Scénario: Données bancaires et segmentation de clientèle

Résumé

Ce scénario décrit le traitement classique, la fouille, de données pour la gestion de la relation client (GRC) ; exploration de données bancaires par des méthodes uni, bi et multidimensionnelles : ACP, AFCM puis segmentation de clientèle par classification non-supervisée k-means, CAH et représentation, interprétation des classes.

1 Introduction

1.1 Présentation

Le travail proposé vient compléter le cours sur l'exploration statistique. Son objectif est double :

1. illustrer l'emploi de chacune des méthodes du cours sur un exemple en vraie grandeur mais relativement simple,
2. aboutir à un objectif classique en Gestion de la Relation Client : une "segmentation" de clientèle.

Il se propose donc de décrire un jeu de données bancaires en utilisant les principales méthodes de statistique exploratoire multidimensionnelle (analyse en composantes principales, analyse des correspondances simple et multiple, analyse factorielle discriminante, classification non supervisée). L'objectif principal est de produire une "segmentation" de clientèle, c'est-à-dire une répartition en classes homogènes des clients en fonction de leur comportement bancaire. C'est aussi la mise en œuvre d'une démarche classique permettant d'affiner sa compréhension des données dans l'idée de construire un score d'appétence pour la carte visa premier. Ce type d'objectif est intégré au module de data mining Apprentissage Statistique. Il s'agit du score d'appétence de la carte Visa Premier mais ce pourrait être un score d'attrition (*churn*) d'un opérateur téléphonique ou encore un score de défaillance d'un emprunteur ou de faillite d'une entreprise ; les outils de modélisation sont les mêmes et sont très largement utilisés dans tout le secteur tertiaire pour l'aide à la décision.

1.2 Travail à réaliser

Ce scénario propose une séquence de commandes SAS permettant d'enchaîner les traitements. Voici la liste des résultats à produire afin d'apporter dans le rapport final des réponses aux questions suivantes.

Exploration élémentaire

Intégrer quelques graphiques, les commenter ; justifier les transformations qui ont été opérées dans le programme SAS sur les variables tant quantitatives que qualitatives avant d'aborder la suite des traitements. Pourquoi certaines variables ont été supprimées, quelles corrections ont été apportées aux données.

Analyse en composantes principales

Justifier le choix de dimension, interprétation des axes et donc des comportements des clients dans chacune des grandes directions. Sur le plan principal, que dire de la répartition des possesseurs/non-possesseurs de carte VP ?

Analyse factorielle des correspondances

Explicitez, justifiez les choix qui ont été faits pour le recodage en classes des variables quantitatives. Nombre et interprétation des deux premiers axes de l'AFCM. Explicitez, commentez quelques rapprochements notables de modalités comme ceux, par exemple, caractérisant le profil des possesseurs de cartes VP. Que pouvez vous dire du 3ème axe ? Quelles remarques vous suggère la représentation des individus (possesseurs / non possesseurs) ?

Classification

Combien de classes retenir ? Caractériser et interpréter de façon détaillée les classes ou segments de clientèle issus de la classification opérée à partir des composantes de l'AFCM. Comparer, par une AFC simple, la classification obtenue avec celle en usage dans les services commerciaux de la banque (variable segv2s).

2 Description des données

2.1 Les variables

La liste des variables est issue d'une base de données retraçant l'historique mensuel bancaire et les caractéristiques de tous les clients. Un sondage a été réalisé afin d'alléger les traitements ainsi qu'une première sélection de variables. Les variables contenues dans le fichier initial sont décrites dans le tableau ci-dessous. Elles sont observées sur 1425 clients.

