MATHÉMATIQUES ET SCIENCES HUMAINES

OLIVIER BARBARY
LUZ MARY PINZON SARMIENTO

L'analyse harmonique qualitative et son application à la typologie des trajectoires individuelles

Mathématiques et sciences humaines, tome 144 (1998), p. 29-54
<http://www.numdam.org/item?id=MSH_1998__144__29_0>

Math. Inf. Sci. hum., (36e année, n°144, *1998*, pp. 29-54)

L'ANALYSE HARMONIQUE QUALITATIVE ET SON APPLICATION À LA TYPOLOGIE DES TRAJECTOIRES INDIVIDUELLES

Olivier BARBARY[1], Luz Mary PINZON SARMIENTO[2]

RÉSUMÉ — *Cet article présente une synthèse théorique et pratique de «l'analyse harmonique qualitative» en tant qu'outil de statistique descriptive de processus aléatoires, et son application à l'étude de la mobilité humaine. Dans la première partie, on s'intéresse, d'un point de vue mathématique, à l'analyse harmonique d'un processus qualitatif et à son approximation par l'analyse de correspondance du tableau des durées de séjour des individus dans les états possibles du processus*[3]*. La deuxième et la troisième partie sont consacrées à une application aux données d'une enquête sur la mobilité résidentielle et professionnelle, et les événements familiaux que connaissent les résidents de l'aire métropolitaine de Bogota. On y montre comment la méthode fournit une typologie des trajectoires intra-urbaines et permet de la mettre en relation avec d'autres éléments de la biographie des individus.*

SUMMARY — Nominal harmonic analysis applied to individual trajectories typology

The article aims at a theoretical and practical presentation «nonominal harmonic analysis» as a tool for descriptive statistic of categorical stochastic processes and its application to human mobility studies. The first part is a mathematical overview of the extension of scalar spectral analysis to a categorical process and how it can be approximate by a particular kind of correspondences analysis based on times that individuals accumulate in the process states. The second and third parts deal with application to a data set taken from a survey on residential mobility in Bogota metropolitan area. We found the method able to provide an interesting typology of urban residential trajectories and to show some relationships between residential mobility and other biographical variables like family events and professional changes.

INTRODUCTION

La collecte de données biographiques est de plus en plus fréquente en sciences sociales. Dans l'étude des formes de la mobilité humaine, elle tente de répondre à des questionnements nouveaux, dans des contextes thématiques, disciplinaires et méthodologiques variés. Qu'il s'agisse en effet de la mobilité spatiale, sociale ou économique, de celle des individus ou des groupes sociaux, d'approches démographiques, géographiques, économiques ou socio-anthropologiques, l'objet de la recherche est à chaque fois un phénomène continu (les vies

---

[1] Institut de Recherche pour le Développement (IRD, ex Orstom), Centro de investigación y Documentación Socio Económica (CIDSE) – Universidad del Valle, e-mail : olibarba@mafalda.univalle.edu.co.

[2] Universidad Nacional de Colombia, e-mail : lpinzon@ciencias.ciencias.unal.edu.co.

[3] Un programme SAS développé par O. Barbary, mettant en œuvre l'ensemble de la méthode, est disponible sur demande à l'auteur.

humaines sont composées d'une infinité d'instants, de lieux, d'événements...). Dès lors, même si aucun système d'observation ne peut prétendre rendre compte de cette continuité "théorique", l'appareil des concepts et des méthodes de collecte doit viser l'observation la plus exhaustive possible du temps et des espaces dans lesquels se déroule la mobilité des hommes. Récemment, ce domaine a connu de grands progrès, depuis la définition de nouveaux concepts d'observation jusqu'à la réalisation d'enquêtes sur des échantillons représentatifs. Mais si l'on sait de mieux en mieux collecter les données biographiques, de nombreuses difficultés subsistent quant à leur analyse, particulièrement dans le champ de la statistique descriptive où font encore défaut des méthodes respectant la richesse et la «continuité» des corpus. Alors que la priorité est généralement d'obtenir une typologie des trajectoires individuelles, on s'en tient souvent à des analyses transversales ou à l'examen d'indicateurs longitudinaux monovariés, nécessairement réducteurs de l'information originale.

Cependant, les acquis méthodologiques sur la statistique descriptive multivariée des processus se sont diversifiés et deviennent prometteurs (Van Der Heijden [1987] est à notre connaissance la synthèse la plus complète à ce jour). Il faut revenir, pour situer le point de départ de ces recherches, au travail pionnier de Deville et Saporta (Analyse harmonique qualitative, [1980]) et à la thèse de Saporta (Université Paris VI, [1981]) ; c'est là sans aucun doute que se fondent au plan théorique les méthodes exploratoires, et plus précisément typologiques, d'analyse des données biographiques. Comme le dit Saporta ([1981], p. 10) dans une introduction clairvoyante à une époque où n'existent que très peu de données individuelles longitudinales sur des échantillons importants : «*le problème ne sera pas tant de découvrir des périodicités comme le fait l'analyse spectrale, que de trouver les traits dominants de différenciation des évolutions des individus*».

Nous partons dans ce travail d'une technique déjà éprouvée, l'analyse harmonique qualitative (AHQ), pour la compléter et l'appliquer à l'étude de la mobilité spatiale. La justification théorique de l'AHQ est l'objet de la première partie : avec cette méthode, l'analyse factorielle des correspondances (Benzecri [1973]) trouve une nouvelle application à la typologie des données longitudinales. La seconde partie présente rapidement l'enquête sur la mobilité des populations de l'aire métropolitaine de Bogota et décrit les étapes et les problèmes de la mise en œuvre de la méthode sur ces données biographiques. Dans la troisième partie, à partir de l'exemple de certains résultats, nous montrons que cette approche est maintenant complètement opérationnelle, et tentons d'illustrer ses avantages, ses limites et ses perspectives.

1. BRÈVE THÉORIE DE L'ANALYSE HARMONIQUE QUALITATIVE

Nous allons décrire maintenant de manière synthétique les fondements théoriques de l'AHQ jusqu'à une propriété fondamentale dans la pratique : l'équivalence avec une analyse de correspondance particulière. Nous passerons ensuite en revue les techniques d'approximation numérique utilisées pour l'exécution de l'AHQ. Notre présentation s'en tiendra à l'énoncé, sans démonstration, des principales étapes de la justification de la méthode. Le lecteur intéressé par une présentation mathématique complète doit s'en remettre aux travaux Saporta [1981] et Deville [1982] ; on trouve dans Saporta [1996] une autre présentation synthétique de la méthode orientée vers l'analyse des données biographiques.

1.1. Notations, définitions et principe

L'analyse harmonique d'un processus qualitatif part de la donnée des éléments suivants :

- un intervalle de temps $T = [0, T]$ munit de la mesure de Lebesgue et β, l'ensemble des boréliens de T

- l'ensemble fini χ des modalités ou états du processus, de cardinal m, $\chi = \{1,..., m\}$
- un ensemble d'unités statistiques Ω (ensemble fini ou non d'individus), munit d'une probabilité P
- un processus X sur $\Omega \times T$ à valeur dans χ :

$$X : \Omega \times T \rightarrow \chi$$
$$(\omega, t) \rightarrow X_t(\omega) = x \ ;$$

1_t^x étant la variable indicatrice de l'événement $X_t = x$, nous noterons :

$P^x(t)$, la probabilité $P(X_t = x) = E(1_t^x)$ et

$P^{x,y}(t,s)$, la probabilité $P(X_t = x \cap X_s = y) = E(1_t^x 1_s^y)$.

Un codage scalaire (ou réel) du processus X, est une fonction f de $\chi \times T$ dans $\Re$:

$$f : \chi \times T \rightarrow \Re$$
$$(x, t) \rightarrow y = f_t(x)$$

Bien sûr, dans l'ensemble de tous les codages scalaires de X, certains sont plus naturels et pratiques que d'autres. Nous considérerons ici les codages de carré intégrable, c'est-à-dire ceux qui vérifient :

$$\int_T f_t^2(x) P^x(t) dt < \infty \quad \forall x \in \chi.$$

Un tel formalisme «probabiliste» peut surprendre s'agissant d'analyse de données où ce que l'on observe en fin de compte ne sont que des fréquences. Saporta ([1981], pp. 10-11) s'en explique : «*si nous avons choisi de parler de 'processus' et de 'probabilité' là où certains ne pourraient voir que 'courbes' et 'fréquences', c'est que ce langage nous a semblé le plus simple à utiliser et le mieux adapté au cas d'une infinité non dénombrable de variables. Il reste cependant que nous avons toujours eu à l'esprit le fait que Ω représente une population d'individus et un processus un ensemble de trajectoires individuelles*».

L'idée directrice du travail de Deville et Saporta est de faire la synthèse de deux démarches. La première consiste à se donner un codage numérique du processus qualitatif et à en faire l'analyse harmonique (d'où l'appellation donnée à la méthode). La seconde est de choisir dans T une suite finie d'instants, $0 \leq t_1 < t_2 ... < t_n \leq T$, pour effectuer l'analyse canonique des variables qualitatives X_{t1}, $X_{t2,...}X_{tn}$. En fait l'unité des deux approches apparaît si l'on traite le problème comme celui de la décomposition spectrale d'un opérateur caractérisant l'évolution temporelle du processus (opérateur analogue à celui de la covariance d'un processus scalaire).

1.2. L'analyse spectrale d'un processus qualitatif

Soit H l'espace de Hilbert des processus aléatoires réels sur $\Omega \times T$ de carré intégrable, c'est-à-dire : $H = L^2(\Omega \times T) = \{Y : \Omega \times T \rightarrow \Re \ / \int_T \int_\Omega Y^2(\omega, t) dp(\omega) dt < \infty \}$. Soit $L^2(\Omega)$ le sous-espace des processus constants dans le temps et $L^2(X)$ le sous-espace engendré par les codages scalaires de X. Pour chaque t, $L^2(X_t)$ est l'ensemble des variables β_t-mesurables, où β_t est la

tribu engendrée par X_t. $L^2(X_t)$ se compose de variables réelles, dépendantes du temps, de la forme :

$$f_t(X_t) = \xi_t = \sum_{x \in \chi} a_t^x \, I_t^x \ , \qquad (a_t^x \in \Re) \tag{1}$$

L'opérateur d'espérance conditionnelle à β_t, $E_t(\xi) = E(\xi / X_t)$, est donc la projection orthogonale de $L^2(\Omega \times T)$ sur $L^2(X_t)$. Les propriétés de E_t (E_t est hermitien, idempotent et de rang au plus égal a m, le nombre d'états) et le fait que, inversement, tout opérateur ayant ces propriétés est une espérance conditionnelle relative à une variable qualitative ayant au plus m modalités, permettent de considérer comme équivalents l'opérateur E_t et la variable X_t. Les dépendances statistiques entre deux instants t et s du processus sont donc résumées par le produit $K(t,s)$ des opérateurs E_t et E_S, fonction de H dans lui-même définie de manière analogue à la fonction de covariance :

$$K: L^2(\Omega \times T) \rightarrow L^2(\Omega \times T)$$

$$\xi \xrightarrow{K} \eta$$

avec : $\eta: (\omega, t) \rightarrow \eta_t(\omega)$

où : $\eta_t(\omega) = \int_T E_t E_s(\xi_s) ds$ pour tout t de T.

Les processus propres et leurs génératrices

Saporta [1981, pp. 101-103] et Deville [1982], pp. 61-62]) démontrent alors que K est compact et de trace finie égale à mT. K admet donc une décomposition spectrale en vecteurs propres ξ^i, appelés «processus propres» :

$$K = \sum_{i=1}^{\infty} \lambda^i \xi^i \otimes \xi^i$$

Les ξ^i sont des processus de variance totale unité, orthogonaux dans H, associés aux valeurs propres positives λ^i (avec $\sum_{i=1}^{\infty} \lambda^i = mT$), satisfaisant l'équation :

$$\lambda^i \xi_t^i = \int_T K(t,s) \xi_s^i ds \qquad \forall \, i \tag{2}$$

Le produit tensoriel $\xi^i \otimes \xi^i$ est l'opérateur de rang 1 de H qui transforme le processus ψ_t en le processus $(\int_T E(\xi_s^i \psi_s) ds).\xi_t^i$.

L'équation (2) s'écrit également :

$$\lambda^i \xi_t^i = \int_T E_t E_s(\xi_s^i) ds$$

$$\lambda^i \xi_t^i = E_t \int_T E_s(\xi_s^i) ds$$

Et puisque : $E_s(\xi_s) = \xi_s$,

on a : $\lambda^i \xi_t^i = E_t \int_T \xi_s^i ds$

On définit maintenant une variable aléatoire z^i, appelée génératrice du processus propre ξ^i par :

$$z^i = \int_T \xi_s^i ds,$$

où il est clair que z^i ne dépend pas du temps et, par conséquent, appartient au sous-espace $L^2(\Omega)$ de $L^2(\Omega \times T)$. On a donc $\lambda^i \xi_t^i = E_t(z^i)$, c'est-à-dire que pour chaque t, le processus propre ξ^i s'obtient par projection orthogonale de z^i sur $L^2(X_t)$, d'où le nom de génératrice.

