naturels (naissance ou décès) : certains individus entrent ou sortent de l'observation à un moment donné de la période d'analyse. C'est ce que les spécialistes appellent des données censurées. En ce qui concerne les trajectoires résidentielles observées sur trente ans avant la dernière arrivée à Cali, les individus qui ont moins de trente ans à leur dernière arrivée à Cali sont censurés à gauche. Pour prendre en compte cette censure, nous ajouterons aux 36 états possibles de la variable résidentielle un état codé 0, pour représenter les durées de censure.

Les équations développées dans la section 2.2 définissent donc le modèle que nous appliquerons. Un modèle de Markov exige en entrée la suite des états de la trajectoire de chaque individu. Nous avons envisagé deux codages :

- En considérant la durée des étapes, on fait apparaître le code d'état un nombre de fois proportionnel à la durée de séjour (Codage A). Dans l'application nous

avons testé l'unité de temps minimale (un an) définie par les données brutes. Chaque code de la trajectoire représente une permanence d'un an dans l'état. Pour la trajectoire de l'individu 1 de l'échantillon, par exemple, on obtient :

`11,15,15,15,15,15,15,15,15,15,15,15,11,11,11,11,11,11,11,11,11,11,11,11,11,10,10,10,10`

Cet individu est donc resté une année dans l'état 11, ensuite douze années successives dans l'état 15, puis est retourné pendant treize années successives à l'état 11 et enfin il passe quatre années successives dans l'état 10 avant d'arriver à Cali.

- En ne considérant que l'ordre des états sans tenir compte de la durée de séjour dans chaque étape (Codage B). Si l'on ne prend, pour le même individu 1, que la succession des états on obtient : `11,15,11,10`

  Cet individu a connu dans sa trajectoire tout d'abord l'état 11, ensuite l'état 15, puis est retourné à l'état 11 et enfin il passe à l'état 10 avant d'arriver à Cali.

3.3. Initialisation du modèle

Dans l'algorithme EM, le maximum local de la vraisemblance obtenue à la fin de la mise en route de l'algorithme dépend du choix initial des probabilités $p(i \in k/D_i, \Theta)$. En effet, la vraisemblance maximale obtenue par l'algorithme n'est pas forcément le maximum global. Autrement dit, des choix différents de probabilités initiales peuvent fournir des maxima locaux de la vraisemblance différents.

Nous avons envisagé deux alternatives pour l'initialisation des paramètres :

L'initialisation aléatoire sur $p(i \in k/D_i, \Theta)$. Les trajectoires individuelles sont distribuées de façon aléatoire dans les K classes. Ainsi chaque individu est affecté aléatoirement à une et une seule des 34 classes.

L'initialisation par le résultat de l'AHQ. Cette initialisation reprend la classification obtenue par l'AHQ pour distribuer les individus dans les classes. Les probabilités $p(i \in k/D_i, \Theta)$ initiales sont donc égales à un pour la classe d'appartenance donnée par l'AHQ, et nulles pour toutes les autres classes. C'est un cas limite puisque l'on injecte le résultat qu'on cherche à obtenir dès l'initialisation. Nous l'utilisons à titre de référence.

En résumé, la démarche que nous avons développée comprend 4 étapes principales :

1. le codage de la variable d'état, lieux de résidence, à partir de l'information biographique collectée.

2. les différents essais de CMD pour les deux types de codages proposés et les différentes initialisations.

3. le calcul des tableaux de fréquences représentant les profils de mobilité moyenne des individus de la classe pour la variable et l'élaboration des graphiques correspondants (Barbary, 1996).

4. l'élaboration des graphiques des matrices de transition [2] .

4. Résultats et discussion

4.1. Comportement de l'algorithme et similarité des classifications

Pour chacun des deux codages, nous avons effectué cinq essais différents d'initialisation : quatre initialisations aléatoires sur les probabilités $p(i \in k/D_i, \Theta)$, et une initialisation par les résultats de l'AHQ. L'algorithme s'arrête dès que la variation du logarithme de la vraisemblance entre deux itérations successives n'excède pas 0.01%, ou lorsque nous atteignons l'itération numéro 500. Pour chaque codage, parmi toutes les différentes classifications résultantes nous conservons celle dont la valeur du logarithme de la vraisemblance finale est la plus élevée.

Draier et Gallinari (2001), proposent de mesurer la similarité de deux classifications (C_A et C_B, issues des deux algorithmes A et B) à partir d'un critère basé sur la théorie de l'information; en calculant d'abord les probabilités *a priori* $P(C_A = k)$ qu'un individu quelconque appartienne à la classe k de la classification C_A (également pour C_B), et ensuite $P(C_A = k_A, C_B = k_B)$ la probabilité jointe qu'un individu quelconque appartienne à la classe k_A de la classification C_A et à la classe k_B de la classification C_B. La quantité d'information mutuelle entre les deux distributions est donnée par (7). Si *MI* est élevée, les deux algorithmes ont identifié des structures similaires dans l'ensemble de données.

$$MI(C_A, C_B) = \sum_{i \in C_A} \sum_{j \in C_B} P(C_A = i, C_B = j) \log \frac{P(C_A = i, C_B = j)}{P(C_A = i) P(C_B = j)} \quad (7)$$

A partir de chaque classification retenue pour chacun des codages et de celle de l'AHQ, nous comparons ces différents codages, en nous servant d'une version normalisée de *MI*, notée MI_N (voir tableau 1) :

TABLEAU 1
Information mutuelle normalisée entre CMD et AHQ pour les codages A et B

MI_N	Cl. codage A	Cl. codage B	Cl. AHQ
Cl. codage A	1	0.49	0.52
Cl. codage B	0.49	1	0.54
Cl. AHQ	0.52	0.54	1

De manière générale, les deux codages conduisent à des écarts de structures similaires vis-à-vis de la classification obtenue par l'AHQ. Nous conservons donc les résultats du codage plus simple : le codage B. Le tableau 2 donne les niveaux

[2] Le programme C qui a été développé pour effectuer la classification avec le mélange de densités est disponible sur demande à A. Estacio-Moreno.