TABLE 1 – Liste des variables et de leur libellé

Identif.	Libellé
matric	Matricule (identifiant client)
depts	Département de résidence
pvs	Point de vente
sexeq	Sexe (qualitatif)
ager	Age en années
famiq	Situation familiale (Fmar : marié, Fcel : célibataire, Fdiv :divorcé, Fuli :union libre, Fsep : séparé de corps, Fveu :veuf)
relat	Ancienneté de relation en mois
pcspq	Catégorie socio-professionnelle (code num)
quals	Code "qualité" client évalué par la banque
GxxGxxS	plusieurs variables caractérisant les interdits bancaires
impnbs	Nombre d'impayés en cours
rejets	Montant total des rejets en francs
opgnb	Nombre d'opérations par guichet dans le mois
moyrv	Moyenne des mouvements nets créditeurs des 3 mois en Kf
tavep	Total des avoirs épargne monétaire en francs
endet	Taux d'endettement
gaget	Total des engagements en francs
gagec	Total des engagements court terme en francs
gagem	Total des engagements moyen terme en francs
kvunb	Nombre de comptes à vue
qsmoy	Moyenne des soldes moyens sur 3 mois
qcred	Moyenne des mouvements créditeurs en Kf
dmvtp	Age du dernier mouvement (en jours)

2.2 Lecture des données

Les donnés sont disponibles dans le répertoire "data" tandis que les programmes sont accessibles dans ce répertoire.

Attention :

- la plupart des programmes écrits en SAS sont fournis mais les commandes contenant des noms des fichiers sont dépendantes de l'environnement de travail. Il faut donc les adapter en fonction de l'emplacement (répertoire de travail) de ces fichiers.
- Les données sont anonymées et datent du siècle dernier, elles n'ont plus d'intérêt "commercial".
- De façon générale, plutôt que de conserver tous les fichiers de données intermédiaires à une étude, ce qui peut nécessiter beaucoup d'espace disque, il est important, voire crucial, d'archiver **tous** les programmes intermédiaires de saisie, sélection, transformation des données. En effet, en cas de problème ou même simplement d'un mauvais choix méthodologique, il faut pouvoir rapidement repartir d'une étape précédente. Les programmes, avec des commentaires, sont ensuite insérer en annexe d'un rapport.

Charger les programmes des fichier `visa_lec.sas`, `visa_trans.sas`, `visa_code.sas`.

Excuter le programme `visa_lec.sas`

Prendre le temps de lire le programme afin d'en comprendre les principaux éléments. Les noms de variables sont choisis de façon à être un peu explicite tout en étant courts (5 caractères). De plus, nous avons anticipé en ajoutant un "`S`" aux variables qui seront supprimées (on ne pouvait le savoir *a priori*),

TABLE 2 – Liste des variables et de leur libellé (suite)

Identif.	Libellé
boppn	Nombre d'opérations à M-1
facan	Montant facturé dans l'année en francs
lgagt	Engagement long terme
vienb	Nombre de produits contrats vie
viemt	Montant des produits contrats vie en francs
uemnb	Nombre de produits épargne monétaire
uemmts	Montant des produits d'épargne monétaire en francs
xlgnb	Nombre de produits d'épargne logement
xlgmt	Montant des produits d'épargne logement en francs
ylvnb	Nombre de comptes sur livret
ylvmt	Montant des comptes sur livret en francs
nbelts	Nombre de produits d'épargne long terme
mtelts	Montant des produits d'épargne long terme en francs
nbcats	Nombre de produits épargne à terme
mtcats	Montant des produits épargne à terme
nbbecs	Nombre de produits bons et certificats
mtbecs	Montant des produits bons et certificats en francs
rocnb	Nombre de paiements par carte bancaire à M-1
ntcas	Nombre total de cartes
nptag	Nombre de cartes point argent
segv2s	Segmentation version 2
itavc	Total des avoirs sur tous les comptes
havef	Total des avoirs épargne financière en francs
jnbjd1s	Nombre de jours à débit à M
jnbjd2s	Nombre de jours à débit à M-1
jnbjd3s	Nombre de jours à débit à M-2
carvp	**Possession de la carte VISA Premier**

un "Q" aux variables qualitatives, tandis que chaque code de variable conservée débute par une lettre différentes pour faciliter des graphiques à venir. Les variables de "nombre de jours de débit" sont regroupées en leur somme. Le programme filtre enfin les clients pas ou peu concernés (comptes professionnels, interdits bancaires) et élimine les variables inappropriées.

Vérifier la bonne lecture du fichier en cliquant successivement dans la fenêtre `Explorer` sur `libraries`, `sasuser`, `visprem`. Après vérification de la bonne lecture du fichier, fermer la fenêtre avec le menu `File`.