En utilisant la nouvelle variable et en intégrant l'équation précédente, on obtient :

$$\lambda^i \int_T \xi_t^i dt = \int_T E_t(z^i) dt$$

$$\lambda^i z^i = \int_T E_t(z^i) dt \qquad (3)$$

Si nous appelons Q l'opérateur $\int_T E_t(.) dt$, l'équation (3) devient :

$$\lambda^i z^i = Q z^i \qquad (4)$$

L'analyse spectrale de K (processus propres ξ^i) dans $L^2(\Omega \times T)$ se déduit donc de l'analyse plus simple de Q (génératrices z^i) dans $L^2(\Omega)$.

Les codages propres

Une autre forme de l'équation aux valeurs propres s'obtient en revenant à l'expression (1) de ξ_t. En abandonnant l'indice i pour simplifier l'écriture, (2) se transforme en :

$$\lambda \xi_t = \mathrm{E}_t \left(\int_T \xi_s ds \right) = \int_T E_t \xi_s \, ds$$

$$\lambda \sum_{x \in \chi} a_t^x 1_t^x = \int_T E_t \left(\sum_{y \in \chi} a_t^y 1_t^y \right) ds$$

$$= \sum_{y \in \chi} \left(\int_T a_s^y E_t \left(1_s^y \right) ds \right).$$

Mais on a également,

$$E_t\left(1_s^y\right) = \sum_{x\in\chi}\left(1_s^y \frac{E\left(1_t^x 1_s^y\right)}{E\left(1_t^x\right)}\right) = \sum_{x\in\chi}\left(1_s^y \frac{P_{(t,s)}^{x,y}}{P_{(t)}^x}\right).$$

Par conséquent,

$$\lambda \sum_{x\in\chi} a_t^x 1_t^x = \sum_{x\in\chi} 1_t^x \sum_{y\in\chi}\left(\int_T a_s^y \left[\frac{P_{(t,s)}^{x,y}}{P_{(t)}^x}\right] ds\right),$$

et, par identification, le système d'équations :

$$\lambda a_t^x = \sum_{y\in\chi}\left(\int_T a_s^y \left[\frac{P_{(t,s)}^{x,y}}{P_{(t)}^x}\right] ds\right) \qquad x = 1,2,...,m \qquad (5)$$

En résumé, les équations (4) et (5) montrent que les z^i et les a_t^x forment deux décompositions spectrales du processus. Dans la première, la série des variables aléatoires (z^i, $i=1,...,mT$) sont indépendantes du temps et dans la seconde, les codages réels non aléatoires (a_t^x, $x=1,...,m$) dépendent du temps. Comme l'analyse canonique généralisée, l'analyse harmonique qualitative est basée sur la recherche de deux vecteurs aléatoires de corrélation maximum. Il s'agit d'identifier l'élément z^i de $L^2(\Omega)$ et le processus scalaire (codage de X_t) $\xi_t = \sum_{x\in\chi} a_t^x 1_t^x$ de $L^2(X_t)$ qui ont la corrélation maximum dans $L^2(\Omega\times T)$, puis d'itérer ce processus sous la condition d'orthonormalité ; c'est-à-dire de faire l'analyse canonique des sous-espaces $L^2(\Omega)$ et $L^2(X)$ dans $L^2(\Omega\times T)$.

1.3. Équivalence avec l'analyse des correspondances et approximation numérique de l'AHQ

Supposons maintenant qu'il existe $p + 1$ instants $0=t_0<t_1...<t_{p-1}<t_p=T$ tels que le processus soit stable sur chaque intervalle $[t_{j-1}, t_j[$, c'est-à-dire qu'aucun individu ne change d'état durant ce temps. Soit ξ_j le codage propre sur $[t_{j-1}, t_j[$, E_j l'espérance conditionnelle à ξ_j et l_j la longueur de l'intervalle j ($l_j = t_j - t_{j-1}$) ; l'équation (2) devient :

$$\lambda\xi_k = E_k \sum_{j=1}^{p} l_j E_j \xi_j ,$$

et puisque $E_j\xi_j = \xi_j$ et $z = \sum_{j=1}^{p} l_j\xi_j$,

$$\lambda\xi_k = E_k(z) \qquad (2')$$

Étant donné que les codages propres associés aux variables ξ_k sont des fonctions constantes sur chacun des intervalles de la partition de T, le système d'équations (5) s'écrit de la manière suivante :

$$\lambda a_k^x = \sum_{j=1}^{p}\sum_{y=1}^{m}\left(l_j \frac{P_{(k,j)}^{x,y}}{P_{(k)}^x} a_j^y\right) \qquad x = 1,2,...,m \qquad (5')$$

Cette nouvelle forme (5') correspond à l'équation fondamentale de l'analyse canonique généralisée. Quand Ω est fini, composé de n individus, on retrouve une analyse de correspondance particulière (voir G. Saporta [1981], pp. 110-111 et J.C. Deville [1982], pp. 68-74). Dans ce cas en effet, les équations (2') et (5') sont les équations de l'analyse des correspondances de l'ensemble des indicatrices des modalités d'états $\{X_j=y,\ j \in T,\ y \in \chi\}$ multipliées par les mesures des intervalles $[t_{j-1},\ t_j[$. Les codages réels a_k^x sont donc les vecteurs propres de l'analyse des correspondances du tableau disjonctif à n lignes (nombre d'individus) et mp colonnes (produit du nombre d'états par le nombre d'intervalles de stabilité du processus), dont la case élémentaire vaut 1 si l'individu est dans cet état durant cet intervalle de temps et 0 sinon. Mais un tel tableau peut rarement être soumis à l'analyse, à cause de sa taille et de sa structure très «clairsemée» due au fait que les changements individuels d'état ne sont pas, en général, synchronisés. Par exemple, l'observation sur 600 individus de 3 changements d'état en moyenne d'une variable à 5 modalités engendre un tableau disjonctif pondéré de 600 lignes et 9000 colonnes, avec seulement un cinquième des cases non nulles ; il est clair que son analyse ne donnera pas de résultat typologique intéressant, chaque individu formant à lui seul un type.

La solution pratique consiste à diviser l'intervalle d'observation du processus en un nombre raisonnable d'intervalles (de durées constantes ou non), sans tenir compte des changements individuels d'état. Le tableau de fréquence est ensuite construit en calculant la proportion de temps passé par les individus dans chacun des états possibles des variables au cours de chaque intervalle de recodage (voir ci-dessous la définition de la densité individuelle de séjour dans les états). Ce tableau est soumis à l'analyse des correspondances et les résultats s'interprètent avec les techniques habituelles. Il faut souligner que cette analyse, si elle conserve intégralement l'information sur les durées de présence des individus dans les états, perd en revanche, en cumulant les temps de séjour au sein des intervalles de recodage, l'ordre de passage dans les états et les retours éventuels lorsque ces changements interviennent dans un même intervalle.

Il existe diverses possibilités pour le découpage de la période d'observation et le calcul des fréquences. Lors de la division de l'intervalle $[0,T]$ en p intervalles $[0,t_1[,...,[t_k,t_{k+1}[,...,$ $[t_{p-1},T]$ de longueurs égales ou non (notons τ cette partition), le choix des t_k doit s'appuyer à la fois, comme nous le verrons par la suite, sur la connaissance préalable du processus (arguments propres à la discipline qui l'étudie), et sur l'étude de la distribution des changements d'état des individus dans le temps (arguments statistiques). Une fois choisie la partition du temps, le codage scalaire du processus recherché ($Y(i,j,k),\ i \in \Omega,\ j \in \chi,\ k \in \tau$), est une fonction à valeur réelle définie sur le produit cartésien $\Omega \times \chi \times \tau$. Notant $\tau(i,j,k)$ le temps de séjour de l'individu i dans l'état j au cours de l'intervalle $[t_{k-1},t_k[$, Deville et Saporta [1980] donnent à $Y(i,j,k)$ la valeur $\tau(i,j,k)/T$; c'est-à-dire la proportion du temps total d'observation du processus (T) que l'individu i a passé dans l'état j durant l'intervalle de temps $[t_{k-1},t_k[$, que nous appellerons aussi la densité de présence de l'individu dans l'état. Dans ce cas, quelle que soit la partition de $[0,T]$, la métrique sur le temps est uniforme et la somme de chaque ligne du tableau vaut 1 (100 % du temps d'observation de l'individu). L'interprétation du tableau est alors exactement celle d'une table de contingence qui distribue la densité de présence des individus entre les états et les intervalles de temps.

Cependant, du point de vue du calcul algébrique effectué lors de l'analyse factorielle, rien n'oblige à ce que les intervalles de codage soient de même durée ou que la métrique sur le temps soit uniforme, bien au contraire. D'une part plusieurs arguments statistiques militent en faveur

d'un découpage plus détaillé dans les périodes où les changements d'état sont nombreux et plus relâché lorsqu'ils sont rares (voir Deville [1982] et Florette [1988]). D'autre part, du point de vue de la problématique qui va nous occuper dans la suite (la stratégie résidentielle des individus), il est naturel de s'intéresser plus spécialement aux changements de résidence qui surviennent à l'âge adulte (plus particulièrement la période de plus forte mobilité, 15-35 ans par exemple), dont la décision revient généralement à l'individu, plutôt qu'aux changements de résidence durant l'enfance ou la vieillesse, plus souvent décidés par des tiers. On peut alors avoir intérêt à faire varier la métrique dans différents segments de la période d'observation.

Ainsi, on adopte souvent la solution qui consiste à calculer la densité de présence des individus dans les états par rapport à la durée de chaque intervalle de temps : $Y(i,j,k)=\tau(i,j,k)/(t_k-t_{k-1})$, proportion de la durée de l'intervalle $[t_{k-1},t_k[$ que l'individu i a passé dans l'état j. Dans ce cas la somme d'une ligne du tableau vaut p, le nombre d'intervalles définis dans $[0,T]$, et si les intervalles $[t_{k-1},t_k[$ sont de longueur variable, la métrique sur le temps n'est plus uniforme. C'est cette méthode que nous avons appliquée aux données de l'enquête de Bogota. Les algorithmes correspondant à l'ensemble des techniques présentées dans cette section ont été programmés sous SAS.

2. L'ENQUÊTE ET L'APPLICATION DE L'AHQ

2.1. La collecte des données

Contexte et problématique

L'enquête et l'analyse statistique présentées ici s'inscrivent dans une recherche entreprise depuis août 1992 par une équipe de chercheurs du C.E.D.E. (Centro de Estudios sobre el Desarollo Económico, de l'Université des Andes - Colombie) et de l'I.R.D.[4] sur la mobilité des populations de Bogota et son impact sur la dynamique de l'aire métropolitaine. Parallèlement à ce programme s'est développé depuis février 1994 un programme de recherche en coopération entre l'I.R.D. et le département de mathématiques et statistique de l'Université Nationale de Colombie sur les méthodes d'analyse statistique des données biographiques[5].

Bogota est la métropole latino-américaine qui a connu la croissance démographique la plus rapide durant les années cinquante et soixante (plus de 6 % par an). Entre 1951 et 1964 la ville a doublé sa population et en 1970 elle comptait 2,5 millions d'habitants. Depuis une vingtaine d'années, le rythme de croissance de la capitale colombienne, comme celui des autres métropoles du sous-continent, s'est ralenti : il était d'environ 2,5 % annuel lors du dernier recensement en 1985. Au moment de l'enquête C.E.D.E./O.R.S.T.O.M. (octobre 1993), Bogota compte près de 5,5 millions d'habitants et croît toujours à un rythme légèrement supérieur à 2 % par an. Ce ralentissement de la croissance s'explique par les effets conjugués de trois phénomènes : la réduction du rythme d'accroissement naturel due à la baisse rapide de la fécondité, une diminution des flux migratoires à destination de Bogota proprement dit et la transformation du schéma géographique de la croissance au profit des municipalités périphériques de l'aire métropolitaine. L'évolution démographique récente de la ville s'accompagne de nouvelles stratégies de localisation résidentielle des habitants qui produisent des changements rapides et importants dans la répartition de la population et les formes de ségrégation sociale au sein de l'agglomération. Ce sont ces recompositions, mal mesurées et peu étudiées jusqu'à présent, qui sont au centre de la thématique de l'enquête.

---

[4] Équipe dirigée par Françoise Dureau (I.R.D.) et Carmen Elisa Florez (C.E.D.E.).

[5] Équipe dirigée par Olivier Barbary (I.R.D.) et Juan Ramos (Universidad Nacional de Colombia).

Population soumise à l'enquête

L'unité d'observation pour l'enquête est le ménage, défini comme le groupe de personnes qui occupe tout ou partie d'un logement et partage les repas. La population des ménages inclut les résidents habituels (personnes vivant la majeure partie du temps dans le ménage même si elles se sont absentées au moment de l'enquête pour une période courte – depuis moins de trois mois) et les résidents non habituels (personnes qui vivent la majorité du temps hors du ménage mais ont habité au moins 30 jours — consécutifs ou non — dans le ménage enquêté au cours de l'année écoulée, que ces personnes soient ou non présentes au moment du passage de l'enquêteur). Ainsi, s'ils accumulent la durée de présence nécessaire, les individus suivants seront considérés comme faisant partie du ménage : militaires du contingent, élèves internes ou travailleurs exerçant leur activité hors du ménage ou de Bogota qui reviennent régulièrement ou périodiquement dans leurs foyers, personnes emprisonnées ou hospitalisées pour des temps courts, employés-domestiques ou autres salariés lorsqu'ils dorment dans le logement, personnes de passage (invitées) ou en pension. En revanche, les personnes qui louent une ou plusieurs pièces du logement et cuisinent séparément («inquilinos»), forment des ménages différents.

