Typologies

Sommaire

- Les phénomènes observés (attitudes, comportements, préférences) ne sont souvent que la itudes parfois très différents.
- -
  la problématique de la segmentation.
  « marketing personnalisé » (one-to-one).
- Ces variables vont jouer le rôle de variables modératrices et influencer (1) le niveau de la moyenne, (2) la sensibilité du comportement à une autre variable.

-
  doivent être
  Pertinents sur le phénomène observé : (1) *homogènes* au sein du groupe, (2) *différents* entre les groupes
  Opérationnels : (3) *identifiables* pour pouvoir et (4) *taille suffisante* pour que cette personnalisation soit rentable

- Les variables de segmentation sont
  Objectives, liés à l'individu, ce sont les déterminants économiques (revenu, richesse, propriété du logement), démographiques (sexe, âge-génération, composition du foyer, présence enfants) ou socio-culturels (pays, région, urbanité) classiques qui dirigent les

  Subjectives,
  pour expliquer les comportements :
  - **liés à l'individu**
  - **liés au phénomène étudié** : implication (dans un achat, une catégorie de

    relation, attachement/fidélité à la

  Comportementales, avec la fréquence des transactions (Frequency), la durée de la relation (première et dernière transaction) (Recency), le montant du panier (Monetary), eprise

- La méthode de segmentation détermine comment la sélection des variables constituant les groupes va être faite
  Segmentation« a priori » : les variables déterminantes sont connues et présélectionnées (variables objectives). On recherche un lien direct entre ces variables (X) et une variable de comportement (Y) dans une démarche *explicative* de sa variance
  : Typologie monothétique ou segmentation par arbre.

Segmentation« a posteriori » : la variable qui va servir de variable de segmentation est construite à partir des informations collectées. Les groupes sont constitués pour restituer la diversité (variance) *sur cette variable* ance à un typologie polythétique ou simplement typologie (nodale, Hiérarchique)

Segmentation « concomitante » on cherche simultanément à obtenir une bonne qualité explicative du comportement (Y) et une bonne cohérence (variance de X) des groupes ainsi constitués.

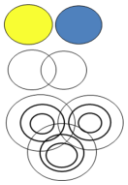
-

Bien choisir les variables, leur distribution et les pré-traitements (valeurs aberrantes, standardisation), leurs relations (corrélation, ACP), leurs transformations (variable nominale en variable quantitative par AFACO).

Choisir une fonction de distance : Comment définir « proches » ? (distances, similarité, dissimilarité)

Choisir la nature de la typologie

- Appartenance exclusive simple (0001)
- Appartenance multiple (0101)
- Appartenance probabiliste (floue) (0.5 0.1 0.2 0.2)



Choisir la méthode :

- hiérarchique)
- Solution « spatiale miné de groupes (typologie nodale)
- Approche séquentielle : hiérarchique sur un sous-échantillon puis nodale.
- Approche itérative ou par simulation :
  - quels points se retrouvent toujours ensembles (nuées dynamiques et formes fortes) ?
  - quels points sont à neutraliser ?
    :
- comment décider de regrouper deux segments ? (hiérarchique)
- quelle solution de départ (seed) ? (nodale)

Combien de groupes faut-il retenir ?

- Les critères internes (sur la variance des variables X) : Maximiser la variance inter-groupes / minimiser la variance intra-groupe, Pseudo R2, Pseudo F (pseudo car la même information est utilisée pour constituer les groupes et mesurer la qualité de la typologie
- Les critères externes : capacité à expliquer une variable comportementale,
- Souvent entre 3 (intérêt de la segmentation) et 8 (capacité à gérer les groupes).

