Scénario: exploration enquête INSEE sur le patrimoine des français

Résumé

Exploration des données de l'enquête INSEE sur le patrimoine des français. Analyses uni, bi et multidimensionnelles : ACP, AFCM, classification non-supervisée k-means, CAH des individus et interprétation des classes.

1 Présentation

1.1 Objectif

Le travail présenté décrit le déroulement classique d'une étude marketing dans un service de Gestion de la Relation Client. Les données sont extraites d'une enquête INSEE sur le patrimoine des ménages et des personnes constituant ces ménages. Le service marketing d'une banque ou d'une société d'assurance souhaite mieux connaître la structure socio-économique des ménages en France en rapport à leur patrimoine. Deux objectifs sont poursuivis sur ce type de données :

- Définition de classes ou segments homogènes des individus, c'est-à-dire des ménages ou des individus constituant les ménages,
- Estimation d'un score ou d'un risque en vue d'un objectif particulier. Il s'agira, dans un deuxième temps, d'estimer un score d'intérêt pour un produit d'épargne particulier comme par exemple l'assurance vie, en fonction des caractéristiques socio-économiques des individus.

Seul le premier objectif est poursuivi dans la présente étude avec, en plus, l'idée d'étudier la structure des données afin d'évaluer la faisabilité du 2ème objectif ; ce sera l'objet d'un autre scénario.

1.2 Les données

Les données sont publiques et accessibles sur le site de l'INSEE. Elles sont issues d'une enquête "Patrimoine" réalisée en 2003-2004 qui faisait suite à des enquêtes "Actifs financiers" de 1986 et 1992. L'enquête "Patrimoine" visait une connaissance globale du patrimoine des ménages ainsi que de ses principales composantes. Elle recense tous les différents types d'actifs financiers, immobiliers et professionnels détenus par les individus ainsi que les différents types d'emprunts (immobilier, à la consommation, achat voiture, biens professionnels, ...). Outre la connaissance de souscriptions à différents types de produits, l'enquête interroge aussi les individus sur leur motivation de détention permettant ainsi d'appréhender le comportement patrimonial des ménages. Certaines questions liées à l'héritage, à la donation, à l'historique professionnelle et familiale, à la situation matrimoniale et financière sont ainsi renseignées. Les données, très volumineuses et relativement complexes sont accessibles sous la forme de 4 archives : ménages (657 variables pour 9 692 observations), individus (223 variables pour 22 821 observations), produits (89 variables pour 82 413 observations), transmission (29 variables pour 9 239 observations). Des variables clef identifiant le ménage, le numéro de l'individu dans le ménage ainsi que le numéro de produit possédé, permettent leur bonne mise en relation.

Dans cette étude, il a été choisi *a priori* de prendre pour unités statistiques les individus car le produit "assurance vie" est nominatif. Cette approche est aussi justifiée par le nombre important de personnes vivant seules et aussi par les difficultés à obtenir des informations patrimoniales quantitatives des ménages. C'est surtout sur les individus que cette enquête finalement renseigne.

Les variables se répartissent en différents groupes :

Sociologiques âge, sexe, situation matrimoniale, nombre d'enfants et petits enfants, études et diplômes, situation professionnelle, localisation géographiques, âge des parents, existence ou non des parent et de grands-parents, occurrence ou non d'événements graves (maladie, décès, divorce...) ;

Actifs financiers Ils ont été regroupés en plusieurs sous catégories : les livrets d'épargne (défiscalisé ou non), l'épargne logement (CEL, PEL), l'épargne retraite hors assurance vie (PEP, complémentaire retraite volontaire, assurance décès volontaire, bons de capitalisation), l'épargne salariale et les stock-options, les valeurs mobilières (compte titre, PEA, SICAV, FCP). Beaucoup de produis financiers considérés par l'enquête ne sont pas pris en compte dans cette étude car bien trop rarement possédés par les personnes interrogées : bons autres que capitalisation, compte à

terme, compte courant d'associés, livret épargne entreprise...

