Les méthodes de classification appliquées aux recensements : réalisation d'une typologie des entreprises ostréicoles à l'aide d'une Classification Ascendante Hiérarchique.

Gabrielle Lesur-Irichabeau, Patrick Point

LES METHODES DE CLASSIFICATION APPLIQUEES AUX RECENSEMENTS : REALISATION D'UNE TYPOLOGIE DES ENTREPRISES OSTREICOLES A L'AIDE D'UNE CLASSIFICATION ASCENDANTE HIERARCHIQUE

Gabrielle LESUR-IRICHABEAU
Doctorante
GREThA UMR CNRS 5113
Université Montesquieu - Bordeaux IV
Avenue Léon Duguit 33608 Pessac Cedex
gabrielle.lesur-irichabeau@u-bordeaux4.fr

Patrick POINT
Directeur de Recherche CNRS
GREThA UMR CNRS 5113
Université Montesquieu - Bordeaux IV
Avenue Léon Duguit 33608 Pessac Cedex
patrick.point@u-bordeaux4.fr

RÉSUMÉ – Cette étude a pour objectif de montrer comment les données issues de recensement peuvent être traitées de façon à réaliser une typologie des individus sans perte d'information. Le problème avec les recensements, c'est qu'ils permettent de dégager des variables qui font référence à des thématiques différentes et qui ne peuvent donc pas être traitées conjointement. La solution consiste à créer, à partir des variables initiales, des variables synthétiques qui font ensuite l'objet d'une Classification Ascendante Hiérarchique. On obtient alors des regroupements des individus en un petit nombre de classes qui tient compte de l'ensemble de l'information initiale.

ABSTRACT – The aim of this study is to show how inventory's data could be treat in order to realize a typology of individuals with complete initial information. The problem of inventory is to create a lot of variables in different sets of themes. These sets of themes could not be treating together. The solution is to treat these sets of themes one by one to create new synthetic variables by factorial analysis and clustering and then, with a new factorial analysis and clustering on news synthetics variables, to obtain regrouping of individuals. The regrouping in few clusters must take into account all initial information.

Mots-clés : Analyse de données, Classification Hiérarchique Ascendante, recensement, typologie.

Keywords: Data mining, Agglomerative Hierarchical Clustering, inventory of population, typology.

1. Introduction

La classification est définie par Gordon (1987) comme la procédure permettant de séparer un ensemble d'individus afin d'en faire un petit nombre de classes dans lesquelles les individus appartenant à la même classe ont des caractéristiques communes qu'ils n'ont pas avec les autres classes. Par cette procédure, on souhaite créer des sous-groupes les plus homogènes possibles de l'ensemble des individus.

Généralement, les méthodes de classification sont utilisées dans le cas de variables relevant de la même thématique ou synthétisant un ensemble d'informations. Un problème se pose dès lors que les données disponibles sont des variables relatives à différentes thématiques. Ce cas de figure est notamment celui des recensements. De par leur volonté de recueillir une information exhaustive sur une population, les recensements offrent une multitude de variables faisant référence à une pluralité de thématiques.

Si l'on souhaite dépasser le stade d'un traitement descriptif global, afin de révéler des sous-catégories de populations et utiliser les méthodes de classification dans le but de créer une typologie, on doit procéder de façon différente. On ne peut en effet raisonnablement pas traiter conjointement l'ensemble des variables issues de recensements sans prendre le risque d'aboutir à une partition de la population dont on ne pourra révéler les caractéristiques dominantes. Il est donc nécessaire auparavant de réduire le nombre de variables proposées à la classification.

La procédure de réduction des variables doit suivre une démarche cohérente. Elle va consister en le regroupement des variables par « thème » à partir desquels sera extraite une nouvelle variable synthétique grâce aux méthodes d'analyse des données (analyse factorielle et classification réalisée avec le logiciel SPAD 7). Ainsi, l'ensemble des nouvelles variables pourra faire l'objet d'une classification conduisant à une typologie de l'ensemble des individus.