Difficultés : faiblesse du cadre théorique

- Est- ?
- Pas de critères explicites pour guider ces différents choix
- Il existe toujours plusieurs solutions (! même avec le même algorithme, par exemple en fonction des valeurs de départ)

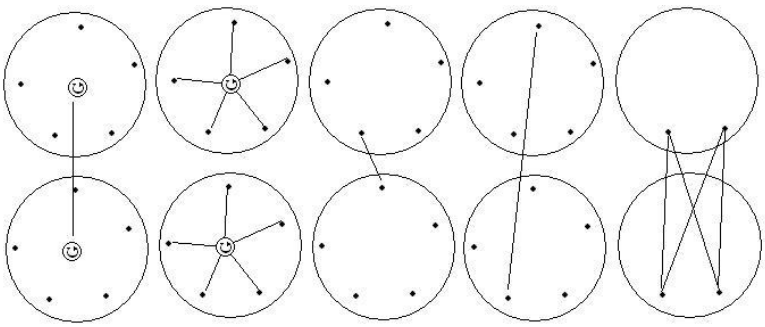
Proc CLUSTER : Typologie hiérarchique

Caractéristiques :
- Peu adapté à de gros volumes de données
- Choix délicat entre les nombreuses méthodes donnant des résultats différents
- Choisir la méthode de Ward si les données sont quantitatives (elle donne des groupes sphériques)
- je suis beau »), des pré-traitements des données errants)
- Statistiques sur les variances inter et intra cluster, et pseudo Anova (pseudo F) car la variable est un découpage construit a posteriori qui maximise la variance inter-groupes

Principe :
- Approche séquentielle offrant un arbre typologique (dendrogramme, tree) de 1 seul groupe à n

  Ascendante : partir de n et regrouper progressivement
  Descendante : partir de 1 groupe et décomposer progressivement.
- où couper oisir le nombre de groupes). Le dendrogramme reporte les distances inter-groupes pour chaque fusion/scission des clusters (R^2 semi-partiel)
- Il les plus pertinents (faible section) et ceux qui « coûtent » le plus en termes
- (ne pas le demander si les individus sont trop nombreux !)

Choix du regroupement
- Source : http://v8doc.sas.com/sashtml/stat/chap23/sect4.htm
- (1CENTROID) Méthode des centroïdes : Regroupement selon la proximité des centres de gravité (centroïdes)
- (2 WARD) Méthode de la variance (Ward) : Minimisation de la variance intra-groupe
- (3 à 5 SINGLE, AVERAGE, COMPLETE) Méthode de chaînage par le plus proche voisin : distances entre les points de chaque groupe : simple (le plus court, le plus long), moyen, complet,



Options usuelles

- create output data set	OUTTREE=
- Copy variables in data set to the OUTTREE=	COPY
- specify clustering method	METHOD=
- suppress normalizing of distances	NONORM
- suppress squaring of distances	NOSQUARE
- standardize variables	STANDARD
- omit points with low probability densities	TRIM=
- cubic clustering criterion	CCC

- pseudo F and t2 statistics PSEUDO
- statistiques simples (my, écart-type,..) SIMPLE
- root-mean-square standard deviation RMSSTD
- R2 and semipartial R2 RSQUARE

Choisir le nombre de clusters

- Général : Les critères ne sont valides que pour des clusters « compacts » (non étirés). Il faut rechercher un consensus entre ces statistiques : une forte valeur pour CCC et pseudo F combinée avec un accroissement important du pseudo t² pour un découpage supplémentaire.
- Les critères globaux reconstitution de la variance totale

  RMSSTD (root-mean-square standard deviations) écart-type

  RSQ (R^2) : % de variance représenté par les clusters (intra) =1-Σwk/VT. A maximiser mais sans trop de groupes

