ANALYSE STATISTIQUE DE DONNEES BIOGRAPHIQUES SUR L'INSERTION DES MIGRANTS EN VILLE. Directions de travail pour une recherche méthodologique et comparatiste

Olivier BARBARY
Statisticien, ORSTOM

Ayant débuté depuis mon entrée à l'ORSTOM un programme de recherche en mathématiques appliquées aux sciences sociales, je me propose de donner ici, a l'aide d'une rapide revue bibliographique, les premières pistes méthodologiques d'un travail qui sera centré sur le traitement statistique des données biographiques. Mais je voudrais d'abord dire comment est né ce programme de recherche et ce qui justifie, dans la dynamique actuelle de certaines recherches urbaines à l'ORSTOM, ce travail méthodologique et comparatiste.

1. CONTEXTE, JUSTIFICATION ET OBJECTIFS D'UNE RECHERCHE

De 1985 à 1988[1], nous avions travaillé avec une équipe de recherche pluridisciplinaire[2] à la mise au point d'une nouvelle technique de sondage pour la collecte de données socio-démographique en milieu urbain, qui grâce à l'utilisation d'une image satellite comme base de sondage et de stratification, s'avère rapide, fiable et adaptée aux moyens économiques et humains des villes des pays en développement. Cette recherche s'est conclue par une enquête réalisée à Quito (Equateur) qui poursuivait un double objectif : au plan méthodologique , il s'agissait de tester la technique de collecte en vraie grandeur, au plan thématique, le questionnaire comprenait un important volet biographique par lequel nous cherchions à rendre compte de l'itinéraire résidentiel et professionnel de 3157 chefs de ménages Quiténiens.

A mon entrée à l'ORSTOM en septembre 1990, l'unité de recherche 5E, à laquelle j'était intégré, tenait une journée de réunion sur le thème : "méthodes de

---

1 : Durant cette période, j'ai bénéficié d'une allocation de recherche du ministère de la recherche afin de réaliser une thèse à l'EHESS, thése soutenue fin 1988.

2 : Cette équipe, animée par F. Dureau (démographe) comprenait O.Barbary (statisticien), B. Lortic (télédétection) et A. Michel (télédétection). Pour un exposé plus détaillé des résultats et des suites de ce programme, on peut se reporter à la communication de F. Dureau.

recueil et de traitement des biographies". Partant du constat que des approches qualitatives et quantitatives des récits de vie sont pratiquées dans nombre des projets de recherche, R. CABANES[1] proposait au chercheurs de l'UR "d'examiner à travers les différences de méthodes, les différences d'objet, de pointer les modifications que la technique fait subir à la définition de l'objet et explorer les possibilités d'un mixage des techniques en fonction de l'objet que l'on se propose d'étudier". C'est dans ce contexte de réflexions méthodologiques que l'intégration à l'UR d'un statisticien développant des recherches sur les méthodes statistiques appliquées aux sciences sociales est apparue utile.

Plus précisément, **l'étude des processus d'insertion des migrants en ville** à l'aide du recueil et de l'exploitation de données biographiques est un thème commun à plusieurs recherches menées dans l'UR, et en particulier une approche partant de l'observation des résidents urbains, qu'on peut qualifier rapidement de "quantitative", c'est à dire menée à partir du recueil d'un questionnaire biographique rétrospectif "fermé" sur un échantillon représentatif des migrants résidant en ville, est développée par deux programmes : "Analyse des pratiques résidentielles et économiques des quiténiens" (responsable : F. DUREAU), programme basé sur l'enquête réalisée à Quito que je viens d'évoquer et "L'insertion urbaine des migrants à DAKAR" (responsable : Ph. ANTOINE). Dans ces deux programmes la phase de collecte est achevée et l'exploitation des enquêtes a commencé, générant un réel besoin de traitements statistiques adaptés à la nature des données et aux objectifs thématiques définis. Ce contexte scientifique justifie donc qu'après avoir travaillé sur les méthode de collecte de l'information dont ont besoin la recherche et la planification urbaine, je m'oriente maintenant vers les techniques d'analyse statistique de cette information

Afin de situer les grandes lignes de la demande méthodologique liée à ces deux recherches et de voir en quoi une approche comparatiste pourrait être enrichissante, je voudrais souligner un certains nombre de similitudes qui existes entre ces deux programmes, tant au niveau des problématiques et des objets de recherches définis, qu'aux niveaux de la nature et des structures de l'information recueillie et des analyses envisagées.

Pour présenter ses objectifs à Dakar, l'équipe animée par Philippe Antoine écrit [ANTOINE et col.1990] :

> " Nous essayons, dans ce travail, de **saisir les stratégies et les moyens mis en oeuvre par les migrants** et les non migrants, à travers les réseaux sociaux, parentaux, culturels, **pour s'insérer, eux et leurs familles, dans l'économie et la vie urbaines**. Le processus de l'insertion en ville doit être abordé en le replaçant dans l'ensemble des cheminements migratoires connus par les individus.

---

[1] : R. CABANES [1990] : Ordre du jour de la réunion d'UR du 13-14 septembre, 8 p

> L'insertion urbaine est un processus dynamique, mais qui n'aboutit pas toujours à une intégration durable ou définitive en ville. Il s'agit de voir comment migrants et non migrants, arrivent à satisfaire un certain nombre de besoins, en particulier travail et logement, alors qu'ils ne disposent peut-être pas ni des mêmes atouts, ni des mêmes exigences. Nous avons retenu trois composantes de l'insertion en ville: l'accès au travail, l'accès au logement, la constitution du ménage et son éventuel éclatement géographique."

