

2 AVRIL 1985.

- 46 O.R.S.T.O.M. Fonds Documentaire
N° : 17.240 ex 1
Cote : B

UNE UTILISATION DE LA MÉTHODE DES NUÉES
DYNAMIQUES LORS D'UNE ENQUÊTE SOCIO-DÉMOGRAPHIQUE

PAR PHILIPPE ANTOINE
ET AGNÈS GUILLAUME

La tentation est toujours grande, lors d'une enquête statistique, de multiplier les questions afin de saisir le mieux possible la réalité sociale. Or il est souhaitable qu'une enquête portant sur un échantillon conséquent d'individus ne comporte qu'un nombre restreint de questions (contraintes de temps, de coût, de fiabilité des réponses, ...). Afin de mieux comprendre et expliquer les données recueillies, il semble préférable de se livrer à des investigations plus approfondies auprès de certains enquêtés. Divers critères peuvent être envisagés : tirer au hasard un individu sur 100 ou 1 000, retenir les personnes les plus communicatives, procéder à un choix raisonné, etc. Ces différents modes de tirage d'un sous-échantillon ne sont guère satisfaisants car aucun moyen ne permet de connaître le degré de représentativité des individus sélectionnés.

Certaines méthodes d'analyse de données peuvent aider dans ce choix et la méthode des nuées dynamiques utilisée dans cet exemple présente de nombreux avantages qui justifieraient son utilisation plus fréquente lors de l'exploitation et de l'analyse des enquêtes démographiques et statistiques portant sur des effectifs importants.

1 - OBJECTIF INITIAL : PASSER DU QUANTITATIF AU QUALITATIF

A partir des données recueillies à Abidjan lors de l'enquête ivoirienne à passages répétés (EPR) (1) nous comptons prolonger l'analyse des relations

(1) Cette enquête a été réalisée et financée par la Direction de la Statistique de Côte d'Ivoire (01 BP V 55 - ABIDJAN 01). Les premiers résultats sont disponibles auprès de ce service. Pour une présentation de l'enquête d'Abidjan, voir aussi ANTOINE Ph. et HERRY Cl. (1983).

Ph. ANTOINE et A. GUILLAUME sont des démographes de l'ORSTOM.

- 47 -

familiales en essayant de mieux cerner l'espace de vie des chefs de ménage tant dans les liens entretenus avec le milieu d'origine que dans le champ des relations familiales en ville. Une démarche plus anthropologique s'imposait, il restait donc à choisir les individus susceptibles d'être enquêtés. L'hypothèse initiale émise est la suivante : il existe des comportements différenciés selon les types d'habitat et il convient donc de distinguer les chefs de ménage selon cette catégorie et également selon le type de famille auquel ils appartiennent. L'habitat nous semble être une variable assez synthétique traduisant bien la diversité du fait abidjanais. Quatre types peuvent être distingués :

- l'habitat de haut et moyen standing (ou résidentiel) qui regroupe les catégories sociales les plus élevées ;
- l'habitat économique moderne, où résident surtout des salariés qualifiés du secteur moderne ;
- l'habitat évolutif, ou cour, constitue l'habitat le plus répandu et le plus ancien de l'agglomération ; chaque cour comprend plusieurs logements ;
- l'habitat sommaire comprend des constructions en bois de palme et surtout des baraques. Il est malgré sa précarité relativement bien structuré et doit être considéré comme un habitat de substitution à l'habitat évolutif.

Les individus retenus doivent donc être représentatifs des différentes modalités d'une typologie combinant catégories d'habitat, formes de ménage (familiale nucléaire, famille élargie, groupement de célibataires et ménage d'une personne) et incluant d'autres variables pouvant influencer la typologie (CSP, âges, ...). Notre but est de sélectionner des individus, a priori représentatifs des différentes classes, que nous devons ensuite retrouver sur le terrain et interviewer pour mieux connaître à la fois l'évolution de la composition de leur famille, leur réseau familial en ville et les modalités de leur installation à Abidjan. Par ailleurs, cette nouvelle investigation amène à appréhender le devenir de la population cinq ans plus tard (1978 - 1983).

2 - LE CHOIX DE LA MÉTHODE

Les méthodes de classification automatique effectuent une partition d'un ensemble de données mais peu d'entre elles permettent de dégager les individus

in: Statéco, Bull. de liaison, n° 38, juin 1984

17.240 ex 1

représentatifs de cette typologie. Différentes approches sont possibles comme rechercher une partition qui minimise un certain critère, selon un certain seuil ; ou bien retenir des techniques basées sur l'agrégation de groupes par itérations successives.

Nous avons retenu les méthodes de partitionnement par agrégation à partir de noyaux (1). En effet, elles permettent de traiter de grands ensembles de données (plusieurs milliers d'individus).