Endettement Cette catégorie tient compte de l'endettement pour des motifs aussi variés que l'achat de la résidence principale, les gros travaux immobiliers, l'achat de voiture, le crédit à la consommation et les emprunts à titre professionnel. Les variables sont binaires : détention d'une dette ou non ;

Patrimoine des ménages Deux variables sont référencées pour évaluer le patrimoine des ménages financier ou immobilier. Même en interrogeant confidentiellement les personnes et sur des tranches de valeurs approximatives, le nombre de valeurs manquantes est exceptionnellement élevé pour ces variables : plus de 90% pour le patrimoine financier et plus de 60 pour le patrimoine globale et immobilier. Ces variables sont inexploitables. Seule celle signalant la détention d'héritage et/ou donation est utilisable ;

Revenu des ménages Il en est de même pour les revenus, les français sont très susceptibles sur ces questions. Malgré une enquête très détaillée par tranche de type de revenu, les informations obtenues sont insuffisamment renseignées et, ne considérer que celles obtenues, introduirait sans doute des biais importants.

Un travail préliminaire important a donc consisté à construire un "entrepôt de données" fiable et suffisamment renseigné à partir des quatre bases de données disponibles sur le site de l'INSEE. Certaines variables seront considérées qualitatives (tranches d'âge), beaucoup de modalités sont regroupées pour éviter les trop faibles effectifs de certaines, des variables quasiment constantes sont éliminées. Obtenir des données correctes et fiables est généralement l'étape la plus longue d'une étude statistique, surtout si les sources sont multiples. Pour gagner du temps, tout ce travail de sélection, nettoyage des données est résumé dans un programme écrit en R : `lec-trans-patrinsee.R` dans le répertoire `http://wikistat.fr/programmes`. Consulter ce programme pour comprendre la logique du déroulement de cette étape.

Finalement, à l'issue de ce travail préliminaire, l'étude débute avec une base contenant 22821 individus décrits par les 36 variables définies dans le Tableau 1. Ces données sont accessibles dans le fichier `bdd-insee.dat` du répertoire `http://wikistat.fr/data`.

Un premier enseignement de cette étude : il aurait été sûrement beaucoup plus efficace, à même coût, d'interroger beaucoup moins de monde avec beaucoup moins de questions mais en prenant le temps d'obtenir des réponses précises à l'ensemble des questions ! C'est malheureusement un comportement excessivement répandu dans beaucoup de disciplines, des Sciences humaines à la Biologie, de viser un niveau de détail beaucoup trop fin au regard de la précision des données ou de la taille de l'échantillon.

1.3 Gestion des données manquantes

Dans beaucoup de sitations et jeux de données, le premier problème rencontré et l'un des plus délicats à prendre en compte, est celui des données manquantes ; absences qui peuvent être la conséquence de différentes causes et dont le traitement va différer justement en fonction de cette cause. Une absence de données peut être due à :

- un refus de répondre lors d'une enquête,
- une question mal formulée ou inappropriée à la personne interrogée,
- une erreur de saisie, de codage,
- une panne de l'appareil de mesure,
- ...

Pour simplifier, deux grands types sont à prendre en compte correspondant à différentes stratégies :

1. L'absence ne dépend pas du contexte, des autres variables, le "trou" est aléatoire (*missing at random*) dû à une erreur indépendamment des variables observées. Différentes stratégies sont applicables :
   - Si le taux de manquants est faible, il suffit de supprimer l'observation correspondante, le caractère aléatoire des trous ne biaisera pas les résultats,
   - dans le cas contraire, il peut être trop pénalisant de perdre des informations, différentes méthodes d'imputation sont envisageables : construire un modèle élémentaire (régression, moyenne ou médiane des valeurs...) de prévision de la ou des données manquantes en fonction des autres variables ; utiliser une méthode acceptant des données manquantes et pour certaines tout en imputant les valeurs (régression PLS, algorithme EM d'optimisation de la vraisemblance...)
2. Le trou dépend des autres variables. Il n'y a pas de méthode statistique pour répondre à cette situation, c'est souvent le simple "bon sens" qui permet de contourner la difficulté.