3 Exploration élémentaire des données

3.1 Outils et objectifs

Cette étape peut apparaître fastidieuse, elle est néanmoins indispensable ; son but essentiel est d'aboutir à un sous-ensemble de données pertinentes. En situation réelle, cette phase d'obtention d'une base fiable est celle prenant souvent le plus de temps ; elle a été considérablement raccourcie dans ce TP par une organisation très directive. Cette phase serait rendue plus facile par l'utilisation d'un outil interactif comme le module SAS/INSIGHT (choix "Interactive data analysis" ou "analyse interactive des données" du choix "Analyse" du menu "Solutions"). Malheureusement de module a disparu avec la version 9.4 de SAS. Des fonctionnalités identiques sont disponibles dans SAS/IML Studio mais... que sous Windows ou alors en utilisant JMP toujours sous Windows. Pour éviter d'avoir à jongler entre les logiciels, l'analyse est faite en utilisant les procédures classiques de SAS/Base, SAS/Stat ou SAS/Graph. Malheureusement les aspects "interactifs" ont disparu ; il n'est alors pas possible d'observer en direct l'effet, par exemple, d'une transformation d'une variable sur sa distribution.

Les transformations sont donc supposées puis valider après transformation des données dans une étape `data` classique.

Une exécution avec R ou python serait plus efficace à ce niveau du travail.

3.2 Exploration univariée

Le but essentiel de cette phase est de préparer le programme de transformation des données en vue des étapes ultérieures. Voici quelques objectifs dont

les modalités d'exécution sont décrites plus loin.

- identifier, en vue de leur suppression ou de leur imputation, les variables présentant trop de données manquantes,
- même chose pour les variables quasi constantes (une même valeur ou une même modalité sur la presque totalité de l'échantillon),
- regrouper les modalités à trop faibles effectifs d'une même variable qualitative,
- rechercher des transformations des variables (log ou racine ou "Cox") dont la distribution est très dissymétrique,
- supprimer les clients dont la majorité des valeurs sont manquantes,
- identifier et s'interroger sur les clients présentant des valeurs atypiques même après transformation,

Variables quantitatives

Résumé global.

```
/* Toutes les variables quantitatives */
proc means data=sasuser.visprem;
run;
```

Repérer les "variables" constantes à supprimer.

Analyser la distribution de chaque variable quantitative.

```
/* Distribution d'une variable */
proc univariate data=sasuser.visprem plot;
var TAVEP RELAT QSMOY;
run;
```

Dans l'ensemble, les variables discrètes sont conservées à l'identique sauf certaines jugées inutiles qui seront supprimées. Pour les autres, exprimant des soldes, une transformation monotone ($\log(a+x)$, $\sqrt{(a+x)}$) susceptible de rendre leur distribution plus "symétrique" et de limiter le nombre de valeurs atypiques est utilisée.

La variable `RELAT` présente un groupe de valeurs atypiques voire douteuses, s'en souvenir.

La moyenne des soldes moyens sur trois mois (`QSMOY`) pose un problème mais elle semble peu discriminante, elle est conservée en l'état. Sinon une transformation par la fonction "argument sinus hyperbolique" peut être utilisée avec ce type de distribution mais ce n'est pas indispensable.

Variables qualitatives

La procédure `freq` s'applique aux variables qualitatives pour décrire les fréquences de chacune de leurs modalités.

```
proc freq data=sasuser.visprem;
tables FAMILQ PCSPQ;
run;
```

On observe ainsi que certaines variables ont des modalités trop peu fréquentes. Elles nécessitent des regroupements qui seront intégrés dans la suite du travail. Par exemple, combien trouvez vous d'agriculteurs et de professions inconnues ?

Résultats

Intégrer les principaux résultats (graphiques, tableaux) dans un traitement de texte avec les commentaires.

3.3 Exploration bivariée

Explorer les relations entre les variables prises 2 à 2 afin de se familiariser avec leurs structures.