Françoise DUREAU résume ainsi son questionnement sur les pratiques résidentielles et professionnelles des habitants de Quito [DUREAU 1990] :

> "Phénomène complexe, la mobilité spatiale vers Quito et au sein de l'agglomération n'a pas bénéficié d'un volume suffisant d'observations et d'analyses pour permettre une bonne compréhension du rôle des migrations dans la dynamique démographique et économique de la capitale équatorienne (.../...).
>
> - Quelles sont les stratégies développées par les quiténiens en matière d'occupation de l'espace géographique et économique de Quito ? Dans quelles logiques individuelles, familiales, sociales s'inscrivent les pratiques résidentielles et professionnelles mises en oeuvre par les migrants à Quito ?
>
> - Quel est l'impact de ces pratiques sur la dynamique de Quito : développement démographique et économique global de la ville, mais aussi structuration interne de l'espace quiténien (dynamique différentielle de certains secteurs), et structuration des échanges entre Quito et certains lieux de l'espace national ?"

Ces deux citations disent assez il me semble combien les deux recherches ont des problématiques parallèles. Quoi de plus naturel d'ailleurs en face des grands problèmes qui traversent la question urbaine dans le tiers monde, même s'ils doivent se poser bien différemment en Afrique de l'ouest et dans l'Amérique andine. Une analyse comparatiste des résultats des deux enquête parait aller de soi et ne pourra être que riche d'enseignements au plan thématique. Encore faut il qu'une telle approche soit facilement réalisable aux plans technique et méthodologique c'est à dire que les conditions de la collecte et le contenu informationnel des questionnaires permettent des objectifs et des méthodes d'analyse homogènes. Et là encore, nous sommes dans des conditions d'expérimentation quasi idéales.

Tout d'abord pour ces deux enquêtes, la sélection par des techniques de sondages probabilistes d'échantillons importants à partir de base de sondage bien

actualisées (respectivement plus de 3000 et plus de 1500 individus à Quito et à Dakar) nous garantie une bonne représentativité statistique des univers étudiés : deux agglomérations urbaines "entières", étant comprises leurs extensions les plus récentes. La taille des échantillon permettra de plus de mener des analyses assez détaillées sur des sous échantillons et donc de produire des résultats sur des sous ensembles des univers ;
- sous-ensembles spatiaux : quartiers, districts urbains, régions particulière de provenance des migrants etc.
- sous-ensembles sociaux ou démographiques : catégories de nivaux socio-économique, de nivaux scolaire, classe d'age, de sexe etc.
Nous aurons aussi la possibilité, fondamentale pour l'analyse démographique, de constituer des cohortes homogènes dans les échantillons.

Du point de vue maintenant des contenus informationnels nous disposons toujours dans les deux cas :
- d'un questionnaire socio-démographique " classique", avec à Dakar une saisie systématique et relativement fine des liens de parenté, souvent complexes dans le contexte africain, au seins d' unités diverses: noyaux familiaux, ménages (mono ou polynucléaires éventuellement éclatés spatialement), réseaux sociaux divers etc..
- du recueil assez complet dans les deux cas des biographies résidentielles et professionnelles de tous les individus de l'échantillon (à Dakar ces données biographiques s'étendent aux événement familiaux : mariages, naissances, décès etc...)[1]
En résumé, pour cette recherche méthodologique et son application comparatiste au cas de Quito et Dakar, je partirais de la donnée des itinéraires résidentiels et professionnels d'un échantillon représentatif d'individus migrants et non migrants, donnée qui inclue la description des parcours aussi bien en dedans qu'au dehors des villes enquêtées.

Enfin il faut parler des similitudes qui existent pour ces deux enquêtes dans certaines particularités des données collectées en vue d'un certain type d'analyses ; ces particularités témoignent je pense dans les deux cas d'une volonté de renouvellement des méthodes classiques de collecte et d'analyse démographiques des phénomènes migratoires et de leurs conséquences sur les dynamiques de l'urbanisation dans les PVD.
A Dakar comme à Quito, les méthodes de collectes et d'analyse employées évitent de limiter l'observation et la réflexion aux seuls migrants pour ne pas segmenter à priori le continuum des différentes formes de mobilité résidentielles et professionnelles, que celles ci soient définies à partir de critères spatiaux (migration vers ou à partir d'une ville, changement de résidence au sein de

[1] L'échantillon est constitué de chefs de ménages à Quito, il comprend des individus ayant tout types de statut dans les ménages à Dakar. Pour plus de précision sur les structures des échantillons et le contenus des questionnaires voir : Dureau et col.. [1988] et Antoine et col.. [1991].

l'agglomération) ou à partir de critères temporels (refus de la dichotomie migration temporaire versus migration définitive).
Autre originalité commune aux deux recherches, la volonté de cerner les phénomènes de multirésidence de fait au sein d'une période de temps qui conduit à une notion de densité de résidence, pour mieux rendre compte des stratégies (résidentielles, économiques etc...) de divers acteurs individuels ou collectifs (groupes familiaux, réseaux sociaux etc...) qui intègrent plusieurs lieux de l'espace régional, national voir international.
Il faudra bien sur que le travail méthodologique du statisticien reprenne à son compte cette volonté novatrice et la traduise dans la construction d'outils d'analyse qui respectent les données et les objectifs "thématiques".

