Création de typologie sous SPSS

À Propos de ce document

Ce document a été créé dans le but d'aider toute personne qui débute dans SPSS, logiciel trè puissant mais très peu sympathique.

Ce document se base sur la version 11.0 Base de SPSS, en version anglaise. La plupart des exemples sont issus des dictatiels du programme SPSS en lui-même.

Toutes les remarques, tant sur le fond que sur la forme, sont les bienvenues. N'hésitez pas à me contacter à l'adresse suivante : <lemoal@lemoal.org> ou à venir visiter mon site internet : http://www.lemoal.org/spss/

Merci.

Introduction

Les méthodes de classification sont très utilisées en marketing. Ce sont notamment grâce à elles qu'une entreprise peut segmenter son marché, selon des critères quantitatifs.

Deux types de classification sont possibles : la « Nuées dynamiques (K-Means Cluster Analysis) » ou la « classification hiérarchique (Hierarchical Cluster Analysis) ».

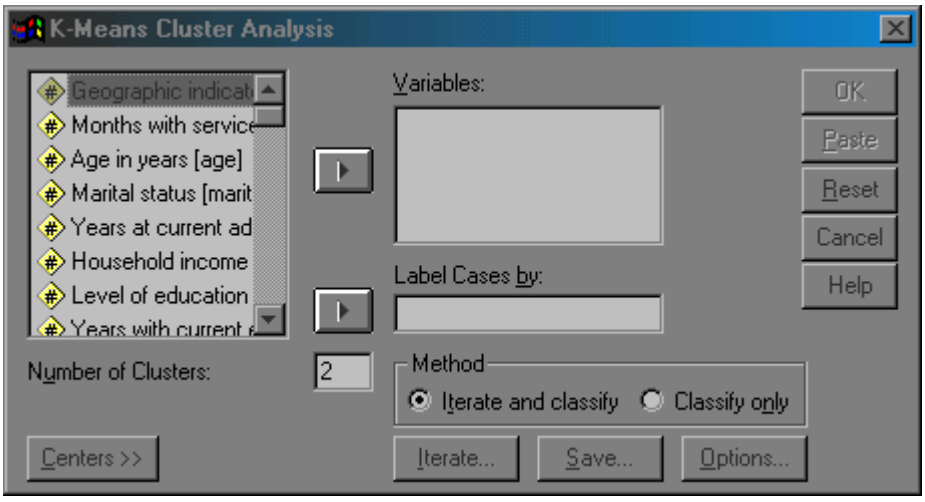
Si le nombre d'observations est supérieure à 100, il est recommandé d'utiliser les nuées dynamiques. C'est d'ailleurs la plus couramment utilisé en marketing et celle que nous étudierons ici. La classification hiérarchique est trop longue au delà de 100 individus (et plus exigeante en terme de mémoire pour le PC) et ne sera pas traitée ici.

Important :

- Il faut noter que la classification en nuées dynamiques nécessitent des **données quantitatives**. Si vous possédez des données qualitatives, l'analyse ne sera pas possible, à moins de passer par une Analyse en Composante Multiple (ACM).
- Il est recommandé d'utiliser des données centrées et réduites pour l'analyse.

La démarche à suivre sous SPSS

Aller dans Analyse > Classify > K-Means Clusters . La boîte de dialogue suivante apparaît alors :



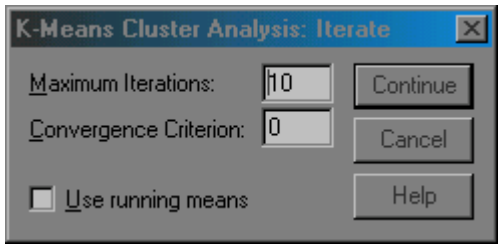
On choisit les variables qui nous paraissent les mieux adaptées à la création de typologies en les sélectionnant dans la partie de droite puis en cliquant sur la flèche qui pointe vers la droite.