Deux quantitatives

Certaines variables sont tellement corrélées, qu'une seule suffit (TAVEP et UEMMTS).

```
proc gplot data=sasuser.visprem;
symbol1 value=dot;
plot TAVEP*UEMMTS=1;
run;
goptions reset=all;
```

Qualitative et Quantitative

```
proc boxplot data = sasuser.visprem;
```

```
plot qsmoy*carvp;
run;
```

Deux qualitatives

```
proc freq data=sasuser.visprem;
tables sexeq*carvp / plots=mosaicplot;
tables NTCAS*CARVP;
run;
```

Détecter des incohérences

Un client de 22 ans en relation depuis 800 mois avec sa banque ! Identifier ce groupe de clients moins âgés que leur ancienneté dans la banque. Construire par exemple un nuage de points en croisant ancienneté et âge. Trouvez-vous une explication à cette bogue ?

```
proc gplot data=sasuser.visprem;
symbol1 value=dot;
plot RELAT*AGER=1;
run;
```

Vérifier que des clients avec un nombre total de cartes (`ntcas` égal à 0) sont néanmoins possesseurs de la carte Visa Premier !

```
proc freq data=sasuser.visprem;
tables NTCAS*CARVP;
run;
```

Les informations sont issues de bases de données mises à jour à des dates différentes ! De plus, comme cette variable “contient” aussi l’information à prévoir c’est-à-dire la possession de la carte VP, elle ne sera pas utilisée par la suite.

`ZOCNB` (nb d’opérations par cartes) présente beaucoup de données manquantes. Vérifier qu’il s’agit des clients sans carte.

```
data vispremna;
set sasuser.visprem;
if ZOCNB = . then ZOCNBQ = "NA";
   else ZOCNBQ = "OK";
proc freq;
tables NTCAS*ZOCNBQ;
run;
```

La valeur manquante peut donc être imputée par 0 et la variable conservée dans la suite du travail (transformée par la fonction racine).

3.4 Vérifications

Exécuter le programme `visa_trans.sas` de transformation des données. Il crée la table `sasuser.vispremt`.

Étudier ce programme afin de comprendre les transformations réalisées, les justifier, en vérifier le bienfondé : distributions des variables quantitatives, correction de l’ancienneté de la relation.

Croiser les variables quantitatives transformées entre elles (matrice des nuages de points) pour détecter d’éventuelles relations non-linéaires.

```
proc corr data=sasuser.vispremt
  plots (MAXPOINTS=200000) =
  matrix (hist nvar=all nwith=all);
var AGER RELAT QSMOY OPGNBL--GAGEML;
with QCREDL--ZOCNBR;
run;
```

Ce graphique est limité à 10 variables en ligne ou colonne (cf. l’avertissement dans le journal), il faut donc le compléter.

Même après transformation, beaucoup de variables ne présentent pas une distribution très satisfaisante. C’est pourquoi dans une section suivante, nous nous intéressons à des transformations rendant qualitatives des variables quantitatives. Il s’agit en fait de coder la possession ou non d’un produit financier indépendamment du montant afin de ne retenir que l’information qui semble la plus pertinente.

4 Analyse en composantes principales

Une ACP permet d'approfondir la compréhension des données et de leur structure de corrélation. Plusieurs macros sont utilisées afin de compléter les résultats de la procédure `princomp` et tracer des graphiques en contrôlant les options.

4.1 Construction des graphiques et résultats

Exécuter dans SAS les fichiers des macros commandes nécessaires avant de les appeler avec les commandes ci-dessous.

```
%acp(vispremt,carvpr,ager relat kvunb
   opgnbl--zocnbr);
%gacpbx;
%gacpsx;
%gacpix(nc=2);
%gacpvx;
```

4.2 Interprétations

Quel choix de dimension semble le plus raisonnable ? Justifiez.