TABLEAU 2
Résultats des initialisations avec le codage B

Initialisation	Logarithme de la vraisemblance initiale	Logarithme de la vraisemblance finale	Nombre d'itérations
Aléatoire*	–6318,97	–5968,95	164,75
AHQ	–6298,20	–6216	112

* Pour l'initialisation aléatoire on a pris la moyenne des valeurs obtenues pour chacune des 4 initialisations effectuées.

de vraisemblance et le nombre d'itérations nécessaire pour atteindre la convergence avec ce codage.

L'initialisation par le résultat de l'AHQ a une convergence plus rapide, mais conduit à un maximum de la vraisemblance inférieur. Cela étant, avec cette initialisation, le pourcentage d'individus retrouvés dans les mêmes classes par les deux méthodes est de 90%, ce qui démontre que la CMD est capable de « conserver » le résultat de l'AHQ. C'est-à-dire qu'il existe un maximum local de la vraisemblance au sens de la CMD qui correspond, à peu de chose près, au résultat typologique de l'AHQ. C'est un premier résultat qui valide, en quelque sorte de manière croisée, les deux méthodes.

4.2. La comparaison du résultat typologique avec celui de l'AHQ

A partir de la matrice de confusion entre la classification fournie par l'AHQ et celle de la CMD avec l'initialisation aléatoire (annexe B [3]), nous avons construit deux indicateurs synthétiques pour chaque classe $(i, j = 1, \ldots, 34)$:

$$1.\ C_1(i) = \frac{\max\{card(C_{CMD,i} \cap C_{AHQ,k}), k = 1, \ldots, 34\}}{card(C_{CMD,i})}$$

$$2.\ C_2(j) = \frac{\max\{card(C_{AHQ,j} \cap C_{CMD,k}), k = 1, \ldots, 34\}}{card(C_{AHQ,j})}$$

Avec l'aide des graphiques de C_1 et C_2 (annexe B) on peut distinguer d'une part les classes de la CMD « similaires » à celles de l'AHQ (les indicateurs sont $C_1 \geqslant 35$ % et $C_2 \geqslant 44$ %) : sur les 34 classes obtenues, 21 sont dans ce cas; et d'autre part les «nouvelles» classes de la CMD, c'est-à-dire celles pour lesquelles la classe majoritaire correspondante de l'AHQ est déjà une classe « similaire » à une autre classe de la CMD (pour la plupart des « nouvelles » classes C_1 est inférieur à 30 %) : 13 classes sur 34 sont dans ce cas [4] .

Mais, comment comparer plus finement les typologies issues des deux méthodes? Quelles différences existent entre les classes « similaires » dans les deux

[3] Les classes de la classification obtenue par la CMD ont été rénumérotées pour faire apparaître les effectifs les plus élevés dans la diagonale de la matrice.

[4] Les classes 14 et 15, apparaissant respectivement dans les types 2 et 3 du groupe des classes similaires du graphique de l'annexe B n'appartiennent pas, par définition, à ce groupe. En effet, pour ces deux classes les deux indicateurs (C_1 et C_2) sont calculés sur des effectifs différents dans la matrice de confusion.

méthodes? Quel type d'information originale fournissent les «nouvelles» classes de la CMD?

Pour caractériser les classes de l'AHQ nous disposons des graphiques des profils de mobilité moyenne des individus de chaque classe; ils montrent la distribution de présence des individus dans les états durant chaque intervalle qui divise la période d'analyse. Pour caractériser les classes de CMD, nous adoptons une approche similaire : nous construisons des profils et leur graphiques donnant le pourcentage de présence des individus dans chaque état pour chacune des trente années de la période précédant l'arrivée à Cali. Aussi bien dans les graphiques des profils de mobilité de l'AHQ que dans ceux de la CMD, l'instant de la dernière arrivée à Cali est le temps zéro (0). L'antériorité par rapport à cet événement est donc indiquée par des chiffres négatifs dans ces graphiques (voir annexe C).

