



HAL
open science

Effets masqués en analyse prédictive

Ganaël Bascoul

► **To cite this version:**

Ganaël Bascoul. Effets masqués en analyse prédictive. Gestion et management. Université de Strasbourg, 2013. Français. NNT : 2013STRAB015 . tel-02572980

HAL Id: tel-02572980

<https://theses.hal.science/tel-02572980v1>

Submitted on 14 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



UNIVERSITE DE STRASBOURG
ECOLE DOCTORALE AUGUSTIN COURNOT (ED 221)

THESE

**EN VUE DE L'OBTENTION DU GRADE
DE DOCTEUR ES SCIENCES DE GESTION**

Présentée par

Ganaël Bascoul

Le 27 juin 2013

Effets Masqués en Analyse Prédictive

Sous la direction du

Professeur Jacques Thépot

Professeur de Sciences de Gestion à l'Université de Strasbourg
Laboratoire de Recherche en Gestion et Économie - LARGE - EA 2364

Membres du jury (par ordre alphabétique)

Professeur Bertrand Koebel

Professeur de Sciences Économiques à l'Université de Strasbourg
Bureau d'Économie Théorique et Appliquée - BETA - UMR 7522

Professeur Ian Ross Macmillan

Professor of Policy Analysis and Public Management - Bocconi
Director of the Dondega Centre for Research on Social Dynamics
Università Bocconi, Milano, Italy

Professeur Bertrand Munier

Professeur émérite à l'IAE de Paris
Ancien Directeur de département à l'École Normale Supérieure de Cachan



UNIVERSITE DE STRASBOURG
ECOLE DOCTORALE AUGUSTIN COURNOT (ED 221)

THESE

**EN VUE DE L'OBTENTION DU GRADE
DE DOCTEUR ES SCIENCES DE GESTION**

Présentée par

Ganaël Bascoul

Le 27 juin 2013

Effets Masqués en Analyse Prédictive

Sous la direction du

Professeur Jacques Thépot

Professeur de Sciences de Gestion à l'Université de Strasbourg
Laboratoire de Recherche en Gestion et Économie - LARGE - EA 2364

Membres du jury (par ordre alphabétique)

Professeur Bertrand Koebel

Professeur de Sciences Économiques à l'Université de Strasbourg
Bureau d'Économie Théorique et Appliquée - BETA - UMR 7522

Professeur Ian Ross Macmillan

Professor of Policy Analysis and Public Management - Bocconi
Director of the Dondega Centre for Research on Social Dynamics
Università Bocconi, Milano, Italy

Professeur Bertrand Munier

Professeur émérite à l'IAE de Paris
Ancien Directeur de département à l'École Normale Supérieure de Cachan

Remerciements

Au professeur Jacques Thépot,

Professeur de Sciences de Gestion à l'Université de Strasbourg

Pour le grand honneur d'avoir été votre élève, et d'avoir pu profiter de vos conseils avisés autant sur le plan scientifique que sur le plan humain, veuillez trouver dans ce manuscrit l'expression de notre plus profond respect.

Au professeur José Allouche,

Directeur de l'école doctorale de l'IAE Paris,

Et au Professeur Bertrand Munier

Professeur émérite à l'IAE de Paris

Ancien Directeur de département à l'École Normale Supérieure de Cachan

Pour leur très honorable soutien lors d'une des périodes les plus dures de notre existence, leur sens aiguë de la morale et de l'éthique. Leur droiture guidera nos pas, et leur courage inspirera toutes nos décisions. Qu'ils trouvent ici l'expression de notre plus sincère reconnaissance.

Aux membres du jury de thèse,

le professeur Bertrand Koebel,

Professeur de Sciences Économiques à l'Université de Strasbourg

et le professeur Ian Ross Macmillan

Professor of Policy Analysis and Public Management – Bocconi

Pour l'attention portée à notre travail, leur contribution à cette réflexion, et la qualité des échanges que nous avons pu avoir et que nous aurons à l'avenir, sachez que votre éclairage marquera définitivement notre ambition d'excellence et d'ingéniosité dans nos futures fonctions d'enseignant et de chercheur.

À tous les professeurs de Mathématiques qui nous ont encouragé depuis le lycée,

Pour Madame Laurent et Monsieur Florac, du Lycée Daudet à Nîmes, Monsieur Alessandri et Monsieur Meunier du Lycée Joffre à Montpellier, qui constituent à eux seuls la preuve que l'éducation est le cœur de notre République. Qu'ils trouvent ici la trace de mon souvenir ému.

À mes Parents, Françoise et Bernard, à ma sœur Marie, et à mon frère Pierre,

Pour leur soutien sans faille, et la discipline intellectuelle qu'ils m'ont transmise.

Sommaire

Introduction générale.....	7
Première Partie: Analyse locale des critères de choix en marketing	22
A. Introduction à l'analyse locale des critères de choix.....	23
A.1. Notre approche	23
A.2. Exemples de différences entre critères de choix	24
A.2.i. Mariages limousins.....	24
A.2.ii. Footing.....	25
A.2.iii. Bière	26
A.2.iv. Interprétation.....	26
A.3. Démarche de recherche	27
B. Positionnement théorique et managérial.....	28
B.1. L'analyse locale du point de vue de la recherche en marketing	28
B.1.i. Études antérieures traitant les cas marginaux dans un objectif de prédiction.....	29
B.1.ii. Études antérieures analysant les comportements de cas marginaux après transition	31
B.1.iii. Études antérieures portant sur les critères de choix de transition des switchers	33
B.1.iv. Positionnement académique de notre recherche	34
B.2. Intérêt managérial de l'analyse locale des critères de choix.....	37
B.2.i. D'un point de vue offensif.....	37
B.2.ii. D'un point de vue défensif.....	39
B.2.iii. Positionnement managérial de notre recherche	41
C. Définitions et principe de l'analyse locale.....	42
C.1. Définitions	42
C.1.i. Définition d'effet local.....	42
C.1.ii. Définition de la frontière de choix	43
C.1.iii. Cas marginaux et cas extrêmes.....	44
C.1.iv. Mode de représentation choisi	44
C.2. Principe de l'analyse locale	46
C.3. Base de données utilisée	47
D. Première méthode : les régressions logistiques emboîtées.....	48
D.1. Régression et poupées russes.....	48
D.2. Description formelle des régressions logistiques emboîtées	49
D.3. Résultats sur les données <i>Orange Juice</i>	50
D.4. Aller plus loin que les régressions logistiques emboîtées	52
E. Méthode d'analyse des effets locaux (AEL).....	53
E.1. Orientation méthodologique	53
E.2. Différentes étapes de la méthode AEL	54
E.3. Première étape de la méthode AEL : Discrimination	54
E.3.i Flexibilité	55
E.3.ii Présentation des Machines à Vecteurs Support (SVM).....	56
E.4. Deuxième étape de la méthode AEL : Allocation au sein des bins	65
E.5. Troisième étape de la méthode AEL : Mesure des effets locaux.....	68
E.5.i. ANOVA fonctionnelle (FANOVA).....	68
E.5.ii. Calcul des effets locaux	72
E.6. Évaluation de la qualité des résultats	73
F. Applications de la méthode AEL.....	77
F.1. Cas du choix d'une marque de Jus d'orange.....	77
F.1.i. Discrimination	77
F.1.ii Calcul des probabilités et des effets locaux	80
F.1.iii. Tests d'égalité sur les effets	86
F.1.iv. Interprétation des résultats.....	94
F.2. Application de la méthode AEL sur la base Référendum	97
F.2.i. Référendum 2005	97

F.2.ii. Données d'enquête IPSOS	97
F.2.iii. Hypothèses	100
F.2.iv. Résultats et interprétation	100
G. Discussion.....	108
G.1. Effets de marges	108
G.2. Black boxes et modélisation marketing.....	109
G.3. Utilisation pour l'optimisation.....	110
G.3.i. Quelques pistes.....	110
G.3.ii. Mais quelques limites aussi	113
G.4. Implications managériales	114
G.5. Limites et futures recherches	115
G.5.i. Spécification du modèle à chaque étape	115
G.5.ii. Limites	120
G.5.iii. Futures recherches	120
Conclusion sur la méthode AEL	122
Transition entre la Partie 1 et la Partie 2.....	124
<i>Seconde Partie: Méthode APC-PLS d'estimation des effets d'âge-période-cohorte (APC)</i>	
<i>appliquée à l'analyse de la fidélité à une marque automobile</i>	
<i>127</i>	
A. Introduction à l'analyse Age-Période-Cohorte en Marketing	129
B. Double positionnement.....	134
B.1. Influences de l'âge et de la cohorte sur la fidélité à la marque	134
B.1.i. Des points de vue complémentaires.....	134
B.1.ii. Choix d'un champ d'application : la fidélité à la marque sur le marché automobile	136
B.2. Modélisation des effets d'âge et de cohorte.....	137
B.2.i. Multicolinéarité parfaite entre les variables étudiées.....	138
B.2.ii. Méthode de modélisation des effets.....	139
C. Influences de l'âge et de la cohorte sur la fidélité à la marque	142
C.1. Influence de l'âge sur la fidélité à la marque	142
C.1.i. Définitions de l'âge.....	142
C.1.ii. Influences de l'âge sur la fidélité à la marque	145
C.1.iii. Hypothèse sur les effets d'âge sur la fidélité à la marque dans le cas de l'automobile.....	151
C.2. Influence de la cohorte sur la fidélité à la marque	152
C.2.i. Définitions des cohortes.....	153
C.2.ii. Influence de la cohorte sur la fidélité à la marque	155
C.2.iii. Hypothèse sur les effets de cohortes dans le cas de l'automobile	160
C.3. Résumé et rappel des hypothèses.....	165
C.3.i. Résumé de la revue de littérature	165
C.3.ii. Résumé des hypothèses	166
D. Méthodes de modélisation des effets d'âge et de cohorte	167
D.1. Définitions et notations.....	167
D.1.i. Définitions de la multicolinéarité	167
D.1.ii. Multicolinéarité en APC	170
D.1.iii. Notations utilisées	176
D.1.iv. Définitions des effets APC	177
D.2. Illustration de la multicolinéarité APC sur la base de données France	178
D.2.i. Nos données : des Courants Commerciaux au pseudo-panel	178
D.2.ii. Description des effets bruts	181
D.2.iii. Description de combinaisons d'effets APC.....	185
D.2.iv. Critique de l'analyse des effets bruts.....	188
D.3. Méthodes de modélisation APC fondées sur les moindres carrés	191
D.3.i. Omission de variable	191
D.3.ii. Contraintes d'égalité.....	197
D.4. Méthodes de modélisation APC non fondées sur les moindres carrés	203
D.4.i. Régression sur composantes principales (PCR)	203
D.4.ii. Parcimonious Cohort Estimation (PACE).....	204