Deux méthodes de "partitionnement" ont été prises en considération :

- la méthode des boules optimisées ;
- la méthode des nuées dynamiques.

Ces deux méthodes sont des algorithmes d'agrégation autour de noyaux variables. La différence essentielle entre elles étant le mode de choix des noyaux à partir duquel on agrège.

Dans le cas des boules optimisées, l'agrégation se fait en fonction d'une distance inférieure à un rayon fixé a priori (mais dont le choix peut être guidé). Si la méthode dégage facilement des typologies, elle ne permet pas d'extraire automatiquement les individus les plus représentatifs. Le nombre de classes de la typologie est très sensible à une variation même minime du rayon.

La méthode nuées dynamiques mise au point par E. DIDAY (1971) est la mieux adaptée à nos objectifs. Elle permet de "générer une partition à partir d'un corps de données sur lequel on ne demande pas de faire d'hypothèses a priori. Cette partition doit réaliser les deux propriétés suivantes :

(1) Les méthodes de classification hiérarchique (méthodes ascendantes) procèdent par agrégations successives (selon un critère de distance) d'éléments isolés au départ, à la différence d'autres méthodes de classification qui agrègent selon un noyau central fixé a priori. Les méthodes hiérarchiques n'ont pu être utilisées pour un fichier aussi important que le nôtre (7 800 individus x 60 modalités).

- les individus de chaque partie se ressemblent le plus possible ;
- les individus de deux parties différentes se ressemblent le moins possible" (DIDAY, p. 20).

Le processus est itératif et à chaque étape la qualité de la partition s'améliore. Le nombre de classes souhaité est déterminé a priori ainsi que le nombre d'éléments centraux désiré, c'est-à-dire le nombre d'éléments au centre du noyau qui seront énumérés. Au départ, un ensemble de points ou noyaux d'une classe peut être tiré au hasard. Autour de ces points se regroupent les éléments les plus proches pour former une partition. La distance calculée par rapport au centre de classe est la distance du Chi-Deux. A partir de cette partition créée, une autre famille de noyaux est définie, elle regroupe les points les plus proches formant une nouvelle classe et ainsi de suite jusqu'à obtention d'un nombre fini de classes. Si, après un certain nombre d'itérations, les classes formées sont stables, les données sont dites "classifiables" et constituent des "formes fortes". Les individus qui changent de classes selon les tirages sont les "individus charnières".

3 - L'APPLICATION

L'application présentée ici a été réalisée à l'aide des logiciels mis à la disposition des utilisateurs par l'ADDAD (1).

Un fichier "chef de ménage" a été établi à partir des données de l'EPR (strate Abidjan). Il comprend environ 7 800 chefs de ménage africains résidents au 1er passage de l'enquête et seulement une dizaine de variables démographiques et socio-économiques parmi les plus discriminantes (âge, situation matrimoniale, instruction, profession, habitat, type de famille, ...) sont retenues pour l'analyse typologique. Les modalités de ces variables sont tout d'abord regroupées et

(1) ADDAD : Association pour le Développement et la Diffusion de l'Analyse des Données.

Laboratoire du Professeur BENZECRI.

Siège social : Laboratoire de Statistique
4 rue André Chenier
92130 ISSY-les-MOULINEAUX.

des classes. Le troisième tableau présente la liste des individus centraux de chacune des classes, correspondant à la sélection désirée (10 individus par classe dans notre cas). Deux tableaux d'aide à l'interprétation sont fournis également (non présentés ici) : une mesure d'homogénéité des classes (variance) et une mesure de la valeur du critère pour chaque partition obtenue. Enfin, quand la convergence est atteinte, un tableau des individus des formes fortes figure en fin de listing.

Il faut bien distinguer entre les éléments centraux obtenus en sortie et l'ensemble des individus composant la classe. Un simple retour au fichier permet de connaître les caractéristiques des individus tirés, qui ont pour une même classe des caractéristiques voisines. Par contre l'ensemble des individus composant une même classe peut être plus hétérogène.

Le tableau 2 donne la répartition des 9 classes selon l'habitat et selon la catégorie socio-professionnelle. Ainsi le type 1 regroupe essentiellement des cadres, indépendamment de l'habitat où ils résident (une minorité logeant dans les cours plus traditionnelles). Dans ce cas, le statut professionnel prime sur le type de logement de résidence. De même, pour la classe 7 qui regroupe exclusivement des chefs de ménage apprentis ou aides familiaux. Il s'agit le plus souvent dans ce cas de jeunes qui logent dans la boutique de leur patron. A l'inverse, la classe 8 rassemble essentiellement des chefs de ménage résidant dans l'habitat sommaire quel que soit le type de famille auquel ils appartiennent. Après analyse des différents tableaux, il s'avère que certains types doivent comporter des intitulés différents. C'est le cas pour les types :

Type 1 : Chef de ménage cadres ou employeur

Type 3 : Chef de ménage famille élargie salarié des secteurs commerce et service moderne

Type 5 : Chef de ménage sans emploi (ce type comprend aussi des hommes, alors que les individus étalons et les individus représentatifs sont des femmes)

Type 6 : Chef de ménage famille nucléaire ou groupement de célibataire résidant dans l'habitat de cour

Type 7 : Chef de ménage apprenti ou aide familial

Type 8 : Chef de ménage résidant dans l'habitat sommaire.