Dans les paragraphes précédents nous constations d'une part qu'un recouvrement important, existe entre les problématiques qui présideront aux exploitations des données recueillies à DAKAR et à QUITO et d'autre part que l'homogénéité des données permet d'envisager un traitement simultané des deux enquêtes qui permettra une analyse comparée des résultats. Plus précisément , si l'on résume la problématique des deux recherches, quatre thèmes importants peuvent faire l'objet d'analyses statistiques "en parallèle".

- l'identification et la qualification des trajectoires spatio-temporelles (résidentielles et professionnelles) des individus depuis leur lieux d'origine jusqu'à leur situation en ville à la date de l'enquête.

- l'accès au logement : itinéraires résidentiel et mobilité spatiale des citadins migrants et non migrants.

- l'accès au travail : modalités de l'insertion professionnelle en ville.

- le rôle des groupes sociaux, souvent éclatés spatialement, auxquels appartiennent les individus : famille, collectivité villageoise, réseaux sociaux, etc.

Une avancée méthodologique significative sur le premier thème conditionne à notre avis une approche pertinente des trois autres. C'est donc l'exploration des pistes méthodologiques pour l'analyse statistique des itinéraires biographiques individuels et collectif qui va nous occuper principalement dans la suite et plus particulièrement, comme nous le verrons, dans la direction de l'analyse exploratoire et typologique de ces données. Pour dire les chose autrement, le projet d'une recherche comparatistes sur les processus d'insertion urbaine, les formes de la mobilité résidentielle et professionnelle et leurs conséquences sur la dynamique démographique et économique des villes trouve dans les deux enquêtes réalisée à Quito et à Dakar les moyens de son ambition puisque les données et les problématiques s'y prêtent ; il s'agit maintenant en priorité de se donner les moyens méthodologique et technique de sa réalisation. Une fois cet objectif atteint et si les résultats sont à la hauteur des espérances..., il me faudra,

compte tenus du calendrier prévu de ce programme (approximativement deux ans), réduire le champ thématique de l'application de ces techniques, c'est la raison pour laquelle je privilégierais le second thème (accès au logement) et ses interaction avec le quatrième (stratégies et rôle des groupes sociaux dans l'insertion urbaine).

2. QUELQUES PISTES POUR L'ANALYSE TYPOLOGIQUE DES DONNEES BIOGRAPHIQUES : REVUE BIBLIOGRAPHIQUE

2.1 Le problème

Avant d'explorer dans la bibliographie ce que propose les recherches actuelles dans le domaine de l'analyse statistique des données biographique, il me semble utile de dire rapidement comment se pose le problème en terme d'exploitation statistique pour ces deux enquêtes. Etant donné les tailles des échantillons les deux enquêtes doivent faire l'objet d'une exploitation "quantitative" des données recueillies, mais toutes les analyses n'utiliseront évidemment pas les mêmes méthodes. Une partie de l'exploitation ne soulève pas de problèmes méthodologiques particulier. C'est le cas lorsque les questions posées relèvent d'une analyse essentiellement **transversale** : étude des différences entre diverses catégories de population, démarche de démographie différentielle. On aura alors recours aux méthodes classiques de l'analyse statistique et démographique. Mais certains des problèmes posés imposent une analyse **longitudinale** des biographies migratoires et professionnelles. C'est ici que les questionnements thématiques des chercheurs en sciences sociales génèrent **un objet de recherche proprement statistique.**

En effet, pour identifier et qualifier les processus d'insertion urbaine des migrants, il faut conserver, tout au long de l'analyse des données, **l'ordre** de la séquence d'événements qu'ils connaissent et **les durées** passées dans chacun des états résidentiels et professionnels. **La caractérisation globale des itinéraires** que nous cherchons ne peut pas découler de l'analyse transversale séparée de chaque type d'événement : premier changement de résidence, premier emploi, première résidence ou premier emploi dans la ville étudiée etc. En effet, on aboutirait ainsi à caractériser les individus indépendamment pour chaque événement considéré; le croisement de ces typologies, outre qu'il risque de générer des classes d'effectifs trop faibles, regroupera des individus ayant certes connu un même ensemble d'événements, mais éventuellement dans des ordres, à des dates et avec des durées de séjour dans chaque état totalement différents. On ne pourra de ce fait rien conclure sur l'homogénéité des itinéraires des individus du groupe, pas plus que sur les types de dépendance statistique existant éventuellement entre ces événements. Comme le disent D. COURGEAU et E. LELIEVRE [1989] : "On voit dès lors que l'unité d'analyse ne doit plus être

l'événement mais la biographie individuelle, considérée comme un processus complexe".

2.2 Deux approche théoriques : concurrence ou complémentarité méthodologique ?

Posée en ces termes, **l'analyse statistique des biographies est un domaine de recherche relativement récent.** Dans son développement actuel, on peut distinguer deux approches théoriques qui sont à mon avis plus complémentaires que concurrentes.