Parmi les classes « similaires » retrouvées par les deux algorithmes, trois types différents peuvent être identifiés :

1. Les classes « presque » identiques dans les deux typologies ($C_1 \geqslant 65$ % et 54 %$\leqslant C_2 \leqslant 78$ %). Il s'agit des classes 1, 2, 6, 8, 9, 12, 22, 23. Les profils de mobilité des classes obtenus par les deux méthodes sont alors très semblables. Dans l'annexe C, par exemple, nous comparons les classes 6 des deux typologies. Pour la CMD, comme pour l'AHQ, la plupart des individus ont connu des étapes dans le nord du Valle urbain (état 26) ou le sud du Valle urbain (état 24), avant leur dernière entrée à Cali. Cette classe contient des individus nés dans le nord et le sud du Valle urbain. La plus forte densité de séjour dans le sud du Valle urbain se situe entre 7 et 11 ans avant la dernière arrivée à Cali. Pour l'AHQ, la plupart des individus sont nés dans le nord du Valle urbain et connaissent une étape migratoire dans le sud du Valle urbain, généralement entre 5 et 11 ans avant leur dernière arrivée à Cali. Les profils des deux classes sont très proches, la seule différence est que la CMD ajoute à la classe de l'AHQ quelques individus nés dans le sud du Valle urbain.
2. La majorité de la classe de l'AHQ est dans la classe de la CMD ($C_2 \geqslant 61$ %), et les individus communs constituent environ la moitié de la classe de la CMD (43 %$\leqslant C_1 \leqslant 59$ %). Les profils de mobilité des classes de la CMD (3, 4, 5, 7, 10, 18 19, 20) identifient les mêmes traits caractéristiques que dans les classes de l'AHQ (c'est-à-dire les transitions les plus fréquentes). L'autre moitié des individus formant les classes de la CMD partagent, pour la plupart, une ou plusieurs transitions avec les trajectoires les plus fréquentes de chaque classe, mais s'en différentient, soit par des lieux de naissances différents, soit encore parce qu'il s'agit de migrants plus âgés avec des durées de trajectoires plus longues.
3. Dans les classes 13, 16, 25, 29 et 31, environ la moitié de la classe de l'AHQ est dans la classe correspondante de la CMD (44 %$\leqslant C_2 \leqslant 66$ %), mais ces individus représentent, en général, moins de la moitié de la classe de la CMD (35 %$\leqslant C_1 \leqslant 42$ %). De même que dans le cas précédent, le trait caractéristique des classes de l'AHQ demeure visible dans les profils de mobilité des classes de la CMD, mais ces dernières classes incorporent des individus atypiques (suivant des trajectoires sans transitions partagées).

Les « nouvelles » classes fournies par la CMD sont de petite taille (8 à 21 individus). Elles peuvent être divisées en deux groupes : les classes marquées (plus de 50 % des individus) par des trajectoires « simples » (entre deux et quatre transitions : classes 14, 15, 21, 28, 32 et 34), et les classes de trajectoires complexes (classes 11, 17, 24, 26, 27, 30 et 33) [5] .

Le premier groupe de classes, facilement interprétable, amène une information « nouvelle » par rapport à l'AHQ. La classe 28 (la « nouvelle » classe 9), par exemple (annexe D), est formée de vieux migrants de Tolima, Huila, Caquetá, Putumayo urbain (72 %, état 32), ayant connu des étapes résidentielles antérieures à Cali (états 3 et 4), durant une partie importante de la période d'analyse, et d'autres étapes de retour à leur lieux d'origine ou ailleurs avant leur dernière arrivée à Cali. La CMD regroupe donc les trajectoires bien définies des vieux migrants du Tolima, Huila, Caquetá, Putumayo urbain, alors qu'elles étaient divisées par la typologie de l'AHQ.

L'interprétation des classes du deuxième groupe est moins facile; néanmoins, d'après les graphiques on arrive à les expliquer partiellement. La classe 24 (la « nouvelle » classe 6), par exemple (annexe E), regroupe des vieux migrants du sud et du nord du Valle urbain, avec des étapes de résidence antérieures à Cali et des retours aux lieux d'origines avant leur dernière arrivée à Cali. Mais on y trouve aussi des individus nés dans l'intérieur du Nariño urbain et des trajectoires vers Cali beaucoup plus complexes avec des destinations migratoires très diversifiées (voir partie supérieure du graphique). Certainement, la difficulté d'interprétation est due à la complexité des trajectoires, et au fait que l'algorithme rassemble des trajectoires générées par un même modèle sous la contrainte d'un nombre fixe de classes.

5.3. Matrices de transition : une aide pour l'interprétation de la typologie.

Dans les cas où l'interprétation de la typologie à partir des profils de mobilité des classes est difficile, on peut : (i) soit retourner aux trajectoires des individus constituant la classe; cependant si la taille de la classe est importante ce travail est fastidieux; (ii) soit utiliser la matrice de transition de la classe pour examiner la dynamique des changements d'état. Dans la pratique, on construit plutôt la matrice de transition jointe, parce qu'elle inclut la vraisemblance de l'état de départ et donne plus d'information. Si on appelle $p(j/i)$ la probabilité de transition à l'état j sachant qu'on est dans l'état i, alors $p(i,j) = p(j/i)p(i)$ est la probabilité jointe : la probabilité qu'un individu ait une résidence dans l'état i suivi d'une résidence dans l'état j.