Compte tenu de la complexité des graphiques (nombre de variables), une classification de celles-ci peut aider, dans certains cas, à l'interprétation. Attention, l'algorithme utilisé dans la procédure `varclus` n'a pas grand chose à voir avec ceux étudiés ultérieurement pour trouver des classes d'individus. La procédure `varclus` construit une classification hiérarchique descendante de *variables* selon un algorithme itératif opérant une dichotomie sur un groupe de variables et donc initialement sur toutes. Un groupe de variables est divisé en deux sous-groupes après le calcul d'une ACP. Un premier sous-groupe est constitué de toutes les variables les plus corrélées au premier axe, les autres constituent le deuxième sous-groupe. Le deuxième groupe est vide (plus de dichotomie) si la deuxième valeur propre de l'ACP est jugée trop petite, par exemple inférieure à 1 pour des variables réduites. Ainsi, chaque groupe de variables obtenu est considéré comme pouvant se résumer par une seule dimension soit une seule variable.

Le programme ci-dessous fournit donc une classification des variables :

```
proc varclus data=sasuser.vispremt
     outtree=tree noprint;
     var ager relat kvunb opgnbl--jnbjdl;
run;
proc tree data=tree graphics hor;
run;
```

L'interprétation des axes d'une ACP de données réelles est un travail délicat, car emprunt de subjectivité, mais il permet de se faire une idée synthétique d'un ensemble de données complexe. Ce sera ensuite très utile pour l'interprétation des classes des clients.

- Premier axe : repérer un effet "taille". Interprétations en une ligne.
- Deuxième axe : Il sépare deux groupes de variables. Comment les caractériser. Ces deux groupes de variables déterminent deux axes de dispersion des clients. Comment peut-on résumer ces deux types de comportements.
- Une fois que la structure du premier plan est assimilée, pouvez vous trouver une interprétation du troisième axe principal ? Il est possible de s'aider de la matrice des corrélations variables $\times$ facteurs affichée dans la fenêtre `output` qui permet de prendre en compte les variables les plus liées au 3ème axe.

```
%gacpvx(x=1,y=3);
%gacpix(x=1,y=3,nc=2);
```

Remarquer que les porteurs de carte VP sont très dispersés sur tous les plans. Quelle remarque peut-on en tirer sur l'objectif de discrimination de ces deux classes.

5 Analyse factorielle multiple de correspondances

5.1 Préparation des données

L'ACP se limitait à l'étude des variables quantitatives. L'AFCM permet de prendre en compte toutes les variables à condition de recoder celles quantitatives en qualitatives par découpage en classes. Certaines règles, surtout de bon sens, ont été appliquées pour aboutir au résultat.

- Le nombre de classes doit être relativement restreint pour limiter les dimensions et faciliter les interprétations,
- il est préférable de choisir, pour des variables distribuées “normalement”, des classes d’effectifs égaux : les bornes sont des quantiles.
- Dans le cas des données bancaires, les variables présentant une distribution très asymétrique se résument souvent en une variable dichotomique : présence ou absence d’un produit financier donné. Dans ce dernier cas, il est inutile de conserver à la fois une variable “nombre” et une variable “montant” du même produit financier.
- Enfin, pour des raisons techniques de lisibilité des graphiques, les codes des modalités associés à une même variable commencent par la même lettre. La macro qui réalise les graphiques représente de la même couleur toutes les modalités d’une même variable pour faciliter les interprétations.

Étudier et **exécuter** le contenu du programme de recodage `visa_code.sas`.

5.2 Sortie des graphiques et résultats

AFCM des caractéristiques sociales

Il est classique de rechercher une première afcm n’utilisant que les variables résumant les caractéristiques sociales (signalétique) des clients en variables principales. Les variables bancaires sont projetées en tant que variables supplémentaires.

```
proc corresp data=sasuser.vispremv observed
    out=resul mca dim=8;
tables famiq sexeq pcspq  ageq relatq
       kvunbq vienbq uemnbq xlgnbq
       ylvnbq zocnbq nptagq carvp endetq
       gagetq facanq lgagtq havefq qsmoyq
       opgnbq moyrvq dmvtpq boppnq itavcq
       jnbjdq tavepq;
supplementary kvunbq vienbq uemnbq xlgnbq
       ylvnbq zocnbq nptagq carvp endetq
       gagetq facanq lgagtq havefq qsmoyq
       opgnbq moyrvq dmvtpq boppnq itavcq
       jnbjdq tavepq;
run;
%gafcix;          /* plan principal 1 x 2 */
```

Dans le cas de ces données bancaires, le graphique ne présente pas un grand intérêt mais permet de souligner quelques particularités de l’échantillon.