L'approche probabiliste que développent principalement les statisticiens démographes à partir de modèles non-paramétriques, paramétriques ou semi-paramétriques (COX [1975], COURGEAU & LELIEVRE [1989]). Leur point de départ est une formalisation mathématique des biographies individuelles comme étant des processus aléatoires : la biographie d'un individu est représentée par une suite de variables aléatoires T_1, T_2,...T_n, qui sont les durées de séjour dans les divers états qui la composent, pris dans l'ordre chronologique. La démarche consiste alors à estimer, à partir des données de l'enquête, un modèle de distribution de chacune de ces variables, pour tenter ensuite de modéliser la distribution plus complexe de l'ensemble de la trajectoire (distribution conjointe). Les méthodes statistiques et les programmes informatiques que produit cette approche permettent de réaliser une **analyse fine des interactions entre phénomènes** démographiques et d'apprécier les effets des caractéristiques individuelles (sociales, économiques, culturelles etc.) et des événements "extérieurs" (contexte économique, socio-politique, catastrophes naturelles etc.) sur les durées de séjour dans des états donnés. Mais cela n'est possible qu'à la condition qu'on s'intéresse à un système d'événements et d'états en nombre limité et précisément définit a priori, **cette approche ne fournit pas d'outil descriptif ou typologique des trajectoires individuelles.**

L'approche par l'analyse des données. Les techniques d'analyse exploratoire des grands tableaux de données - analyses factorielles, classifications automatiques - ont souvent été appliquées aux séries chronologiques synchrones (principalement en économie) dans un but descriptif ou même prévisionnel. Mais le problème est en fait totalement différent ici : contrairement aux séries chronologiques habituelles, les données chronologiques des biographies individuelles sont d'une part asynchrones -il ne s'agit pas d'une série de "mesures" effectuées à des dates fixe pour tous les individus mais au contraire d'un ensemble de séquences d'événements datés ayant chacune son "horloge" propre, d'autre part les nombres d'événements et l'ordre dans lequel ils surviennent sont a priori différents d'un individu à l'autre. Du fait de ces caractéristiques particulières des données, c'est relativement tardivement, il y a environ dix ans, que les théoricien et les praticiens de l'analyse exploratoire des

données se sont attaqué à l'élaboration, au test et à l'utilisation d'outils spécifiques pour traiter le temps passé dans des situations successives.

2.3 Une rapide revue bibliographique et quelques principes méthodologiques

Ma recherche bibliographique initiale portait sur les deux approches théoriques, mais, outre qu'elle offre un champs de recherche beaucoup plus vierge que l'approche modélisatrice probabiliste, l'approche exploratoire typologique me semble plus adaptée à une phase de la recherche sur les processus d'insertion urbaine où la priorité est selon moi de décrire ces processus et d'en identifier les déterminants possibles dans un contexte ou les successions d'événements observées sont très complexes et ou l'univers des variables potentiellement explicatives est foisonnant. Ici, il ne me semble pas que l'on soit déjà en mesure de formuler des hypothèses suffisamment précises pour les valider (ou les invalider) à l'aide de modèles et de tests statistiques. Comment, de surcroît, dans un cadre de "plan d'expérience" si peu contrôlé peut on être raisonnablement sur que les hypothèses probabilistes qui fondent l'inférence sur le modèle ne soit pas partiellement voir totalement contredite par les données ? (doit on même employer le terme de plan d'expérience s'agissant de données de sciences sociales lorsqu'on songe aux protocoles d'échantillonnage et de mesure utilisés dans les sciences physiques ou biologiques, domaines qui ont fait le développement et le succès de la théorie statistique de l'estimation et du test). Pour toute ces raisons, j'ai choisi de commencer l'exploitation de ces deux enquête par une démarche exploratoire à l'aide d'outils statistiques typologiques qui visera entre autre à définir plus précisément des système d'événements et des sous populations pouvant faire l'objet d'une approche modélisations avec un risque "épistémologique" raisonnable[1].

Je vais donc me limiter maintenant à une revue rapide (non exhaustive) de quelques articles récemment parus dans le domaine de l'analyse exploratoire de données chronologiques individuelles : théorie ou techniques d'analyse factorielle (ACP ou AFC) ou de classification automatique qui directement ou indirectement (certains articles ne traite pas uniquement ou spécifiquement de calendriers biographiques) me semble apporter une contribution au problème de l'analyse typologique des données biographiques.

---

[1] Pour toute ces raisons, je n'ai pas voulu ici détailler, même partiellement, la bibliographie sur l'approche modélisatrice, très riche puisqu'elle couvre en fait tous les développements théoriques fait sur les processus stochastiques appliqués aux données de survie ("faillure time data" ou "faillure time series" dans la littérature anglo-saxonne). On trouvera en annexe quelques références d'articles et d'ouvrages importants au plan théorique et des exemples d'application démographiques.

J. PICARD [1987] introduit ainsi la méthode de classification de profils individuels évolutifs qu'il a développée :

> "L'analyse de l'évolution multivariée d'individus se ramène à l'étude des structures d'un tableau tridimensionnel dont les trois indices correspondes respectivement aux individus, aux variables et au temps. Les méthodes utilisées pour cette étude reposent essentiellement sur deux types d'approches factorielles :
>
> - la première approche a été proposée par LEBART [1966, 1969] sous le nom d'analyse factorielle locale (AFL) qu'il a appliquée à des contiguïtés spatiales. Indépendamment, LE FOLL [1982] introduit l'analyse factorielles des évolutions (AFE) puis en fait une généralisation sous le nom d'analyse factorielle pondérée (AFP). dans un article récent, CARLIER [1985] reprend l'AFE et montre que c'est une ACP sur un tableau d'observations associé à un graphe descriptif des contiguïtés, comme l'a fait LEBART pour l'AFL."

De cette première classe de méthodes d'analyse factorielle nous présenterons les idées directrices de l'article de LEBART [1969] intitulé "analyse statistique de la contiguïté".
PICARD poursuit :

> "-la deuxième approche est celle de la méthode STATIS développée par L'HERMIER DES PLANTES et ESCOUFIER [1976, 1978, 1980] dont un prolongement très intéressant de FOUCART [1983] aboutit à l'analyse factorielle de opérateur."