D.5. Méthodologie APC-PLS.....	210
D.5.i. Principe et étapes de la méthode APC-PLS.....	210
D.5.ii. Etape 1 : Régression PLS	211
D.5.iii. Etape 2 : Introduction de l'information a priori	215
E. Application de la méthode APC-PLS sur deux exemples.....	228
E.1. Application au cas automobile en France	228
E.1.i. Première étape de la méthode APC-PLS	228
E.1.ii. Deuxième étape de la méthode APC-PLS	229
E.2. Application au cas automobile en Allemagne.....	234
F. Retour sur les hypothèses.....	237
F.1. Rappel des hypothèses.....	237
F.2. Validation des hypothèses dans le cas automobile en France	237
F.3. Validation des hypothèses dans le cas automobile en Allemagne	238
G. Discussion.....	240
G.1. Comparaison des méthodes PCR, PACE, et APC-PLS.....	240
G.2. Comparaison des résultats obtenus.....	241
G.3. De l'usage de l'information a priori	243
G.4. Exemple complémentaire de la méthode de report des écarts.....	245
G.5. Point sur l'interprétation de v dans la méthode de report des écarts	249
H. Limites et futures recherches.....	250
H.1. Information a priori utilisée.....	250
H.2. Non comparabilité des résultats.....	250
H.3. Age, période et cohorte : « facteurs » observables	252
H.4. Extension des résultats à des biens de consommation courante	256
H.5. Substitutions possibles de la technique PLS.....	256
Conclusion sur la méthode APC-PLS	257
Conclusion générale.....	258
Bibliographie	263
Annexes.....	280
Annexe 1. Exemple complémentaire dans l'industrie du disque.	281
Annexe 2. Extrait du questionnaire Courants Commerciaux	282
Annexe 3. Diagramme de Lexis des données sur le ré-achat automobile (France).....	290
Annexe 4. Code MatLab pour la méthode PACE (Rust and Yeung, 1997).....	292
Annexe 5. Tableau récapitulatif des résultats des méthodes PCR, PACE et PLS.....	293
Annexe 6. Intervalles de sensibilité complémentaires	296
Annexe 7. Section du code MatLab pour la méthodologie AEL	301
Annexe 8. Rédaction de la démonstration concernant la méthode de report des écarts	306

Introduction générale

« A ship in port is safe, but that is not what ships are for.

Sail out to sea and do new things. »

Grace Murray Hopper, Mathématicienne, 1906-1992.

L'objectif de cette thèse est l'élaboration de deux méthodologies visant à révéler des effets jusqu'alors masqués dans la modélisation marketing. Plus précisément, dans la première partie, nous cherchons à mettre en œuvre une méthode d'analyse locale des critères de choix dans un contexte de choix binaires. Dans une seconde partie, nous mettons en avant les effets de générations dans l'étude des comportements de choix. Ces méthodologies, respectivement AEL et APC-PLS, sont toutes deux appliquées sur des cas réels afin d'en illustrer le fonctionnement et la pertinence.

Notre constat de départ a été que ces deux types d'effets, effets locaux des critères de choix et effets de génération, ont en commun de ne pas être souvent pris en compte dans les techniques de modélisation habituelles.

Concernant les effets locaux, nous pouvons prendre l'exemple d'une élection présidentielle. Au début d'une campagne, les candidats à une élection au suffrage universel se positionnent suivant une première série d'arguments, souvent proches des arguments traditionnels de leurs partis politiques, en les adaptant au contexte socio-politique de la période électorale. Cette première série d'arguments leur permet de rassembler les membres traditionnels de leurs partis respectifs et d'assurer une première base de vote. Puis dans un second temps, il s'agit de convaincre les individus n'ayant pas encore fait leur choix, que les journalistes politiques ont l'habitude de nommer les « indécis ». Si ces individus n'ont pas encore fixé leur choix, c'est parce que les arguments traditionnels des deux partis en compétition n'ont pas eu autant d'effets sur leur choix de vote que sur les choix des partisans traditionnels. Dans ce second temps, les candidats utilisent alors une seconde série d'arguments, différents de ceux de la première série d'arguments, afin de convaincre un maximum d'indécis et de remporter la campagne. Ceci signifie qu'un même argument n'aura pas le même poids auprès d'un partisan (forte probabilité de choisir un candidat) qu'auprès d'un indécis (faible probabilité de choisir

un candidat). Ainsi, en fonction de la probabilité de vote d'un électeur potentiel pour tel ou tel candidat, le pouvoir de conviction des arguments politiques diffère. C'est à partir de ce constat que nous avons décidé de définir et de mesurer les différences d'effets en fonction de la probabilité de choix à travers la notion d'effets dits locaux, en complément des effets généraux habituellement mesurés dans les fonctions de choix.

Ce même raisonnement est tout à fait applicable dans des choix plus quotidiens, comme le choix pour un consommateur entre deux marques leader sur le marché, à la nuance près que l'enjeu ne sera alors plus de remporter une élection pour un candidat, mais d'augmenter sa part de marché pour un industriel.

Dans l'état actuel des analyses sur ces questions, il y avait une différence notable entre les analyses des experts, prenant compte de ces différences locales, et les analyses statistiques. En effet, les experts en analyse politique, font souvent référence à ces nouveaux arguments progressivement mis en avant au cours des campagnes électorales afin de convaincre les indécis. Les chefs de produits savent aussi user de leviers locaux pour faire progresser la part de marché de leur produit en faisant évoluer le marketing mix de leurs produits. Pourtant les analyses statistiques usuelles sont davantage fondées sur l'hypothèse d'une fonction de décision globale. Par exemple, la régression logistique, pour expliquer un choix binaire (un candidat contre un autre, un produit A contre un produit B) propose de décrire l'effet global de chaque variable, et ce de manière indifférenciée en fonction de la probabilité de choix des individus analysés.

C'est afin de concilier ces écarts entre la pratique et la modélisation, que nous proposons d'adapter les techniques de modélisation actuelles. Comme nous le détaillerons un peu plus avant dans cette introduction, notre approche ne consiste pas à contredire les techniques de modélisation actuelles, mais de les compléter par une nouvelle étape d'analyse permettant d'en tirer un plus grand bénéfice en matière d'analyse des effets locaux, et d'ainsi pouvoir établir un dialogue plus riche entre pratique et modélisation. Les deux exemples utilisés précédemment, concernant une campagne politique et un choix entre deux marques nationales, seront repris, détaillés et analysés sur la base de données réelles dans ce travail de thèse, ce qui nous permettra d'illustrer cette approche par des exemples et de mettre en évidence l'utilité de la prise en compte de tels effets locaux en complément des analyses traditionnelles.

Concernant les effets de génération, nous partageons aussi ce premier constat. Les effets de génération sont assez peu pris en compte en modélisation, souvent masqués par les techniques de modélisation habituelles, par les effets d'âge. En effet, une partie des choix de consommation d'une personne ayant aujourd'hui 60 ans peut s'expliquer par le fait que cette personne a aujourd'hui 60 ans, une autre partie peut aussi être expliquée par le fait que cette personne est née il y a 60 ans, c'est à dire qu'elle a eu 20 ans il y a 40 ans, 25 ans il y a 35 ans... L'ensemble des influences reçues dans différents contextes historiques (maturité des marchés, principales marques présentes à chaque période, standard de consommation à chaque époque...) au cours de différentes périodes de sa vie constitue ce que l'on appelle un effet de génération. Or le choix de l'attribution de comportements de consommation à l'effet d'âge ou à l'effet de génération a des conséquences très importantes en matière de prévision des comportements à venir au sein de la population.

Illustrons ces conséquences par un exemple simple. Si les personnes âgées lisent aujourd'hui davantage la presse que ne le font les plus jeunes, est-ce du fait de leur âge ou du fait de leur génération ? Est-ce qu'il s'agit d'une activité plus adaptée aux personnes âgées qu'aux jeunes, ou est-ce que parce que dès leur jeunesse, il y a 40 ans, ces personnes ont pris l'habitude de souvent lire la presse et qu'elles conservent cette habitude jusqu'à aujourd'hui ? En d'autres termes, la lecture régulière de la presse est-elle liée à un effet d'âge ou à un effet de génération ? Dans le premier cas de figure, les jeunes d'aujourd'hui liront autant la presse que leurs aînés d'ici quelques décennies, parce qu'ils auront atteint la soixantaine et que cette habitude de consommation s'explique par l'âge. Dans le second cas de figure, les jeunes d'aujourd'hui ne liront pas davantage la presse d'ici quelques décennies, car ils conserveront leurs habitudes actuelles de consommation, qui sont celles de leur génération. Les conséquences en termes d'évolution des ventes de groupes de média à long terme sont radicalement différentes.

Beaucoup d'autres caractéristiques des modes de consommation peuvent être analysées (ou ré analysées) sous cet angle. Ceci permettrait de comprendre quelle est le poids de la composante générationnelle dans les comportements d'achat. Dans cette thèse, cette problématique est illustrée par l'étude d'une composante essentielle de la consommation en marketing : la fidélité à la marque. Plusieurs études citées dans notre revue bibliographique ont mis en évidence le fait que les personnes âgées avaient aujourd'hui une plus grande fidélité à la marque par rapport aux personnes plus jeunes. Un certain nombre de facteurs

matériels et psychologiques permettent de fonder qu'une part de ce phénomène peut s'expliquer par un effet d'âge. Cependant, comme dans l'exemple précédent des habitudes de lecture, on peut s'interroger sur le poids relatif de la composante générationnelle pour expliquer cette fidélité à la marque. Cette étude de cas sur les facteurs de fidélité à la marque dans le secteur automobile, menée dans deux pays, la France et l'Allemagne, permet de révéler qu'une partie de ce phénomène peut aussi s'expliquer par une composante générationnelle.

Si une grande majorité des études des phénomènes liés à l'âge ne prennent pas en considération la composante générationnelle, ceci peut s'expliquer en partie par le fait que dans une optique de segmentation, l'âge et la génération ne constitue qu'une seule et même variable (à savoir l'âge aujourd'hui) et qu'en l'absence d'utilisation de cette segmentation à des fins prédictives, il n'est pas nécessaire de faire une distinction entre ces deux facteurs explicatifs. Par contre, dès que l'on souhaite donner une explication et une interprétation des phénomènes observés, le fait d'ignorer la composante générationnelle constitue un risque de fortement biaiser l'évaluation du poids de la variable d'âge. De même, dès que l'on souhaite utiliser cette segmentation à des fins prédictives, la qualité des prédictions peut être largement altérées par l'absence d'une prise en compte de la composante générationnelle. Plusieurs recherches, citées dans notre revue de littérature, insistent donc sur l'importance de l'intégration des générations en modélisation.

Cependant, un second enjeu limite une telle intégration, qui est celui de la difficulté intrinsèque de modéliser simultanément les composantes d'âge et de génération, car ces deux variables forment une relation de multicollinéarité parfaite avec une troisième variable essentielle qui est la période. En effet, l'âge est égal à la période actuelle (année en cours), moins la période de naissance qui définit une génération (année de naissance). Cette relation de multicollinéarité parfaite ne permet donc pas d'intégrer telles quelles ces trois variables comme variables explicatives dans les modèles statistiques traditionnels qui ne peuvent pas gérer la multicollinéarité parfaite. Il est donc beaucoup plus compliqué de chercher à prendre en compte simultanément ces deux phénomènes, que de chercher à les considérer séparément. Cette difficulté a souvent amené à ne prendre en compte que les effets d'âge, avec comme variable de contrôle les effets de période (qui correspond la plupart du temps à l'année de collecte des données) et ainsi à négliger, pour des raisons techniques, les effets potentiels liés à la génération.