TABLE 1 – Signification des variables retenues et liste des modalités

Identif.	Libellé	Modalités
Asvi	Possession ou non assurance vie	AsO, AsN
AsviR	idem	1, 0
Sexe	Genre	Sh, Sf
Age	Age	Quantitatif
Tage	Tranches d'âge	T10 à T90
Couple	Vie ou non en couple	CouO CouN
Vmatri	Statut matrimonial	Vcel Vmar Vveu Vdiv
Nation	Nationalité	Nfra Nnat Netr
Diplome	Niveau de diplôme	Dsan, Dcep, Dtec (cap, bep), Dbepc, Dbact Dbacg, Db+2, Db+5
Occupa	Type d'occupation	Oact, Oina (foyer, chom, other), Oret
Work	Niveau professionnel	WctA (cadre, catA), WctB (agent, catB, tech), Wemp, WctC (osp, ouv)
statut	Statut professionnel	spri, spub, sind
Herit	Bénéfice ou non d'un héritage	HerO, HerN
Pere	Présence ou non du père	PerO PerN PerI
Mere	Présence ou non de la mère	MerO, MerN MerI
Gparp	Grands parents paternels	GppO GppN GppI
gparm	Grands parents maternels	gpmO gpmN gpmI
Jgrav	Evènement grave dans la jeunesse	JgvO JgvN
Livep	Livret d'épargne	LivO LivN
Epalo	Epargne logement	EplO EplN
qpep	Plan d'épargne populaire	qppO qppN
vmob	Valeurs mobilières	vmoO vmoN
asdecv	Assurance décès volontaire	asdO asdN
Retrait	Epargne retraite	RepO RepN
livdf	Livret défiscalisé	ldfO, ldfN
pel	Plan épargne logement	pelO, pelN
cel	Compte épargne logement	celO, celN
xcapi	Bons de capitalisation	xcpO xcpN
fepsal	Epargne salarial ou stock options	fesO fesN
Qpea	Plan épargne action	QpeO QpeN
Urbani	Niveau d'urbanisation	U1 à U5
Zeat	Région de résidence	Zso Zpar Zoue Zne Zmed Zidf Zcen
Nbenf	Nombre d'enfants	Quantitatif
Nbenfq	Nombre d'enfants	Nbe0, Nbe1, Nbe2, Nb>3
Iogoc	Type d'occupation du logement	Iloc, Iprp (usufruit)
terre	Possession de terres	terO terN
dette	Dettes ou emprunts	detO detN
bdetre	Emprunt achat maison	bemO bemN
hdetvo	Emprunt voiture	hevO hevN

C'est ce dernier cas qui est très présent dans les données étudiées.

2 Exploration élémentaire

2.1 Lecture

Le fichier contenant les données et à télécharger a été créé par dans R par l'instruction

```
write.table(bdd, "bdd-insee.dat", row.names=FALSE)
```

Pour l'importer dans SAS, il suffit d'exécuter le programme :

```
/* attention au répertoire*/
proc import datafile="bdd_insee.dat"
   out=sasuser.bddinsee dbms=dlm
   replace;
getnames=yes;
run;
```

2.2 Analyse univariée

L'analyse élémentaire qui va suivre va montrer que les données de patrimoine contiennent beaucoup de données manquantes mais pas *missing at random*. Les absences dépendent des autres variables, les conserver biaiserait ou bloquerait les analyses. Comme l'effectif de l'échantillon est important ($n = 22821$), le plus raisonnable est, dans ce cas, de supprimer des observations tout en s'assurant qu'aucun biais n'est ainsi introduit dans les objectifs de l'analyse.