Échantillonnage

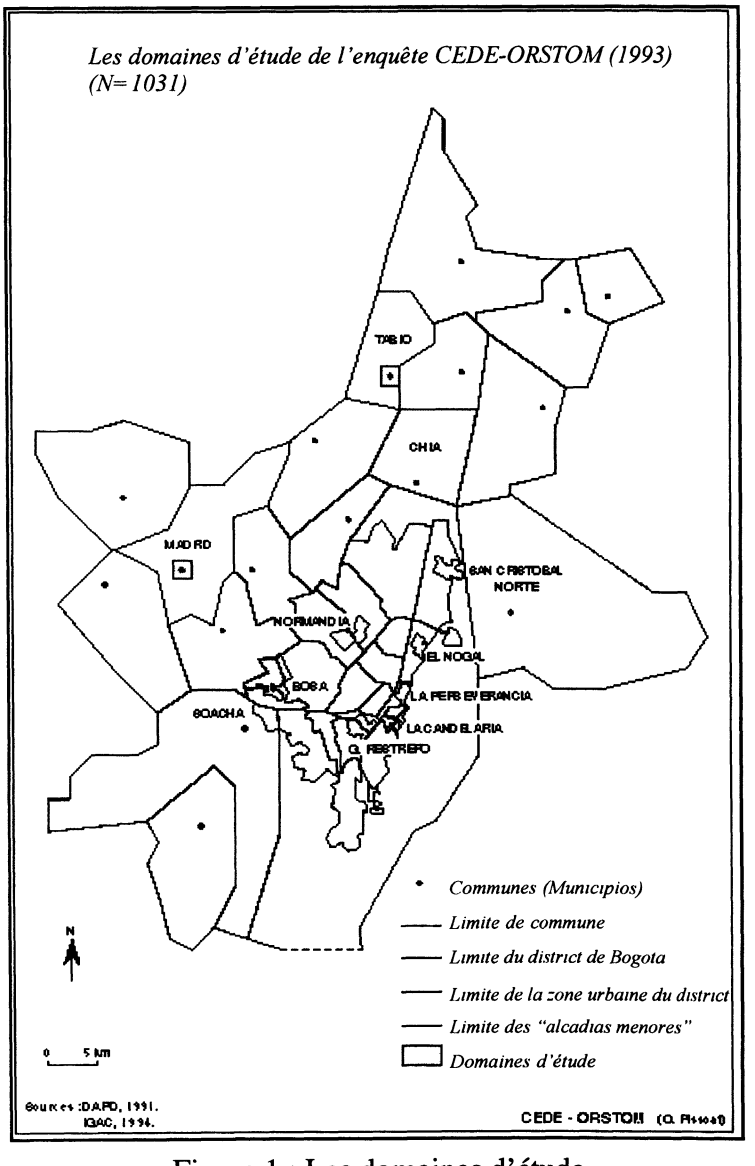


Figure 1 : Les domaines d'étude

Pour des raisons de coût, le plan de sondage adopté ne vise pas la représentativité de l'ensemble de l'aire métropolitaine de Bogota mais seulement l'observation fiable de 11 domaines d'études (4 communes ou parties de communes de l'aire métropolitaine et 7 quartiers de Bogota), ces zones ayant été choisies *a priori* comme ayant une valeur heuristique pour la problématique du programme (Figure 1). Dans chacun des 11 domaines préalablement stratifiés, le plan d'échantillonnage consiste en un sondage aréolaire à trois degrés avec probabilités inégales de sélection des unités primaires (U.P.). Au premier degré les aires sélectionnées dans chaque strate sont des «manzanas» (pâtés de maison), unités spatiales que définissent le réseau de voirie et les limites naturelles ou autres : rivières, «quebradas» (fossés), clôtures etc. On assure une bonne répartition géographique de l'échantillon en sélectionnant les U.P. au moyen d'une grille de points dont la taille de la maille est fonction du taux de sondage dans la strate (tirage spatial systé-

matique) ; la probabilité de sélection de chaque U.P. est donc proportionnelle à sa superficie. Au second degré, après avoir établi la liste complète des logements du pâté de maison, on sélectionne dans chaque U.P. cinq logements (unités secondaires) par tirage systématique équiprobable dans la liste. Enfin au troisième degré, les unités d'observation sont tous les ménages des logements sélectionnés.

Questionnaire

Le questionnaire de l'enquête comprend un formulaire d'information socio-démographique couvrant l'ensemble des individus des ménages sélectionnés (1031 ménages), ainsi qu'une série de modules visant à recueillir, sous forme de calendriers, des données rétrospectives sur la biographie résidentielle, professionnelle et familiale d'un sous échantillon de la population des ménages appelé "sous-échantillon biographique". L'échantillon auquel est soumis la partie biographique du questionnaire est composé d'un individu de plus de 18 ans par ménage enquêté (1031 individus) et, sa structure, contrôlée par des quotas de sexe, d'âge, de relation de parenté avec le chef de ménage et de statut migratoire. Les conclusions de l'analyse n'ont donc de portée qu'à l'intérieur de cet univers, pour cette population particulière, et non pour l'ensemble de la population de Bogota et son aire métropolitaine[6].

Variables soumises à l'analyse

Visant en premier lieu une typologie des trajectoires résidentielles en ville, l'analyse se base sur la variable d'état construite à partir de l'observation des localisations dans l'aire métropolitaine de l'ensemble des résidences de plus d'un an connues par les individus (variable active)[7]. Le niveau d'agrégation géographique de la variable doit fournir la précision maximum dans l'analyse de la mobilité spatiale intra-urbaine tout en conservant des effectifs suffisants dans chaque modalité. La nomenclature géographique qui convient le mieux à cet objectif est celle des "alcadias menores" de Bogota : 19 unités que nous appellerons désormais arrondissements. À ces modalités, qui décrivent les résidences à l'intérieur de la ville, s'ajoute une modalité pour les lieux de Bogota non spécifiés, quatre modalités pour les quatre zones d'étude de la périphérie de l'aire métropolitaine (communes de Chia, Tabio, Madrid et Soacha), une modalité pour les autres communes de l'aire métropolitaine (autres mun. a.m.) et enfin une modalité pour les résidences situées hors de l'aire métropolitaine. Au total, les individus peuvent donc se trouver dans 26 modalités d'état.

Pour mettre à jour les relations existant entre les types de mobilité résidentielle et les autres composantes de la biographie, nous introduisons 7 variables longitudinales illustratives qui résument des «chapitres» de la biographie que l'on peut supposer être déterminés et/ou déterminants par rapport à la trajectoire spatiale : relation de parenté avec le chef de ménage, statut matrimonial, co-résidence avec les enfants, composition du ménage, accès au logement, carrière d'éducation et mobilité professionnelle. Chaque variable ayant son calendrier propre, le nombre d'étapes vécues par les individus de l'échantillon est donc différent pour chacune d'elles, et indépendant du nombre d'étapes résidentielles.

Enfin une caractérisation socio-économique des individus pratiquant chaque type de mobilité et des ménages auxquels ils appartiennent sera obtenue grâce à un second ensemble de variables illustratives transversales, caractéristiques individuelles et collectives à la date de l'enquête. On reprend dans ce bloc de variables, les descripteurs socio-démographiques

---

[6] À propos du contrôle des quotas, voir aussi le premier paragraphe de la troisième partie ; le lecteur intéressé par des informations plus détaillées sur la méthodologie de l'enquête et la représentativité des résultats peut consulter Dureau, Florez, Barbary, Garcia, Hoyos [1994] et Dureau et Florez [1996].

[7] Une présentation de la structure informatique des données biographiques se trouve dans Morales [1996].

classiques des individus — sexe, âge, statut migratoire, niveau d'instruction, catégorie socioprofessionnelle etc., des ménages — taille du ménage, caractéristiques du logement (taille, statut d'occupation, indice de promiscuité) et du chef de ménage — sexe, âge, statut migratoire, âge moyen des enfants.

2.2. Les étapes d'exécution de l'AHQ jusqu'à la classification des trajectoires

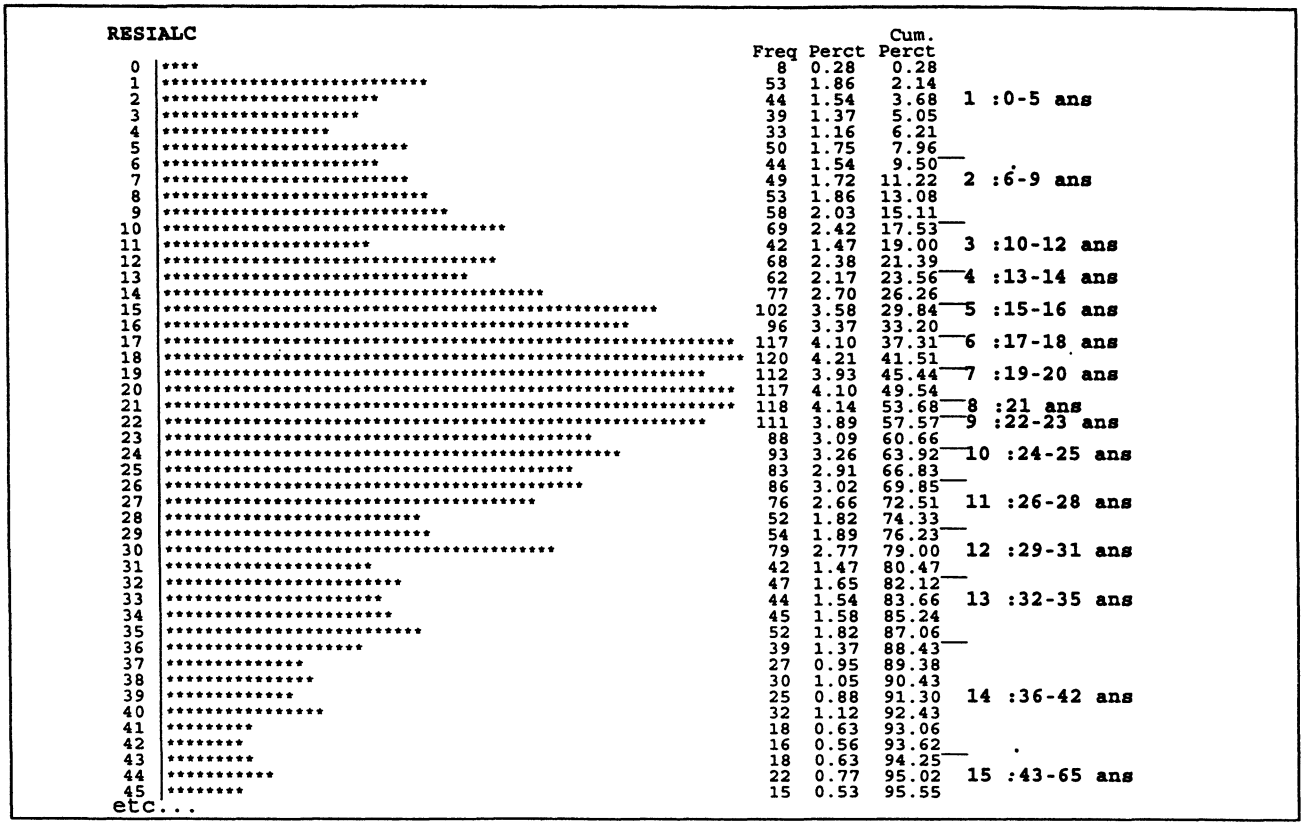
Le temps de l'analyse : période d'analyse, données censurées, discrétisation

L'analyse des données présentées dans la section précédente va être menée selon ce qu'on peut appeler le temps biographique individuel, c'est-à-dire en suivant les individus depuis leur naissance jusqu'à l'âge atteint à la date de l'enquête. D'autres options sont possibles, on pourrait par exemple mener l'analyse en temps historique (en suivant les individus entre deux dates) ou encore selon une horloge démarrant à un événement autre que la naissance, mais le choix du temps biographique correspond à une orientation problématique déterminée. Il permet en effet de mettre en relation, au niveau individuel et au niveau statistique dans l'ensemble de l'échantillon ou dans chacune des classes de la typologie, les différents itinéraires résidentiels, familiaux, éducatifs ou professionnels, pour que s'exprime la cohérence des stratégies individuelles et collectives. L'option du temps biographique étant prise, il faut, pour coder les données, choisir une période de temps commune à tous les individus quel que soit leur âge à la date de l'enquête. Si l'on souhaite conserver la totalité des étapes vécues, c'est l'âge atteint à la date de l'enquête par l'individu le plus vieux (92 ans dans notre cas) qui fixe l'étendue de la période. Afin d'éviter que la partie finale du tableau soit presque totalement vide («censure à droite», voir Figure 2), on préfère arrêter l'analyse à 65 ans pour la variable active et 70 ans pour les variables illustratives, sauf la carrière éducative arrêtée, celle-ci, dès 45 ans.