AFCM de toutes les variables

Calculer l’AFCM de toutes les variables qualitatives incluant les quantitatives recodées en classes. Obtenir les graphiques à l’aide des macros.

```
proc corresp data=sasuser.vispremv observed
    out=resul mca dim=8;
tables famiq sexeq pcspq kvunbq vienbq
       uemnbq xlgnbq ylvnbq zocnbq
       nptagq carvp endetq gagetq facanq
       lgagtq havefq ageq relatq qsmoyq
       opgnbq moyrvq dmvtpq boppnq itavcq
       jnbjdq tavepq;
run;
%gafcix;          /* plan principal 1 x 2 */
%gafcix(x=1,y=3); /* plan principal 1 x 3 */
```

5.3 Interprétations

Axes et modalités

Combien d’axe retiendriez-vous ? Sont-ils faciles à interpréter ? Montrer que l’interprétation des deux premiers axes est similaire à celle de l’ACP. Les coordonnées listées dans la fenêtre `output` permettent de différencier les libellés superposés. Caractériser les possesseurs de carte visa premier.

Individus

Pour obtenir une représentation simultanée des individus, le programme ci-dessous calcule l’équivalent de l’AFCM par AFC du tableau disjonctif complet (attention à la virgule). Il fournit donc des coordonnées pour les individus. Elles permettent à ce niveau d’en contrôler la dispersion.

```
proc corresp data=sasuser.vispremv  out=resul
```

```
    dim=8;
tables matric,famiq sexeq pcspq kvunbq vienbq
       uemnbq xlgnbq ylvnbq zocnbq
       nptagq carvp endetq gagetq facanq
       lgagtq havefq ageq relatq qsmoyq
       opgnbq moyrvq dmvtpq boppnq itavcq
       jnbjdq tavepq;
run;
```

La même macro permet la représentation graphique mais la coloration des individus dépendant de leur matricule n'a aucun intérêt. Pour qu'elle dépende de la possession de la carte Visa premier un petit traitement est nécessaire : fusion selon le matricule de la table initiale avec celle des résultats après tri. La variable `carvpr` recodée sert alors d'identificateur et la taille des points est paramétrée (`tp=1`). Ce type de gymnastique demande un peu de compétences avec SAS et s'avère très utile.

```
proc sort data=sasuser.vispremv out=vispremr;
by matric;
proc sort data=resul out=resul;
by _name_;
data resul;
merge vispremr (keep=matric carvp
      rename=(matric=_name_)) resul ;
by _name_;
select(carvp);
when('Coui') _name_= '0000';
when('Cnon') _name_= '1111';
otherwise;
end;run;
%gafcix(tp=1);
%gafcix(x=1,y=3,tp=1);
```

Apprécier sur les différents plans la plus ou moins bonne séparation des deux classes, possession ou non, de la carte visa premier. Commentaires.

6 Classification non supervisée

6.1 Introduction

L'objectif principal, lors de cette première étude des données, est de fournir une typologie ou segmentation des clients. C'est-à-dire de définir des classes les plus homogènes au regard des comportements bancaires. Les algorithmes de classification étudiés sont adaptés à des variables *quantitatives* ou des matrices de distances. L'intégration d'informations qualitatives peut se faire par un recodage préalable (scoring) à l'aide d'une analyse des correspondances multiples. Ce TP se propose donc de comparer deux approches : classification à partir des seules variables quantitatives ou classification à partir des scores issues d'une afcm. D'autres approches sont envisageables sur des variables qualitatives qui nécessitent la définition d'une distance ou dissimilarité entre individus adaptée aux variables qualitatives. Mais, nécessitant la construction de la matrice $n \times n$ des distances des individus deux à deux, elles ne sont pas adaptées aux très grands tableaux.