Notre second constat a été que ces types d'effets étaient peu pris en compte en gestion, notamment parce qu'ils étaient masqués par les techniques de modélisation habituelles. Comme nous l'avons partiellement abordé dans les paragraphes qui précèdent, les techniques de modélisation habituelles ne permettant pas de détecter facilement ou directement ces effets, qu'ils s'agissent des effets locaux ou des effets de générations, ils ont souvent été ignorés pour des raisons simplement techniques, et ont été masqués par d'autres effets, qui étaient eux plus facilement ou directement mesurables par les techniques disponibles. Ainsi, les effets locaux, valables sur des sous-populations précises, sont « masqués » par des effets généraux, valables sur la majorité de la population, lors de l'analyse des critères de choix.

Dans l'exemple de la campagne politique, les clivages traditionnels entre les deux principaux partis politiques ressortiront ainsi d'avantage que les critères de choix des personnes caractérisées comme étant indécises, alors que les critères de choix de ces personnes indécises sont pourtant les plus cruciaux pour remporter une élection, car ce sont ces personnes qui détermineront l'issue de l'élection. De même, lorsque l'on cherche à comprendre les évolutions des comportements de consommation dans le temps, les effets de génération sont très souvent « masqués » par les effets d'âge, qui sont plus immédiatement mesurables et interprétables, et auxquels on accorde généralement de ce fait, un trop grand pouvoir explicatif.

Ceci justifie l'appellation de « masqués » pour les effets considérés dans la présente thèse. Il s'agit donc ici non seulement de remettre en évidence l'importance de ces effets masqués, mais aussi de proposer des moyens techniques de mesurer l'importance relative de l'ensemble des effets permettant d'expliquer les phénomènes abordés.

Pour autant, nous ne souhaitons pas nous priver dans cette démarche de la puissance des méthodes de modélisation habituelles, et notre démarche consiste donc, dans les deux cas, à apporter des outils complémentaires aux outils existants afin d'enrichir les résultats habituels de résultats complémentaires sur les effets masqués que constitue les effets locaux dans les fonctions de choix binaires et les effets de génération dans l'analyse temporelle des comportements de consommation.

La question de recherche de ce travail de thèse est donc la suivante : Comment compléter avantageusement les techniques de modélisation actuelles afin de révéler et de mesurer les effets masqués que sont les effets locaux dans les fonctions de choix binaires et les effets de génération dans l'analyse temporelle des comportements de consommation ?

Pour atteindre cet objectif, notre approche a été de partir des techniques existantes pour cerner précisément le rouage technique qui ne permettait pas d'aller plus loin dans l'analyse des effets, et d'adopter un autre point de vue sur le rouage en question pour pouvoir débloquer le potentiel explicatif des modèles traditionnels.

Pour l'analyse des effets locaux dans l'analyse des choix binaires, nous avons constaté que la probabilité de choix n'était utilisée que de manière transitoire afin de mesurer les effets généraux. Notre changement de point de vue a donc consisté à ne plus considérer ces probabilités de choix comme une simple étape de calcul, mais aussi comme une source d'information en soi. Alors il devient possible de réutiliser cette précieuse information que constitue la probabilité de choix pour affiner l'analyse des résultats. Cette réutilisation de l'information sur les probabilités de choix permet d'approfondir l'analyse des résultats par tranche de population, les tranches étant déterminées par les probabilités de choix.

Pour l'analyse temporelle des comportements de consommation, nous avons constaté que toutes les méthodes disponibles pour contourner le problème de multicollinéarité parfaite consistaient à détériorer la qualité des données d'origine (les données sur l'âge, la période et la cohorte) afin de passer d'une multicollinéarité parfaite à une collinéarité forte. Pour détériorer les données initiales de manière optimale, la pratique courante consiste à utiliser une information exogène, dite information a priori, afin de réduire au minimum l'impact de la détérioration des données initiales sur le résultat final. Cependant, la multicollinéarité forte (vs. parfaite) à laquelle on aboutit ainsi comporte elle-même un certain nombre de problèmes, et peut donc constituer un frein important à l'adoption de techniques de modélisation intégrant la composante générationnelle.

Notre changement de point de vue a donc été de considérer que la politique optimale à adopter ne consistait pas à détériorer les données initiales, mais à l'utilisation d'une méthode d'estimation des effets pouvant tolérer la multicollinéarité parfaite sans que les données soient détériorées, à savoir la méthode de régression PLS. Cette méthode autorise en effet le

modélisateur à prendre comme variables explicatives un ensemble de variables parfaitement colinéaires sans que cela pose de problème d'estimation. Libre ensuite au modélisateur d'introduire, après estimation des effets, une information a priori lui permettant d'enrichir son analyse avec des informations exogènes qu'il aurait à sa disposition. Notre changement de point de vue s'illustre aussi à travers les termes employés, car l'information *a priori*, est à présent introduite *a posteriori*, et n'est donc pas un prérequis à la modélisation, mais la possibilité d'enrichir l'analyse a posteriori.

Une fois ces points clefs détectés et les nouveaux points de vue choisis, ce travail de thèse se concentre sur la mise en œuvre technique de ces apports méthodologiques. Dans les deux types d'effets abordés, les mises en œuvre ont beaucoup de points communs. Tout d'abord, nous avons fait appel à des solutions techniques encore peu utilisées en modélisation marketing : l'ANOVA fonctionnelle et les SVMs (pour Support Vector Machines) en analyse des choix binaires, et la régression PLS dans l'analyse temporelle des comportements de consommation. Issues d'autres écoles de modélisation que l'analyse économétrique traditionnelle, ces méthodes ont des atouts complémentaires aux techniques habituelles, et permettent donc d'extraire de nouvelles informations à partir des données et des techniques actuellement disponibles. Leur principes sont présentés dans le cours du texte, avec des renvois à des manuels de référence pour une présentation exhaustive, et sont ensuite appliqués et adaptés pour correspondre à la nature des problématiques marketing abordées.

Concernant l'analyse des effets locaux, la procédure méthodologique se décompose en plusieurs étapes. D'abord modéliser les choix des individus à travers un modèle prédictif, puis décomposer la population par tranches de sous-populations en fonction des probabilités de choix et enfin calculer l'impact de chacune des variables de choix pour chaque sous-population. Pour la première étape, nous avons utilisé parallèlement la régression logistique et les SVMs, issus du champ de l'apprentissage statistique, champ encore éloigné de la littérature marketing traditionnelle, afin de montrer que les deux méthodes pouvaient être utilisées dans le cadre de notre méthodologie, sachant que le modélisateur pourra ensuite choisir la technique de modélisation lui fournissant les meilleurs résultats de prédiction (en termes de validation croisée). La phase de calcul des effets relatifs sur chaque sous-population fait elle appel à une version fonctionnelle de l'ANOVA, elle aussi issue des techniques d'apprentissage statistique. Elle est appelée ANOVA fonctionnelle ou FANOVA. Dans notre cas de figure, cette méthode issue des techniques d'apprentissage statistique permet d'obtenir,

simplement à partir de la probabilité de choix et de l'ensemble des variables de choix, une expression de la probabilité de choix comme la somme de fonctions monovariées des variables de choix. Cette méthode, malgré la qualité de ses caractéristiques techniques, n'a encore été que très peu utilisée en sciences de gestion en général, et notre méthodologie constitue la première application de FANOVA à l'analyse des choix de consommation. Pour compléter cette phase de calcul, un travail sur la normalisation des effets estimés, les intervalles de confiance sur les estimations des effets, et la présentation synthétique des résultats d'une telle étude a été réalisé pour rendre la méthodologie d'analyse des effets locaux aussi accessible et ergonomique que possible.

Concernant l'analyse temporelle des comportements de consommation, notre approche consiste tout d'abord à discrétiser les variables d'âge, de période et de cohorte, puis d'estimer simultanément les effets de ces variables grâce à la régression PLS, et enfin de permettre au modélisateur d'introduire l'information a priori dont il dispose pour l'introduire a posteriori (c'est-à-dire à la suite de la phase d'estimation) dans les résultats obtenus. Concernant la régression PLS, nous l'avons privilégiée par rapport aux autres techniques de régression plus traditionnelles parce qu'elle permet de prendre en compte simultanément, malgré leur multicolinéarité parfaite, l'ensemble des données d'âge, de période et de cohorte sans avoir à les détériorer. A l'issue de cette estimation, est obtenu un premier jeu d'effets d'âge, de période et de cohorte, dans lequel aucune information a priori n'a été introduite. Afin de permettre l'introduction d'une telle information, a été développé spécifiquement dans le cadre de cette thèse, une méthodologie dite de « report des écarts successifs » permettant d'obtenir un nouveau jeu d'effets fournissant exactement les mêmes prédictions que le premier jeu d'effets initialement obtenu. L'information a priori peut prendre la forme de n'importe quelle expression fonctionnelle ayant pour paramètre les effets d'âge, de période et de cohorte. Une démonstration par récurrence de la validité mathématique de cette méthode est proposée dans le texte, et une version détaillée de cette régression est fournie en annexe. A ce jour, aucune autre technique disponible à notre connaissance ne permettait de passer d'un jeu d'effets initial à un autre sans modifier les prévisions du modèle de régression à l'origine du jeu d'effets initial. Nous ne pensons pas qu'il s'agisse de la seule manière d'effectuer une telle opération, mais celle-ci a le double avantage : de ne pas dégrader la qualité des prévisions fournies par le modèle (puisque les effets sont modifiés par l'introduction d'information a priori, mais les prédictions sont identiques) et de pouvoir introduire une grande diversité d'information a priori (toute information pouvant s'exprimer sous forme d'une expression

fonctionnelle à minimiser). Enfin, une première analyse de la robustesse des résultats par rapport à l'information a priori est proposée pour compléter cette méthodologie en permettant de tenir compte de l'incertitude que l'on peut avoir sur la qualité de cette information a priori.

Toutes ces innovations méthodologiques proviennent de la nécessité d'ouvrir le champ des possibles afin d'étendre les possibilités en modélisation marketing, tout en capitalisant sur les outils déjà disponibles en modélisation économétrique ou en apprentissage statistique. Ainsi, le lecteur tirera un plus grand profit de l'application de méthodes qu'il utilisait auparavant.

Dans le texte ces deux méthodologies portent les noms suivants :

- AEL pour Analyse des Effets Locaux, concernant l'analyse des effets locaux dans les fonctions de choix binaires
- APC-PLS pour Age-Période-Cohorte et PLS pour régression PLS, concernant l'analyse temporelle des comportements de consommation.

La mise en œuvre de ces méthodologies a nécessité la définition d'un certain nombre de concepts et de notations dont les définitions sont apportées au début de chaque section technique. A chaque fois que cela a été possible, nous avons rapproché ces méthodologies des standards utilisés en économétrie, pour ne pas définir d'outils ou de concepts de manière redondante avec l'existant. Enfin, dans le cours du texte, nous avons mis l'accent sur les raisons des choix méthodologiques effectués, et nous avons essayé de ne pas noyer le lecteur sous les notations techniques et les algorithmes, tout en préservant le niveau de rigueur scientifique qu'exige la recherche dans ce domaine. Sur ces détails les plus techniques, le lecteur expérimenté est invité à consulter les annexes et les références statistiques proposées.