Les individus n'ayant pas atteint ces âges sortent d'observation, donc des modalités prévues pour les variables longitudinales, à partir de leur âge à la date de l'enquête (phénomène appelé "censure à droite" dans le jargon de l'analyse longitudinale). On note au passage que le type de censure que l'on doit prendre en compte dépend du type de temps choisi pour l'analyse : avec un temps historique par exemple, les données seraient censurées à gauche pour les individus n'étant pas encore nés à une date donnée. La solution retenue pour cette première analyse sur l'ensemble de l'échantillon est l'ajout d'une modalité supplémentaire à chacune des variables longitudinales dans laquelle l'individu entre dès qu'il est censuré. Ainsi l'ensemble des individus de l'échantillon est "présent" dans une modalité d'état tout au long de la période d'analyse (de 0 à 65 ans). Ce choix peut être critiqué puisque la durée de "présence" dans la modalité de censure influence le résultat typologique : dans le cas du temps biographique individuel et de la censure à droite, cet effet se ramène, comme on le verra, à un effet d'âge (la structure par âge des classes étant nécessairement homogène). On peut limiter cet effet en traitant les modalités de censure en supplémentaire dans l'AFC ou en réalisant des analyses séparées pour différentes cohortes d'individus. Nous n'avons retenu aucune de ces solutions car nous cherchons une typologie globale de la mobilité, pour l'ensemble de l'échantillon et prenant en compte l'ensemble de la trajectoire résidentielle. Dans cette perspective l'effet d'âge nous paraît au contraire devoir être conservé : un individu de 20 ans ne peut pas appartenir à la même classe qu'un individu de 50 ans (c'est-à-dire avoir la même trajectoire au même âge sur l'ensemble de la période d'analyse).

Comme on l'a dit en présentant la méthode, la mise en œuvre de l'AHQ repose sur un découpage de la période d'analyse en un nombre "raisonnable" d'intervalles de recodage. Nous avons d'abord testé un découpage uniforme de la périodes 0-65 ans en segments quinquennaux, les fréquences étant calculées en proportion du total de la période analysée. Cette analyse fournit des grosses classes d'individus jeunes dont les trajectoires sont peu homogènes et, à l'inverse, une description trop détaillée des groupes de population plus âgée. L'examen de la distribution

des changements d'état selon l'âge des individus (Figure 2) permet de définir un découpage, mieux adapté aux données, en 15 périodes d'amplitude variable correspondant à peu près aux quantiles de la distribution. La précision est bonne entre 13 et 25 ans, elle diminue avant et après. La même démarche a été appliquée aux variables longitudinales illustratives. Pour l'ensemble des variables, les fréquences sont calculées en proportion de la durée de chaque intervalle de recodage. Le tableau soumis à l'AFC comprend 1031 lignes, 398 colonnes actives (15×27 moins 7 colonnes vides) et 625 colonnes illustratives.



RESIALC	Freq	Perct	Cum. Perct	
0	8	0.28	0.28	
1	53	1.86	2.14	
2	44	1.54	3.68	1 :0-5 ans
3	39	1.37	5.05	
4	33	1.16	6.21	
5	50	1.75	7.96	
6	44	1.54	9.50	
7	49	1.72	11.22	2 :6-9 ans
8	53	1.86	13.08	
9	58	2.03	15.11	
10	69	2.42	17.53	
11	42	1.47	19.00	3 :10-12 ans
12	68	2.38	21.39	
13	62	2.17	23.56	4 :13-14 ans
14	77	2.70	26.26	
15	102	3.58	29.84	5 :15-16 ans
16	96	3.37	33.20	
17	117	4.10	37.31	6 :17-18 ans
18	120	4.21	41.51	
19	112	3.93	45.44	7 :19-20 ans
20	117	4.10	49.54	
21	118	4.14	53.68	8 :21 ans
22	111	3.89	57.57	9 :22-23 ans
23	88	3.09	60.66	
24	93	3.26	63.92	10 :24-25 ans
25	83	2.91	66.83	
26	86	3.02	69.85	
27	76	2.66	72.51	11 :26-28 ans
28	52	1.82	74.33	
29	54	1.89	76.23	
30	79	2.77	79.00	12 :29-31 ans
31	42	1.47	80.47	
32	47	1.65	82.12	
33	44	1.54	83.66	13 :32-35 ans
34	45	1.58	85.24	
35	52	1.82	87.06	
36	39	1.37	88.43	
37	27	0.95	89.38	
38	30	1.05	90.43	
39	25	0.88	91.30	14 :36-42 ans
40	32	1.12	92.43	
41	18	0.63	93.06	
42	16	0.56	93.62	
43	18	0.63	94.25	
44	22	0.77	95.02	15 :43-65 ans
45	15	0.53	95.55	

etc...

Figure 2 : Distribution des changements de résidence (RESIALC) selon l'âge

Pondération de l'échantillon

Compte tenu du fait que le plan de sondage n'est pas auto-pondér_ et que le sous-échantillon biographique n'est pas probabiliste (méthode des quotas), le problème se pose de savoir si l'analyse factorielle doit être faite sur les données pondérées par les facteurs d'extrapolation (inverse de la probabilité de sélection des ménages). Les deux AFC (pondérée ou non) donnent des résultats tout à fait semblables en ce qui concerne l'interprétation des facteurs de plus haut rang (dix premiers facteurs équivalents, à quelques inversions près). Cependant lors de la classification dans l'espace des premiers facteurs, la pondération a pour effet de rendre très "volatiles" les individus enquêtés à Chia, Madrid et Tabio (trois communes à faible population), alors que l'inertie des individus des autres domaines d'étude, beaucoup plus peuplés, est au contraire très forte. Dans cette étape de mise à plat des différents types d'itinéraires, nous préférons que tous les individus soient à égalité (AFC non pondérée). En revanche, lors de la phase de description et caractérisation des classes, il nous a semblé important de tenir compte du poids démographique (même approximatif) de chaque type de mobilité ; nous avons donc pondéré les données pour analyser la structure de chaque classe.

Analyse factorielle

L'analyse des correspondances (CORRESP de SAS ou CORBI de SPADN) fournit un histogramme de valeurs propres très plat (Figure 3), ce qui ne surprend pas étant donné la structure du tableau : grand nombre de colonnes au regard du nombre de lignes et abondance de colonnes presque vides. Rappelons ici que le recodage adopté, s'il conserve la totalité des durées de séjour des individus dans les états, perd en revanche l'ordre chronologique des étapes puisque toute permutation des colonnes est indifférente pour le résultat de l'AFC. D'autre part deux individus ayant des itinéraires strictement semblables mais simplement décalés d'une ou deux années, sont très éloignés sur les plans factoriels. Le tableau analysé est donc très "bruité" par rapport à la structure de proximité qui nous intéresse, "bruitage" que l'on retrouve dans l'allure de l'histogramme des valeurs propres.

```
SPAD.N  Sistema Portable para el Analisis de Datos
  Copyright (C) CISIA, 1987, 1991 - Version 2.52

| NO | VALOR PROPI | PCENT | PCENT ACUMU. |
|  1 |  .7099 | 3.17 |  3.17 | ********************************************************************************
|  2 |  .6291 | 2.81 |  5.98 | ************************************************************
|  3 |  .5973 | 2.67 |  8.65 | ********************************************************
|  4 |  .5800 | 2.59 | 11.24 | *******************************************************
|  5 |  .5562 | 2.48 | 13.72 | *****************************************************
|  6 |  .5291 | 2.36 | 16.09 | ***************************************************
|  7 |  .5046 | 2.25 | 18.34 | *************************************************
|  8 |  .4848 | 2.17 | 20.51 | ***********************************************
|  9 |  .4834 | 2.16 | 22.67 | ***********************************************
| 10 |  .4602 | 2.06 | 24.72 | *********************************************
etc...
```

Figure 3 : Histogramme des dix premières valeurs propres de l'AFC (tableau 1031 × 398)

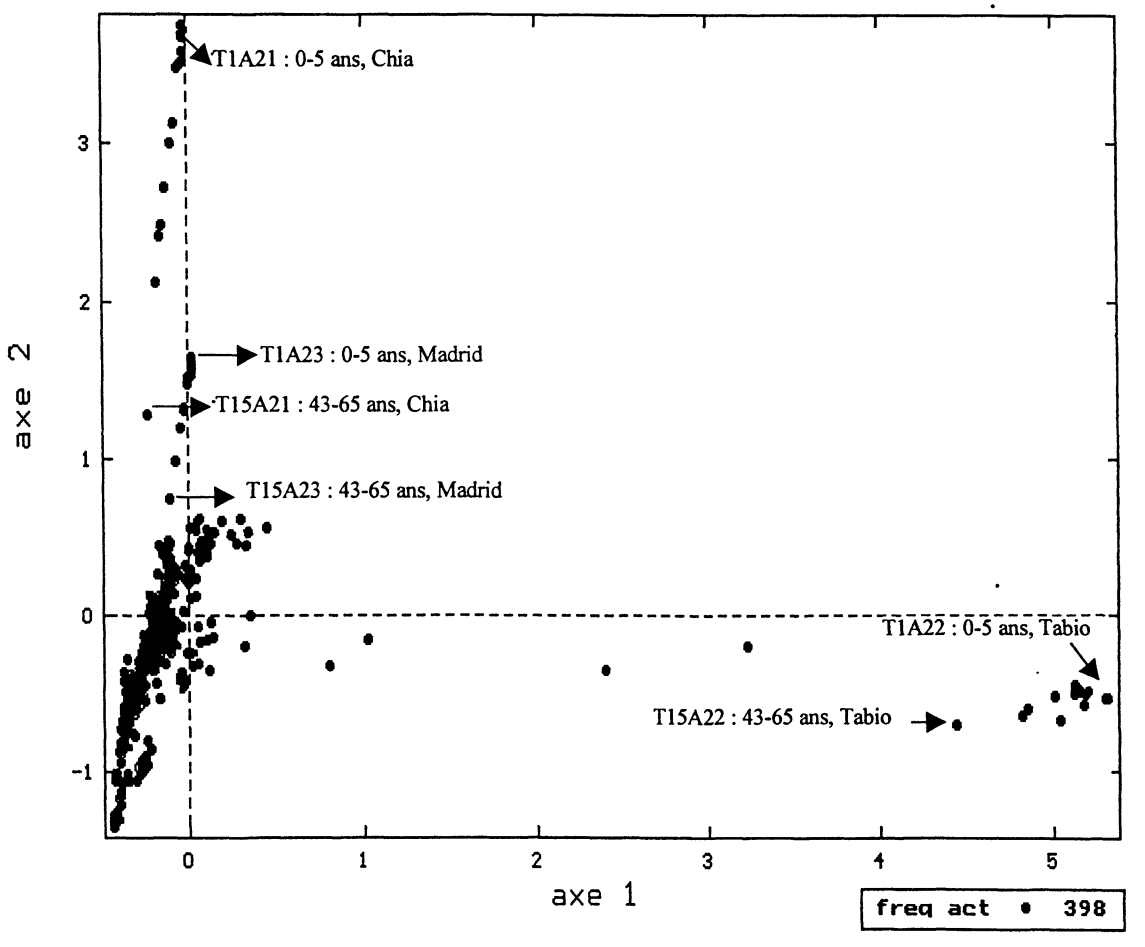


Figure 4 : Variables actives sur le premier plan factoriel (1 × 2)

Malgré cela, l'interprétation des premiers facteurs est aisée et même si la chronologie est formellement perdue lors du codage, cette méta-information structure si fortement les données qu'elle ressort sur tous les axes factoriels utiles. Le plan 1 × 2 (Figure 4) montre le type de structure mise en évidence par les premiers axes : les séquences de variables concernant un même lieu de résidence sont regroupées, ordonnées chronologiquement le long des axes, elles correspondent à des sous-ensembles d'individus minoritaires mais très stables dans ces lieux : dans le cas des deux premiers facteurs, trois groupes ayant vécu toute leur vie respectivement à Tabio (A22), Chia (A21) et Madrid (A23). Pour compléter la démarche classique d'interprétation de l'AFC, on peut projeter sur les plans les variables illustratives longitudinales et faire apparaître les points représentant les centres de gravité des caractéristiques transversales. A partir des premiers facteurs de l'analyse, on parvient ainsi sans difficulté à identifier et caractériser l'ensemble des groupes stables. Ces individus représentent à peu près 22 % de l'échantillon et environ 52 % de l'inertie expliquée par les 10 premiers facteurs ; ces facteurs, dont l'interprétation est claire, totalisent 25 % de l'inertie totale du nuage. Les schémas de mobilité qui caractérisent le reste des individus (78 % de l'échantillon) sont moins faciles à mettre en évidence à partir de la seule interprétation des facteurs. Ce qui caractérise les axes, ce sont des associations entre modalités qui témoignent de transitions fréquentes, à certains âges, entre les lieux qu'elles représentent. Nous n'avons plus, comme pour les individus stables, d'axes entièrement déterminés par un ou deux groupes aux trajectoires semblables, mais seulement la mise en évidence de groupes ayant en commun une transition donnée à un âge donné mais dont le reste de la trajectoire peut différer. Ces résultats, typiques de l'application de l'analyse harmonique qualitative aux données de calendriers (voir Deville [1982], Béret [1988] et Degenne, Lebeaux et Mounier [1995]), restent insuffisants du point de vue de notre objectif typologique.