Repérer les effectifs des données manquantes pour chaque variable ; repérer également les modalités à faible effectif :

```
proc freq data =sasuser.bddinsee;
run;
```

2.3 Analyse bivariée

Dans le module "analyse interactive des données (module SAS/Insight)", ouvrir la table générée puis représenter un diagramme barre de la variable et un histogramme de la variable `Age` ; Observer l'histogramme une fois que les

observations avec données manquantes de `Couple` sont sélectionnées. Même chose avec la variable `Work`. Justifier le choix opéré ci-dessous :

```
data test;
set sasuser.bddinsee;
if Tage='T10' or Tage='T20' then delete;
run;
```

Tracer toujours avec Insight des diagrammes boîtes parallèles, de la variable `Age` en fonction des deux modalités de la variable `Asvi` (possession ou non d'une assurance vie) ; un premier diagramme avec toutes les observations et un autre sans les observations avec données manquantes dans `Work`. Comparer les deux graphiques, quelle conclusion avant d'exécuter :

```
data test;
set sasuser.bddinsee;
if Tage='T10' or Tage='T20' then delete;
run;
```

Les résultats du programme ci-dessous révèlent encore des soucis :

```
proc freq data = test;
table diplome*work/chisq;
run;
```

Réaliser l'analyse des correspondances simples entre les variables `Work` et `Diplome` avec et sans les données manquantes.

```
/* avec les valeurs manquantes */
proc corresp data=test out=resul;
table Diplome,Work ;
run;
%gafcix;
/* Sans les valeurs manquantes */
data test;
set sasuser.bddinsee;
if Tage='T10' or Tage='T20' then delete;
if Diplome='NA' or Work='NA' then delete;
run;
proc corresp data=test out=resul;
table Diplome,Work ;
run;
%gafcix;
```

Que pouvez vous en déduire ?

Les variables décrivant les ressources et le patrimoine n'ont même pas été prises en compte car elles présentent beaucoup trop de données manquantes : 90 % pour le patrimoine financier, 60 % pour l'autre ; même chose pour les variables relatives aux petits enfants. C'est la faiblesse majeure de cette enquête ; cf. la remarque introductive sur la stratégie adoptée pour cette enquête trop "exhaustive" sur certains points ; trop d'informations tue l'information.

Le même type de démarche est entrepris sur les autres variables :

```
proc freq data = test;
table Mere Pere Gparp gparm Urbani Zeat;
run;
```

Pour aboutir finalement à l'exécution du programme ci-dessous. Justifier les choix opérés pour construire l'entrepôt de données `sasuser.patriminsee` qui servira de base aux analyses à venir.

```
data sasuser.patriminsee;
set sasuser.bddinsee;
if Diplome="NA" then delete;
if Tage = "T10" or Tage="T20" then delete;
if Work="NA" or Work="Waut" then delete;
if statut="NA" then delete;
if Jgrav="NA" then delete;
if Gparp="Gppi" or Gparp ="NA" then delete;
if Mere="MerI" or Pere="PerI" then delete;
if Tage="T90" then Tage="T80";
if Vmatri="Vdiv" or Vmatri="Vveu" then Vmatri="Vsep";
if Diplome="Dbacg" or Diplome="Dbact"
   then Diplome="Dbac";
if Urbani="U6" then Urbani="U5";
if Zeat="Znor" or Zeat="Zest" then Zeat="Zne";
drop Nation xcapi ;
```

```
run;
proc freq;
run;
```

2.4 Bilan de la première étape

Dresser le bilan de ce long travail de préparation visant à obtenir des données fiables, représentatives et suffisament renseignées pour décrire l'ensemble des observations sans perdre de vue l'objectif qui est de s'intéresser au patrimoine des individus, notamment la possession de certains produits financiers comme une assurance vie.

Sauf erreur, il reste alors 11887 observations et 37 variables :
- 3 quantitatives : `AsviR, Age, Nbenf`,
- 34 qualitatives : `Asvi, Sexe, Tage, Couple, Vmatri, Diplome, Ocupa, Work, statut, Herit, Pere, Mere, Gparp, gparm, Jgrav, Livep, Epalo, fepsal, vmob, livdf, pel, cel, qpep, asdecv, Retrai, Qpea, Urbani, Zeat, Nbenfq, Iogoc, terre, dette, bdetre, hdetvo`.