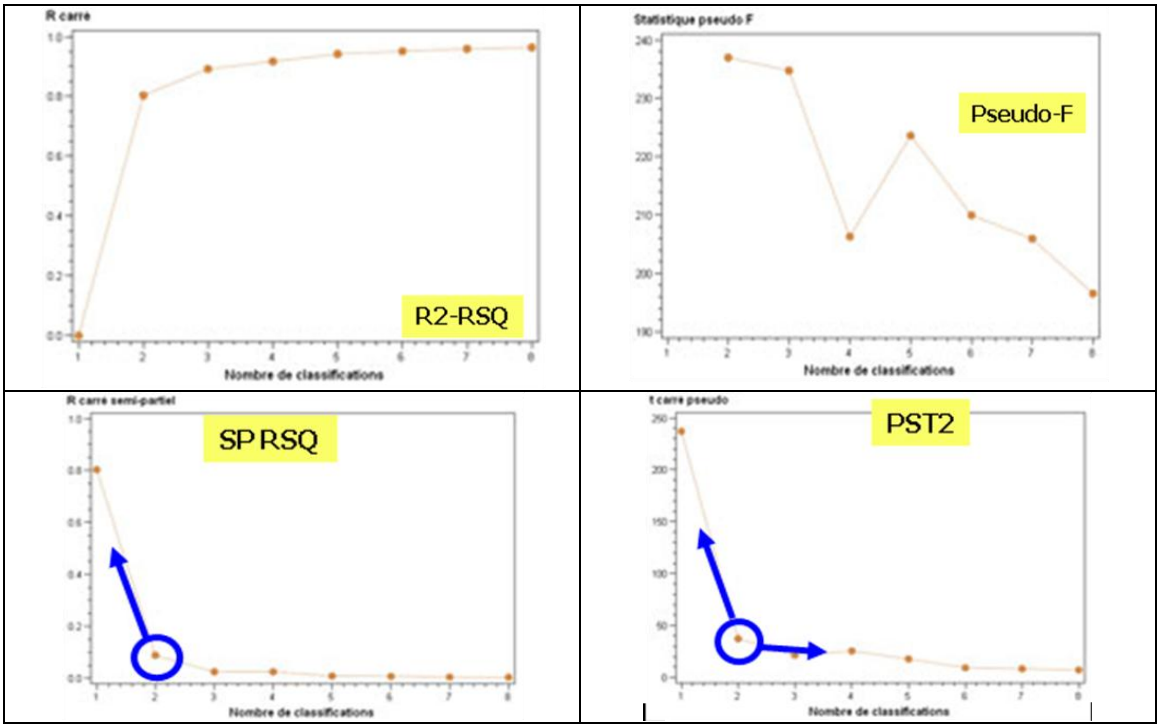
  PSF (Pseudo F) écart entre tous les groupes = {(T-Σwk) / (G-1) } / {(Σwk) / (n-G)}. A maximiser

-

  commence le « plat » juste après une chute

  SPRSQ (Semipartial R-Squared) : % de variance expliquée gagnée par le regroupement de 2 clusters. Choisir juste avant une forte hausse

  PST2 (Pseudo t2) écart entre les deux derniers clusters regroupés (marginal). Le meilleur découpage a une forte différence de pente avant (petite) et après (forte)



- CCC (Cubic Clustering Criterion) http://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf

  distribution

  - H0 : distribution uniforme
  - s multivariées normales sphériques aux variances égales

  efficacité des différents niveaux de regroupement

  - > à 2 ou 3 = OK; 0-2 à valider ;

- largement <0 (-30) : problème avec des individus très atypiques à retirer

Interprétation en fonction du nombre de clusters

- Si clusters ex
- Si CCC<0 et en baisse pour >2 groupes, distribution unimodale ou asymétrique
- Si CCC augmente toujours, les données sont granulaires ou manquent de finesse (arrondies)

Root-Mean-Square Total-Sample Standard Deviation = 10.89784

Root-Mean-Square Distance Between Observations = 30.82375

Historique des classifications

NCL	Classifications jointes		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
8	CL16	CL24	8	0.0037	.964	.944	3.50	197	7.2	
7	CL15	CL18	13	0.0047	.959	.936	3.65	206	8.4	
6	CL9	CL8	18	0.0078	.951	.926	3.54	210	9.3	
5	CL13	CL11	20	0.0090	.942	.911	3.71	224	17.8	
4	CL7	CL12	16	0.0250	.917	.888	1.87	206	25.4	
3	CL6	CL10	24	0.0253	.892	.842	2.55	235	21.4	
2	CL4	CL3	40	0.0884	.803	.709	3.23	237	37.3	
1	CL5	CL2	60	0.8034	.000	.000	0.00	.	237	