Classification du nuage des individus

Pour parvenir à une typologie complète des trajectoires, nous avons procédé à une classification des individus dans l'espace vectoriel des premiers facteurs de l'analyse des correspondances[8]. Après plusieurs essais, en faisant varier le nombre de facteurs de 7 à 15 et en explorant les partitions jusqu'à 30 classes et plus, on constate que jusqu'à 10 facteurs la typologie gagne en précision. A partir du onzième facteur, la taille de la classe la plus importante augmente, même si l'on considère des partitions comprenant un grand nombre de classes dont beaucoup ont par conséquent des effectifs trop faibles. Sur cette base empirique, nous avons retenu l'espace vectoriel constitué par les dix premiers facteurs de l'AFC, que nous considérons comme le sous-espace conservant l'information utile (25 % de l'inertie totale du nuage). La part considérable du "bruit" dans l'information originale (75 %) s'explique par le fait que les 27 modalités d'état croisées avec les 15 intervalles de temps produisent un tableau très clairsemé dont une grande part de la variabilité, qui provient des petits décalages temporels entre trajectoires équivalentes, n'est pas interprétable. Le nuage des individus dans cet espace est ensuite soumis à des algorithmes de classification ascendante hiérarchique ou semi-hiérarchique (CLUSTER — critère de Ward sous SAS, SEMIS de SPADN) ; les deux procédures donnent des résultats très proches et nous avons conservés ceux de SPADN qui présentent l'avantage d'optimiser la partition une fois choisi le nombre de classes (Figure 5)[9].

La partition en 15 classes constitue la typologie que nous allons décrire. Elle est obtenue après coupure de l'arbre de classification et optimisation par affectation des individus à la classe dont le centre de gravité est le plus proche (PARTI de SPADN). Après cette maximisation de

---

[8] La même démarche est appliquée par Degenne, Lebeaux et Mounier [1995] aux données sur l'insertion professionnelle des jeunes en France.

[9] Rappelons qu'une partition en un nombre donné de classes, issue d'un arbre de classification ascendante hiérarchique, n'est pas optimale au sens du critère de la maximisation de l'inertie inter-classe.

l'inertie inter-classe, la partition explique 82 % de l'inertie totale du nuage dans l'espace des dix premiers facteurs, ce qui montre que la démarche prend correctement en compte, hors le "bruit", l'information significative fournie par l'enquête.

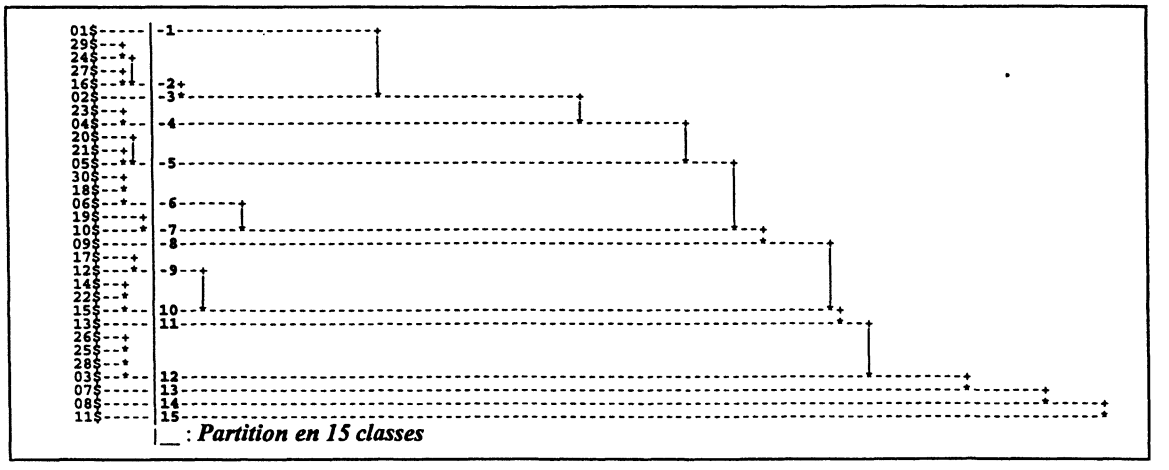


Figure 5 : Arbre de classification dans l'espace des dix premiers facteurs de l'AFC

Caractérisation et interprétation du résultat typologique

Le caractère longitudinal des données impose l'emploi de techniques particulières pour caractériser et interpréter la typologie. Classiquement en analyse typologique, les coordonnées des centres de classes sur les facteurs et les valeurs tests associées, ainsi que les individus les plus proches du centre de chaque classe, fournissent la base d'une interprétation "indirecte" des classes à partir des facteurs. Pour décrire une typologie de trajectoire, il est à la fois plus direct, plus précis et plus riche de revenir à la donnée originale codée. En affectant sa classe d'appartenance à chaque individu et en calculant les fréquences de séjour moyennes des individus de la classe dans les modalités d'état au cours de chaque intervalle de codage, on obtient le profil de mobilité de la classe que l'on peut traduire graphiquement (Figures 9 à 23 en annexe). La même technique fournit les profils des classes correspondant à chacune des variables longitudinales illustratives (Figures 6 à 8). Tous ces profils peuvent être comparés entre eux ainsi qu'au profil d'ensemble de l'échantillon[10]. Enfin, on peut éditer la série de tableaux croisés qui mettent en relation la typologie avec chacune des caractéristiques transversales des individus et des ménages retenues comme illustratives ; le pouvoir "explicatif" de chaque caractéristique transversale peut être résumé par la statistique du χ^2 associée au tableau et l'on structure le commentaire en repérant les cellules du tableau ayant les contributions les plus fortes au χ^2 global.

A partir de l'ensemble de ce matériel, on peut donc, pour chaque classe, décrire le comportement résidentiel des enquêtés et dégager leur trajectoire spatiale spécifique. Cette trajectoire est mise en relation avec la succession des événements du cycle de vie que permettent d'appréhender les variables longitudinales illustratives. On signale également les caractéristiques démographiques et socio-économiques qui complètent la "carte d'identité" de la classe. Des

[10] Pour permettre ces comparaisons, il faut éliminer l'effet des structures par âge, différentes dans chaque classe, effet directement lisible dans l'importance prise par la modalité de censure à droite au fur et à mesure que l'on progresse dans l'âge. Pour ce faire on calculera les fréquences par modalité d'état pour chaque âge sur l'ensemble des individus de la classe ayant atteint cet âge (i.e. hors individus censurés) ; le total de chaque colonne est ainsi normalisé à 100%.

hypothèses sur les dépendances et les interrelations qui structurent la biographie des individus peuvent être formulées et soumises à vérification. Des régressions logistiques permettent d'associer ensuite des tests de signification statistiques aux régularités ou associations repérées[11].

3. EXEMPLES DE RÉSULTATS

Avant d'examiner la typologie et les profils des variables longitudinales, il faut remarquer que la structure de l'échantillon biographique selon le sexe, l'âge et la relation de parenté avec le chef de ménage ne s'écarte pas beaucoup de celle de l'ensemble de la population de 18 ans et plus observée dans l'échantillon. Les seuls écarts à noter sur les variables contrôlées par les quotas concernent le statut des individus dans le ménage : une légère sur-représentation des chefs de ménage (c.m.) (42 % vs 36 %) et des conjoints (33 % vs 27 %) au détriment des enfants du c.m. (17 % vs 25 %). Cette distorsion se répercute logiquement sur le statut matrimonial (célibataires sous-représentés : 25 % vs 32 %) et sur le statut d'activité (femmes au foyer sur-représentées : 28 % vs 23 %). Ces différences d'amplitude sont raisonnables et prouvent que le contrôle des quotas a été correctement appliqué ; elles permettent d'aborder sans crainte l'analyse des résultats dans leur ensemble. Toutefois lorsque l'on considère les structures de l'échantillon au sein de certains domaines de l'étude, ces distorsions peuvent s'accentuer et il faut être vigilant dans l'interprétation des classes qui contiennent beaucoup d'individus provenant de ces zones. Ainsi par exemple, les enquêtés biographiques de Bosa et Soacha comptent trop de femmes, respectivement 65 % vs 51 % et 68 % vs 53 %, tandis que ceux de Tabio et Madrid sont au contraire trop masculins, respectivement 62 % d'hommes vs 47 % et 58 % vs 50 %. Enfin c'est bien entendu dans les classes à faibles effectifs (en général les groupes de population stable) que les distorsions les plus fortes apparaissent et compliquent l'interprétation des résultats.

3.1. Le schéma d'ensemble de la mobilité résidentielle de l'échantillon dans l'aire métropolitaine

Globalement, la typologie en quinze classes est le reflet de deux phénomènes principaux : la mobilité intra-urbaine à l'échelle des arrondissements et la migration provenant de l'extérieur de l'aire métropolitaine (a.m) — plus précisément l'âge et le lieu d'arrivée des migrants dans l'a.m. Si l'on considère les deux indicateurs synthétiques du tableau 1 (colonnes 8 et 5), on distingue clairement quatre groupes.

1. Très stables (8 % des individus, Figures 9 à 11 en annexe)

Ce groupe, formé par les classes dont la mobilité intra-urbaine a été la plus faible, est entièrement localisé hors de Bogota : stables à Chia (classe 14, 5 % de l'échantillon), stables à Madrid (classe 13, 2 %) et stables à Tabio (classe 15, 1 %). On trouve logiquement dans ces classes une très large majorité de natifs de l'a.m. mais les migrants ne sont pas totalement absents (35 % à Chia).

2. Stables (18 % des individus, Figures 12 à 15)

Un peu plus mobile que le précédent, ce groupe est constitué de noyaux de population stable dans divers quartiers anciens de Bogota : Usaquen (classe 10, 3 %), Rafael Uribe et Antonio Nariño (classe 11, 4 %), La Candelaria et Santa Fé (classe 12, 6 %), auxquels se joignent de

[11] Faute de place, nous n'en faisons pas état ici. Des exemples de résultats de ce type figurent dans Degenne, Lebeaux et Mounier [1995].

jeunes migrants stabilisés dans le centre également ancien de Madrid (classe 2, 1 %) et à Usaquen (classe 9, 4 %).

3. Mobilité moyenne (44 % des individus, Figures 17 à 21)

Le groupe le plus important de la typologie est formé par des individus qui ont connu environ deux changements d'arrondissement depuis leur entrée dans l'a.m. et que caractérisent des longs séjours dans certains quartiers, sans pour autant que les résidences à la date de l'enquête soient toujours très concentrées : classe 8 (longs séjours à l'extérieur de Bogota, dans d'autres communes de l'a.m., 3 %), classe 7 (longs séjours à Candelaria et Teusaquillo, 2 %), classe 5 (longs séjours à Rafael Uribe, 9 %), classe 4 (longs séjours dans la périphérie Ouest de Bogota, 8 %) et classe 1 (vieux migrants arrivés entre 20 et 40 ans dans tous les quartiers, 22 %).

4. Mobilité forte (30 % des individus, Figures 22 et 23)

Le groupe le plus mobile comprend la classe 6 (7 % de l'échantillon), composée de vieux migrants arrivés jeunes et de natifs qui ont tous connu une forte mobilité dans les quartiers du péri-centre nord et de la périphérie nord et ouest de la ville ; et la classe 3 (23 %) qui regroupe les jeunes migrants arrivés entre 10 et 25 ans, dont les itinéraires se jouent souvent dans la périphérie ouest et sud et les résidences à la date de l'enquête sont assez concentrées dans des quartiers de développement récent : Bosa, Soacha, Gustavo Restrepo. A titre d'exemple, nous poursuivons de manière un peu plus détaillée l'analyse de cette classe dans les paragraphes suivants.

CLASSES (1)	NINDIV (2)	POIDS (3)	PCTPOIDS (4)	PCTMIG (5)	DURMRES (6)	IMOBRES (7)	IMOBARR (8)
1	217	36156	22,32	95,9	21,3	12,7	7,86
2	25	2326	1,44	69,0	21,8	9,32	4,02
3	196	36557	22,57	83,3	15,8	18,5	14,23
4	72	13890	8,58	39,9	27,6	9,57	7,08
5	65	14706	9,08	30,1	30,0	9,33	6,67
6	89	11889	7,34	54,7	38,6	14,0	11,05
7	32	3039	1,88	12,2	47,4	8,69	6,13
8	34	4244	2,62	24,8	40,8	9,16	6,00
9	37	6515	4,02	40,3	23,6	10,2	5,56
10	25	4294	2,65	11,3	34,8	5,91	2,83
11	26	6157	3,80	25,8	35,6	4,98	3,15
12	90	9486	5,86	19,2	39,3	7,23	3,60
13	39	3529	2,18	13,6	29,7	5,20	1,56
14	48	7956	4,91	35,2	29,2	7,44	1,16
15	36	1234	0,76	6,6	43,2	5,41	2,17
Total	1031	161978	100	58,4	26,4	10,9	7,23

(1) : Rang de la classe dans la typologie
(2) : Nombre d'individus de l'échantillon biographique
(3) : Population extrapolée de la classe
(4) : Pourcentage de la population totale extrapolée
(5) : Pourcentage d'individus nés hors de l'a.m.
(6) : Durée moyenne de résidence dans l'a.m. (années)
(7) : Fréquence des changements de résidence dans l'a.m. (100 x nb. de chgt./nb. d'années de résid. dans l'a.m.)
(8) : Fréquence des changements d'arrondissement dans l'a.m. (100 x nb. de chgt./nb. d'années de résid. dans l'a.m.)