REPARTITION DE LA TYPOLOGIE SELON L'HABITAT ET LES CATEGORIES SOCIO-PROFESSIONNELLES

TABLEAU 2

HABITAT	1	2	3	4	5	6	7	8	9	TOTAL
Typologie N. DYN.										
Résidentiel	163	5	154	27	13	23	2	1	17	405
Economique Moderne	76	5	971	215	42	86	3	7	44	1 449
Cour	25	106	390	1 905	230	1 618	90	0	623	4 987
Sommaire	1	8	14	14	22	0	19	731	144	953

CATEGORIES SOCIO-PROFESSIONNELLES

Sans emploi	4	61	9	4	280	2	0	6	27	393
Cadres - Employeur	261	2	14	41	0	24	0	3	7	352
Exploitant indépendant	0	24	68	579	27	563	0	206	296	1 763
Salarié qualifié commerce - service	0	3	1 157	161	0	105	0	35	93	1 554
Salarié non qualifié commerce - service	0	5	15	139	0	255	0	133	83	630
Salarié qualifié industrie	0	0	250	383	0	194	0	68	121	1 016
Salarié non qualifié industrie	0	6	5	802	0	537	0	240	186	1 776
Aide familial - Apprenti	0	0	0	0	0	0	114	0	0	114
Travailleur de l'agriculture	0	23	11	52	0	47	0	48	15	196
TOTAL	265	124	1 529	2 161	307	1 727	114	739	828	7 794

Ainsi pour certains types la définition doit être revue et l'hypothèse avancée doit donc être modifiée. Il n'en demeure pas moins, dans notre cas, que les types retenus se stabilisent rapidement, 91 % des individus se positionnent dès la première itération et dès la troisième itération chaque individu a rejoint sa partition. A titre d'exemple nous donnons les effectifs de chacune des combinaisons possibles (tableau 3). Le code correspond au type affecté à chacune des trois itérations : le code 463 par exemple correspond à un individu qui rejoint le type 4 à la première itération, le type 6 à la seconde et le type 3 à la dernière itération. Ce tableau permet d'évaluer le nombre des individus charnières, ainsi par exemple 31 individus passent de la classe 9 (ménage 1 personne) à la classe 7 (apprenti ou aide familial). Il n'y a aucune difficulté à lister ces individus afin de les identifier, et éventuellement les retrouver ensuite sur le terrain.

Notre objectif premier était d'obtenir une liste d'individus appartenant à chacun des types. La méthode de sélection a répondu à notre attente, toutefois dans notre cas le temps écoulé entre l'enquête initiale et l'investigation sociologique est trop long : cinq ans, mais malgré ce handicap, nous avons pu retrouver de nombreux chefs de ménage. Reste à s'interroger sur la notion d'individu central d'une classe, d'individu moyen. Effectivement, pour une classe donnée, les principales caractéristiques retenues sont semblables (par exemple pour les vieux : l'âge, l'habitat, la date d'installation à Abidjan, ...), mais derrière ces indications ponctuelles se cachent des vies bien singulières. Il est évident que la méthode ne permet de sélectionner les individus qu'en fonction des variables introduites et si les personnes sont interrogées sur un autre sujet (les conditions de leur insertion à Abidjan par exemple) les réponses divergent. Mais même si le passé diffère, des traits saillants communs apparaissent comme l'importance du tuteur au moment de l'adolescence, le rôle joué par l'épouse, etc.

4 - INTERET ET LIMITES DE LA METHODE

Cette méthode permet d'obtenir une représentation synthétique de grands ensembles de données en utilisant un temps de calcul relativement limité. Elle fournit de façon plus lisible le classement de chaque individu au sein de la typologie, ce qui constitue, pour cet aspect, un avantage par rapport aux méthodes d'analyse factorielle des correspondances. En effet, la projection des individus