Dans ma première recherche bibliographique, je n'avait pas trouvé les articles théorique cité par PICARD sur les méthodes STATIS. Seul un article de UBERTALLI et PERNIN [1990] présentait une application de ces méthodes à l'étude longitudinale des carrières d'infirmières. Cet article ne contenant qu'un exposé méthodologique très succin, nous présenterons seulement ici en quelques lignes le principe de ces méthodes.

Enfin j'ajouterais deux autres approches factorielles de ces problèmes à l'inventaire de PICAR. La première est l'analyse harmonique qualitative, une application particulière de l'analyse des correspondances multiples, développée et justifiée par DEVILLE et SAPORTA dans deux articles théoriques [1980, 1982] et qui a été appliquée depuis par exemple par BERET [1988]. La seconde, qui pour être certainement la plus simple de toutes ne doit pas pour autant être oubliée est celle qui consiste à appliquer l'analyse des correspondance multiples classique à un tableaux ou l'on a résumé les trajectoires individuelles a étudier en un certains nombre de variables quantitatives ou qualitatives qui les décrirons synthétiquement. Tout repose alors sur le choix de ses résumés de trajectoire.

Etant donné son caractère très classique, il n'y a pas lieu de présenter cette méthodologie ici. Un excellent exemple de résultats intéressants obtenus grâce à ces méthode se trouve dans MARPSAT [1984].

L'analyse statistique de la contiguïté de LEBART

> " Les statisticiens ont souvent affaire à des ensembles de mesures ou d'observations qui ne peuvent être considérées comme des réalisations indépendantes de variables ou de vecteurs aléatoires. En économie, dans les Sciences Humaines ou Biologiques, les répétition d'épreuves identiques sont extrêmement rares.
> Cependant, le champs des observation suggère souvent la forme des liaisons entre observations. c'est le cas des série chronologiques et des modèles à erreurs liées dans le temps des économètres. Nous étudierons ici le cas plus général où les observations se réalise sur un graphe (i.e. le cas où un ensemble de couples d'observations est privilégié : couple d'observation successives pour les séries temporelles, couples d'observations contiguë pour les variables régionales, etc...)"

C'est ainsi que LEBART pose le problème auquel il apporte des éléments de solution dans son article de 1969 "Analyse statistique de la contiguïté". l'idée de départ est donc que, dans beaucoup de cas, l'évolution temporelle et/ou spatiale des données observées n'est pas indépendante de la structure du graphe sur lequel ces données se réalisent (graphe des contiguïtés spatiale ou temporelles). Comme c'est précisément l'hypothèse que l'on peut faire à propos des données biographiques individuelle, les développements de LEBART peuvent certainement nous fournir des pistes méthodologiques intéressantes.

Dans cet article, l'auteur a deux objectifs principaux :
- éprouver la validité (à l'aide d'un test statistique) de cette hypothèse de dépendance des données vis à vis du graphe et décrire cette dépendance
- fonder sur cette idée une nouvelle technique d'analyse factorielle qui permette la mise en évidence de l'échelle et de la localisation des liaisons spatiales ou temporelles entre les données.

Au départ de sa démarche, LEBART reprend la notion de coefficient de contiguïté définit par GEARY [1933] :

$$c = \frac{n-1}{2n_l} \frac{\sum_l (Z_i - Z_j)^2}{\sum_i (Z_i - Z)^2}$$

où : $n_l = \sum_{i=1}^{n} k_i$ est le nombre de sommets "voisins" dans le graphe,

$\sum_l ( )$, la somme pour tous les couples (i,j) de sommets "voisins" et

$\sum_i ( )$, la somme pour tout les sommets.

Ce coefficient peut donc être vu comme le rapport de deux estimations de la variance de la variable z : la première au numérateur est une estimation tenant compte de la position contiguë ou non des observations sur le graphe, tandis que la seconde est l'estimation classique de la variance empirique. Si les observations varient indépendamment de la structure du graphe, le coefficient de GEARY vaudra 1. Le calcul des moments de la distribution des coefficient de contiguïté permet donc de construire un test de l'hypothèse de dépendance par rapport au graphe.

Ayant atteint son premier objectif, LEBART généralise le coefficient de GEARY à la notion de contiguïté au niveau ß dans laquelle la variance du numérateur est calculée en prenant en compte les couple de sommet distant de ß sur le graphe. Cette généralisation lui permet tout d'abord de tester non plus seulement la dépendance vis à vis des voisinage strictes sur le graphe mais également les dépendances vis à vis de voisinages à 2, 3, etc... sommets de distance. Ainsi parvient t-on à identifier l'échelle spatiale ou temporelle à laquelle s'exerce l'autocorrélation des données. L'exemple d'application donné est celui d'une dépendance spatiale de données socio-économiques (6 variables) concernant 88 département français, pour lesquels on test 9 niveaux de contiguïté :

> "Nous obtenons donc six graphiques de neuf point chacun, qui caractérisent les niveaux auxquels se disperse la variable étudiée et qui précisent l'échelle géographique des phénomènes économiques. (.../...) Toutes les variables considérées ont un caractère de contiguïté très marqué, mis à part la densité du réseau routier dont l'influence se limite aux départements immédiatement voisins. Les revenus ont une zone d'influence s'étendant sur un rayon de trois départements, autour d'un

département donné. La contiguïté des dépenses est constamment moins marquée" etc...