D'un point de vue plus théorique, chaque méthode ouvre une brèche en modélisation : l'étude de comportements « à la marge » avec la méthodologie AEL d'une part, et l'introduction après estimation de l'information a priori pour la méthodologie APC-PLS d'autre part. Ces brèches permettent de considérer les problèmes habituels suivant d'autres priorités et d'approcher parfois différemment les choix de modélisation. Par exemple pour l'AEL, dans le cas d'une fonction de classification, on s'intéresse à sa performance à la marge, plus qu'à sa performance globale, et l'on préférera, à performance globale égale, la fonction de classification qui aura la meilleure performance à la marge, c'est-à-dire pour les personnes ayant une probabilité de choix rendant leurs choix plus difficilement prévisibles. De même

pour l'APC-PLS, le choix de la régression PLS ne s'effectue pas simplement suivant l'interprétabilité de coefficients mais davantage sur sa capacité à fournir un premier jeu d'effets qui pourra ensuite être modifié. Enfin, dans les deux cas de figure, nous avons tenté de combiner deux écoles de pensée en statistiques : l'approche économétrique et l'approche d'apprentissage statistique. Chacune de ces approches est fondée sur des principes théoriques et sur des modes de calculs différents. Le lecteur habitué à l'approche économétrique pourra d'abord être surpris par l'usage de techniques de ré-échantillonnage ou d'algorithmes a priori plus complexes que les méthodes de calcul traditionnel en économétrie. Cependant, il lui sera nécessaire de garder à l'esprit que l'apprentissage statistique constitue aujourd'hui un pan entier de la théorie statistique et que ses fondements théoriques et algorithmiques sont aussi solides que ceux de l'économétrie traditionnelle. Notre but n'est jamais d'opposer ces deux écoles, mais au contraire de mettre en avant leurs complémentarités et de démontrer la possibilité de réaliser des méthodologies hybrides, tirant le meilleur parti de chaque école. Nous espérons que ce travail de thèse contribuera, à sa très humble échelle, à de tels rapprochements qui nous semblent en tous points bénéfiques. Ces brèches méthodologiques ouvrent une réflexion sur la manière d'aborder les problèmes de modélisation. Ces points seront abordés dans le chapitre de discussion.

Les présentations des méthodologies AEL et APC-PLS sont proposées respectivement dans les Parties 1 et 2 de la thèse. Chaque partie comporte une introduction, une présentation du contexte, un positionnement théorique et managérial de la recherche, une revue bibliographique, la question de recherche, une définition des principes et des définitions, une description des méthodologies antérieures, une présentation complète de la méthodologie et de ses démonstrations, des études de cas détaillées, une discussion et une conclusion.

Concernant la méthodologie AEL, la partie débute par une introduction à l'analyse locale des critères de choix (Section A), illustrée de trois exemples concrets et une présentation de la démarche. S'ensuit en section B une description du positionnement académique de la méthodologie AEL par rapport aux recherches antérieures en matière d'analyse locale des choix binaires ; ainsi qu'un positionnement du point de vue managérial, pour expliquer comment la problématique des effets locaux peut être utilisée par les entreprises soit pour conquérir des parts de marché en optimisant leur performance commerciale d'un point de vue coût/bénéfice, soit pour protéger leur part de marché actuelle vis-à-vis de la concurrence.

La troisième section (notée C) fournit l'ensemble des définitions et des notations nécessaires pour aller plus loin dans l'analyse des effets locaux. Seront notamment abordés dans cette section, les définitions d'un effet local, d'une frontière de choix, de laquelle découle les définitions des cas marginaux et des cas extrêmes, correspondant dans l'exemple d'une campagne électorale aux indécis et aux partisans. C'est aussi dans cette section que sera décrite la base de données Orange Juice utilisée tout au long de la Partie 1 pour illustrer les développements techniques. En effet, afin d'illustrer chaque étape de la méthodologie AEL par des éléments concrets, un cas général sera traité au fil de la présentation de la méthodologie.

Dans la section D, une première approche des effets locaux, que nous avons appelée régressions logistiques emboîtées, met en évidence de manière concrète et chiffrée l'analyse des effets locaux sur la base de données Orange Juice. Cette première approche permet aussi de révéler les limites d'une approche uniquement basée sur la régression logistique et souligne ainsi la nécessité d'innover pour mesurer les effets locaux.

Le cœur de l'étude est exposé dans la section E qui débute par une présentation de la méthode AEL et de ses différentes étapes. La première étape de discrimination (prévision des choix des individus selon un choix binaire) est faite par l'analyse discriminante de Fisher, mais aussi les Support Vector Machines, méthodologie issue de l'apprentissage qui nécessite un développement spécifique. S'ensuivent une deuxième étape d'allocation des individus selon des tranches de probabilités appelées bins, et une troisième étape de mesure des effets locaux par la méthodologie FANOVA. Cette section se conclut par des calculs de mesure de la fiabilité des résultats. La section F présente des applications de cette méthodologie sur deux problématiques déjà présentées plus avant dans cette introduction : l'analyse du choix d'achat entre deux marques leader à partir des données de la base Orange Juice, qui est une base de données publique, publiée dans un manuel de référence américain, et l'analyse des résultats d'une campagne politique à partir des données IPSOS sur les votes au référendum de 2005 en France. Pour tester la significativité des résultats, certains développements spécifiques aux méthodes d'apprentissage statistique sont inclus dans ces études. En effet, contrairement à la plupart des approches en économétrie, les méthodes d'apprentissage statistique ne font pas d'hypothèses aussi fortes sur les lois de distribution sous-jacentes des phénomènes étudiés, et il est donc ici nécessaire de réaliser des tests de normalité des distributions, avant d'obtenir les paramètres de ces distributions empiriques utilisés pour les tests d'égalité. Si cette

démarche peut sembler plus longue, elle permet de sécuriser les résultats obtenus, car tous les calculs sont effectués à partir de distributions qui ont été étudiées en détail, ce qui n'est pas toujours le cas en économétrie traditionnelle où de très fortes hypothèses sur les distributions sont trop souvent considérées comme validées a priori sans être vérifiées. Il s'agit d'une des questions abordées dans la discussion de la section G.

Les résultats de ces applications illustrent comment utiliser cette méthodologie. Ils prouvent aussi que ces effets locaux existent, et que la méthodologie AEL permet de les mesurer et de les tester, pour savoir s'il existe des différences significatives entre les importances relatives des critères de choix de plusieurs sous-populations caractérisées par leur probabilité de choix. La méthodologie AEL, satisfait une des conditions exposées dans la question de recherche de cette thèse, c'est-à-dire qu'elle peut s'adapter à toutes les techniques actuelles d'analyse de choix binaires. Nous prenons ici l'exemple de l'analyse discriminante de Fisher, mais toute technique de discrimination ayant pour sortie une probabilité de choix par individu peut être combinée avec l'AEL, qui fournira des résultats complémentaires sur l'importance locale des critères de choix pour les individus étudiés. Cette flexibilité sera mise en avant plusieurs fois dans le texte parce qu'elle constitue un des atouts de la méthodologie et de sa généralisation.

Différentes questions, comme celle de l'étude des hypothèses sur les distributions statistiques, ou celle des forts préjugés de la plupart des économétristes envers les « black boxes », nous éclairent sur les complémentarités que l'on peut espérer du rapprochement entre l'école d'économétrie et celle de l'apprentissage statistique, autant en termes d'outils, que d'atouts méthodologiques. Elles sont exposées dans la section G qui contient la discussion de la méthodologie AEL. Avant de conclure, cette partie aborde les implications managériales de la prise de conscience des effets locaux, les limites de la méthodologie proposée, ainsi que les perspectives de recherche pour aller plus loin dans la découverte et l'analyse des effets locaux.

La méthodologie APC-PLS est présentée dans la Partie 2 de cette thèse. Le sigle APC correspond à âge (A), période (P) et cohorte (C), qui sont les trois variables utilisées dans cette approche pour effectuer une analyse temporelle des comportements de consommation. La cohorte, qui regroupe toutes les personnes ayant la même année de naissance, est utilisée pour mesurer les effets de génération. La génération regroupe en général les personnes nées suivant une tranche de plusieurs années successives (une décennie, par exemple les années 80,

ou une période historique, comme l'après-guerre). Du fait que ces découpages par génération soient arbitraires, et que la cohorte donne un niveau de précision plus fin, c'est en général la cohorte qui est retenue comme base de travail dans les études économétriques sur ce thème, et c'est sur cette base que nous travaillerons.

Le focus de cette étude est autant managérial, par l'étude de la fidélité à la marque, phénomène temporel par nature, que méthodologique. L'exemple précisément étudié sera la fidélité des consommateurs à une marque automobile lors d'un nouvel achat. Cet exemple est développé sur les marchés français et allemands. Cependant les enjeux méthodologiques abordés, et les solutions proposées, sont applicables à l'étude de tout comportement de consommation, pour lequel les données de comportement et d'âge sont disponibles sur plusieurs périodes successives (en effet, à partir de l'âge et de la période, on peut déduire la cohorte, utilisée pour mesurer les effets de génération).

La section A présente les enjeux généraux des modélisations APC, autant sur le plan de leur importance managériale pour prévoir les comportements à venir, que sur la difficulté de mesurer simultanément les effets de ces trois variables du fait de leur multicolinéarité parfaite.

La Section B positionne cette recherche au sein des deux champs auxquels elle appartient : celui de la modélisation des effets d'âge et de cohorte, et celui de l'étude des facteurs temporels déterminant la fidélité à la marque.

La section C présente une revue de bibliographie sur les effets de l'âge et de la cohorte sur la fidélité à la marque. Pour les raisons que nous avons évoquées plus haut, l'âge est une variable davantage étudiée en marketing. Plusieurs définitions de l'âge existent dans la littérature : l'âge chronologique, l'âge subjectif ou encore la période du cycle de vie, et nous retiendrons l'âge chronologique qui est le plus objectif. Les effets de l'âge sur la fidélité tels que le vieillissement biologique, le déclin cognitif, l'aversion au changement ou la sélectivité socio-émotionnelle seront présentés. Nous déduisons de cette revue de littérature des hypothèses à vérifier sur l'effet de l'âge sur la fidélité à une marque automobile. Concernant la cohorte, une courte présentation théorique présentera quelles formes peuvent prendre les effets de la cohorte dans les représentations graphiques d'études des comportements, après avoir repréciser les différences de définitions entre cohorte et génération. Puis, plus spécifiquement, les principaux effets de la cohorte sur les comportements de fidélité seront

présentés à travers les concepts de conservatisme, de nostalgie, et de self-consistency. Nous déduirons là encore de cette autre revue de littérature des hypothèses à vérifier sur l'effet de l'âge sur la fidélité à une marque automobile. A la fin de cette section C un résumé des hypothèses à tester est proposé.

Après une section C portant sur les hypothèses d'impacts des variables APC sur la fidélité, la section D aborde l'autre pendant de cette recherche, à savoir la complexité méthodologique des modélisations de type APC. Cette section aborde les sujets de la multicollinéarité parfaite et les méthodes qui peuvent être utilisées pour gérer cette multicollinéarité.