2. Matériel et méthode

L'application que nous réalisons ici a pour but de dresser une typologie des entreprises ostréicoles arcachonnaises. Selon Landais (1996), deux grands types de méthodes existent pour construire une

typologie : celles basées sur des enquêtes de terrain et des entretiens, et celles résultant d'un traitement analytique et statistique d'une base de données existante. Nous avons fait le choix ici d'une analyse multivariée pour réaliser notre typologie.
Les données utilisées sont celles du premier Recensement national de la Conchyliculture de 2002. Il a permis de recueillir tout un ensemble d'informations relatives à l'activité conchylicole et aux entreprises qui la pratiquent.
Initialement, le questionnaire comportait 160 questions principales structurées autour de grands thèmes (informations relatives aux huîtres creuses, aux palourdes, à la main d'œuvre familiale…) ; la saisie des réponses a donné lieu à l'obtention d'environ 640 variables. Comme le précise Pagès (2002), les questionnaires utilisés pour les recensements sont généralement construits de telle sorte que les questions sont regroupées en thématiques, thématiques qui définissent alors autant de groupes de variables. Dans notre cas, 42 thèmes apparaissaient.
Finalement, concernant les seules entreprises arcachonnaises, il a été retenu 96 variables – plus 20 concernant la production, la commercialisation et les mortalités – parmi les 361 qui étaient relatives au Bassin d'Arcachon. Cependant, la structure thématique initiale du questionnaire n'a pas été retenue. Il apparaissait plus opportun de traiter certaines informations conjointement alors qu'initialement elles étaient dissociées car destinées à un traitement descriptif plus large. 9 thèmes ont donc été dégagés (Tableau 1).

Tableau 1 - Thèmes issus du recensement

GES (11) [1]	Gestion	EQBAT (5)	Equipement en bâtiment
INFCE (6)	Informations relatives au chef d'entreprise	EQMAT (13)	Equipement en matériel
		STRU (21)	Structure
FORMACE (4)	Formation du chef d'entreprise	SURF (8)	Surfaces
MO (22)	Main d'œuvre	CAPT (6)	Captage du naissain d'huîtres creuses

Pour réaliser la classification, le regroupement des variables au sein de thèmes n'est pas le seul travail préliminaire. Il est notamment nécessaire de procéder à une analyse factorielle des données. C'est l'objet du point suivant.

2.1. Type de variables et choix des techniques d'analyse.

Fréquemment, parmi les variables dont on dispose après un recensement, certaines sont qualitatives, d'autres quantitatives. Leur analyse conjointe doit alors faire appel à une méthode spécifique d'analyse des données, l'Analyse Factorielle de Données Mixtes (AFDM).
Cependant, cette technique est essentiellement utilisée lorsque le nombre d'individus est faible. En effet, lorsque cela est le cas, l'Analyse des Correspondances Multiples (ACM) est rendue instable et est donc délaissée au profit de l'AFDM. Selon Escofier et Pagès (2008), on considère qu'en-deçà de 100, le nombre d'individus n'est pas suffisant pour assurer la stabilité de l'ACM.
En ce qui nous concerne, nous disposons de 369 individus ; l'utilisation de l'AFDM n'est donc pas une nécessité. De plus, nous ne cherchons pas à mettre en évidence des liaisons linéaires entre les individus de notre échantillon comme dans le cas d'une Analyse en Composantes Principales (ACP) mais plutôt à observer l'existence ou non de proximités entre individus afin de réaliser une typologie de ces derniers. L'utilisation de l'ACM est donc tout à fait justifiée ici.
Pour ce faire, il est nécessaire d'effectuer sur les données quantitatives continues[2] des manipulations afin de les présenter sous forme qualitative, l'ACM ne supportant en variables actives que ce type de données. Cela permet d'homogénéiser les variables soumises ensuite aux ACM. D'après Escofier et Pagès (2008), il est d'ailleurs très courant que les variables qualitatives étudiées dans une ACM résultent d'une transformation de variables numériques. Ainsi, les variables quantitatives continues seront transformées en variables qualitatives par un codage en classes. Ce codage va néanmoins réduire l'information traitée – l'appartenance à une classe ou un intervalle étant moins précise

[1] Le nombre entre parenthèses correspond au nombre de variables composant le thème en question.
[2] Pour les variables quantitatives discrètes le codage a été effectué relativement à la valeur de l'observation.