LEBART passe ensuite à l'application de ces notions à l'analyse factorielle d'un ensemble d'observations contiguë. Il considère alors plusieurs variables X_1, X_2,...X_p se réalisant sur un même graphe (graphe des régions, des communes , des cellules ou des dates d'observations chronologiques) et il introduit alors la matrice de contiguïté des observations : "matrice de co-variance des accroissements des variables correspondants à des observations ayant une certaine distance sur le graphe. L'idée est ici que s'il y a indépendance des évolution spatiales ou temporelles par rapport à le structure de graphe, l'analyse factorielle (ACP) de la matrice de contiguïté ne sera pas significativement différent de l'analyse factorielle de la matrice de variance-covariance classique des variables. Au contraire, si l'hypothèse d'indépendance est fausse, les plans factoriels issus de l'analyse de la matrice de contiguïté montrerons des contraction des nuages de points vers l'origines dues au fait que lorsque les observations contiguë sont liées, la quantité du numérateur : $\frac{1}{2n_l}\sum_{l}(X_i-X_j)^2$, sous estime la variance.

En pratique il suffira donc de réaliser conjointement les analyse en composantes principales de la matrice de variance-covariance et des matrices de contiguïtés aux différent niveaux ß auxquels on s'intéresse à priori, puis de comparer les résultats et d'interpréter les éventuelles différences.

La méthode STATIS de L'HERMIER DES PLANTES et ESCOUFIER

L'outil statistique principal qu'utilise les méthodes multi-tableaux STATIS est l'analyse en composante principale de différents tableaux de distances construit à partir des différentes "dimensions" des tableaux de données de départ : distances inter-individus, inter-variables, inter-dates etc... Lorsqu'on veut traiter des calendriers d'événements, l'utilisation de ces méthodes impliquent des transformations des données de départ. Nous donnerons un aperçu schématique d'une telle démarche en prenant comme exemple l'application présentée dans l'article de UBERTALLI et PERNIN [1990].

UBERTALLI et PERNIN observent les parcours professionnels des premières promotions de l'école d'infirmières de Roanne. S'agissant d'ACP de tableaux de distance, il s'agit tout d'abord de faire en sorte que les trois dimensions du problème (individus, variables, dates) puissent être représenté dans des espaces euclidiens (nous verrons plus tard de quelle distance sont munit ces espaces). Pour se placer dans ce cadre, les auteurs décident tout d'abord de discrètiser et de synchroniser le temps d'observation des individus : elles choisissent, à partir de la date d'obtention du diplôme, des "instants clefs" d'observation des individus (T_1 = diplôme + 1 an, . T_2 = diplôme + 3 ans, T_3 = diplôme + 5 ans, T_4 =

diplôme + 10 ans). Ensuite, pour chacune de ces dates, on résumera la carrière individuelle par un certains nombre de variables métriques :

- le temps d'exercice infirmier en mois de travail
- le temps d'études complémentaires en mois
- le nombre d'arrêts d'activité
- le nombre de changements de secteurs
- le nombre de changement d'établissements

Une fois opérée cette "médiatisation" des informations recueillies sur les carrières individuelles, l'application de la méthodes STATIS se déroule en quatre principales étapes.

1. Analyse de "l'interstructure"

Cette étape a pour but la description des relations globales entre les différentes dates. Pour chaque date on calcule la matrice des distances euclidiennes entre individus dans l'espace des variables. Cette matrice est celle d'un opérateur linéaire qui traduit la "structure individus" à chaque date (d'ou le nom d'interstructure). Le produit scalaire de HILBERT-SMITH entre opérateurs linéaires permet de construire une distances entre date et donc une matrice de distances inter-date que l'on va soumettre à l'ACP. L'interprétation des plans factoriels fournit une typologie des structure globales individus x variables aux différentes dates et permet de se faire une idée de l'évolution temporelle moyenne des individus.

2. Analyse des "distances compromis"

Repartant des matrices des distances inter-individuelles à chaque date, on construit maintenant une matrice de distances compromis, c'est à dire la somme des distances à chaque date pondérée par les composantes du premier vecteur propre de l'interstructure. C'est en somme une distance inter-individuelle moyenne sur l'ensemble des dates, dans laquelle une date intervient d'autant moins qu'elle est atypique (au sens de l'ACP) par rapport aux autres. L'ACP de ce tableau de distances décrit donc la structure individus commune à l'ensemble des dates, fournissant une sorte de résumé temporel des trajectoires individuelles.

3. Analyse de l'intrastructure des variables

considérons maintenant le grand tableaux rectangulaire obtenu en "mergeant" les différents tableaux individus x variables de chaque date. De plus, pour permettre la superposition des plans factoriels avec ceux obtenus dans l'ACP des distances compromis, pondérons à nouveau chacun des tableaux élémentaires individus x

variables par les racines carrées des composantes du premier vecteur propre de l'interstructure. L'analyse factorielle de ce tableau et plus précisément la projection du nuage des variables x dates permet en fait d'observer les corrélations entre ces variables x dates et les composantes principales du compromis. C'est à dire que l'intrastructure des variables permet d'interpréter les axes du compromis à partir des variables de départ, descriptives des carrières individuelles. Les positions des points variables x dates décrivent les relation qu'ont entre elles les variables à chaque date et l'importance de chaque variable dans la discrimination des diverses formes de carrières.

4. Analyse de l'intrastructure des individus

Enfin l'analyse de l'intrastructure des individus se fera par la mise en éléments supplémentaires des tableaux individus x variables à chaque date dans l'ACP du tableau des distances compromis. On obtiendra sur les plan factoriels du compromis un point par individu et par date que l'on peut relier dans l'ordre chronologique pour obtenir la trajectoire de chaque individus. Pour interpréter les formes et les positions relatives de ces trajectoires il faudra revenir à l'interprétation des axes du compromis donné par l'intrastructure des variables.