Pour montrer comment cette aide à l'interprétation permet de lever certains doutes et conduit à une description plus précise des trajectoires, nous détaillerons un exemple. Le profil de mobilité de la « nouvelle » classe 6 est, comme on l'a dit, complexe; il y a en particulier deux points dans le temps où l'on ne peut pas suivre avec certitude la trajectoire la plus commune : 11 ans et 6 ans avant la dernière arrivée à Cali. Dans le premier cas (11 ans), on peut se demander si la transition majoritaire, celle des individus partant de l'Antioquia et du Viejo Caldas Urbain (état 34), s'effectue vers le reste de la Colombie (état 35) ou vers le nord du Valle rural (état 25). Dans le deuxième (6 ans), on s'interroge sur la transition des individus

[5] Nous renommons les nouvelles classes de la CMD selon leur ordre d'apparition dans la matrice de confusion. Ainsi, la classe 11 est la « Nouvelle classe 1 », la classe 14 est la « nouvelle classe 2 », et ainsi de suite.

partant du reste de la Colombie vers le sud du Valle urbain (état 24) ou vers le nord du Valle urbain (état 26). D'après le graphique de la matrice de transition jointe (annexe F), on peut d'abord confirmer que, pour presque tous les individus de la classe, la première migration vers Cali a lieu depuis le sud du Valle urbain (transition de l'état 24 vers l'état 3), qu'il s'agisse donc de migrations directes depuis les villes du sud du département ou d'une séquence : Nord du Valle urbain, Sud du Valle urbain, Cali. Il est clair, d'autre part, que la transition depuis l'Antioquia et le Viejo Caldas Urbain se réalise vers le nord du Valle urbain (transition de l'état 34 vers l'état 26), et que la transition depuis le reste de la Colombie existe vers l'ensemble des villes du Valle (transitions de l'état 35 vers les états 26 et 24), plus fréquente cependant vers le sud du Valle urbain. On peut alors conclure, chez ces vieux migrants du sud et du nord du Valle urbain ayant tous connu au moins deux migrations à Cali, à l'existence de deux trajectoires migratoires principales vers Cali : la première passant par l'Antioquia et le Viejo Caldas Urbain, puis le nord du Valle rural, la deuxième par des migrations plus lointaines dans le reste de la Colombie suivies de retours aux lieux d'origines du Valle urbain.

5. Conclusions

La CMD repose sur un codage simple et l'initialisation aléatoire conduit à une vraisemblance un peu supérieure à celle atteinte avec l'initialisation AHQ. Peut-on supposer que le maximum global est proche du maximum local atteint par l'initialisation aléatoire ? Quoi qu'il en soit, au sens de la maximisation de la vraisemblance des trajectoires observées par un mélange de chaînes de Markov, on peut conclure que l'initialisation aléatoire fournit un meilleur résultat que l'initialisation par les classes de l'AHQ.

Il y a donc, du point de vue de l'analyse typologique, une « logique » différente d'agrégation des trajectoires par les deux algorithmes. On peut avancer une hypothèse qui explique probablement en grande partie cette différence. Dans le codage et l'algorithme de l'AHQ, la durée de présence dans les états joue un rôle très important dans la formation des classes. C'est en particulier le cas de la durée de la première étape, généralement longue, qui s'écoule dans le lieux d'origine entre la naissance de l'individu et sa première migration (le plus souvent à l'âge adulte). De ce fait, les classes de l'AHQ présentent pour la plupart une très forte homogénéité des lieux de naissances et des âges à la première migration, au détriment parfois de l'homogénéité de la séquence des lieux de migration qui viennent ensuite. En revanche, l'algorithme de la CMD avec le codage retenu (*cf.* 4.1) ne tient pas compte de ces durées mais exige que la majorité des transitions entre états (lieux) soit commune aux trajectoires d'une même classe. Il tend donc à constituer des classes plus mélangées en terme de lieux d'origine que celles de l'AHQ, mais plus homogènes si l'on considère la trajectoire spatiale dans son ensemble. D'un point de vue thématique, donc, la CMD donne un résultat typologique suffisamment proche de celui obtenu par l'AHQ pour affirmer sa validité en tant que technique de classification de trajectoires résidentielles : les principales trajectoires migratoires « simples » sont identifiées dans des classes cohérentes avec celles obtenues par l'AHQ. Cependant, le résultat typologique obtenu est également suffisamment différent de la classification de l'AHQ pour apporter

une information supplémentaire et donc valider la CMD comme un outil d'analyse typologique complémentaire de l'AHQ.

Au plan théorique et méthodologique, la CMD permet non seulement de concevoir, d'une façon probabiliste, l'appartenance des trajectoires aux classes, mais aussi de choisir les modèles composant le mélange, et de trouver les paramètres et les poids optimaux de chaque classe d'une façon itérative à partir des données observées. De plus, grâce à la représentation graphique des matrices de transition jointe, la CMD facilite l'interprétation des classes, et contribue ainsi à diminuer le temps nécessaire à cette étape.

Les matrices de transition sont un résultat important de la CMD. Elles peuvent être utilisées pour générer des trajectoires. Plusieurs modèles de simulation des réalités urbaines, comme la mobilité résidentielle et d'autres dynamiques sociodémographiques, ne se basent pour l'instant que sur la prise en compte de l'offre et de la demande d'un certain nombre de facteurs jugés déterminants : marché du logement, marché de l'emploi, etc. Le fait de leur intégrer des modèles qui permettent de générer les trajectoires des agents pourrait certainement leur permettre de mieux approcher la réalité.

Actuellement nos travaux visent à mieux prendre en compte le temps passé dans un état, par les individus, en le modélisant explicitement. Pour cela nous considérons des modèles semi markoviens, où l'on ne considère que les transitions entre états tout en introduisant une densité de durée dans chaque état. Une étape postérieure est d'étendre l'approche CMD à l'analyse de trajectoires multivariées, où chaque individu est décrit par plusieurs types de trajectoires, résidentielles, socioprofessionnelles, familiales, par exemple.