Après une définition et des illustrations de la multicollinéarité parfaite dans le cas des variables d'âge, de période et de cohorte, notamment à travers les diagrammes de Lexis, sont données les définitions des effets bruts et nets des variables APC. Pour illustrer ces effets sur des données concrètes, notre base de données issue de l'enquête sur les achats de voitures neuves est décrite, avec une attention particulière portée aux variables APC et la variable de ré-achat d'une même marque automobile. S'ensuit une description des effets bruts, et une critique d'une analyse temporelle des comportements basée uniquement sur les effets bruts, ne permettant pas de distinguer quelle part de l'effet d'âge est en réalité liée à la cohorte, et inversement. Après cette présentation de la modélisation des effets APC suivant leurs effets bruts, l'étude présente en détail, et à l'aide de plusieurs exemples, les méthodes de modélisation APC fondées sur les moindres carrés, qui imposent une dégradation des données en amont, et enfin des méthodes de modélisation APC non fondées sur les moindres carrés, à savoir la méthode PACE (pour *Parcimonious Cohort Estimation*), qui ne permet pas l'introduction d'information a priori, et enfin la méthode APC-PLS qui permet l'introduction d'information a priori sans toutefois dégrader les données d'origine. Cette méthode comprend deux étapes majeures décrites plus haut, à savoir une première estimation des effets sans détérioration des données initiales, puis une introduction a posteriori, c'est-à-dire après une première estimation des effets, de l'information a priori dont nous disposons.

Ensuite, cette méthodologie APC-PLS est appliquée aux marchés de l'automobile français et allemand au sein de la section F. Ces applications se décomposent toujours suivant la même procédure comprenant une estimation des effets d'âge, de période et de cohorte, suivie d'une étude de la sensibilité des résultats à l'information a priori utilisée à travers des intervalles dits de sensibilité. Ces applications mènent à des conclusions robustes, et cohérentes entre les

marchés français et allemand, quant à la dimension temporelle des comportements de fidélité sur ces deux marchés. Il est alors possible de revenir dans la section F sur les hypothèses de recherche résumées à la fin de la section C, et de conclure que ces hypothèses sont validées tout en relevant certaines nuances intéressantes entre les résultats des marchés français et allemands, qui peuvent s'expliquer par des contextes historiques différents entre les industries dans ces deux pays.

Enfin la section G est consacrée à une discussion portant sur la comparaison des résultats obtenus suivant les différentes méthodes abordées et sur une réflexion à propos de l'usage de l'information à priori. La comparaison des résultats fait apparaître une cohérence des résultats, quelle que soit les méthodes employées, et malgré les différences théoriques de fond entre ces méthodes, ce qui permet d'à nouveau consolider la validation des hypothèses formulées. Quant à la discussion sur l'information a priori, elle met en avant les avantages méthodologiques à n'introduire l'information a priori qu'après l'estimation des effets, car cette introduction permet à la fois de ne pas avoir à détériorer les informations initiales dont nous disposons, et de ne pas avoir à se fixer des contraintes arbitraires, telles que la parcimonie des effets à mesurer, ce qui interdit toute introduction d'information exogène autre que la dite parcimonie. Pour conclure, la section H met en avant les limites de ce travail et les pistes de futures recherches. Ces limites sont intrinsèques à toute modélisation sur des phénomènes APC, du fait de la multicollinéarité parfaite et des problèmes d'identification qui en découlent, mais cela n'empêche pas pour autant d'essayer constamment d'améliorer le traitement de ces questions qui sont d'une importance primordiale d'un point de vue autant théorique que managérial, c'est pourquoi d'autres pistes de futures recherches sont évoquées en ouverture.

A travers ces deux parties, nous avons cherché à révéler des effets jusqu'alors souvent masqués par les techniques de modélisation actuelles, en développant des méthodologies complémentaires aux méthodologies existantes, dans une posture d'ouverture et d'innovation, et ce, afin de répondre à notre question de recherche à l'aide de tous les outils disponibles, y compris dans des champs connexes à la modélisation marketing traditionnelle. Notre but est ici en effet de contribuer à l'évolution des pratiques de modélisation en marketing, en gardant à l'esprit qu'une méthodologie doit proposer, au-delà du simple outil, une manière d'aborder les questions que pose le marketing aujourd'hui.

Première Partie: Analyse locale des critères de choix en marketing

A. Introduction à l'analyse locale des critères de choix

La modélisation marketing porte en grande partie sur les critères de choix dans les comportements d'achat dans le but de mieux comprendre les attentes des consommateurs et d'améliorer l'efficacité de l'action marketing, depuis la conception des produits ou services jusqu'à leur mode de distribution et de vente.

Nous allons ici présenter la manière dont nous avons choisi d'étudier ces critères de choix. Elle consiste à donner une *interprétation plus approfondie d'un modèle* de classification par l'analyse locale des caractéristiques du modèle. Par analyse locale, nous entendons l'analyse suivant le caractère plus ou moins prévisible du choix des individus, c'est-à-dire suivant le degré de certitude que nous estimons à propos de l'issue de leurs choix.

A.1. Notre approche

Cette étude porte en effet sur un aspect particulier de l'étude des critères de choix qui est l'analyse de ces critères en fonction de la probabilité de choix dans le cadre d'une décision binaire, telle qu'adopter ou non un nouveau produit, ou choisir entre une marque A ou une marque B. En effet, les techniques statistiques actuelles nous permettent souvent d'estimer la *probabilité de l'issue d'un choix binaire*. Ainsi il est possible de distinguer les individus dont le choix est très prévisible (dont la probabilité de choix est proche de 0 ou 100%), des individus dont le choix est plus difficile à prédire (dont la probabilité de choix ne nous permet pas d'avoir une prévision très tranchée de l'issue de leur choix).

Cette seconde catégorie d'individus a un grand intérêt en marketing, mais constitue aussi un défi en modélisation des critères de choix. Son intérêt vient, comme nous le verrons plus en détail par la suite, du fait qu'ils constituent la première cible marketing pour augmenter la part de marché d'une marque à la marge. En effet, les individus dont les choix sont très prévisibles sont déjà très convaincus par une des options proposées par le marché et seront donc a priori beaucoup plus difficiles à convaincre. Alors que les individus encore indécis sont les plus susceptibles d'être convaincus si l'on parvient à comprendre plus précisément leurs critères de choix. Et c'est alors qu'intervient le défi méthodologique consistant à ne plus considérer la population dans son ensemble, mais à passer à un niveau d'analyse plus local dans lequel il

sera possible de distinguer l'importance relative de chaque critère suivant les individus, et en particulier suivant les individus étant encore aujourd'hui les plus indécis.

C'est tout le but de cette recherche sur la détection de critères de choix locaux par *l'analyse a posteriori* de modèles de classification. L'expression « locale » faisant référence au fait que l'on s'intéresse à des groupes d'individus catégorisés suivant leurs probabilités de choix.

Nous allons donner quelques exemples montrant comment les poids des critères de choix peuvent varier entre les individus dont les choix sont les plus prévisibles, appelés *cas extrêmes*, et les individus dont les choix le sont moins, appelés *cas marginaux* car ils se situent à la marge de la frontière de décision entre les deux options du choix binaire considéré.

A.2. Exemples de différences entre critères de choix

Ces trois exemples ont pour but de donner une idée plus précise des différences d'influence de différents critères de choix suivant les individus.

A.2.i. Mariages limousins

Dans une société traditionnelle comme le Limousin rural du XIX^{ème} siècle, l'âge était sans doute le meilleur critère pour prédire si un homme était marié ou non, puisque la plupart des hommes de plus de trente ans étaient mariés, et que très peu d'hommes de moins de 18 ans ne l'étaient pas. L'âge est donc un critère de choix important pour les cas extrêmes, dont les choix sont facilement prévisibles.

Cependant, sur la tranche d'âge 18-30, les hommes mariés représentaient la moitié de la population masculine, et la variable « âge » n'avait aucun réel pouvoir prédictif pour cette sous-population. Par contre, savoir si un homme a été enrôlé dans l'armée pour sept ans de service (il s'agissait d'un processus de conscription aléatoire) est un critère explicatif déterminant: en effet, les hommes de cette tranche d'âge n'étaient pas mariés s'ils avaient été conscrits, et généralement mariés sinon. La variable d'enrôlement est donc un bon critère

explicatif pour les cas plus marginaux, dont le comportement est a priori moins déterminé par la variable « âge ».

A.2.ii. Footing

Un autre exemple serait de tenter de prédire si un individu choisira de faire ou non un footing le week-end prochain.

Le meilleur critère de prévision pour les cas extrêmes sera très probablement sa fréquence de footing sur les six derniers mois – une mesure proche de celle de la fidélité mesurée suivant Guadagni et Little (1983). Ceux qui ont déjà fait un footing tous les week-ends sur les six derniers mois en feront très probablement un à nouveau le week-end prochain, alors que ceux qui n'en ont jamais fait sur cette même période auront très peu de chances d'en faire un. La variable de fréquence est donc un très bon critère d'explication pour ces cas extrêmes.

Cependant, pour ceux qui ont fait un footing en moyenne un week-end sur deux sur les six derniers mois, le meilleur critère possible ne sera probablement pas de savoir si leur fréquence est plutôt de 40% ou 60%. D'autres critères, inefficaces sur la prévision des cas extrêmes, pourraient s'avérer pertinents pour comprendre les choix des cas marginaux.

Prenons par exemple le temps qu'il fera le week-end prochain : si le temps est bon, ils feront leur footing, si le temps est mauvais, ils ne le feront pas, ou du moins avec une bien moins grande probabilité.

Inversement, les personnes ne faisant jamais de footing, n'en feront a priori pas un le week-end prochain, quelle que soit la météo. Et les habitués d'un footing ne seront pas perturbés dans leurs habitudes par une météo peu clémente.

Le temps qu'il fera ce week-end est donc un bon critère d'influence à prendre en compte pour les cas marginaux, mais pas pour les cas extrêmes.

A.2.iii. Bière

Notre dernier exemple porte sur l'achat de bière lors d'un passage au supermarché. Ici encore, les choix des cas extrêmes (gros consommateurs de bière et individus n'en consommant jamais) peuvent être expliqués par leurs comportements passés, c'est-à-dire la fréquence d'achat lors des derniers passages au supermarché.

Parmi ceux dont la fréquence d'achat est autour de 50%, c'est-à-dire les consommateurs occasionnels, appelés ici cas marginaux, cette fréquence sera moins déterminante.

Par contre, la présence d'une offre promotionnelle lors de leur passage au supermarché pourrait être un bon critère de choix pour les cas marginaux, alors qu'il n'aurait eu qu'une faible influence dans le choix des cas extrêmes. En effet, les individus ne buvant pas de bière n'en achèteront pas, même s'il y a une promotion exceptionnelle. Et les gros consommateurs de bières en achèteront dans tous les cas, et s'il y a une promotion lors de leur passage, ils profiteront d'un effet d'aubaine.

A.2.iv. Interprétation

Remarquons ici que, dans ces trois exemples, il semble y avoir schématiquement deux types de critères.

Ceux qui peuvent aider à interpréter les choix des cas extrêmes, mais qui sont assez peu explicatifs concernant les choix des cas plus marginaux. Et inversement, les critères qui a priori n'auraient pas aidé à comprendre les choix des cas extrêmes (ici le service militaire, le climat ou l'offre promotionnelle), mais qui s'avèrent intéressants dans l'interprétation des choix des cas marginaux.

Les critères de choix se révèlent donc ne pas être aussi pertinents pour tous les membres de la population. Ainsi il s'avère ici intéressant d'étudier les critères de choix plus localement, c'est-à-dire suivant des sous-populations que nous définirons par rapport à la prédictibilité de l'issue de leurs choix.