6.2 Classification sur variables quantitatives

Attention à un piège

La variable `relat` a une variance nettement plus grande que les autres variables. Sans précaution élémentaire, celle-ci prend une importance prépondérante dans toute tentative de classification :

```
%nudnc(vispremt,carvp,ager relat
    opgnbl--zocnbr,6);
```

Vérifier plus précisément les résultats numériques fournis par cette procédure. En particulier, le tableau `Statistics for variables` (ouvrir la fenêtre `Results` qui permet d'accéder directement à ce tableau) montre l'importance prépondérante prise par la variable `relat`. On obtient ainsi une classification unidimensionnelle triviale dont les classes sont fixées par les valeurs de cette seule variable. Ceci se vérifie directement par un graphique dans sas/insight. Ouvrir le fichier `sasuser.ndclasse` colorier les individus par la variable classe et afficher le nuage de points `RELAT x AGER`.

Ceci montre l'importance de considérer, dans les techniques multidimensionnelles, des variables homogènes en variance. Il est important, dans le cas contraire, de réduire (standardiser) les variables avant une classification. Cette option est prise par défaut en ACP mais pas en classification, elle doit être

calculée.

```
/* standardisation (réduction) des variables
   numériques*/
proc standard data=sasuser.vispremt
    out=sasuser.visprems mean=0 std=1;
run;
```

Choix du nombre de classes

Pour les données bancaires, le nombre de lignes (de clients) est relativement faible et il est possible de calculer directement une CAH, ce qui ne serait pas possible avec un fichier plus gros. Par souci de généralité, on adopte ici une stratégie applicable même à des très gros fichiers. Les différentes étapes permettant de choisir un nombre pertinent de classes sont alors les suivantes :

1. *Restriction du nombre des individus à classer* par réallocation itérative (procédure `fastclus`) des individus dans L classes, où L est choisi arbitrairement égal au dixième de l'effectif de départ.
2. *Classification Ascendante Hiérarchique* (procédure `cluster`) des barycentres des L classes obtenues précédemment. Le poids d'un barycentre est égal à la somme des poids des individus de sa classe. Le saut de Ward est utilisé par défaut. On rappelle que le saut de Ward est utilisé quand on veut maximiser l'inertie inter de la partition.
3. *Représentation graphique du R^2 semi-partiel* (cas du saut de Ward) pour aider l'utilisateur dans le choix du nombre de classes.

Les trois premières étapes sont effectuées à l'aide de la macro-procédure `choixnc`, qui prend en arguments le tableau de données, la liste des variables, le nombre L de classes, et l'indice d'éloignement utilisé dans la CAH (Ward par défaut) :

```
%choixnc(visprems,ager relat opgnbl--zocnbr,100);
```

Le deuxième graphique, est obtenu en exécutant la macro ci-dessous en précisant un nombre max de classes :

```
%critere(10);
```

Classification

Une fois déterminé le nombre K (5, 6 ou 7 ? car 3 est jugé trop petit) de classes de la partition finale au vu des graphiques précédents, il est possible de lancer les nuées dynamiques sur l'ensemble des clients en choisissant comme pôles de départ les barycentres des K classes obtenues en sortie de la `cah` calculée pour le choix du nombre de classe (cette étape peut être vue comme une amélioration de la partition obtenue après exécution de `cah`). La macro `nudnc` réalise cette partition et fournit en sortie une table SAS `sasuser.ndclasse` contenant : les variables initiales, le numéro de classe `classe` de chaque individu, et une variable `distance` donnant la distance de chaque individu au barycentre de sa classe. Cette procédure peut également être utilisée indépendamment en ne précisant pas le paramètre `init=pole` d'initialisation des groupes.

```
%nudnc(visprems,carvp,ager relat
    opgnbl--zocnbr,5,init=poles);
```

Cette partition finale est représentée sur les plans factoriels de l'ACP. Après exécution des commandes suivantes ; commentez les qualités visuelles de la discrimination obtenue. Comment interpréter les classes.

```
%acp(ndclasse,classe,ager relat opgnbl--zocnbr);
%gacpixm;
%gacpvx;
proc gplot data=coorindq;
plot prin2*prin1=ident;
run;
quit;
```