qu'une valeur numérique – mais paradoxalement, comme l'indiquent Escofier et Pagès (2008), cela va permettre d'augmenter la richesse du résultat par la mise en évidence de liaisons non linéaires entre les variables. On va donc procéder au découpage de l'intervalle de variation des variables quantitatives en sous-intervalles définissant alors autant de modalités de la variable. Le principe qui a été retenu pour le découpage de l'intervalle de variation en classes est celui de l'amplitude[3]. Ce principe a tout de même été couplé, pour certaines variables, avec celui de l'effectif[4] notamment pour les extrémités de la distribution, afin de limiter le nombre de modalités de faibles effectifs.
Parmi les 96 variables retenues, 23 étaient des variables quantitatives continues et ont donc été recodées.

S'agissant de la méthode de classification utilisée, notre choix s'est porté sur la Classification Ascendante Hiérarchique (CAH) essentiellement parce qu'elle permet de ne pas avoir à choisir *a priori* des centres de classes provisoires ni le nombre de classes. De plus, la CAH révèle les « vraies » classes si elles existent puisqu'elles sont déterminées automatiquement et ne dépendent pas d'un choix *a priori* comme dans le cas de la méthode de Classification autour des Centres Mobiles (CCM).

2.2. ACM et CAH préliminaires des différents thèmes.

Pour chaque ACM un seuil d'apurement a été choisi. L'apurement permet de s'affranchir, artificiellement, des modalités de faibles effectifs qui n'ont pas pu être évitées précédemment et qui pourraient avoir des effets perturbateurs sur l'analyse. Par exemple, pour le thème *Gestion*, un seuil de 0,5% a été choisi, ventilant les modalités à effectif unique. Pour gérer le problème des modalités de faibles effectifs il existe des techniques de traitement dont quelques unes ont fait l'objet de deux publications de Benali (1986) et Benali et Escofier (1987). Pour notre part, nous avons conservé la technique du seuil d'apurement présente par défaut dans SPAD, et nous avons fait varier le seuil d'apurement en fonction des modalités que nous voulions conserver. C'est pour cela d'ailleurs que tous les thèmes analysés ne l'ont pas été sur la base d'un même seuil d'apurement. Dans le cas où des modalités ont été ventilées, il est tout même prudent de vérifier si les coordonnées des modalités actives sont proches de celles de ces mêmes modalités placées en illustratif à partir des données non-apurées. Si ces coordonnées ont des valeurs différentes, cela signifie que la ventilation a notablement affecté la répartition des individus.
Quant aux CAH, la méthode d'agrégation utilisée est le critère de Ward de minimisation de la variance intra-classes.
On obtient au final 9 classifications, une pour chaque thème, avec des partitions de 3 à 7 classes. Le Tableau 2 est un extrait des résultats des CAH (pour les thèmes *Gestion* et *Informations relatives au chef d'entreprise*) déduits de l'analyse des plans factoriels[5] des modalités actives et des classes de la partition.

Tableau 2 - Extrait du tableau des résultats des CAH des différents thèmes

GES	ges1 (3)[6]	Entreprises de très grande taille, non spécifiques
	ges2 (224)	Entreprises relativement grandes, relativement spécifiques
	ges3 (5)	Petites entreprises, peu spécifiques
	ges4 (137)	Très petites entreprises, relativement spécifiques
INFCE	inf1 (325)	Entrepreneur conchylicole quadragénaire travaillant à temps plein, marié
	inf2 (16)	Entrepreneure conchylicole quadragénaire travaillant à temps plein, mariée
	inf3 (28)	Retraité quinquagénaire travaillant moins de 30 heures hebdomadaire, marié

[3] On découpe l'intervalle de variation de sorte que chaque classe ait la même amplitude.
[4] On découpe l'intervalle de variation afin d'obtenir des classes de même effectif.
[5] Ces plans ne sont pas présentés ici. Pour un exemple de plans factoriels, voir Figure 1 et Figure 2.
[6] Le nombre entre parenthèses correspond au nombre d'entreprises de chaque classe.