3 Exploration multidimensionnelle

3.1 Analyse des correspondances multiples

L'objectif est une prise en compte de toutes les variables retenues afin de mieux apprécier la structure globale des données.

```
proc corresp data=sasuser.patriminsee
     out=resul mca dim=4;
tables Asvi Sexe Tage--Zeat Nbenfq--hdetvo;
run;
%gafcix(tp=1);
%gafcix(x=1,y=3,tp=1);
%gafcix(x=2,y=3,tp=1);
%gafcix(x=1,y=4,tp=1);
```

La lecture des graphiques met en évidence certains artefacts ou correspondances triviales entre certaines modalités. Les rechercher, les expliquer pour ensuite ajouter une commande les éliminant :

```
proc corresp data=sasuser.patriminsee
     out=resul mca dim=4;
tables Asvi Sexe Tage--Zeat Nbenfq--hdetvo;
supplementary Tage Zeat Couple Mere Pere;
run;
%gafcix(tp=1);
%gafcix(x=1,y=3,tp=1);
%gafcix(x=2,y=3,tp=1);
%gafcix(x=1,y=4,tp=1);
```

3.2 Représentation quantitative des individus

Expliquer quelle analyse est calculée sur quelle matrice et avec quels objectifs.

```
data sasuser.patriminsee;
set sasuser.patriminsee;
ident=put(_N_,z5.);
run;
proc corresp data=sasuser.patriminsee
     out=resul dim=6;
tables ident , Asvi Sexe Tage--Zeat Nbenfq--hdetvo;
supplementary Tage Zeat Couple Mere Pere;
run;
proc sort data=sasuser.patriminsee out=inseetemp;
by ident;
proc sort data=resul out=resul;
by _name_;
data resul;
merge inseetemp (keep=ident Asvi
      rename=(ident=_name_)) resul ;
by _name_;
select(Asvi);
when('AsN') _name_= '1111';
when('AsO') _name_= '0000';
otherwise;
```

```
end;run;
%gafcix(tp=1.5);
%gafcix(x=1,y=3,tp=1);
%gafcix(x=2,y=3,tp=1);
```

3.3 Classification non supervisée

Expliquer la démarche permettant d'aboutir à une classification ou "segmentation" des individus de l'enquête. Pourquoi utiliser une CAH ou un algorithme de réallocation ($k$means) ? Justifier la stratégie mise en œuvre et les options retenues pour chacune de ces méthodes de classification. Ne pas hésiter à consulter le code des macros commandes !

Choisir le nombre de classes :

```
proc corresp data=sasuser.patriminsee out=resul dim=6;
tables ident , Asvi Sexe Tage--Zeat Nbenfq--hdetvo;
supplementary Tage Zeat Couple Mere Pere;
run;

proc sort data=sasuser.patriminsee out=temp;
by ident;
proc sort data=resul out=resul2;
by _name_;
data sasuser.varprinc;
merge temp (rename=(ident=_name_)) resul2 ;
by _name_;
if _TYPE_='OBS';
run;

%choixnc(varprinc,dim1-dim6,800);
%critere(20);
```

Construire les classes et les représenter.

```
%nudnc(varprinc,_name_,dim1-dim6,5,init=poles);
proc gplot data = sasuser.ndclasse;
plot dim2*dim1=classe;
run;
```

3.4 Interprétation des classes

La dernière étape vise à construire une interprétation des classes. Une démarche systématique consisterait alors à calculer les moyennes (variables quantitatives) ou les effectifs des modalités pour chacune des variables et chacune des classes. C'est plutôt fastidieux. Une autre approche, globale, consiste à recalculer une analyse des correspondances en intégrant la variable `Klasse` obtenue.

```
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='K1';when(2) klasse='K2';
        when(3) klasse='K3';when(4) klasse='K4';
    when(5) klasse='K5';
otherwise;end;
run;
proc corresp data=varprind out=resul mca dim=5;
tables Asvi Sexe Tage--Zeat Nbenfq--hdetvo klasse;
supplementary Tage Zeat Couple Mere Pere ;
run;
%gafcix;
%gafcix(x=1,y=3);
```