A.3. Démarche de recherche

Les méthodes de prévision habituelles fournissent une évaluation de l'importance globale de chaque critère de choix, quel que soit le niveau de probabilité de choix. Notre approche vise à approfondir l'interprétation des résultats de ces méthodologies de classification en tenant compte de ce niveau de probabilité pour évaluer l'importance des critères.

Il s'agit ainsi d'apporter un diagnostic *a posteriori*, comme il peut en exister sur d'autres enjeux de modélisation. Ainsi d'autres méthodologies étudient l'influence d'individus pris isolément sur une fonction de discrimination linéaire (Critchley et Vitiello, 1991), ou quadratique (Kung 1996), avec une illustration graphique de ces diagnostics (Weiss et Cook, 1992). Notre démarche permet elle, d'analyser, plus en profondeur les modèles de classification afin de mettre en évidence *des différences d'influence entre critères de choix*.

Cette démarche est assez innovante dans le sens où, malgré son importance potentielle, elle est peu abordée en modélisation marketing. La section B situe cette problématique par rapport à la recherche en marketing et aux pratiques managériales. Une fois les nouvelles définitions nécessaires à cette étude posées dans la section C, nous présenterons deux approches développées pour analyser localement les critères de choix :

- La technique des régressions logistiques emboîtées, consistant à réutiliser l'outil traditionnel de régression logistique en se focalisant sur les cas marginaux, présentée dans la section D.
- Et, l'analyse des effets locaux, dont l'ambition est d'utiliser des innovations statistiques récentes pour aborder cette question sous un nouvel angle, présentée dans la section E.

Dans la section F, deux cas d'application complets seront présentés pour illustrer notre démarche, l'un portant sur le choix d'une marque de jus d'orange, l'autre sur les choix de vote lors d'un référendum. Enfin les apports, limites et pistes de recherche futures seront discutés dans la dernière section.

B. Positionnement théorique et managérial

Cette recherche peut être située suivant deux aspects majeurs :

- la recherche académique en marketing,
- l'intérêt de l'analyse locale d'un point de vue managérial.

B.1. L'analyse locale du point de vue de la recherche en marketing

Notre revue de littérature sur le thème nous a permis de relier notre problématique à celle de l'opposition souvent faite en marketing entre les comportements des *switchers*, terme anglo-saxon pour désigner les cas ayant déjà migrés d'une décision à l'autre durant plusieurs choix consécutifs, et les *stayers*, terme relatif aux cas qui ont tendance à reproduire leurs choix précédents et donc à maintenir leurs positions passées.

L'étude de ces recherches donne un premier aperçu de la nécessité d'une analyse *locale* des critères de choix. Ces recherches peuvent être regroupées en deux approches, suivant deux objectifs. L'un serait de prédire les transitions potentielles des individus, l'autre d'analyser les comportements et les critères de choix des *switchers*, sans nécessairement chercher à les prédire. Nous présentons ces deux courants avant d'expliquer comment notre étude se situe par rapport à ces recherches précédentes.

Signalons dès à présent que les *switchers* (resp. *stayers*) ne sont qu'un exemple particulier de nos cas marginaux (resp. cas extrêmes), car nos définitions ne nécessitent pas d'avoir observé d'autres occurrences de choix individuel par le passé, et peuvent donc s'appliquer à des choix uniques, comme nous le verrons dans l'étude du choix de vote lors d'un référendum.

En effet, la différence méthodologique de notre approche par rapport aux recherches antérieures consiste bien à se fonder sur l'observation de choix individuels uniques et non de choix répétés. Comme nous le verrons, cette approche a l'avantage de pouvoir être appliquée à un plus large ensemble de données (les observations de choix répétés, ou données de panel, étant moins fréquentes que les observations de choix uniques, notamment dans les sondages). Par contre, notre approche se situe davantage dans une optique explicative que dans une optique directement prédictive, justement parce que nous n'avons pas accès à plusieurs occurrences pour un même choix individuel. L'application de notre approche à des bases de

données contenant l'observation de choix individuels répétés pourra faire l'objet de recherches ultérieures, dans un démarche de généralisation, comme nous le proposons dans la discussion de cette section.

B.1.i. Études antérieures traitant les cas marginaux dans un objectif de prédiction

Nous allons présenter plusieurs études analysant les cas marginaux dans un objectif de prédiction de transition et positionner notre recherche par rapport à ce courant d'études.

Dans leur ouvrage de synthèse sur la segmentation en marketing, Wedel et Kamakura (2000) expliquent en quoi les cas marginaux sont plus difficiles à catégoriser dans un segment précis, et amènent donc à voir la segmentation de manière plus dynamique, en tenant compte des transitions possibles d'un segment à l'autre. Ils évoquent notamment l'étude de Kamakura, Kim et Lee (1996) portant sur les changements de décision (et donc de segment) au cours du temps grâce à l'analyse de données de scanner sur une période de deux ans.

Dans la même optique, plusieurs recherches élaborent des modèles bayésiens à classes latentes afin d'analyser les taux de transition entre segments. Ainsi, Poulsen (1996) utilise différentes formulations de modèles à classes latentes pour étudier les transitions entre deux marques de thon en boîte et comprendre si le phénomène étudié est stationnaire ou non. Il conclut que le phénomène est stationnaire au sens où les probabilités de choix et les taux de transition sont constants au cours du temps, et donc que malgré un mouvement de va-et-vient autour de la frontière de décision entre les deux marques, les tailles des deux segments restent constantes. Dans notre cadre d'étude, ceci s'interprète par le fait qu'au cours du temps, la position des cas marginaux (définie par leur probabilité de choix) ne varie pas. Cependant ceci ne nous permet pas de comprendre à quels critères ils sont sensibles.

Ramaswamy (1997) analyse l'impact de l'introduction de deux nouvelles marques sur le marché des lessives à l'aide d'un modèle dans lequel chaque segment est composé d'une part de switchers et d'une part de stayers. Cette analyse lui permet d'expliquer les taux de transition entre les anciennes et les nouvelles marques entre deux périodes (avant et après l'introduction de ces deux nouvelles marques) en calibrant correctement la part de switchers dans chaque segment. On peut ainsi différencier les segments entre les plus solides (qui ont

une faible part de switchers et sont donc moins touchés par l'introduction de nouvelles références) et les segments plus vulnérables. Là encore, l'étude permet une explication des taux de transition entre segments, à l'aide de la part de switchers évaluée au sein de chaque segment, mais ne nous explique pas les raisons pour lesquelles ces transitions ont eu lieu, c'est-à-dire les critères de choix des switchers repérés.

Notre recherche se différencie de ces précédentes recherches sur trois points principaux :

- Alors que ces études ont été réalisées par rapport à des choix répétés, notre étude ne porte pas seulement sur des choix répétés, mais peut permettre d'analyser des choix uniques. Ainsi la dénomination de switchers contraint à l'étude de transition d'un choix à un autre, d'un segment à un autre, alors que nos cas marginaux ne sont pas nécessairement envisagés suivant leurs choix passés, mais plutôt dans le cadre d'un choix unique bien précis. C'est pourquoi les études précédentes se focalisent sur des taux de transition calculés à partir de *données longitudinales*, alors que notre approche est applicable sur des *données transversales* ;
- L'objectif principal de ces études est de prédire ces choix de manière stochastique, sans faire appel à des variables explicatives pour prévoir ces choix, alors que notre étude porte précisément sur *l'analyse et l'interprétation de ces variables explicatives*, que nous nommons critères de choix. Notre objectif principal est de mieux comprendre la fonction de classification permettant de prédire ces choix, en examinant les poids respectifs des critères de choix ;
- Enfin, notre approche ne consiste pas à se substituer aux modèles de classification existants, mais d'en approfondir a posteriori l'interprétation en les examinant plus localement. Elle vise donc à faire un *diagnostic local d'une analyse préalable globale*.

En résumé, nous pouvons dire que de Colombo et Morrison (1989) jusqu'à Wedel et al. (1995), le but a été d'analyser les transitions entre marques. L'optique générale de ce type d'étude est de comprendre la structure et l'évolution d'un marché suivant ses différents segments ou ses différentes marques à travers les transitions observées, sans tenir compte des critères de choix caractérisant ces segments et ces marques. Ce courant est suivi de nos jours, comme dans les deux articles détaillés précédemment. Cependant notre optique est, elle, axée

sur l'analyse des critères en eux-mêmes et de leur importance relative pour comprendre les probabilités de transition observées.

B.1.ii. Études antérieures analysant les comportements de cas marginaux après transition

Plusieurs études dans le cadre de services en souscription caractérisent les switchers comme les clients dernièrement acquis par un fournisseur, par rapport aux stayers, clients de plus longue date. Dans ce cadre, l'analyse des switchers ne porte plus sur la prédiction de leur transition, déjà observée, et très peu sur leurs critères de choix, mais avant tout sur leurs comportements une fois leur choix réalisé.

Dans le cadre de l'industrie des services¹, Ganesh, Arnold et Reynolds (2000) catégorisent la clientèle d'une entreprise en deux groupes : les switchers, qui proviennent d'un autre fournisseur de services, et les stayers, dont le premier fournisseur est encore le fournisseur actuel. Les switchers sont encore subdivisés entre switchers ayant migré pour cause d'insatisfaction et les autres (e.g. en raison d'un déménagement ou d'un changement d'emploi). Ces auteurs cherchent à étudier les attitudes et comportements de ces deux types de switchers pris séparément. Ainsi, alors que les switchers sont souvent considérés comme moins importants et que la plupart des recherches précédentes conseillent d'y accorder moins d'efforts qu'à la rétention des stayers, les auteurs de cette étude établissent que des efforts de rétention les concernant seraient pourtant judicieux pour la première catégorie de switchers (les consommateurs insatisfaits par leur précédent fournisseur). En effet, ces switchers sont en général plus satisfaits que les autres switchers (et même que les stayers), mais ont surtout un plus grand niveau de fidélité que les autres switchers.

Dans le cadre de notre recherche, cette étude a un double intérêt.

D'abord elle démontre que les switchers ayant des critères non satisfaits par leur précédent prestataire, peuvent être conquis par un autre prestataire. Le repérage de ces critères est donc crucial pour augmenter sa part de marché.

De plus, ces cas marginaux ne sont pas nécessairement très volatils, une fois mieux ciblés, et peuvent même constituer des clients très fidèles à l'avenir.

¹ Les deux études de l'article concernent le secteur bancaire.

Dans la même perspective que Ganesh et al. (2000), Wangenheim et Bayon (2004) étudient les différences de comportement entre les switchers et les stayers au sein de la base clients d'un fournisseur d'énergie allemand. Ils distinguent eux aussi deux types de switchers: ceux recrutés par un système de parrainage et ceux recrutés à l'issue de campagnes de publicité ou de marketing direct. Les variables étudiées pour différencier ces différents sous-groupes sont la satisfaction, la fidélité et le bouche-à-oreille.