Bibliographie

ANTOINE P., BONVALET C., COURGEAU D., DUREAU F., LELIEVRE E., Groupe de réflexion sur l'approche biographique (1999), *Biographies d'enquêtes, bilan de 14 collectes biographiques.* INED, IRD, Réseau Socio-Economie de l'habitat, Collection Méthodes et savoirs n° 3, Paris, 336 p.

BARBARY O. (1996), *Análisis tipológico De Datos Biográficos En Bogotá.* Universidad Nacional de Colombia, Bogotá, 254 p.

BARBARY O., PINZON SARMIENTO L. M. (1998), *L'analyse harmonique qualitative et son application à la typologie de trajectoires individuelles*, in « Mathématique, informatique et sciences humaines », n° 144, E.H.E.S.S. Paris, pp. 29-54.

BARBARY O. (1999), *Observar Los Hogares Afrocolombianos En Cali, Problemas teóricos Y metodológicos Ilustrados.* En documentos de trabajo 38 CIDSE. Universidad del Valle, Cali, 98 p.

BARBARY O. (2001), *Mesure et réalité de la segmentation socio-raciale : une enquête sur les ménages afrocolombien à Cali*, Population, vol. 56 n° 5, Paris, pp. 773-810.

BARBARY O. (coord.), DUREAU F., HOFFMANN O. (2002), *Systèmes de lieux et mobilités*, chap. 3 in «Recompositions urbaines en Amérique Latine : une lecture structurée à partir du cas colombien», ouvrage coll. coordonné par F. Dureau, O. Barbary, V. Goueset, O. Pissoat, à paraître en 2003, ed. Anthropos-IRD, Paris.

BISHOP C. M. (1995), *Neural Networks For Pattern Recognition*, Oxford University Press, New York, 482 p.

CADEZ I., GAFFNEY S., PADHRAIC S. (2000), *A General Probabilistic Framework for Clustering Individuals and objects*, In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, pp. 140-149.

COURGEAU D. et LELIÈVRE E. (1989), *Analyse démographique des biographies*, INED, Paris, 268 p.

COX D. R. et OAKES D. (1984), *Analysis of survival data*, Chapman y Hall, Londres, 201 p.

DAVIES R., CROUCHLEY R. (1984), *Calibrating Longitudinal Models Of Residential Mobility And Migration. An Assesment a Non-Parametric Marginal Likelihood Approach*. En Regional Science and Urban Economics 14, North-Holland, pp. 231-247.

DEMPSTER A. P., LAIRD N. M., and RUBIN D. B. (1977), *Maximum likelihood from incomplete data via the EM algorithm*. Journal of the Royal Statistical Society, Series B, n° 34, pp. 1-38.

DEVILLE J.C., SAPORTA G. (1980), *Analyse harmonique qualitative*, in Data Analysis and Informatics, E. DIDAY *et al.* éditeurs, North Holland Publishing Compagny, pp. 375-389.

DEVILLE J.C. (1982), *Analyse des données chronologiques qualitatives, comment analyser les calendriers ?* in Annales de l'INSEE, n° 45, pp. 45-104.

DRAIER T., GALLINARI P. (2001), *Characterizing Sequences of User Actions for Access Logs Analysis*, in User Modeling 2001, pp. 228-230.

DUREAU F., DUPONT V., LELIEVRE E., LEVY J.P., LULLE T. (2000), *Métropoles en mouvement. Une comparaison internationale*, Paris, Anthropos-IRD, 656 p.

FENELON J.P. (1997), *Le traitement des données longitudinales : quelques réflexions sur les méthodes*. 4ème journée d'études Céreq - Lasmas - IdL - Laboratoire d'Économie Social « les politiques de l'emploi », pp. 251-263.

GAFFNEY S., SMYTH P. (1999), *Trajectory Clustering with Mixtures of regression Models*, in Proceedings of the ACM 1999 Conference on Knowledge Disovery and Data Mining, S. Chaudhuri and D. Madigan (eds.), New York, NY : ACM, pp. 63-72.

HARTIGAN J. A. (1975), *Clustering Algorithms*. Ed. John Wiley & Sons. 346 p.

KASS R. E., RAFTERY A. E. (1995), *Bayes Factors*, J. Am. Stat. Assoc., vol. 90, n° 430, pp. 773-795.

LAVINE M., WEST M. (1992), «*A Bayesian method for classification and discrimination*». The Canadian Journal of Statistics, 20, pp. 451-461.

LEBART L., MORINEAU A., PIRON M. (2002), *Statistique exploratoire multidimensionnelle*. Paris, 2002, éd. Dunod, 437 p.

MALINVAUD E. (1981), *Méthodes statistique de l'économétrie*, Dunod, Paris, 846 p.

MCLACHLAN G.J. (1987), *On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture*. Applied Statistics 36, pp. 318-324.

MCLACHLAN G. J., KRISHNAN T. (1997), *The EM Algorithm and Extensions*. John Wiley and Sons, New York.

PINZON L. M. (1998), *Manejo del tiempo en el análisis armónico cualitativo. Aplicación al análisis tipológico de datos biográficos*. Universidad Nacional de Colombia. 128 p.