INDIVIDUS STABLES ET INDIVIDUS CHARNIERES

TABLEAU 3

CODE	FREQ	EFF. DEF.	CODE	FREQ	EFF. DEF.	CODE	FREQ	EFF. DEF.
111.	262	265	564.	1		855.	10	
133.	1		566.	2		866.	10	
155.	3		599.	2		877.	1	
222.	124	124	611.	1		886.	1	
333.	1 386	1 529	633.	33		888.	718	739
336.	4		636.	2		911.	2	
344.	30		644.	43		933.	11	
355.	7		646.	2		939.	1	
363.	2		655.	16		944.	4	
366.	11		663.	2		949.	1	
399.	1		664.	74		955.	4	
433.	82		665.	2		959.	1	
444.	1 979	2 161	666.	1 528	1 727	966.	7	
446.	15		686.	1		969.	2	
455.	36		688.	3		977.	31	
463.	1		699.	1		988.	6	
464.	26		766.	1		993.	3	
466.	142		777.	82	114	995.	1	
488.	7		833.	8		998.	5	
544.	1		844.	3		999.	819	828
555.	228	307	846.	1				

sur le plan factoriel est bien souvent d'une lecture difficile lorsque de grands tableaux de données sont pris en considération, la superposition des points constituant une perte d'information.

La partition peut être réalisée, on l'a vu, de deux façons :

- soit à partir d'éléments tirés au hasard, ce qui est utile lorsque aucune information n'est disponible a priori sur le corps des données ; cette procédure semblant un bon préalable à un dépouillement classique ;

- soit à partir d'éléments étalons : la typologie obtenue vérifie ou non les hypothèses préalablement établies. Cette deuxième solution est préférable car, en principe, tout utilisateur a un corps d'hypothèses qu'il désire tester. La typologie peut être réalisée en introduisant des individus fictifs possédant des caractéristiques bien déterminées de façon à obtenir ainsi les individus réels qui s'agrègent à chacun de ces modèles.

Un autre intérêt de la méthode réside dans la possibilité d'extraire la classe affectée à chacun des individus et l'adjoindre au fichier initial. Cette typologie devenant elle-même une nouvelle variable, plus synthétique, permettant d'autres analyses.

A notre sens, la possibilité d'extraire nominalement les individus centraux de chacun des types constitue l'avantage principal de la méthode. Ces individus centraux possèdent en principe les caractéristiques moyennes de la classe. Le nombre de ces éléments est fixé a priori par l'utilisateur. Une identification rapide de ces individus à l'issue de l'enquête permet de réaliser conjointement une investigation plus approfondie sur un sous-échantillon représentatif de chacune des classes obtenues. A l'inverse, les cas marginaux, révélés par leur plus faible capacité à se stabiliser rapidement dans le modèle, méritent aussi d'être étudiés : leur marginalité statistique traduit peut-être une marginalité sociale.

Comme toute méthode d'analyse des données, celle des nuées dynamiques présente quelques limites.

Les modalités comprenant des effectifs peu nombreux tendent à constituer des classes parfois non pertinentes ; mais cette limite est aussi un avantage dans la mesure où certains cas aberrants peuvent être décelés de cette façon, surtout si un fichier de données brutes est utilisé.

Une modalité composant majoritairement une variable occulte les autres possibilités : ainsi nous avons dû constituer, lors du travail préparatoire, un second fichier ne comprenant pas l'habitat de cour, afin de mieux dégager les types intégrant les populations résidant dans les trois autres formes d'habitat.

Des variables trop redondantes risquent de masquer l'influence des autres caractéristiques.

Le choix des étalons laisse une grande part d'arbitraire dans la mesure où certaines variables peuvent être privilégiées à tort. Mais l'utilisation dans un premier temps d'un tirage au hasard des étalons, puis l'élaboration de la typologie à partir des premiers individus tirés limite les risques de manipulation inconsidérée des données.

Cette méthode comporte une part d'arbitraire mais il faut surtout concevoir son utilisation comme le moyen de tester un corps d'hypothèses. Si le choix des étalons est judicieux, ces derniers se retrouvent en grand nombre comme éléments centraux ou pour le moins dans la classe qu'ils ont aidé à déterminer. Si l'hypothèse est fautive, ces individus étalons se dispersent dans diverses classes à l'issue de la partition.

--- ---

BIBLIOGRAPHIE

- ANTOINE Ph., HERRY Cl. - 1983
Quelques éléments pour l'évaluation des résultats de l'enquête à passages répétés menée à Abidjan
STATECO n° 34, juin 1983, INSEE, pp. 15-38.
- DIDAY E. - 1971
Une nouvelle méthode en classification automatique et reconnaissance des formes. La méthode des nuées dynamiques
Revue de Statistiques Appliquées, vol. XX, n° 2, pp. 19-33.
- FLAMENBAUM G., THIERY J., BENZECRI J.P. - 1979
Agrégation en boules de rayon fixe et centre optimisés (boules optimisées)
Les Cahiers de l'Analyse des Données, n° 3, pp. 357-364.
- LEBART L., MORINEAU A., TABARD N. - 1977
Techniques de la description statistique
Dunod, Paris, 351 pages.
- VOLLE M. - 1981
Analyse des données
Economica, Paris, 317 pages.