Master Marketing Paris-Dauphine

Source : Analyse des données appliquée au marketing

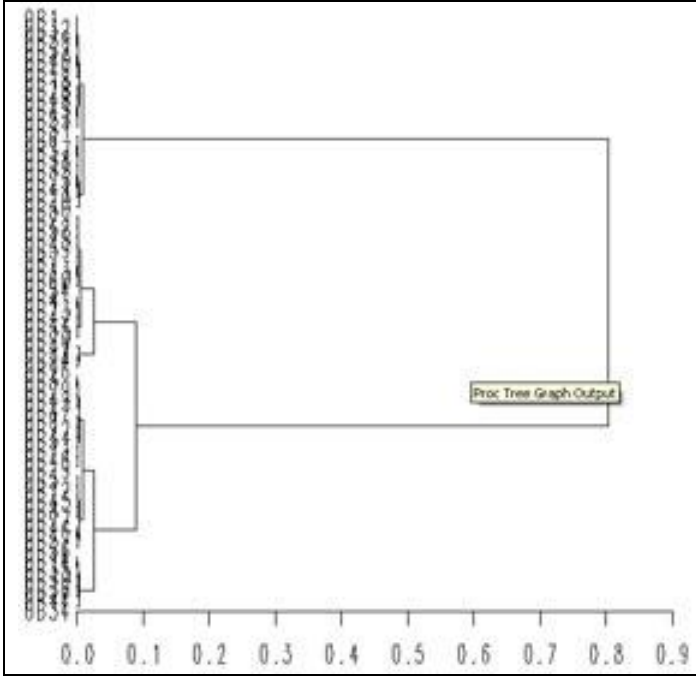
Choisir 3 groupes

car passer à 4

- ne contribue que peu à faible)
-
- fait beaucoup baisser le CCC
- et le PSF

Passer à 2

- donne un meilleur CCC
- ajustement global (détériore le RSQ)



Exemple 1 : distances intercités USA (entrée directe des distances)