3. Résultat

Nous avons donc créé une nouvelle base de données à partir des 9 CAH. Les nouvelles variables correspondent aux thèmes et leurs modalités, aux classes mises en évidence. Chaque variable synthétise donc l'information d'un thème. Nous introduisons par ailleurs une variable supplémentaire – la production apparente[7] (*pa*) – qui sera intégrée en tant que variable illustrative. Elle servira en quelque sorte à valider la démarche initiale. Cette variable était au départ quantitative et nous avons fait le choix de la recoder en variable qualitative, même s'il était possible de l'intégrer en variable illustrative continue. En dehors des extrémités de la distribution de la production apparente, le découpage de l'intervalle en 7 modalités a été fait en conservant la même amplitude.
Après avoir procédé à une ACM et une CAH sur cette nouvelle base de données, on obtient les plans factoriels suivants.

Figure 1 - Modalités actives de l'ACM globale

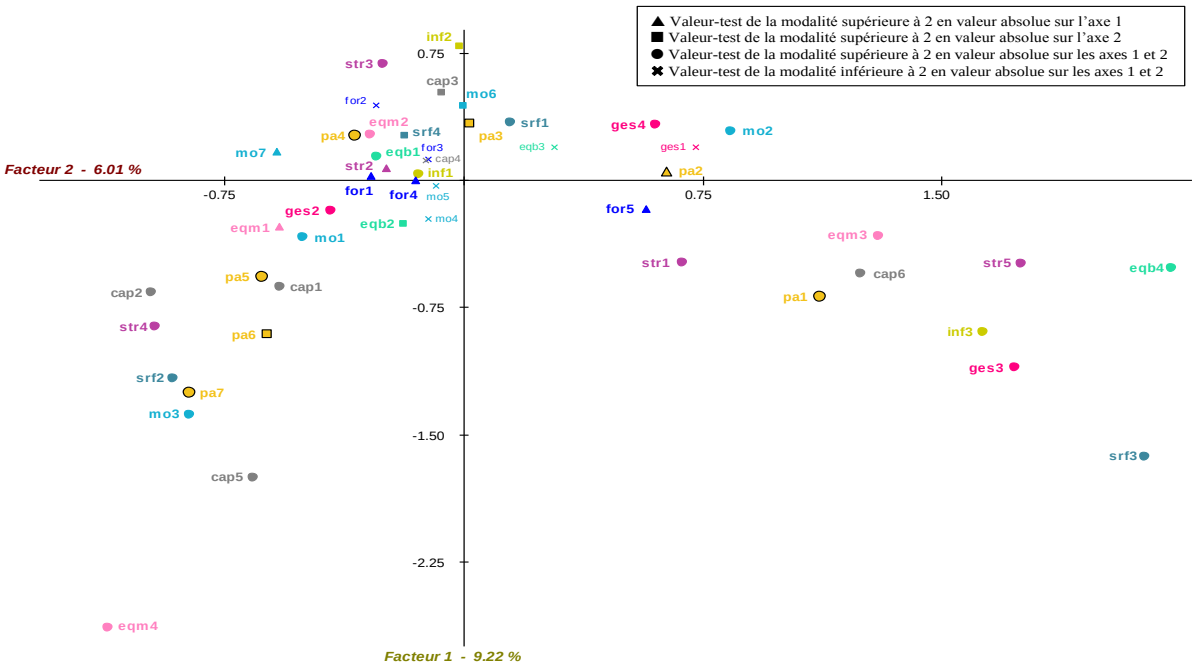
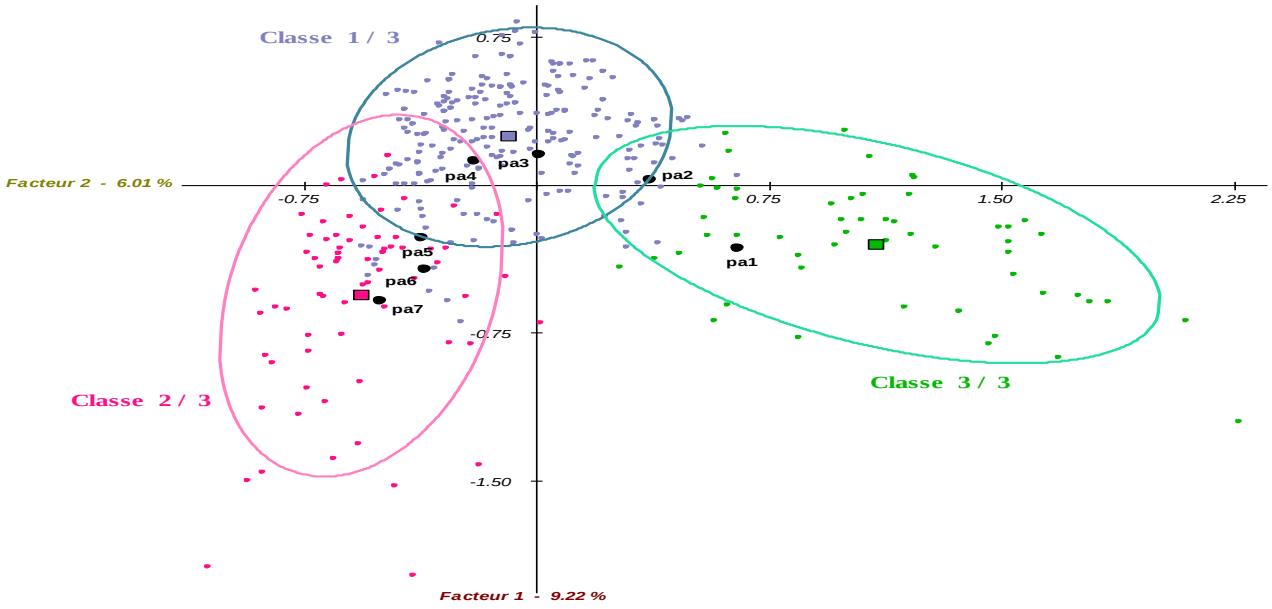