L'analyse harmonique qualitative de DEVILLE et SAPORTA

L'analyse harmonique quantitative a été présentée une première fois dans un court article théorique par DEVILLE et SAPORTA en 1980, puis à nouveau plus longuement par DEVILLE, dans un article de 1982, où après avoir présenté complètement les fondements théorique de la méthode, il l'applique à des données biographique concernant la vie matrimoniale d'un échantillon de femme mariées trois fois ou plus. DEVILLE le dit lui même, plus que l'intérêt même du sujet, c'est la nature très particulière des données et leur caractères à la fois accessibles et pédagogique qui l'a orienté dans ce choix. Or précisément la nature des données considérées par DEVILLE et SAPORTA est exactement identique à celle des données des deux enquêtes de Dakar et Quito. Qu'on en juge d'après la présentation par les auteurs du problème qu'ils vont aborder :

> "De nombreuses données d'enquêtes permettent de retracer l'histoire d'individus pendant une certaine période de temps : évolution de l'activité professionnelle, de la situation matrimoniale, de la résidence. Le but de cet article est de montrer que l'on peut analyser de telles données de façon analogue à ce que l'on ferait pour un processus scalaire (analyse harmonique) et pour un ensemble fini de variables qualitatives non temporelles (analyses canonique généralisée ou analyse des correspondances multiples)."

La réalisation de cet objectif entraîne DEVILLE et SAPORTA dans des développements théoriques assez difficiles. Je ne ferais ici que résumer leur démarche.

Le point de départ est de considérer ce type de données biographiques comme une réalisation d'un processus qualitatif en temps continu $X_t(i)$ à m modalités qui sont donc les m états dans lesquels peuvent se trouver les individus ($X_t(i)$ est donc l'état, parmi les m états possibles, dans lequel se trouve l'individus i à l'instant t). Puisqu'on a affaire à un processus qualitatif que l'on ne sait pas analyser, on va rechercher un codage scalaire de ce processus : $Y_t = f_t(X_t)$ ou f_t est une fonction qui transforme un élément de l'espace produit temps x états en un nombre réel (en d'autre termes on associe à toutes les modalités possibles du croisement des états avec les instants d'observation un codage réel). On peut ensuite faire l'analyse harmonique du processus Y_t, c'est à dire le décomposer en une suite de variables aléatoires indépendante du temps et une suite de fonction du temps (dites harmoniques du processus).

> "L'interprétation pratique peut se faire comme celle des composantes et des harmoniques en analyse harmonique : graphiques des codages de chaque état en fonction du temps, corrélation des génératrices avec certaines variables connues (cercle des corrélations)."

Mais DEVILLE et SAPORTA font ensuite une remarque de grande importance pratique :

> "Cela dit, l'analyse harmonique qualitative est une généralisation de méthodes familières d'analyse des données et s'y ramène d'ailleurs sur le plan numérique. Toutes les techniques d'interprétations habituelles dans ces méthodes sont donc encore valables dans notre cas."

Suit dans l'article la démonstration de ces propriétés de l'analyse harmonique qualitative, puis l'exposé d'une méthode d'analyse en temps discret opérationnelle. C'est cette méthode qui nous intéresse pratiquement. Le principe en est simple :

On commence par discrètiser l'intervalle de temps d'observation du processus : les dates t_0 (début de l'observation), $t_1,\ldots t_n$ (fin de l'observation) sont les bornes des intervalles de temps durant lesquels le processus reste stable, c'est à dire durant lesquels aucun individus ne change d'état. La décomposition de l'analyse harmonique en temps continu devient en temps discret la décomposition en facteurs obtenus par l'analyse des correspondances multiples d'un tableau disjonctif particulier à n lignes et mp colonnes (où n est le nombre d'individus, m le nombre d'état possibles et p le nombre de périodes durant lesquelles le processus est stable). La case courante $k_{i\,(jk)}$ du tableau vaut 1 si l'individu i

est dans l'état j durant la période k, 0 sinon. Le problème pratique est que dans les cas concrets d'application les nombre n, m et p génèrent un tableau de dimension pharamineuse. DEVILLE et SAPORTA proposent alors plusieurs solution d'approximation du problème complet par recodage du tableau de départ. La plus simple et efficace est celle de la division de l'intervalle d'observation en un nombre raisonnable de périodes (de durée constante ou non) sans tenir compte des changements d'états individuels. Le tableau est ensuite construit en calculant le temps passé par chaque individus dans chacun des états possibles au cours de chaque période. On pondérera ensuite chaque colonne par la durée de la période. Il suffit ensuite de soumettre ce tableau à l'analyse des correspondances multiples et d'utiliser les techniques d'interprétations habituelles.

CONCLUSION : UNE METHODOLOGIE D'ANALYSE DES BIOGRAPHIES, POUR QUOI FAIRE ?

On ne peut évidement pas maintenant préjuger des résultats que fournira ou ne fournira pas, au plan thématique, l'exploitation statistique des biographies collectées à DAKAR et à QUITO. Du point de vue de la méthodologie statistique en revanche, les paragraphes précédents ont signalé quelques pistes qu'on peut résumer ainsi :

- la mise au point d'outils d'analyse "exploratoire" des données biographiques pour l'étude typologique des itinéraires individuels et collectifs

- la recherche d'une démarche d'exploitation des enquêtes biographiques alternant les phases d'identification et de caractérisation des itinéraires, par l'approche "analyse des données", et les phase de modélisation voir "d'explication", par l'approche "processus aléatoires", lorsqu'on étudie un système d'événements particulier (étapes migratoires et professionnelles "charnieres" par exemple).