SAPORTA G. (1981), *Méthodes exploratoires d'analyse de données temporelles*, Paris, Cahiers du bureau universitaire de recherche opérationnelle n° 37-38, Université Pierre et Marie Curie, 194 p.

SMYTH P. (1997), *Clustering Sequences with Hidden Markov Models*, In Advances in Neural Information Processing 9.

SMYTH P. (1998), *Model selection for probabilistic clustering using cross-validated likelihood*, Statistics and Computing, in press.

SMYTH P. (1999), *Probabilistic Model Based Clustering of Multivariate and Sequential Data*, In Proceedings of Artificial Intelligence and Statistics, Morgan Kaufman, San Mateo CA., pp. 299-304.

VAN DER HEIJDEN P. G. M. (1987), *Correspondence analysis of longitudinal categorical data*, leidon, dswo press.

YANG Z. R., ZWOLINSKI M. (2001), *Mutual Information Theory for Adaptive Mixture Models*. In IEEE Transactions on Patt. Anal. and Mach. Int., Vol. 23 No. 4. pp. 396-403.

Annexe A. Nomenclature et codes de lieux de résidence

1. Zones de strate socio-économique basse et très basse des quartiers populaires de l'Est de Cali
2. Zones de strate socio-économique basse et très basse des quartiers populaires de l'Ouest de Cali
3. Zones de strate socio-économique moyen de Cali
4. Zones de strate socio-économique moyen - haut et haut de Cali (Résidentielle)
5. Zone rurale de Tumaco
6. Zone des quartiers populaires à Tumaco
7. Zone des quartiers résidentiels de Tumaco
8. Zone urbaine de Tumaco dont le quartier est inconnu
9. Zone rurale de Buenaventura
10. Zone des quartiers populaires à Buenaventura
11. Zone des quartiers résidentiels de Buenaventura
12. Zone urbaine de Buenaventura dont le quartier est inconnu
13. Zone rurale de Barbacoas
14. Zone urbaine de Barbacoas
15. Zone rurale des autres municipes sur la côte pacifique du département de Nariño
16. Zone urbaine des autres municipes sur la côte pacifique du département de Nariño
17. Zone rurale des autres municipes sur la côte pacifique des départements de Cauca y VPta
18. Zone urbaine des autres municipes sur la côte pacifique des départements de Cauca y VPta
19. Zone rurale du Nord du département du Cauca
20. Zone urbaine du Nord du département du Cauca
21. Zone rurale du département du Chocó
22. Zone urbaine du département du Chocó
23. Zone rurale du sud du département du valle
24. Zone urbaine du sud du département du valle
25. Zone rurale du Nord du département du Valle
26. Zone urbaine du nord du département du Valle
27. Zone rurale du reste du département du Cauca
28. Zone urbaine du reste du département du Cauca
29. Zone rurale du reste du département de Nariño
30. Zone urbaine du reste du département de Nariño

31. Zone rurale des département du Tolima, Huila, Caqueta, Putumayo
32. Zone urbaine des département du Tolima, Huila, Caqueta, Putumayo
33. Zone rurale des département Antioquia et Viejo Caldas
34. Zone urbaine des département Antioquia et Viejo Caldas
35. Reste de la Colombie
36. Etranger

Annexe B.
Matrice de confusion AHQ/CMD, distribution des individus de l'AHQ dans les nouvelles classes et graphiques de C1 et C2