Figure 2 - CAH globale



[7] La production apparente permet en fait d'approximer la production « réelle ». Elle se calcule en retranchant les achats de coquillages sur une période aux ventes de coquillages sur la même période.

On a pu mettre en évidence trois classes qui sont assez bien marquées même s'il existe des intersections entre les ellipses de concentration. Par ailleurs, on observe également un effet Guttman. Selon Lebart et *al.*(2000), cette situation est due à une redondance de deux variables et toute l'information est quasiment donnée par le premier facteur. On aboutit alors à la typologie des entreprises ostréicoles arcachonnaises sur la base essentiellement de 8 des 9 thèmes[8].

Tableau 3 – Typologie des entreprises ostréicoles arcachonnaises

	Type 1 (240)	**Type 2 (73)**	**Type 3 (56)**
Gestion	X[9]	Grande taille et spécifique	Petite taille et spécifique
Informations chef d'entreprise	40-50 ans, marié, entrepreneur conchylicole, à temps plein	X	50-60 ans, marié, retraité, travaillant moins de 30h/sem.
Main d'œuvre	X	Abondante	Peu ou pas en dehors du chef d'entreprise
Equipement bâtiment	Bien équipées	X	Peu équipées
Equipement matériel	Peu équipées mais bien dotées en véhicules	Très bien équipées mais peu en véhicules	Peu équipées
Structure	Diversifiée	Très grandes moyennement diversifiées	Petites, plus ou moins diversifiées
Surface	Taille moyenne, DPM[10]	Très grande taille, DPM	Petite taille, DPM et Dp[11]
Captage	Nombre de collecteurs moyen	Nombre de collecteurs important	Peu ou pas de captage
Production apparente	10 à 30 tonnes	Plus de 30 tonnes	Moins de 10 tonnes

Cette typologie conduit au classement des entreprises en trois ensembles hiérarchisés :

Type 2 > Type 1 > Type 3

Les effectifs de chacun des types nous permettent de dire que les types 2 et 3 sont plus ou moins des cas particuliers d'entreprises arcachonnaises, la majorité des entreprises étant de type 1. Ce classement se retrouve par ailleurs au niveau des variables qui étaient initialement quantitatives. Nous avons regroupé ces principales variables quantitatives dans le Tableau 4. Les chiffres présentés – à l'exception de l'effectif et de la part des entreprises de chaque type dans l'effectif total – correspondent à une moyenne des entreprises de chaque type.