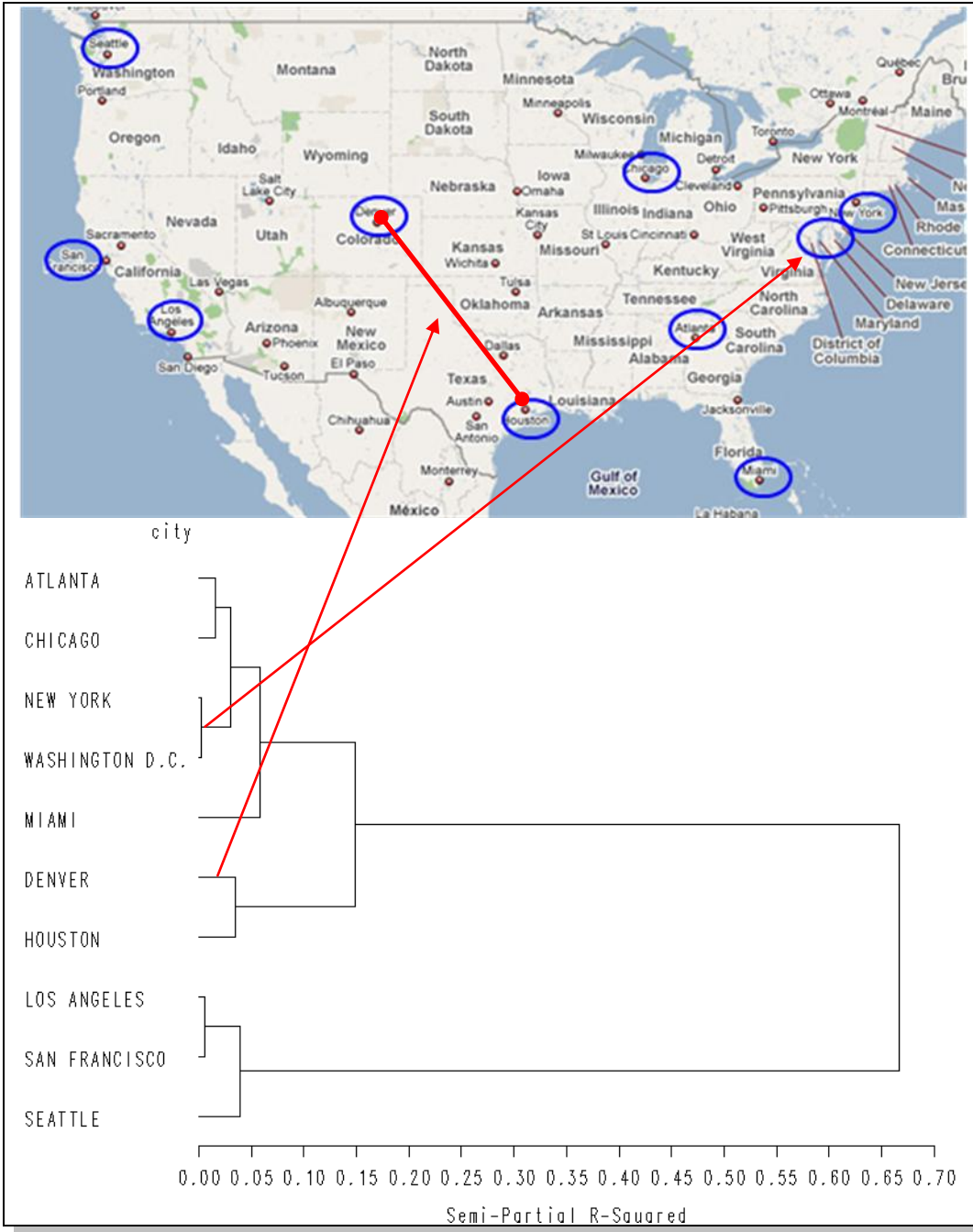
```
                                                      /.
Data                    '   -(                                              ( -

                                                    '  5.'
             55       "/3,

    .
  365    .
 /0/0   70.    .
  5./   72.  657    .
 /714 /523  61/ /152    .
  4.2 //66 /504  746 0117    .
  526  5/1 /41/ /20. 023/ /.70    .
 0/17 /636  727 /423  125 0372 035/    .
 0/60 /515 /.0/ /67/  737 0512 02.6  456    .
  321  375 /272 /00. 01..  701  0.3 0220 0107    .             , ,

-(                                                                ( -

Proc Cluster

                              -(              ( -
                              -(                                        ( -
                              -(              ( -

run
Proc Tree
                              2

    ;
```

Résultats

Historique des classifications

NCL	Classifications jointes		FREQ	SPRSQ	RSQ	PSF	PST2	Ti
9	NEWYORK	WASHINGTON D.C.	2	0.0019	.998	66.7	.	

- Premier regroupement NY et W DC : il apporte 0.0019 en explication de la variance totale. Il reste donc (RSQ = 1-
  intéressant.
- A la fin le pseudo t² indique que le dernier regroupement est peu pertinent (bruque augmentation du t² partiel.


city
ATLANTA
CHICAGO
NEW YORK
WASHINGTON D.C.
MIAMI
DENVER
HOUSTON
LOS ANGELES
SAN FRANCISCO
SEATTLE
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70
Semi-Partial R-Squared

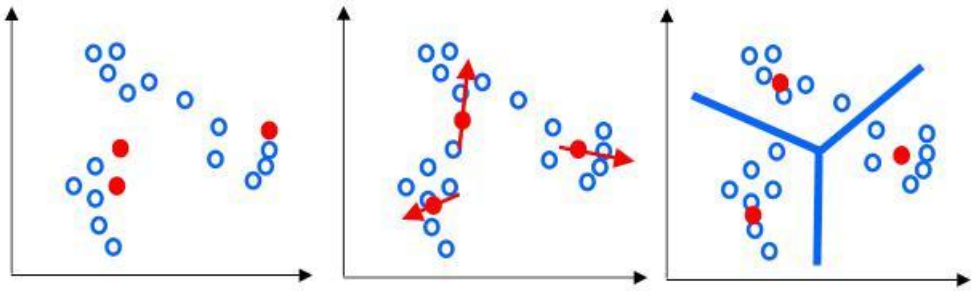
Proc FASTCLUS : Typologie nodale

Caractéristiques :