Classe de l AHQ

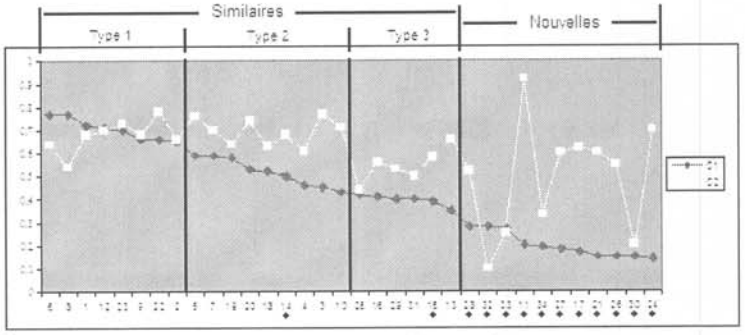
Classe de la CMD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	N_{CMD}	C_1(%)
1	53	0	7	1	0	3	0	2	1	0	0	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	74	72
2	1	64	0	0	0	0	0	0	0	0	0	0	1	0	2	0	25	1	0	0	0	0	0	1	0	1	0	0	0	0	0	2	0	0	98	65
3	0	1	51	0	1	9	0	0	0	0	0	0	0	17	3	0	2	0	0	1	24	0	0	0	1	0	2	0	1	0	0	1	0	0	114	45
4	0	0	1	28	0	0	0	1	1	0	0	0	0	1	18	1	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	7	0	0	61	46
5	1	1	0	0	34	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	18	0	0	0	2	0	0	58	59
6	0	3	0	0	0	48	0	3	1	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	2	1	0	0	0	1	0	0	62	77
7	1	0	0	0	0	0	23	0	0	1	0	0	0	0	1	1	1	0	2	0	1	0	0	1	0	2	0	3	0	0	0	2	0	0	39	59
8	0	0	0	0	0	0	0	20	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	2	0	0	26	77
9	0	2	0	1	0	0	0	1	27	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	3	0	0	41	66
10	1	0	0	0	0	1	0	2	0	24	0	1	0	0	0	2	2	2	0	0	0	0	0	19	0	1	0	0	1	0	0	0	0	0	56	43
11	0	0	0	1	0	1	0	0	0	0	0	0	0	0	2	0	0	0	1	0	1	1	0	1	0	0	0	0	0	1	0	2	0	0	10	20
12	0	0	2	0	0	0	0	0	1	0	1	35	0	1	0	0	0	0	0	0	1	0	0	2	0	1	0	1	0	0	1	2	1	0	49	71
13	1	1	1	0	0	0	0	0	0	1	11	0	23	0	0	1	0	0	0	0	0	0	0	0	0	0	16	0	10	0	0	0	0	1	66	35
14	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	4	0	0	8	50
15	7	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0	18	39
16	1	10	0	2	0	2	0	0	0	0	0	1	0	0	15	0	0	0	0	0	0	0	0	2	0	0	1	1	0	0	0	2	0	0	37	41
17	2	2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	2	0	0	0	0	0	2	0	0	1	0	0	1	0	0	12	17
18	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	12	0	0	0	0	0	1	2	0	1	0	0	0	0	4	0	0	23	52
19	0	0	0	0	0	0	0	1	1	1	0	0	2	0	0	1	0	0	14	0	0	0	0	0	1	0	1	0	0	0	1	1	0	0	24	58
20	2	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	20	1	0	0	1	0	0	0	0	0	7	1	3	0	0	38	53
21	2	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	1	0	2	0	0	0	2	0	0	0	0	1	0	0	0	0	0	0	13	15
22	2	0	1	1	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	1	21	0	0	0	0	0	0	1	0	0	1	0	0	32	66
23	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	16	0	0	0	0	0	0	0	0	1	0	0	23	70
24	0	2	0	0	1	1	0	1	2	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0	1	0	0	14	14
25	0	4	0	0	1	2	1	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	1	3	0	0	26	42
26	1	3	1	2	2	1	1	1	0	1	0	0	0	1	0	1	2	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	20	15
27	1	1	0	3	1	0	0	1	1	0	0	1	0	1	1	0	1	0	0	1	1	1	0	0	0	1	0	0	0	0	1	1	1	0	17	18
28	0	2	0	0	0	1	6	1	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	2	0	0	2	0	4	0	0	21	28
29	0	0	0	2	0	3	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	8	1	0	0	0	20	40
30	0	0	1	0	1	0	1	0	0	2	0	0	1	0	0	0	0	0	0	0	1	0	0	0	2	1	0	0	1	0	0	2	0	0	13	15
31	1	0	1	1	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	8	2	1	0	20	40
32	0	1	0	1	1	0	0	0	0	0	0	4	0	1	0	0	1	0	0	0	2	0	1	0	0	1	0	0	0	0	0	1	0	0	14	28
33	1	0	1	0	0	0	0	1	0	0	0	0	3	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	1	0	11	27
34	0	1	1	0	0	1	0	0	2	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	4	0	1	0	0	0	1	4	0	1	21	19
N_{AHQ}	78	98	69	46	45	75	33	37	40	34	12	50	35	25	31	27	40	19	22	27	40	27	22	27	25	29	30	19	15	10	16	69	4	3		
C_2(%)	68	66	77	61	76	64	70	54	68	71	92	70	66	68	58	56	62	63	64	74	60	78	73	70	44	55	60	62	53	20	50	10	25	33		

N_{CMD} = Nombre total d individus dans la classe de la CMD

N_{AHQ} = Nombre total d individus dans la classe de l AHQ

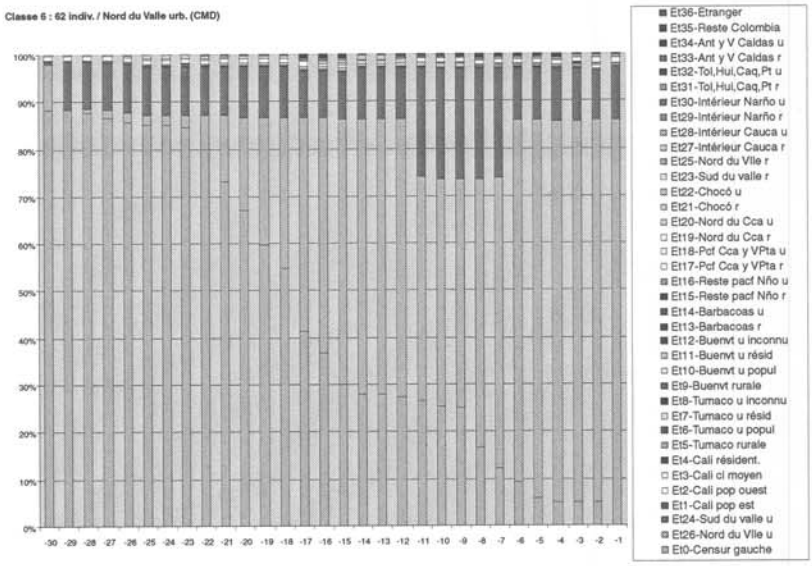
$$C_1(i) = \frac{\max\left\{card\left(C_{CMD,i} \cap C_{AHQ,k}\right) k = 1,\ldots,34\right\}}{card\left(C_{CMD,i}\right)}$$