D'après leur étude, les switchers diffèrent des stayers par un bouche-à-oreille plus positif, et un niveau de fidélité active plus élevé, mais un niveau de fidélité réactive moins élevé². Les switchers recrutés par recommandation diffèrent des autres switchers par des niveaux de satisfaction et de fidélité active plus élevés, et un bouche-à-oreille encore plus positifs. Ceci fait d'eux de meilleurs ambassadeurs de l'entreprise auprès de ses autres potentiels prospects que ne le sont les stayers, à condition de rester vigilant à conserver son avantage compétitif par rapport aux concurrents (à cause du niveau de fidélité réactive plus faible de ces switchers).

Dans une politique de conquête de part de marché, et à condition d'être capable de conserver sa position dans le jeu concurrentiel, l'étude des switchers s'avère donc encore plus déterminante que celle des stayers.

Comme le font remarquer Payne and Frow (1997) sur le cas anglais cette fois, la dérèglementation récente des marchés de l'électricité a amené les fournisseurs à engager deux types de stratégies complémentaires : l'une consistant à retenir les clients existants, l'autre à en acquérir de nouveaux. Pour ces deux stratégies, les cibles à analyser en priorité sont les cas marginaux, qui sont soit les clients du fournisseur les plus susceptibles de passer à la concurrence, soit les clients de la concurrence, les plus susceptibles d'être acquis par les autres fournisseurs. Là encore, les cas marginaux sont au centre de la stratégie marketing.

Par rapport aux recherches présentées dans la sous-section précédente, ces études comparent ce qui différencie les switchers des stayers en termes de comportements post-transition, et mettent en avant le potentiel atout marketing que constitue l'acquisition des switchers. On

² Les auteurs font p.216 une différence entre fidélité réactive, consistant à rester fidèle malgré une modification de la donne concurrentielle (e.g. hausse des tarifs du fournisseur ou baisse des tarifs d'un concurrent), et la loyauté active consistant à ne pas prévoir de changer de fournisseur et à promouvoir l'offre de son propre fournisseur.

peut en déduire l'importance de l'étude des critères de choix de ces switchers en les considérant comme une population aussi stratégique que celle des stayers.

B.1.iii. Études antérieures portant sur les critères de choix de transition des switchers

Ce courant est celui le plus proche de notre approche, car il ne considère ni la probabilité de transition en tant que telle (sous-section B.1.i), ni les différences de comportement après transition (sous-section B.2.ii.), mais les critères de choix et donc de potentielle transition des cas marginaux. Nous allons cependant constater que si ces premières études constituent une très bonne entrée en matière, il existe une différence fondamentale entre leur approche et la nôtre. Ceci nous permettra de situer précisément notre recherche par rapport aux trois types d'études présentés et de passer à la présentation de l'intérêt managérial d'une telle recherche.

Dès les années soixante-dix, Calantone et Sawyer (1978) mettent en évidence le fait que les phénomènes de changement de décision restent une dimension sous-étudiée en marketing. Un des premiers articles, à notre connaissance, traitant des critères de choix des switchers est Zufryden (1986). Son modèle explique les taux de transition d'une matrice de Markov de premier ordre à partir de variables explicatives externes. L'estimation demande d'utiliser des moindres carrés pondérés, mais l'utilisation générale du modèle est très accessible. Il s'agit d'une manière très pertinente d'analyser les critères de choix des switchers.

Peng et Wang (2006) soulignent le manque persistant de recherches sur les cas marginaux, et ce en particulier sur leurs critères de choix alors que d'après eux :

« Theory suggests that switchers and stayers differ significantly in their perceived level of impact of the various marketing tactics on their choice of service providers. » (p.25)

Les différences qu'ils mettent en avant à l'issue de leur revue de littérature entre les critères de choix des switchers et des stayers sont les suivantes. Les stayers accordent plus de poids aux critères de qualité de service, de valeur du service fourni, de réputation du fournisseur ;

alors que les switchers en accordent davantage aux critères du prix et des communications marketing³.

Une autre étude de Choi et Ahluwalia (2006) approfondit ces résultats en indiquant que dans des catégories pour lesquelles le risque est jugé élevé, même les switchers potentiels ne seront que peu influencés par des réductions de prix faibles, et ne pourraient changer de décision qu'au prix d'une forte réduction de prix.

D'un point de vue plus stratégique, Koças et Bohlmann (2008) expliquent à l'aide de modèles théoriques les stratégies de pricing des distributeurs en fonction de la répartition de leur clientèle entre stayers et switchers⁴, et concluent sur l'importance de prendre en compte la nature et la taille relative des segments de switchers dans le jeu concurrentiel pour fixer les niveaux de prix et de réductions de prix. Là encore, prix et réductions de prix semblent apparaître comme des critères de choix plus importants pour les switchers que pour les autres.

Nous allons voir dans la section suivante que si notre méthodologie peut être appliquée pour réaliser des études de ce type sur les critères de choix des switchers, notre approche reste plus générale et se différencie de celles-ci sur deux points que nous allons détailler.

B.1.iv. Positionnement académique de notre recherche

Notre recherche est plus générale que les recherches issues de ce courant dans le sens où :

- a. elle ne nécessite pas l'étude d'une transition effective passée, et peut donc être appliquée sans qu'il y ait encore eu transition ;
- b. elle ne se limite pas à considérer seulement deux catégories de consommateurs, switchers et stayers, mais considère l'ensemble des clients de manière continue.

Revenons sur ces deux différences essentielles.

³ Les mécanismes par lesquels ces techniques marketing peuvent influencer les choix des consommateurs, allant de motivations économiques à hédoniques, ont été mises en avant par Chandon, Wansink et Laurent (2000).

⁴ Le paramètre clé de leur modélisation étant le switcher-to-loyal ratio (SLR).

a. *Probabilité de transition vs. transition observée*

Contrairement à notre approche, toutes les méthodes précédentes nécessitent d'observer deux choix consécutifs pour les mêmes individus, afin de repérer les individus ayant migré et d'analyser leurs critères de choix. La méthode AEL qui sera présentée plus loin permet de travailler sur des données transversales, bien plus accessibles que les données de panels. Dans notre cas, l'effet d'un critère sera donc mesuré sans qu'une transition soit effectivement observée. De plus nous pourrions ainsi analyser l'influence des critères de choix sur les individus pour lesquels aucune transition n'a été observée, comme par exemple les stayers étudiés par Chan, Narashimhan et Zhang (2008).

b. *Une distinction plus fine que switchers vs. stayers*

Comme indiqué dans les limites de la recherche Wangenheim et Bayon (2002, p.219), les différences entre switchers et stayers sont souvent très larges, et une plus grande différenciation entre les groupes de consommateurs pourrait enrichir les résultats de telles études. Ceci est particulièrement important dans l'étude des impacts des critères de choix sur les sous-groupes étudiés.

Nous avons vu que dans les recherches précédentes, il était souvent fait des distinctions entre les différents types de switchers, ce qui amène à affiner leur caractérisation. Mais d'autres études choisissent aussi de préciser la catégorisation des stayers. Yim et Kannan (1999) constatent qu'avec la multiplication des alternatives proposées aux consommateurs lors de ces deux dernières décennies, il est très difficile de parler de consommateurs totalement fidèles. Ainsi, ils caractérisent les stayers suivant deux catégories:

« the first is hard-core loyalty, when consumers exclusively repeat purchase one product alternative, and the second is reinforcing loyalty, when consumers may switch among product alternatives, but predominantly repeat purchase one or more product alternatives to a significant extent. »

Si à la fois les switchers et les stayers peuvent être caractérisés plus finement, cela suppose qu'il est pertinent de dépasser la dualité entre stayers et switchers quand on catégorise les consommateurs par rapport à leur probabilité de choix d'une marque ou d'un produit.

Dans leur étude des consommateurs suivant leur choix de point de vente en Irlande du Nord, Doherty et Nelson (2008) tentent de segmenter les consommateurs suivant leur fidélité au magasin à la fois par leur comportement (par les historiques d'achat) et par leur attitude vis-à-vis des enseignes (par des questionnaires et entretiens). Leur segmentation se fait suivant la probabilité de switch⁵ entre le point de vente qu'ils fréquentent le plus souvent et les autres points de vente disponibles.

Cette étude les conduit à distinguer trois groupes sur le marché considéré⁶ :

- les devoted loyals (25%) : ayant une très faible propension à changer de point de vente, même s'ils y sont fortement sollicités par la concurrence ;
- les susceptibles (47%) : ayant une probabilité plus élevée de changer de point de vente, particulièrement s'ils sont exposés à des offres promotionnelles intéressantes ;
- les promiscuous switchers (28%) : n'ayant pas de point de vente de référence, changeant de point de vente suivant les opportunités, et correspondant davantage à ce que nous appellerons ici les cas marginaux.

Les devoted loyals correspondent à nos cas extrêmes dont le choix n'est que très difficilement influençables. D'après cette étude, les critères de choix de ces cas sont les attributs du point de vente en termes de localisation, de service et d'offre (sans tenir compte de l'attachement plus émotionnel ou affectif au magasin, mesuré par une autre échelle dans cette étude). Les susceptibles donnent moins de poids à ces critères et prennent aussi en compte l'activité promotionnelle dans leur choix, spécialement les bons d'achat. Ils apparaissent aussi très sensibles au fait qu'un distributeur s'intéresse aux questions environnementales et propose des produits *eco-friendly*. Enfin, la troisième catégorie considère que la praticité du point de vente, suivant les circonstances du moment, reste leur principal critère de choix. Ils réagissent très négativement à des défaillances de types rupture de stock ou mauvaise qualité d'un produit. Certes, ils sont sensibles à l'activité promotionnelle, mais des réductions de prix comme des bons d'achat les obligent à revenir sur le même point de vente et sont donc moins appréciées par cette catégorie de consommateurs que par les susceptibles.

⁵ Note : « Likelihood to switch » dans le texte, qui est une mesure équivalente à notre probabilité de choix dans l'étude. Si ce n'est que notre probabilité de choix n'est pas mesurée directement (à travers des analyses d'historiques d'achat ou des questionnaires), mais estimée à partir des choix observés sur un échantillon.

⁶ Étudié à partir d'un échantillon de 157 personnes.

Cette étude et les distinctions précédemment citées entre différents niveaux de switchers ou de stayers prouvent qu'il est possible de faire une distinction plus fine que stayers/switchers. De plus, Doherty et Nelson (2008) poussent à envisager que l'importance relative des critères de choix peut fortement varier *suivant les différentes probabilités de faire un choix* (ici de choisir son point de vente). Ce qui justifie notre approche méthodologique consistant à apporter une technique permettant d'obtenir une *analyse locale des critères de choix*.

B.2. Intérêt managérial de l'analyse locale des critères de choix

D'un point de vue managérial, la détection des critères de choix des cas marginaux constitue le meilleur moyen d'augmenter une part de marché (switchers), ou plus généralement son taux de pénétration (switchers et occasionnels), car ils permettent de mieux comprendre les critères de choix des consommateurs. Mais ceci ne doit pas nous conduire à négliger le fait que les cas marginaux, c'est-à-dire ceux les plus susceptibles de modifier leur choix à l'avenir, peuvent aussi faire partie de la clientèle actuelle de l'entreprise. C'est pourquoi l'intérêt managérial de l'analyse locale des critères de choix sera considéré dans cette sous-section à la fois d'un *point de vue offensif et défensif*.