- Méthode des « K-means »
- Adaptée à de très gros volumes de données (-ne stocke pas la matrice des distances)
- Apporte une solution unique en fonctiond es choix initiaux donc
  - tester différents nombre de clusters pour choisir le meilleur
  - tester différentes valeurs de départ
  - ou donner des valeurs initiales issues de Proc cluster

Principe :

- Choix initial des paramètres et des « points de départ »
  - un nombre de groupes et
  - un seuil de regroupement (S)
  - les points de départ des groupes (seed)
  - le calcul des centres de gravité en lot ou immédiat (Drift)
- uantité de déplacement :DAB = 2.dAB . (NA.NB/(NA+NB)
  - si DAB <= Seuil regroupement de A et B
  - sinon nouveau cluster ou individu « non classé »
- Le calcul du centre de gravité peut être effectué
  - après chaque affectation
  - après le traitement du lot (batch) p



Options usuelles:

- SEED=SAS-data-set specifies an input data set from which initial cluster seeds are to be selected.
- REPLACE=RANDOM : selects a simple pseudo-random sample of complete observations as initial cluster seeds.
- OUT=SAS-data-set creates an output data set to contain all the original data, plus the new variables CLUSTER and DISTANCE
- OUTSEED=SAS-data-set is another name for the MEAN= data set, provided because the data set may contain location estimates other than means.
- OUTSTAT=SAS-data-set creates an output data set to contain various statistics
- INSTAT=SAS-data-set : reads a SAS data set previously created by the FASTCLUS procedure using the OUTSTAT= option. Only cluster assignment and imputation are performed as an OUT= data set is created.
- DISTANCE computes distances between the cluster means.
- DRIFT : les centres de classes sont ajustés après chaque observation
-

  un « outlier »)

- MAXITER=n specifies the maximum number of iterations for recomputing cluster seeds.

Exemple 2

```
     /   2
     0   1
     1   (

     2  /
        /   0.5
        0   0.5
 0./0   '
data
              /  2              (
                                 (
                        '


 3. 11 /2 .0 / 42 06 34 00 1 43 06 24 /3 0
 45 1/ 34 02 1 41 06 3/ /3 1 24 12 /2 .1 /
 47 1/ 3/ 01 1 40 00 23 /3 0 37 10 26 /6 0
 24 14 /. .0 / 4/ 1. 24 /2 0 4. 05 3/ /4 0
 43 1. 30 0. 1 34 03 17 // 0 41 1. 33 /6 1
 36 05 3/ /7 1 46 10 37 01 1 3/ 11 /5 .3 /
 35 06 23 /1 0 40 12 32 01 1 55 16 45 00 1
 41 11 25 /4 0 45 11 35 03 1 54 1. 44 0/ 1
 27 1/ /3 .0 / 55 04 47 01 1 4. 00 3. /3 1
 32 17 /5 .2 / 44 07 24 /1 0 30 05 17 /2 0
 4. 12 23 /4 0 3. 12 /3 .0 / 22 07 /2 .0 /
 3. 0. 13 /. 0 33 02 15 /. 0 36 05 17 /0 0
 25 10 /1 .0 / 24 1/ /3 .0 / 3/ 12 /3 .0 /
 3. 13 /1 .1 / 27 1/ /3 ./ / 45 1/ 25 /3 0
 32 15 /3 .0 / 34 1. 2/ /1 0 41 03 27 /3 0
 4/ 06 25 /0 0 42 07 21 /1 0 32 17 /1 .2 /
 3/ 13 /2 .1 / 23 01 /1 .1 / 3/ 15 /3 .2 /
 30 13 /3 .0 / 31 15 /3 .0 / 45 1. 3. /5 0
 41 11 4. 03 1 36 06 3/ 02 1 35 03 3. 0. 1
 46 1. 33 0/ 1 42 05 31 /7 1 43 10 3/ 0. 1


((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                       (
(                                                                       (
(                                                                       (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
Proc Cluster
                                        -(                                  ( -
                                        -(                ( -
                                        -(                ( -
                                        -(                ( -
run
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                       (
(                                  '                                    (
(                                                1                      (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
Proc Tree
                                   2
                                                   3      -(                        ( -
             run
```

```
proc print                          20'   run

((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                        '       (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
proc sort                                  run
proc sort                                  run
data     0
                                                    '


run

((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                   (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
      2
proc gplot             0
              1(  0
       run

proc tabulate            0

             0   1
                      *    0   1'(
       run

((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                          (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
proc sort           0                  run
proc means          0
             0   1
                                       0   1
       run                                    .
proc print
       run

((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                        (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
      2
proc fastclus
                   3 .             -(         1             1           ( -
                     10            -(           /.                                 ( -
                     /             -(                                          ( -
                                   -(                          ( -
                   0   1
run

((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                          (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
proc print                      run

proc gplot              /
              1(  0
       run

((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                          (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
      2
```

```
proc fastclus

                                -(                                                      (-
                 3              -(          1           1           (-
                    10          -(            /.                                   (-
                    0           -(                                             (-
                                      -(                        (-
                 0   1
run

proc tabulate         0
           0   1

             0   1 *         (             '
run

proc gplot           0        1(  0         run

proc gplot           0        /(  2         run
proc gplot           0         /   2'(        run

proc glm           0

             /   2
      run
```