Tableau 4 - Données réelles pour chaque type d'entreprises

	Type 1	**Type 2**	**Type 3**
Effectif	240	73	56
Part dans l'effectif total	65,04%	19,78%	15,18%
Age du chef d'entreprise	43,7 ans	42,8 ans	52,6 ans
Nombre d'établissements dans l'entreprise	1,05	1,054	1,035
Volume moyen des bassins tout béton ou PVC (en m^3)	80,80	130,16	31,07
Nombre de bateaux dans l'entreprise	1,42	1,73	1,09
Nombre de parcelles disponibles (DPM et Dp)	3,39	4,60	2,64
Surface disponible (ares)	180,95	525,62	129,48
Nombre de collecteurs utilisé pour le captage	16 845	63 314	10 078
Nombre d'équivalents temps plein	2,10	3,77	1,23
Production apparente totale (en kg)	17 560	39 873	9240

4. Conclusion

D'une base de données destinée initialement à décrire une population, il a pu être déduit une typologie des entreprises ostréicoles arcachonnaises. Le fait de regrouper les variables en thèmes, de traiter les thèmes un à un puis de réaliser une nouvelle CAH à partir des classifications réalisées sur les thèmes n'a pas eu pour effet d'approximer une réalité mais bien de la mettre en évidence. Et cela a été confirmé par les données réelles, c'est-à-dire celles qui ont été recodées en variables

[8] Le dernier thème n'est pas suffisamment caractéristique (modalités dont la valeur-test est inférieure à 2 en valeur absolue).
[9] Variables dont les modalités ont des valeurs-test inférieures à 2 en valeur absolue.
[10] Domaine Public Maritime
[11] Domaine privé

qualitatives.
Loin de perdre en information du fait du traitement des variables quantitatives en variables qualitatives et de l'utilisation d'une ACM, cela nous a permis de mettre en évidence des faits qui initialement, même s'ils étaient supposés, n'apparaissaient pas clairement. Cela a permis en outre de regrouper de façon optimale des individus qui n'avaient comme ressemblance commune, que le fait d'être une entreprise ostréicole.
Grâce à cette typologie, les entreprises vont pouvoir être étudiées de façon encore plus précise afin de réaliser un profil de performance de ces dernières et mettre en évidence ce qui, dans une activité telle que celle concernée, va être déterminant dans la productivité.

Bibliographie

[1] Benali, H. (1986). Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs. *Rapports de Recherche n°528, Publication Interne de l'IRISA*, 16 p.

[2] Benali, H., & Escofier, B. (1987). Stabilité de l'analyse factorielle des correspondances multiples en cas de données manquantes et de modalités à faibles effectifs. *Revue de Statistique Appliquée , tome 35* (1), pp. 41-51.

[3] Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology , Vol. 3* (3), pp. 166-185.

[4] Escofier, B., & Pagès, J. (2008). *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation* (éd. 4ème). Dunod.

[5] Girard, S., Pérez Agúndez, J. A., Miossec, L., & Czerwinski, N. (2005). Recensement de la conchyliculture 2001. *Agreste Cahiers n°1* , Ministère de l'Agriculture, de l'Alimentation, de la Pêche et de la Ruralité, Paris, 89 p.

[6] Gordon, A. (1987). A Review of Hierarchical Classification. *Journal of the Royal Statistical Society. Series A , Vol. 150* (n°2), pp. 119-137.

[7] Landais, E. (1996). Typologie d'exploitations agricoles. Nouvelles questions, nouvelles méthodes. *Economie Rurale* (236), pp. 3-15.

[8] Lebart, L., Morineau, A., & Piron, M. (2000). *Statistique exploratoire multidimensionnelle* (éd. 3ème). Dunod.

[9] Pagès, J. (2002). Analyse factorielle multiple appliquée aux variables qualitatives et aux données mixtes. *Revue de Statistique Appliquée , tome 50* (4), pp. 5-37.