Dans « Number of Clusters » (nombre de classes), indiquer le nombre de classes qu'on a à priori repéré dans l'analyse multi-variée (ACP, AFC ou AFCM) qui aura précédé. Il est recommandé de rajouter une classe supplémentaire, au cas où (quitte à en enlever une par la suite). Au niveau de la méthode, il convient de choisir « Iterate and classify » (Itérer et classer)

Plusieurs options sont maintenant possibles : 1. Iterate… 2. Save… 3. Options…

1. « Iterate… »

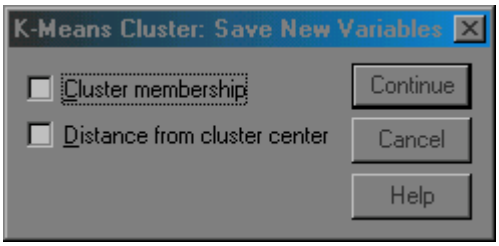
Cliquer sur « Itérate… ». La boîte de dialogue « K-Means Cluster Analysis : Iterate » apparaît alors.



Cette boîte de dialogue sert à indiquer le nombre maximum d'itération. Au départ, il ne faut toucher à rien et laisser la valeur par défaut (10). Si, lors de l'anlyse des résultats, le nombre d'itérations s'avert insuffisant, c'est ici qu'il faudra changer la valeur.

2. « Save… »

Cliquer sur « Save… ». La boîte de dialogue « K-Means Cluster : Save New Variables » apparaît alors.



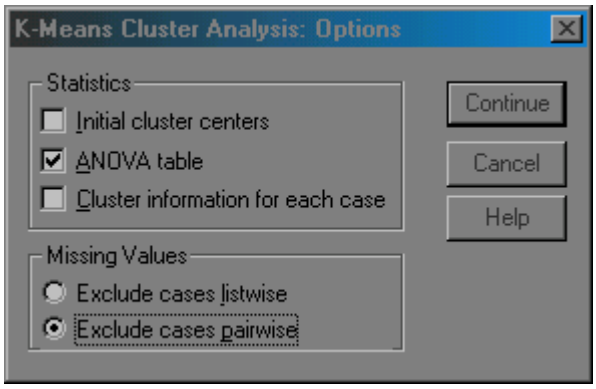
Deux cases à cocher sont possibles :

- Classe d'affectation (c'est-à-dire le groupe) : si on veut donner à chaque individu sa classe d'affectation. Il faut le faire une fois qu'on est satisfait du résultat obtenu mais pas avant.
- Distance au centre de la classe : pour mesurer la distance du centre…

Pour une première analyse, il n'est pas utile de cocher ces options.

3. « Options… »

Cliquer sur « Options… ». La boîte de dialogue « K-Means Cluster Analysis : Options » apparaît alors.



Plusieurs changements sont à opérer :

- Dans « Statistics », cliquer sur « ANOVA Table ». Cela sert à déterminer quelles sont les variables les plus discriminantes dans la constitution des groupes et ne pas conserver « Centres de classe initiaux »
- Dans « Massing Values » (Valeurs manquantes), choisir « Exclude cases pairwises » (exclure seulement les classes non valides).

Analyse des résultats

L'analyse des résultats commence par valider l'analyse en elle-même. Cette première phase passe par l'observation de l'historique des itérations et du nombre d'observations dans chaque classe. L'analyse en elle-même peut ensuite se poursuivre.

1. Historique des itérations :

Dans la plupart des cas, les classes convergent avant la dixième itération. Il n'est donc pas nécessaire de recommencer l'analyse.

Par contre, dans le cas suivant, le nombre d'itérations initiales est trop faible (10). Aucune classe ne converge. Il y a convergence quand ,000 est atteint dans chacune des classes identifiées. Dans le cas présent, il faut donc recommencer l'analyse avec un nombre de classe plus importante.