Proc MODECLUS : Typologie non paramétrique

Caractéristiques :

- Bien adapté pour détecter des clusters très différents en terme de taille, de forme et de dispersion.
- recherche de clusters de tailles et de dispersions identiques (Nécessite un nombre assez important de données pour les retrouver).
- Estime les fonctions de densité des groupes de manière non paramétrique (recherche de maxima sur la densité). Recherche des voisins dans un espace sphérique.

Principe :

- Estimation des densités
-
- Tests de plusieurs radius (rayons) : plus le radius est élevé moins il y a de groupes

```
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(                                                                         (
(                                                                   (
(                          /    4   /                         (
(                                                             (
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((
(
     --         ,    ,    -             -    -   -      -41.11-    -        -       ,   !
                   ./0,
data



/6 /6   0. 00   0/ 0.   /0 01   /5 /0   01 03   03 0.   /4 05
0. /1   06 00   6. 0.   53 /7   55 01   6/ 04   33 0/   42 02
50 04   5. 13   53 1.   56 20   /6 30   05 35   2/ 4/   26 42
37 50   47 50   6. 6.   1/ 31   3/ 47   50 6/
```

```
proc sgplot

   run

     0
(                                    (
proc modeclus
             1
        10 15 35


   run

proc sgplot


            -
   run
```