Tableau 1 : Indicateurs de mobilité moyenne pour les quinze classes de la typologie

3.2. Quelques éléments sur l'insertion des jeunes migrants à Bogota

Les modalités d'insertion résidentielle à l'arrivé en ville

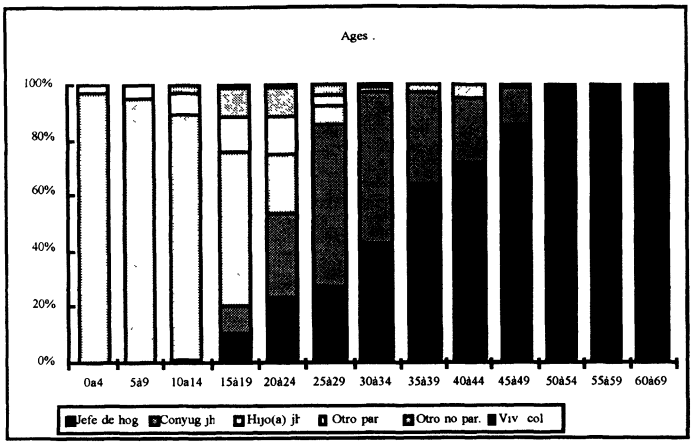


Figure 6 : relation de parenté avec le c.m. (classe 3)

Dans la classe 3 qui regroupe les jeunes migrants, le fait marquant dans le profil longitudinal de la relation de parenté au chef de ménage (Figure 6) est la sortie très rapide du ménage parental et surtout le fait que ces sorties ne correspondent pas toujours à des entrées directes dans les statuts de chefs de ménage ou de conjoints : entre 15 et 24 ans on note des fréquences importantes dans les statuts d'autres parents et d'autres non parents du c.m. (23 % entre 15 et 19 ans, 24 % entre 20 et 24 ans vs 14 % et 12 % aux mêmes âges dans l'ensemble de l'échantillon). De plus le "roulement" est important dans ces statuts de transition qui correspondent à diverses formes d'hébergement gratuit ou non : 54 % des individus de la classe ont connu un épisode de ce type à la date de l'enquête, le fait étant nettement plus fréquent chez les femmes (65 %) que chez les hommes (37 %). A ce niveau de fréquence, le phénomène apparaît donc comme une des modalités principales de l'insertion résidentielle des jeunes migrants à leur arrivée à Bogota.

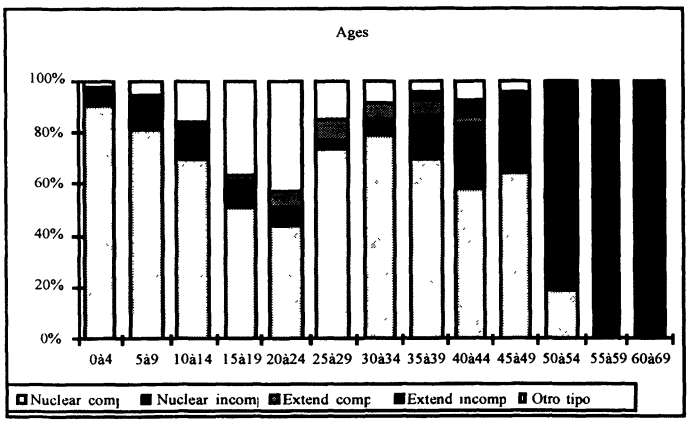


Figure 7 : composition du ménage (classe 3)

Le profil de la composition (comp.) des ménages dans lesquels ont vécu les enquêtés (Figure 7) résume bien la spécificité du cycle de vie des jeunes migrants. Aux âges correspondant à l'arrivée dans l'a.m. (entre 10 et 25 ans) c'est la fréquence des "autres types" de ménages qui frappe (ménages unipersonnels et ménages non familiaux). Là encore, à cause des rotations, la fréquence réelle de ces épisodes est supérieure à ce que fait apparaître le graphique : 72 % des individus de la classe sont concernés avec cette fois une fréquence supérieure chez les hommes (84 % contre 64 % chez les femmes). Si on retourne à la donnée individuelle on constate que pour les femmes la quasi totalité des épisodes de vie en ménage sans noyau familial correspondent à des hébergements sous des statuts dépendants (autres parents ou autres non parents du c.m.), tandis que pour les hommes une fraction importante correspond à un statut de chef de ménage et au partage du logement avec des amis ou des parents collatéraux.

Le schéma d'insertion résidentielle des jeunes migrants entre 15 et 25 ans est donc d'une part très spécifique par rapport à celui des autres individus de l'échantillon aux mêmes âges — fortement marqué par l'hébergement et la vie dans des ménages non familiaux. D'autre part il est très différencié selon le sexe — hébergement très fréquent pour les femmes et épisodes en ménage unipersonnel ou non familial presque systématique pour les hommes. Cependant il

faut noter que, aussi bien pour les hommes que pour les femmes, ces situations spécifiques sont éminemment transitoires. Comme le montrent les Figures 6 et 7, à partir de 25 ans l'accès au statut de chef de ménage ou de conjoint est rapide, ainsi que la formation de ménages familiaux (nucléaires complets ou incomplets).

Une mobilité professionnelle forte et particulière

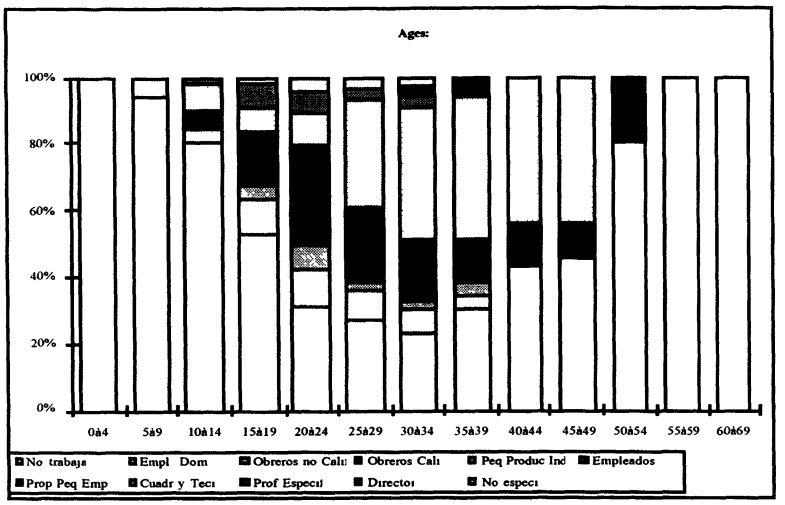


Figure 8 : catégories socioprofessionnelles (classe 3)

Les jeunes migrants ont également des modalités d'insertion professionnelle spécifiques (Figure 8), en particulier en ce qui concerne la mobilité entre emplois salariés et emplois de producteurs indépendants. La distribution des emplois occupés est assez semblable à celle de l'ensemble de l'échantillon jusqu'à 24 ans, on note seulement une proportion un peu supérieure d'employés domestiques (11 % entre 15 et 24 ans vs 8 % dans l'ensemble). Après 25 ans le profil se signale par la proportion très supérieure à la moyenne des petits producteurs indépendants (sans équivalent dans la typologie) : ils représentent 44 % des emplois occupés entre 25 et 29 ans (vs 19 % dans l'ensemble de l'échantillon) et jusqu'à 77 % entre 40 et 44 ans (vs 33 %). S'agissant d'une classe qui regroupe près de 23 % de l'échantillon, on est donc en présence d'un fait d'importance majeure. L'explication réside dans une mobilité forte et surtout particulièrement concentrée dans le temps depuis les emplois salariés vers la catégorie des petits producteurs indépendants ; on peut démonter précisément ce mécanisme.

Sur l'ensemble des individus ayant exercé un emploi à la date de l'enquête (84 % de la classe), 35 % ont occupé au moins un emploi de petit producteur, mais leurs itinéraires professionnels sont de deux types. Pour 22 % d'entre eux, l'activité de petit producteur a été la première et dans ce cas, tous l'ont débutée avant 28 ans et 90 % avant 25 ans. Les autres (13 %) y accèdent après d'autres emplois et cette transition s'effectue pour 21 % d'entre eux avant 22 ans mais pour 74 % entre 22 et 25 ans (5 % seulement après 25 ans), ce qui cause la modification brutale du profil d'activité de la classe à cet âge. Il faut aussi noter que la transition inverse (petits producteurs vers autres emplois) n'est pas moins fréquente, au contraire : 15 % des individus ayant eu un emploi à la date de l'enquête l'ont connue, mais tous avant 22 ans ; ce qui contribue donc également à limiter les densités de présence dans l'état de petit producteur avant 24 ans. Ainsi, ce n'est pas le solde de la mobilité entre les emplois de petits producteurs et les autres qui explique la recomposition observée (au sein de la classe, il est plutôt favorable aux autres emplois), mais le décalage dans le temps des deux types de mobilité : on sort des emplois de petits producteurs vers d'autres emplois entre 15 et 22 ans alors que, dans la majorité des cas, on y entre en venant d'autres emplois entre 22 et 25 ans. Bien sur, étant donnée l'hétérogénéité de cette catégorie, on peut penser que le type d'activité exercée et le revenu ne sont pas les mêmes dans les deux cas...[12]

[12] Le lecteur intéressé par les résultats complets de l'analyse peut consulter : Barbary O., *Análisis tipológico de datos biográficos en Bogotá*, Bogotá, Universidad Nacional de Colombia, Col. Textos n° 24, (1996), 254 p.

CONCLUSIONS

Aux plans théorique et pratique, lorsqu'on envisage son application à des données biographiques complexes, l'analyse harmonique qualitative (AHQ) présente d'importants avantages.

Tout d'abord la discrétisation et la synchronisation des temps biographiques individuels est modulable et permet diverses stratégies de codification des données pour s'adapter à différentes problématiques d'analyse. On peut adopter d'autres logiques de découpage de la période d'analyse que celle présentée ici. Par exemple, celle où les bornes des intervalles de codification sont choisies en fonction de dates qu'on estime déterminantes dans le contexte conjoncturel, social ou politique de la période ; l'analyse prend alors un tour historique et fournit un autre point de vue sur les données (voir par exemple Barbary [1993], pp. 30-64). Une troisième possibilité consiste à adopter un temps biographique plus collectif découpé en fonction d'événements-clé du cycle de vie : les itinéraires individuels ne sont plus synchronisés selon l'âge des individus mais selon le temps passé depuis (ou avant) un événement donné de la biographie : première arrivée des migrants dans la zone d'enquête, première sortie du domicile des parents, premier accès à la propriété du logement etc. Des analyses complémentaires de ce type seraient très certainement instructives dans le cadre problématique qui nous intéresse ici.

D'autre part, pour chacune de ces logiques, la question du découpage optimal du temps afin de conserver aux mieux l'information asynchrone collectée dans l'enquête, peut être posée en termes théoriques et ouvrir le champ à des recherches futures : optimisation, au sens d'un critère à définir, du nombre et des limites des intervalles de codification du processus. En s'en tenant au point de vue empirique que nous avons adopté dans ce travail, on a vu que rien n'oblige à choisir des périodes de codage de durée constante ou une métrique uniforme sur le temps. Ce sont là deux paramètres importants pour adapter la méthode à la structure particulière des données et à la problématique de l'analyse. Mais il reste à mener une étude rigoureuse des conséquences qu'ont ces choix sur la stabilité des résultats typologiques.

En troisième lieu, si l'on collecte de plus en plus dans une même enquête plusieurs types de biographies en parallèle (résidentielles, professionnelles, familiales etc.), c'est évidemment que l'on présume des relations étroites entre elles ; on attend alors de la méthode d'analyse qu'elle puisse les mettre en évidence et les décrire précisément. Pour ce faire, l'AHQ offre deux possibilités. Une première voie, que nous n'avons pas adoptée ici, consiste à analyser la variable d'état complexe qui résulte du croisement des différentes situations résidentielles, professionnelles, familiales etc. (analyse longitudinale multivariée). Elle parait séduisante puisqu'elle donne à toutes les variables de la biographie un poids équivalent dans le résultat typologique. Il ne faut pourtant pas croire qu'on fera un bon usage de la méthode en croisant à tout va les informations biographiques de l'enquête, laissant au calcul le soin d'exhiber une typologie "complète", rendant compte de toutes les relations qui les structurent. Il y a des limites de plusieurs ordres à ce genre d'approche. Un préalable "épistémologique" d'abord : même si les méthodes d'analyse de données ont une *vocation exploratoire*, leur emploi ne dispense jamais d'avoir une *problématique définie* lorsqu'on analyse une enquête : il importe de savoir à l'avance ce que l'on cherche (au moins un peu...) et sur quelle(s) variable(s) se basent les principales hypothèses. Par ailleurs, du point de vue de la fiabilité statistique, la taille de l'échantillon ne permet en général pas de complexifier, au delà d'une limite vite atteinte, l'espace des états biographiques. Cette remarque s'applique d'ailleurs aussi au découpage de la période de temps de l'analyse et donc au degré de précision temporelle auquel on peut prétendre. Enfin, s'agissant d'une méthodologie récente, dans un domaine où l'accumulation d'expérience est fondamentale, on manque encore beaucoup de pratique dans l'interprétation des résultats.

La technique des variables illustratives nous semble mieux convenir à l'analyse descriptive conjointe de différents itinéraires. Tout d'abord elle affirme un *a priori* problématique — la

priorité donnée à la variable active — qui est l'expression d'une démarche expérimentale structurée autour de questions précises. C'est dans ce cadre que l'interprétation trouvera son fil directeur et que le statisticien pourra, en se donnant un peu de mal, proposer des hypothèses et des conclusions intéressantes à ses interlocuteurs de sciences sociales. D'un autre côté cette technique permet de maintenir des normes de représentativité statistique acceptables pour des échantillons qui ne peuvent pas, sauf exception, être très grands (la collecte biographique est complexe et coûte chère !).