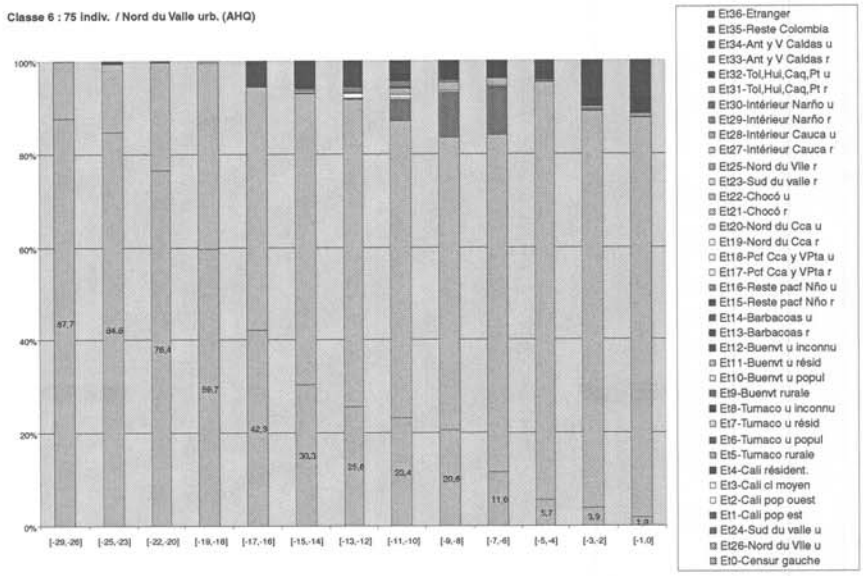
$$C_2(j) = \frac{\max\left\{card\left(C_{AHQ,j} \cap C_{CMD,k}\right) k = 1,\ldots,34\right\}}{card\left(C_{AHQ,j}\right)}$$



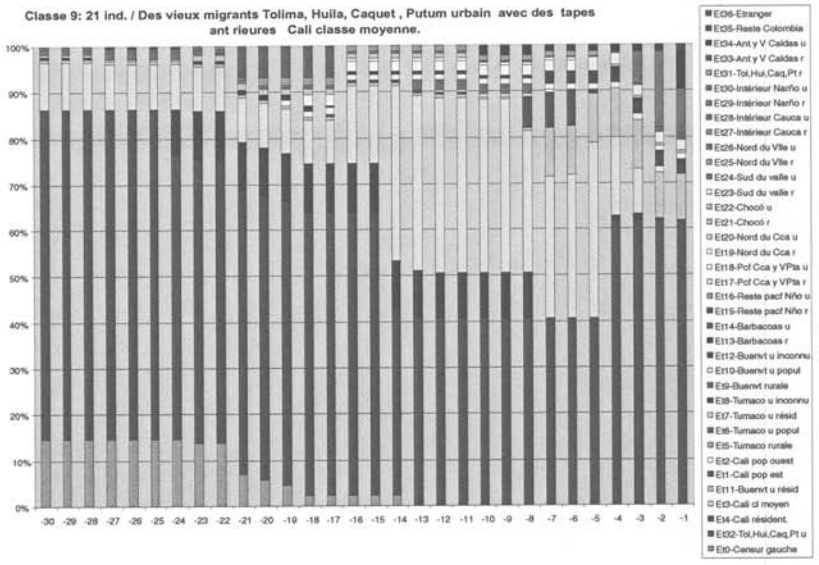
- ◆ Classes pour lesquelles C_1 est d fini avec $i \neq j$ et C_2 est d fini avec $j \neq k$.

Annexe C.
Profil de mobilité moyenne de la classe 6, pour la CMD et l'AHQ

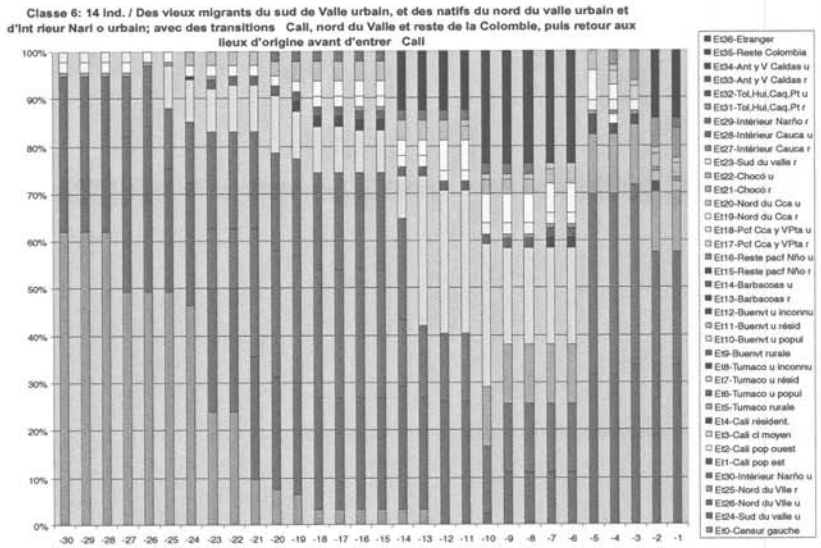




Annexe D.
Profil de mobilité moyenne de la classe 28 (la « nouvelle » classe 9)



Annexe E.
Profil de mobilité moyenne de la classe 24 (la « nouvelle » classe 6)



Annexe F.
Matrice de transition jointe de la « nouvelle » classe 6