Proc VARCLUS : Typologie des variables

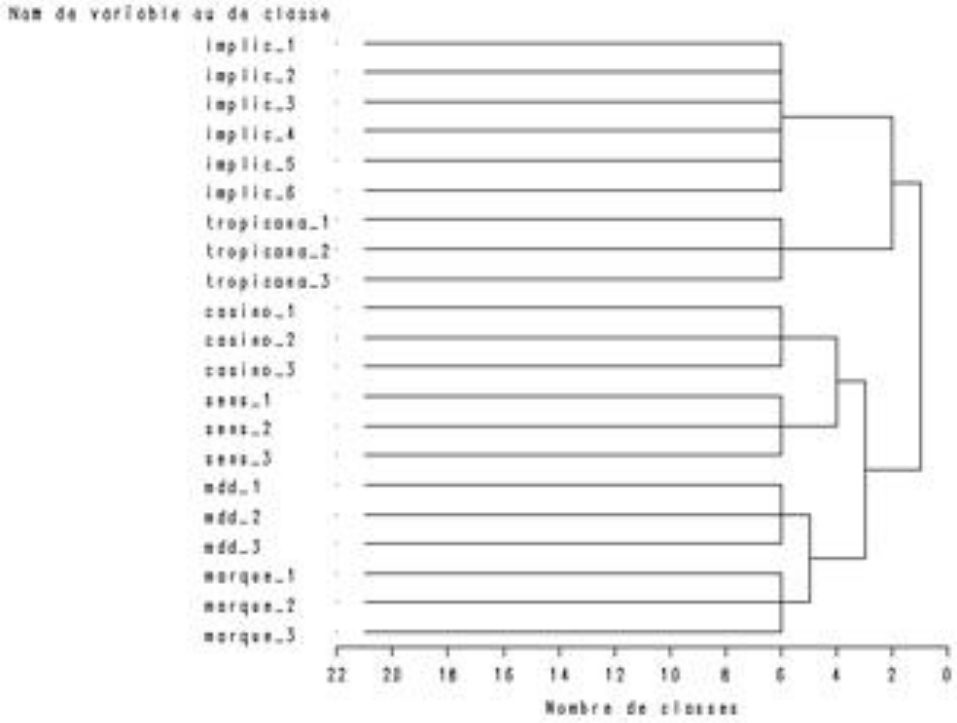
Caractéristiques :

- Partition automatique de nouvelles variables (composante) obtenues par le regroupement, sans recouvrement, des variables quantitatives (ou binaires) corrélées
- Options
  - ACP avec classification hiérarchique « single linkage »
  - ACP oblique quartimax et classification hiérarchique (Proc Varclus)
- Utilisé pour construire des échelles de mesure
  - Sans option : analyse des corrélations et facteur
  - Centroid cov : analyse de la variance et moyenne arithmétique

```
proc corr
           /  2
run
proc varclus
           /  2
run
proc varclus
     /  2
run
```

Exemple sur Orange

Nombre de classes	Variation totale expliquée par classe	Proportion de variation expliquée par classe	Proportion min. expliquée par une classe	Seconde valeur propre maximum dans une classe	R carré minimum pour une variable	Ratio 1-R**2 maximum pour une variable
1	4.421405	0.2105	0.2105	3.963702	0.0023	
2	7.947717	0.3785	0.3247	2.245303	0.1577	0.8024
3	10.140951	0.4829	0.3247	2.062240	0.1577	0.8908
4	11.978383	0.5704	0.4746	1.438379	0.2367	0.8224
5	13.387183	0.6375	0.4811	1.129198	0.3294	0.7092
6	14.512005	0.6910	0.6285	0.834567	0.4724	0.5418



Total variation expliquée = 14.512 Proportion = 0.6910

6 classes		R carré avec			
Classe	Variable	Propre classe	Le plus proche	Ratio 1-R**2	Libellé de variable
Cluster 1	implic_1	0.6527	0.0649	0.3715	Le jus d'orange frais est un produit qui compte vraiment beaucoup pour moi
	implic_2	0.6697	0.0562	0.3500	En matière de jus d'orange, il y a beaucoup à perdre si l'on choisit la mauvaise marque
	implic_3	0.4724	0.0034	0.5294	Il est possible de faire un mauvais choix lors de l'achat d'un jus d'orange frais
	implic_4	0.7293	0.0272	0.2783	Le choix de son jus d'orange frais a une forte valeur symbolique
	implic_5	0.7695	0.0503	0.2427	Je prends du plaisir à boire du jus d'orange frais et j'apprécie particulièrement en boire
	implic_6	0.4777	0.0060	0.5255	On peut dire que le choix de mon jus d'orange frais m'intéresse
Cluster 2	casino_1	0.8141	0.1599	0.2213	Si j'achète ce jus d'orange casino, je vais probablement l'aimer
	casino_2	0.7649	0.1049	0.2626	Je pense que la majorité des personnes qui achètent ce jus d'orange Casino est satisfaite
	casino_3	0.7910	0.1050	0.2335	D'une manière globale, je décrirais ce produit comme attrayant

Les arbres de segmentation

Caractéristiques :

-
- Voir Entreprise miner.