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	3,298	3,590	3,491
2	1,016	,427	,931
3	,577	,320	,420
4	,240	,180	,195
5	,119	,125	,108
6	9,282E-02	8,262E-02	2,654E-02
7	6,882E-02	9,375E-02	3,196E-02
8	5,858E-02	5,080E-02	1,817E-02
9	3,461E-02	8,501E-02	6,318E-02
10	2,489E-02	,359	,333

a. Iterations stopped because the maximum number of iterations was performed. Iterations failed to converge. The maximum distance by which any center has changed is ,188. The current iteration is 10. The minimum distance between initial centers is 6,611.

2. Nombre d'observations dans chaque classe :

Number of Cases in each Cluster

Cluster	1	232,000
	2	288,000
	3	480,000
Valid		1000,000
Missing		,000

Il est recommandé de ne garder que les classes qui représentent 10% ou plus des observations. Dans le cas présent, chaque classe représente plus de 10% des personnes interrogées. Il n'y a pas lieu de recommencer l'analyse.

Si par exemple, le groupe 1 n'aurait eu que 96 individus, l'analyse aurait dû être recommencée avec un groupe de moins, c'est-à-dire 2.

Cette méthode permet également de quantifier chaque segment.

3. Analyse de la variance

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Standardized log-long distance	16,843	2	,968	997	17,395	,000
Standardized log-toll free	45,470	2	,812	472	56,027	,000
Standardized log-equipment	103,643	2	,464	383	223,367	,000
Standardized log-calling card	5,726	2	,986	675	5,808	,003
Standardized log-wireless	52,747	2	,647	293	81,554	,000
Standardized multiple lines	41,641	2	,918	997	45,337	,000
Standardized voice mail	249,971	2	,501	997	499,383	,000
Standardized paging	295,683	2	,409	997	723,187	,000
Standardized internet	122,869	2	,756	997	162,626	,000
Standardized call waiting	282,576	2	,435	997	649,371	,000
Standardized call forwarding	303,329	2	,394	997	770,805	,000
Standardized 3-way calling	282,978	2	,434	997	651,501	,000
Standardized electronic billing	111,455	2	,778	997	143,180	,000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Le F sert à identifier les variables qui sont utiles pour l'identification des différents segments. Attention à ne pas interpréter la signification de F qui, le cas présent, n'est pas très utile.

Les variables avec les plus grands F sont les variables les plus discriminantes des groupes entre eux.

Dans notre exemple, les variables les plus discriminantes sont les variables « Standardized call forwarding » et « Standardized paging ».

4. Centre de classes finaux

Final Cluster Centers

	Cluster		
	1	2	3
Standardized log-long distance	,06	,25	-,18
Standardized log-toll free	,23	,13	-1,07
Standardized log-equipment	,79	-,08	-,76
Standardized log-calling card	,14	,05	-,17
Standardized log-wireless	,40	-,68	-1,20
Standardized multiple lines	,52	-,23	-,11
Standardized voice mail	1,28	-,27	-,46
Standardized paging	1,40	-,36	-,46
Standardized internet	,82	-,55	-,06
Standardized call waiting	,72	,72	-,78
Standardized call forwarding	,76	,74	-,81
Standardized 3-way calling	,69	,75	-,78
Standardized electronic billing	,72	-,60	,01

La lecture des centres de classes finaux permet de donner une signification aux différents groupes déterminés.

L'analyse en elle-même se passe comme pour une analyse multivariée, c'est-à-dire par recherche lexicale à partir des opposés. Par exemple, la classe 1 se caractérise par les variables « Standardized paging » et « Standardized voice mail ».

Les méthodes de classification peuvent donner des résultats très différents suivants les variables utilisées ou les méthodes utilisées. Pour s'assurer de résultats pertinents, il convient de tester plusieurs typologies. pour choisir la meilleure, il convient d'effectuer une analyse discriminante pour chaque typologie créée et ne retenir que celle qui possède le meilleur pouvoir de reclassement. Pour cela, et pour chaque typologie retenue, il faut enregistrer les classes d'affectation pour chaque individu, en recommençant l'analyse et en cochant « Classes d'affectation » dans « Enregistrer… »