Enfin, outre les éléments descriptifs qu'elle permet de dégager, la démarche que nous venons de présenter nous semble également intéressante dans une perspective ultérieure de modélisation. En effet les résultats typologiques permettent, lorsqu'on cherche à mettre en évidence l'effet de variables indépendantes sur les durées de séjour, d'ajuster les modèles sur des sous-ensembles de populations plus homogènes que l'échantillon entier. Dès que les classes atteignent un effectif suffisant (en général 30 à 50 individus), les régularités quasi fonctionnelles observées sur les graphes suggèrent très fortement des hypothèses quant aux formes et aux paramètres des fonctions de séjour dans les états biographiques. Ces hypothèses, dont la justification fait souvent défaut dans les démarches de statistique inférentielle, peuvent d'abord être confirmées, en éliminant l'effet des données censurées, à l'aide d'estimations non-paramétriques (Kaplan et Meier [1958]), pour être ensuite introduites dans la construction de modèles paramétriques ou semi-paramétriques qui estiment directement l'effet de certaines variables exogènes (Cox et Oakes [1984], Courgeau et Lelièvre [1989]). Ainsi il nous semble particulièrement vrai, dans le contexte de l'analyse des données longitudinales, que les approches de la statistique descriptive et inférentielle ne sont pas concurrentes mais au contraire complémentaires ; il serait donc logique, comme le souhaitait B. Riandey lors de la conclusion des journées d'études C.E.R.E.Q./L.A.S.M.A.S. de 1995 «*que ces deux modes d'analyse ne renvoient pas tant aux traditions internes d'une équipe qu'à des phases successives de la recherche*».

Remerciements : Nous tenons à remercier Jeanne Fine (Professeur de l'Université Paul Sabatier) et Leonardo Bautista (Professeur de l'Université Nationale de Colombie) pour leur lecture très attentive de la première version de cet article et leurs suggestions.

BIBLIOGRAPHIE

BARBARY O., *Análisis tipológico de datos biográficos en Bogotá*, Bogotá, Universidad Nacional de Colombia, Col. Textos n° 24, (1996), 254 p.

BARBARY O. Ed. Cient., *Recolección y Análisis de Datos Longitudinales*, Memorias del seminario de capacitación e investigación "Recolección y análisis de datos longitudinales", Bogotá, 9-13 de diciembre de 1996, Universidad Nacional de Colombia, P.R.E.S.T.A. et O.R.S.T.O.M. eds, 1996, 425 p.

BARBARY O., *L'insertion urbaine : le cas de Dakar*, Dakar, O.R.S.T.O.M. et I.F.A.N. eds, (1993), 213 p.

BENZÉCRI J.P. et al., *L'analyse des données. Tome 2 : L'analyse des correspondances*, Paris, DUNOD, (1973), 632 p.

BERET P., "Analyse de données chronologiques relatives à l'insertion professionnelle", in *Les Cahiers de l'Analyse des Données* vol. XIII, n° 2, (1988), pp. 159-174.

COURGEAU D., LELIÈVRE E., *Analyse démographique des biographies*, Paris, éditions de l'INED, (1989), 268 p.

COX D.R., OAKES D., *Analysis of Survival Data*, Londres, Chapman and Hall, (1984), 201 p.

DEGENNE A., LEBEAUX M.O., MOUNIER L., "Construction d'une typologie de trajectoires à partir de l'enquête de suivi des jeunes des niveaux V, Vbis et VI.", Communication aux Journées C.E.R.E.Q.-L.A.S.M.A.S.-I.D.L. sur l'analyse longitudinale du marché du travail, Caen, 28-29 juin 1995, C.E.R.E.Q. et C.N.R.S. eds, (1995), pp. 27-42.

DEGENNE A., MANSUY M., WERQUIN P., Ed. Cient., *Trajectoires et insertions professionnelles*, Deuxièmes journées C.E.R.E.Q.-L.A.S.M.A.S.-I.D.L. sur l'analyse longitudinale du marché du travail, Caen, 28-29 juin 1995, C.E.R.E.Q. et C.N.R.S. eds, (1995), 364 p.

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

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

DUREAU F., FLOREZ C.E., BARBARY O., GARCIA L., HOYOS M.C., *La movilidad de las poblaciones y su impacto sobre la dinámica del área metropolitana de Bogotá*, metodología de la encuesta cuantitativa. Documento de trabajo n°2 : C.E.D.E./O.R.S.T.O.M., Bogotá, C.E.D.E. et O.R.S.T.O.M. eds, (1994), 393 p.

DUREAU F., FLOREZ C.E., *Dos ejemplos de cuestionarios y de operatividad de encuestas longitudinales*,, in Memorias del seminario de capacitación e investigación "Recolección y análisis de datos longitudinales", Bogotá 9-13 dec. 1996, Universidad Nacional de Colombia, O.R.S.T.O.M., P.R.E.S.T.A. eds, (1996), pp. 35-58.

FLORETTE A., *Approximation et choix du découpage dans le cadre de l'analyse harmonique qualitative*, Mémoire de DEA, ENSAE, Paris, 1988.

KAPLAN E.L. MEIER P., "Nonparametric estimation from incomplete observations", in *J. Am. Statist. Assoc.*, n° 53, (1958), pp. 457-481.

MORALES A., "Estructuración, captura y control de bases de datos longitudinales", in Memorias del seminario de capacitación e investigación "Recolección y análisis de datos longitudinales", Bogotá 9-13 dec. 1996, Universidad Nacional de Colombia, O.R.S.T.O.M., P.R.E.S.T.A. eds, (1996), pp. 59-70.

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

SAPORTA G., "L'analyse harmonique qualitative, une synthèse de la théorie", in Memorias del seminario de capacitación e investigación "Recolección y análisis de datos longitudinales", Bogotá 9-13 dec. 1996, Universidad Nacional de Colombia, O.R.S.T.O.M., P.R.E.S.T.A. eds, (1996), pp. 111-120.

VAN DER HEIJDEN P. G. M., *Correspondence analysis of longitudinal categorical data*, Leiden, DSWO PRESS, 1987.

ANNEXES

Les groupes stables et stabilisés dans les communes périphériques (Tabio, Chia, Madrid)