Ces avancées méthodologiques et leur application aux données des enquêtes doivent permettre d'aborder, dans un dialogue regroupant statisticiens et thématiciens, la critique et la réflexion sur l'information collectée dans les enquêtes biographiques sur les processus d'insertion des migrants en ville. Etant donné un questionnement précis, quelles sont les étapes résidentielles et professionnelles "déterminantes" dans les itinéraires que l'on observe ? Quelle est donc l'information "optimale" à collecter (ni trop ni trop peu) pour analyser ces phénomènes?

A ce propos, je ne peut terminer cette rapide revue bibliographique sans tenter de répondre, partiellement bien sur, à l'opinion que Bourdieu formule dans un article intitulé "l'illusion biographique" que j'ai lu au tout début de la rédaction de cette communication.

S'il s'agit de reconnaître que le récit biographique, par convention tacite entre l'enquêteur et l'enquêté ou par commodité chronologique, est orienté - c'est à dire, comme le dit Bourdieu, "comporte un commencement (<< un début dans la vie >>), des étapes et une fin, au double sens de terme et de but (<< il fera son chemin>> signifie il réussira, il fera une belle carrière), une fin de l'histoire" - donc que tout récit biographique est chargé artificiellement de sens, bref que les objets décrits par la "données" biographique sont en partie artefactuels, j'en convient aisément comme d'une évidence. Mais dans cette construction éminemment subjective qu'est le récit biographique, il y aussi, du point de vue de l'enquêté comme de celui de l'enquêteur, du sens "objectif", des événement vécus, c'est a dire du factuel susceptible d'interventions scientifiques : descriptions, interprétations analyses. Doit on adopter la position intransigeante de Bourdieu, et faire de l'analyse sociologique des mécanismes de construction de cet artefact un préalable indispensable à tout traitement, statistique ou autre, de données biographiques? Je pense au contraire, au nom de la nécessaire prise en compte scientifique de la complexité chère à Edgar Morin, qu'il faut continuer à collecter et à analyser des données biographiques, de préférence avec des méthodologies et des grilles d'analyse sans cesse renouvelées. Et, comme en témoigne la rapide exploration bibliographique qui précède, le champs ouvert des recherches méthodologiques est vaste.

NOTES BIBLIOGRAPHIQUES

ANTOINE Ph, BOCQUIER Ph, FALL A. S., GUISSE Y. M., NANITELAMO J. [1990].: L'insertion urbaine des migrants à Dakar, présentation de l'étude réalisée par l'IFAN et l'ORSTOM; *multigr.* 8 p.

BERET P. [1988] : Analyse de données chronologiques relatives à l'insertion professionnelle; " Les Cahiers de l'Analyse des Données" vol. XIII, n° 2, pp 159-174.

COURGEAU D., LELIEVRE E. [1989] : Analyse démographique des biographies; édition de l'INED, Paris, 268 p.

COX D.B. [1972] : Regression models and life tables (with discussion); Journal of Royal Statistical Society, B 34, pp 187-220.

DEVILLE J.C., SAPORTA G. [1980] : Analyse harmonique qualitative ; 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 ? Annales de l'INSEE, n° 45, pp 45-104.

DUREAU F. [1990] : Recueil et analyse de biographies migratoires et professionnelles à Quito (Equateur); in "Travail et Travailleurs", ORSTOM, 10 p.

ESCOUFIER Y., L'HERMIER DES PLANTES [1978] : A propos de la comparaison graphique des matrices de variance, Biometrical Journal, Vol 20, 5, pp 491-497.

FOUCART T. [1983] : Une nouvelle approche de la méthode STATIS, Revue de Statistique Appliquée, Vol. 31, 2, pp 61-95.

LEBART L [1969] : Analyse statistique de la contiguïté, Publications de l'Institut de Statistique de l'Université de Paris, 1969, Vol. XVIII pp 81-112.

LE FOLL Y.[1982] : Pondération des distances en analyse factorielle. Statistique et analyse des données, 1982, 1, pp 13-31.

MARPSAT M. [1984] : Chômage et profession dans les années trente; Economie et Statistique n° 170, pp 53-69.

PICARD J. [1987] : Classification des profils évolutifs incomplets et asynchrones; Revue de Statistique Appliquée, vol. XXXV (2) pp 27-38.

UBERTALLI B., PERNIN M.O. [1990] : Utilisation d'une méthode multitableaux en sciences sociales. Une étude longitudinale de carrières : les 12 premières promotions de l'école d'infirmières de Roanne. Paris, Population n° 6, pp. 1092 - 1100.

PRATIQUES SOCIALES ET TRAVAIL EN MILIEU URBAIN

les cahiers

n° 16 - 1991

MIGRATIONS, TRAVAIL, MOBILITES SOCIALES : METHODES, RESULTATS, PROSPECTIVE.

Séminaire ORSTOM - Garchy 24-27 Septembre 1991
Communications des séances 1 et 2

Editeurs scientifiques
Véronique DUPONT et Françoise DUREAU

UR : Migrations, Travail, Mobilités Sociales
Département SUD - ORSTOM -
72 Route d'Aulnay - 93143 BONDY Cedex - Tél : 48 47 31 95