Classification sur composantes principales

Une deuxième approche, pas nécessaire ici mais utile lorsque le nombre de variables est très important, consiste à enchaîner une technique de classification sur les variables principales d'une ACP. la variance des variable est alors contrôlée par l'intervention de l'ACP. Cette démarche se retrouve dans d'autres circonstances : régression ou réseau de neurones sur composantes principales. Voici comment la réaliser à fin "pédagogique".

```
%acp(vispremt,carvp,ager relat opgnbl--zocnbr,q=6);
```

```
data sasuser.varprinc;
set coorindq; run;
%choixnc(varprinc,prin1-prin6,100);
%critere(10);
%nudnc(varprinc,ident,prin1-prin6,5,init=poles);

proc gplot data=sasuser.ndclasse;
plot prin2*prin1=classe;
run;
quit;
```

Autre possibilité : ouvrir dans sas/insight la table `sasuser.vpclas2` pour visualiser avec un "rotating plot" les clients dans les coordonnées de l'ACP (prin1, prin2, prin3). Colorer les individus par classe : edit/window/tools).

Deux classifications ont ainsi été obtenues. La deuxième semble offrir des classes mieux discriminées et évite une très grosse classe moyenne. Donner une interprŕation des classes.

Jusque là, nous avons négligé les variables qualitatives. La prise en compte de ces variables est l'objet de la section suivante, c'est cette approche qui sera privilégiée avec ce type de données.

6.3 Classification de variables qualitatives et quantitatives

Classification sur scores d'une afcm

Il suffit d'exécuter les procédures de classification lorsque les clients sont repérés par leurs valeurs sur les composantes principales de l'AFCM. Les données sont contenues dans la table `work.resul`.

```
proc corresp data=sasuser.vispremv
out=resul dim=5;
tables matric,famiq sexeq pcspq kvunbq vienbq
       uemnbq xlgnbq ylvnbq zocnbq
       nptagq carvp endetq gagetq facanq
       lgagtq havefq ageq relatq qsmoyq
       opgnbq moyrvq dmvtpq boppnq itavcq
       jnbjdq tavepq;
run;
/* extraction des observations */
proc sort data=sasuser.vispremv out=vispremr;
by matric;
proc sort data=resul out=resul2;
by _name_;
data sasuser.varprinc;
merge vispremr (rename=(matric=_name_)) resul2 ;
by _name_;
if _TYPE_='OBS';
run;
%choixnc(varprinc,dim1-dim5,100);
%critere(10);
%nudnc(varprinc,_name_,dim1-dim5,4,init=poles);
```

Ouvrir ensuite dans sas/insight la `sasuser.vpclas3`, visualiser avec un "rotating plot" les clients dans les coordonnées de l'AFC (dim1,dim2,dim3), colorer les individus par classe. Il reste à interpréter les classes mais c'est un autre problème. On peut déjà essayer avec une AFCM intégrant à la fois les variables initiales ainsi que cette nouvelle variable `classe`. Cela nécessite le tri préalable par matricule de la table `sasuser.ndclasse` pour sa fusion avec la table `work.vispremr`.

```
proc sort data=sasuser.ndclasse out=resul3;
by _name_;
data varprind (drop=classe);
merge sasuser.varprinc resul3 (keep=_name_ classe) ;
by _name_;
select (classe);
when(1) klasse='W1';when(2) klasse='W2';
        when(3) klasse='W3';when(4) klasse='W4';
    when(5) klasse='W5'; when(6) klasse='w6';
otherwise;end;
run;
proc corresp data=varprind  out=resul dim=5 mca;
tables famiq sexeq pcspq kvunbq vienbq
       uemnbq xlgnbq ylvnbq zocnbq
```

```
       nptagq carvp endetq gagetq facanq
       lgagtq havefq ageq relatq qsmoyq
       opgnbq moyrvq dmvtpq boppnq itavcq
       jnbjdq tavepq klasse;
supplementary klasse;
run;
%gafcix;
%gafcix(x=1,y=3);
```

Les proximités des modalités des différentes variables avec celles `W1` à `W5` permettent alors, au moins grossièrement, de caractériser les comportements bancaires des clients dans chacune des classes. C'est l'objectif principal de ce projet.