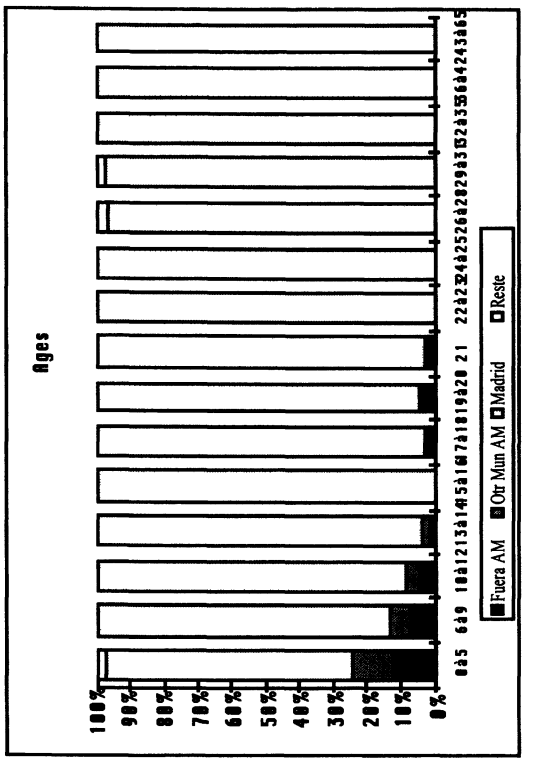


Figure 11 : Classe 13, stable à Madrid

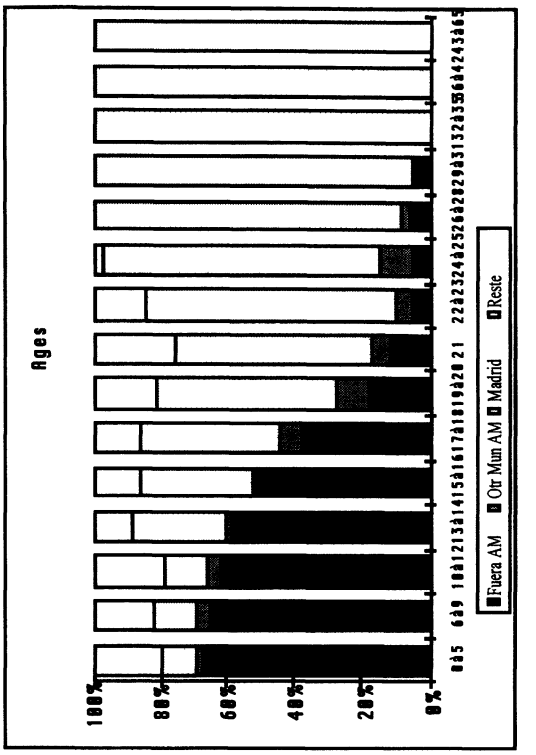


Figure 12 : Classe 2, stabilisée à Madrid

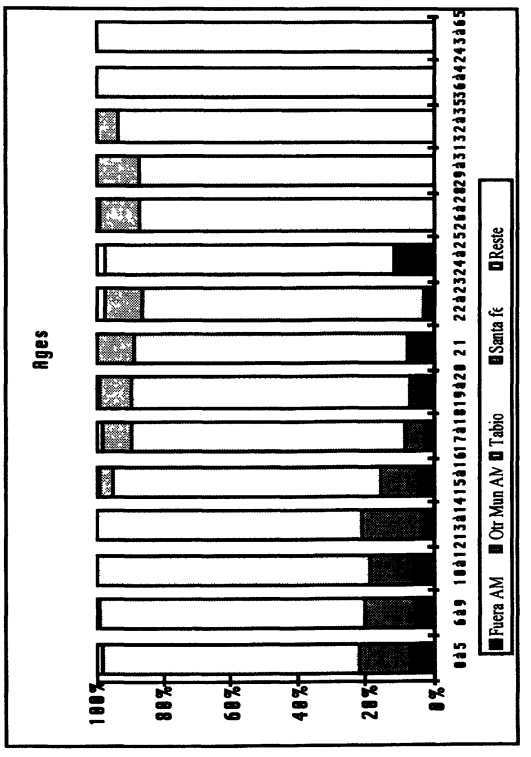


Figure 9 : Classe 15, stable à Tabio

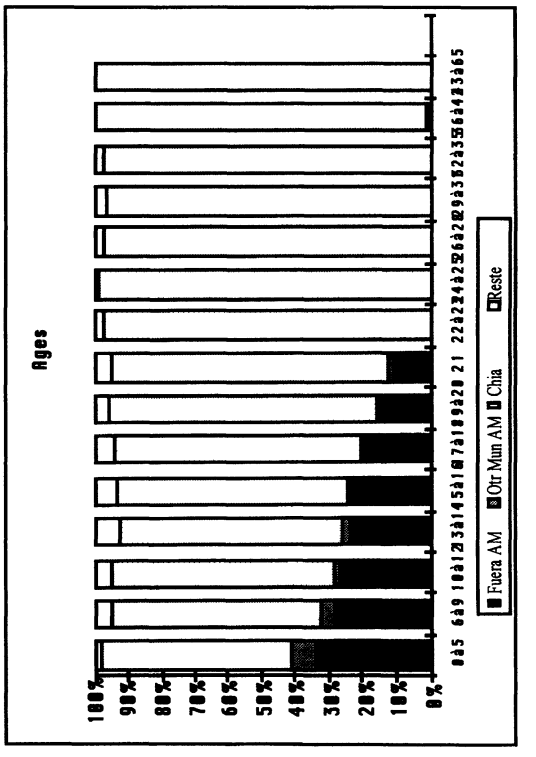


Figure 10 : Classe 14, stable à Chia

Les groupes stables et stabilisés à Bogota

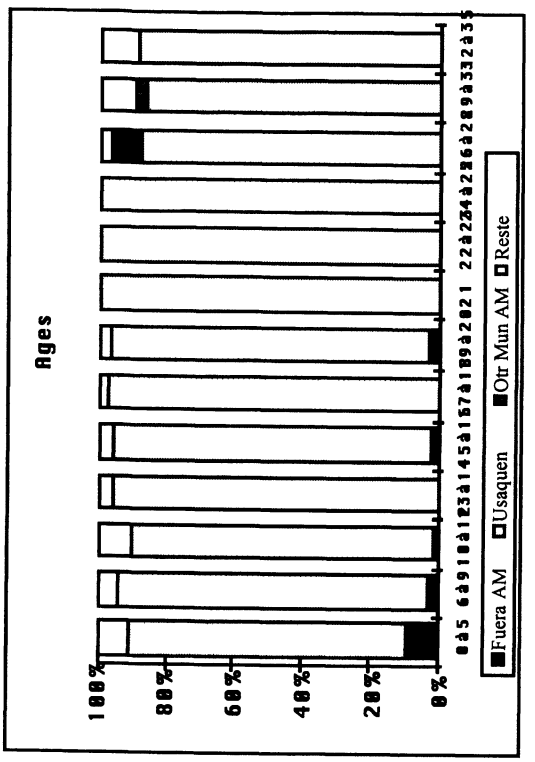


Figure 15 : Classe 10, stable à Usaquén

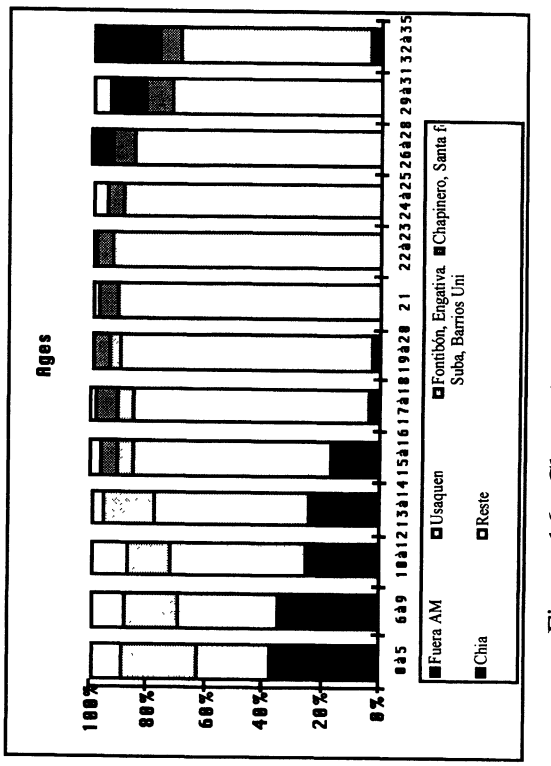


Figure 16 : Classe 9, stabilisée à Usaquén

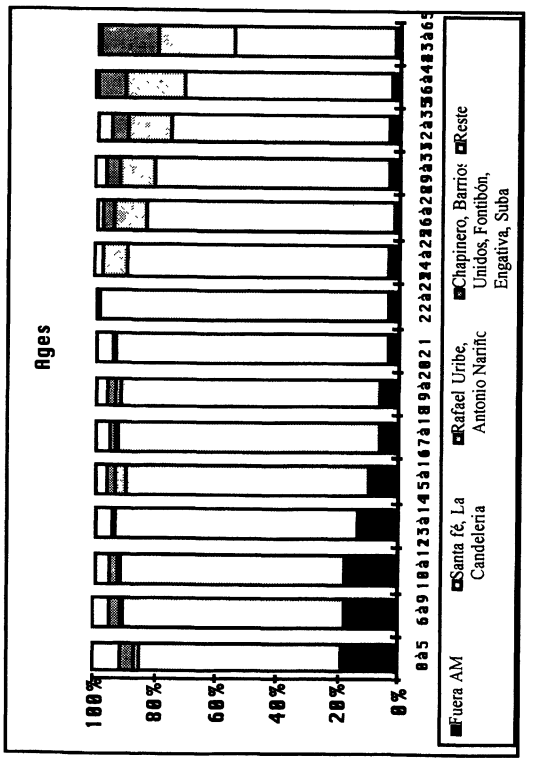


Figure 13 : Classe 12, stable à Santa Fé et La Candelaria

Figure 14 : Classe 11, stable à Antonio Nariño et Rafael Uribe

Les groupes à mobilité intra-urbaine moyenne

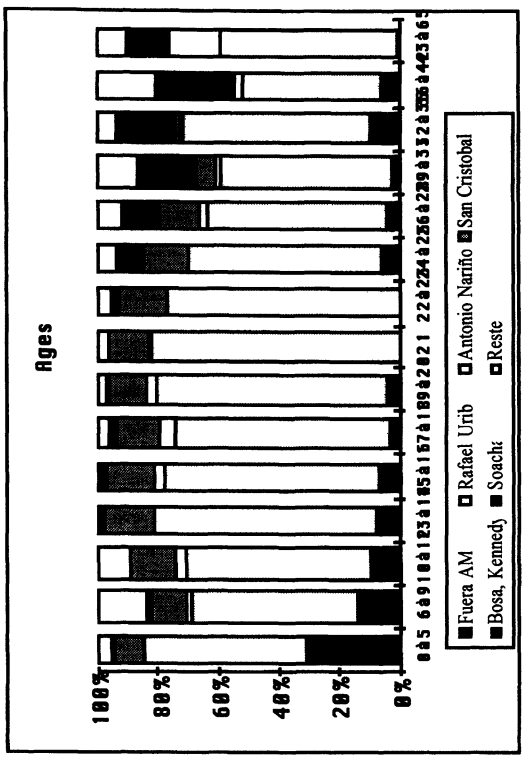


Figure 19 : Classe 5, longs séjours à Rafael Uribe et San Cristobal

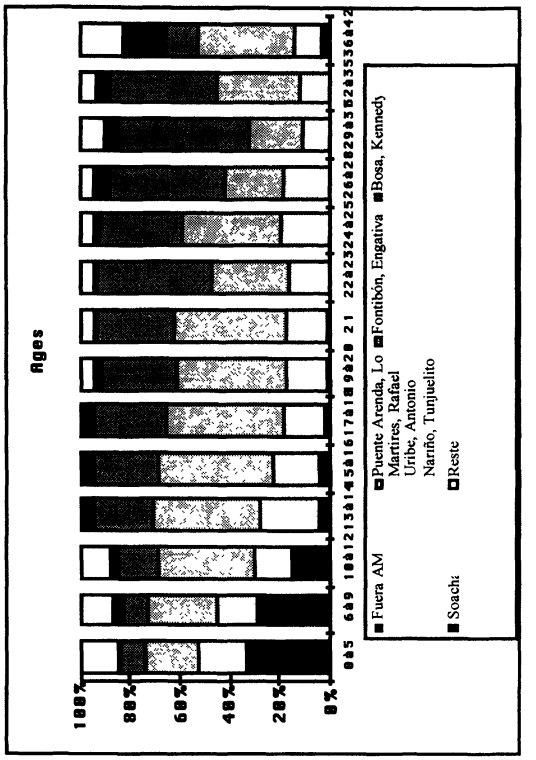


Figure 20 : Classe 4, long séjours dans la périphérie Ouest

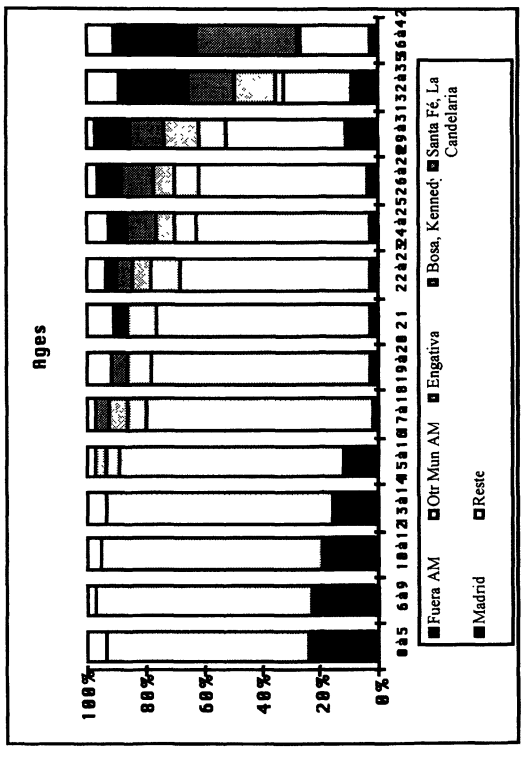


Figure 17 : Classe 8, longs séjours d'autres communes de l'a.m.

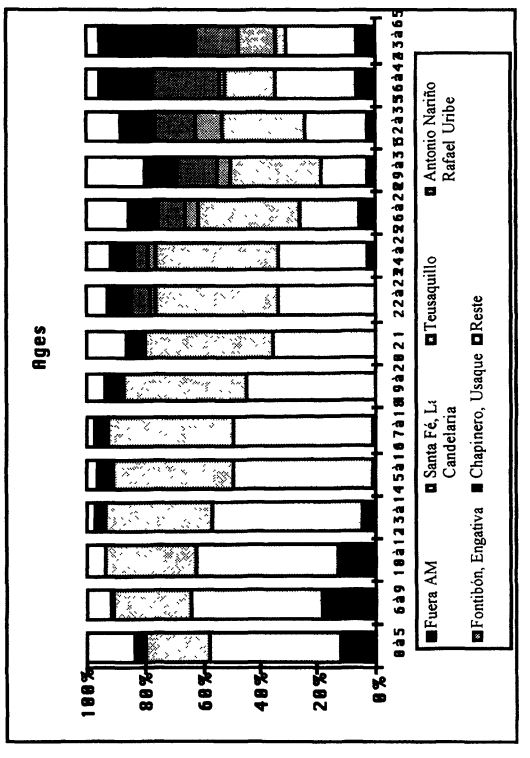


Figure 18 : Classe 7, longs séjours à La Candelaria et Teusaquillo

Les migrants et les groupes à mobilité intra-urbaine forte

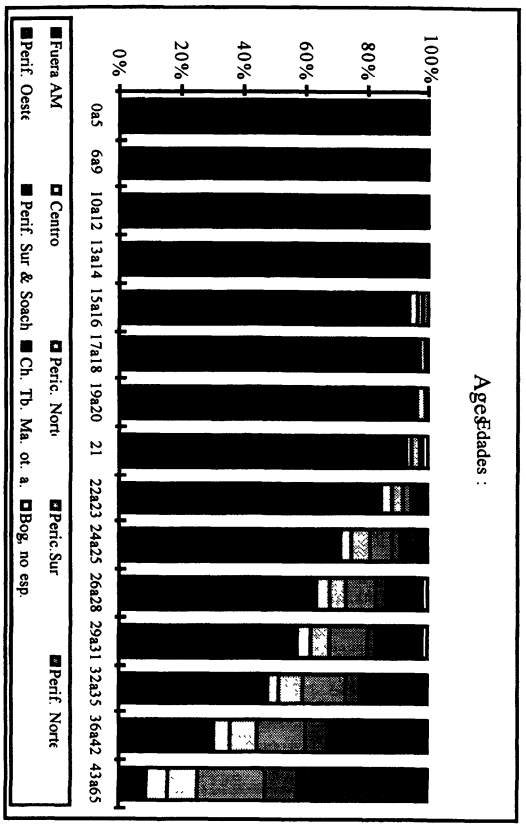


Figure 21 : Classe 1, vieux migrants arrivés entre 20 et 40 ans

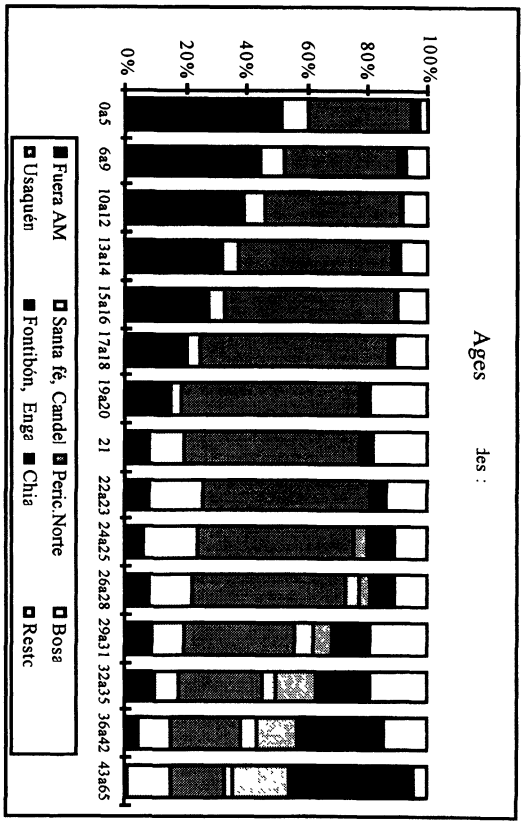


Figure 22 : Classe 6, vieux migrants arrivés avant 20 ans

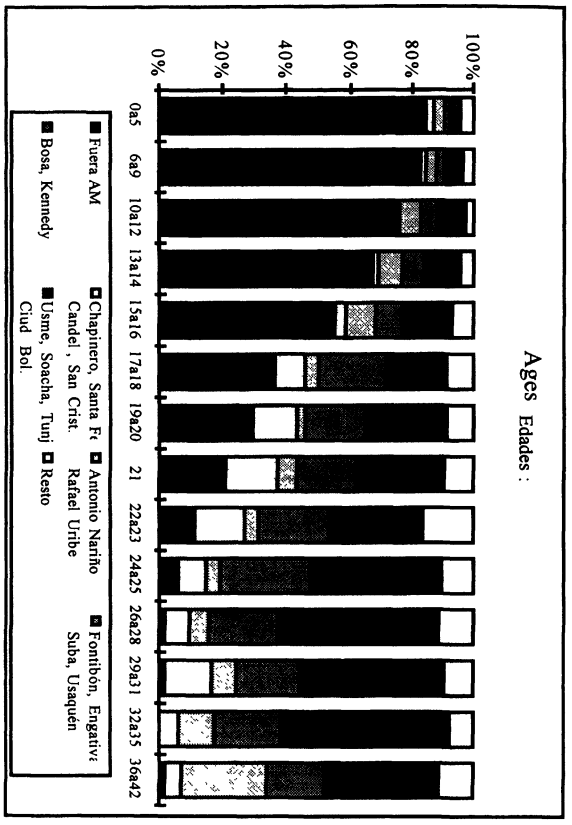


Figure 23 : Classe 3, jeunes migrants arrivés entre 10 et 25 ans