B.2.i. D'un point de vue offensif

Les cas marginaux sont les plus susceptibles à court terme de choisir une option, car ils ne sont pas encore liés fortement à une autre option. Dans le cas d'un choix entre deux marques, l'acquisition de cas marginaux (switchers) signifie gagner de la part de marché sur la concurrence. Dans le cas d'un choix entre adoption ou non adoption, ou consommation et non-consommation, l'acquisition de cas marginaux (occasionnels) revient à augmenter ses ventes en augmentant la demande primaire pour le produit ou le service considéré.

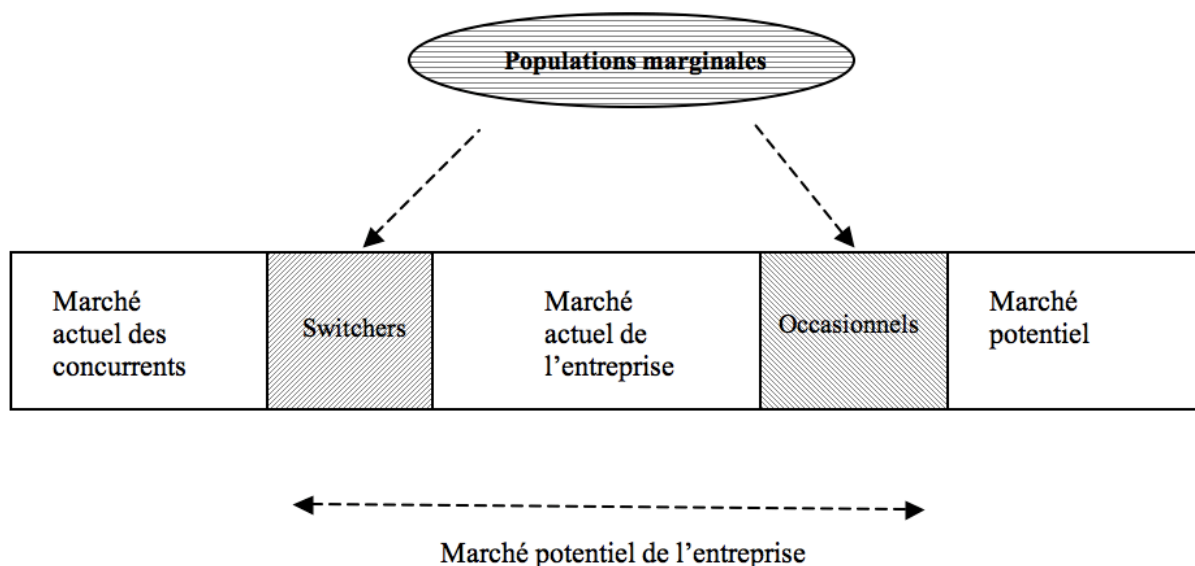


Figure B.2.1. Utilisation offensive de l'analyse locale des critères de choix (1/2)

Ces deux cas de figure sont décrits dans la figure B.2.1. ci-dessus. Pour chaque cas, appliquer une telle stratégie demande de connaître quels sont les critères de choix caractéristiques de ces sous-populations marginales (critères de choix locaux) pour pouvoir les utiliser.

La connaissance de ces critères de choix locaux permettra de mieux ajuster le mix marketing envers ces cibles. Par exemple pour détecter les caractéristiques du produit à mettre potentiellement en avant dans les communications marketing les ciblant (Publicité), ou pour lancer de nouvelles versions des produits existants satisfaisant mieux leur attente (Produit). Une bonne illustration de cette politique d'acquisition est fournie par Farano, Romberger, et Meer (2009), et restituée dans la figure B.1.2. ci-après. Ils démontrent dans leur article que les stayers représentent déjà 90% des ventes de Starbucks et qu'il semble difficile d'augmenter les ventes auprès de cette sous-population (bien qu'il soit probablement possible d'améliorer à la marge leur satisfaction pour consolider leur rétention comme nous le verrons dans la sous-section suivante). Par contre les cas marginaux constituent une marge de progression des ventes six fois supérieure à celle des cas extrêmes, et ne sont pas encore acquis à la concurrence. Il s'agit donc là d'une forte opportunité de croissance. S'ils ne sont pas encore convaincus par une des deux options, c'est que leurs critères de choix n'ont pas été complètement satisfaits. Une analyse locale des critères de choix permettra d'identifier ces critères encore mal connus et d'en tirer parti.

The Real Opportunity for Starbucks

Starbucks already has almost 90% of the business of its most loyal customers. Not much room for growth there. And it's not likely to get much business from people who are loyal to competitors. So it needs to focus on the sizable group of "switchers" – those who go both to its shops and to others. By giving these customers more of what they need, Starbucks can dramatically turn around its business, even if its most loyal customers are cutting back.

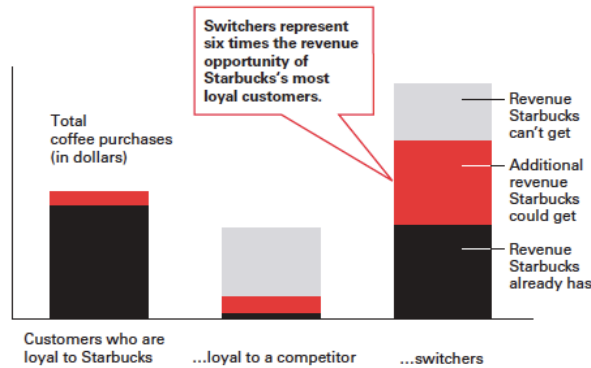


Figure B.2.2. Utilisation offensive de l'analyse locale des critères de choix (2/2)

Mais, l'analyse locale des critères de choix peut aussi être utilisée à des fins défensives, et s'avérer donc particulièrement utile en période de difficulté ou de crise. En effet, les cas marginaux peuvent aussi être les clients actuels les plus susceptibles de passer à la concurrence et l'analyse locale des critères de choix permettrait alors de renforcer sa part de marché face aux attaques de la concurrence.

B.2.ii. D'un point de vue défensif

Suivant leur position par rapport à la frontière de choix, les cas marginaux peuvent être perçus comme appartenant à une cible à conquérir, ou un capital à défendre. D'un point de vue défensif, identifier les critères de décision des cas marginaux peut donc être utilisé pour prévenir le départ des clients les plus susceptibles d'être séduits par la concurrence.

Cette problématique a déjà été identifiée par des praticiens tels que Clive Humby (2004) :

« We work with Tesco to target the right customers for each campaign. One of those priority targets are customers vulnerable from competition, especially during periods of high competitor activity. »

Farano, Romberger, et Meer (2009), dans leur article portant précisément sur les moyens à la disposition des distributeurs pour traverser la crise économique actuelle, établissent comme deuxième règle (parmi cinq) de mieux comprendre les consommateurs moyennement fidèles, afin de mieux les servir.

Ils mettent aussi en garde contre la pratique consistant à trop se focaliser sur les stayers en analysant leurs historiques d'achat. Ceci amène en effet à « trop » servir ces consommateurs qui dépensent déjà quasiment tout leur budget chez l'enseigne, et à négliger l'analyse des attentes des clients qui consomment certaines catégories de produits dans d'autres points de vente (achats qui par définition ne sont pas dans les historiques d'achat).

« Rule 2: Close the Needs-Offer Gap. To capture more business, you must entice those customers spending elsewhere to spend with you instead. That means closing the gap between what they want and what you offer, not merely ordering more of what's already selling well. »

Et savoir ce que veulent ces consommateurs moins réguliers demande de connaître les critères de choix qui leur sont spécifiques et qui sont les critères locaux de choix des cas marginaux. Ils peuvent être différents des critères locaux des cas extrêmes, qui sont ceux ayant déjà fait leurs preuves auprès d'une population qui ne peut consommer plus.

Dans ce cadre défensif, nous pouvons aussi nous intéresser aux cas extrêmes⁷ déjà acquis par l'entreprise, mais desquels elle souhaiterait améliorer la satisfaction pour garantir leur rétention. En utilisant les critères de décisions locaux liés à cette sous-population, il serait possible de comprendre comment orienter les évolutions de l'offre qui leur est destinée, mais aussi de sélectionner les bonnes actions commerciales pour les amener à augmenter le nombre de produits ou services consommés (cross-selling) ou à monter en gamme (up-selling).

Ces problématiques managériales renforcent l'intérêt du positionnement académique de cette recherche, présenté dans la sous-section B.1.

⁷ En effet, il serait périlleux de considérer ces cas extrêmes dans un cadre offensif, car il s'agit d'individus ayant très peu de chance de changer d'avis et donc d'être convaincus par une autre option. Les efforts financiers visant à les convertir risquent donc de s'avérer très peu rentables.

B.2.iii. Positionnement managérial de notre recherche

Notre positionnement managérial consiste à proposer aux managers une vue plus nuancée de la structure du marché, à la fois au niveau des groupes (switchers vs. stayers) et des critères de choix (important ou non).

Pour reprendre les mots de Little (1970) :

« Managers know only two probabilities, zero and one, whereas model builders know all probabilities, except zero and one. »

Supposer implicitement que tous les critères jouent de la même façon à travers toute la population, ce qui est le cas habituellement dans l'interprétation d'une analyse discriminante, permet seulement de considérer des critères globaux. De ce point de vue, un critère a un impact significatif ou n'en a pas. Or nous verrons qu'en examinant de plus près les modèles de classification, le poids d'un critère peut varier d'un sous-groupe à l'autre.

Notre méthodologie se justifie par la nécessité de pouvoir identifier les critères de choix des cibles jugées prioritaires. D'un point de vue pratique, nous espérons que cette méthode, en apportant une vue alternative de la structure de la demande, permettra aux managers d'enrichir la perception intuitive de leur marché.

C. Définitions et principe de l'analyse locale

Nous commençons cette section par la définition des termes clefs de cette recherche. Ensuite nous en exposons le principe, et concluons par une présentation de la base de données qui nous servira de premier exemple d'application.

C.1. Définitions

Rappelons ici que le problème consiste à trouver une manière de mesurer les effets d'une variable, ou un critère de choix sur une décision binaire pouvant prendre les valeurs à 0 ou 1 (e.g. achat vs. non-achat, marque A vs marque B) pour un individu donné. Nous ne souhaitons pas seulement connaître si la décision a changé ou pas, mais aussi dans quelle mesure la décision a été influencée par les critères de choix. Aussi nous avons besoin de mesurer l'effet de chaque variable sur la probabilité de faire le choix 1, c'est-à-dire l'impact de chaque variable sur la fonction de probabilité notée f .

C.1.i. Définition d'effet local

Nous appelons *effet* d'un critère de *choix* la part de variance de la probabilité de choix expliquée par le critère en question. Le niveau de variance expliquée est aussi communément appelé pouvoir explicatif.

Habituellement, l'effet d'un critère est supposé constant à travers la population, indépendant de la sous-population que l'on étudie. Cependant, nous pensons que l'on pourrait tirer avantageusement partie de cet effet en le calculant localement, sur des sous-groupes de la population globale. Nous proposons de définir ces sous-groupes à partir des probabilités de choix.

Nous appelons donc *effet local* d'un critère de *choix* la part de variance expliquée de la probabilité de choix pour une sous-population définie par un intervalle de valeurs de cette probabilité.

L'*effet local relatif d'un critère* est lui, défini comme le ratio entre la variance expliquée par un critère au sein du sous-groupe, par rapport à la somme des variances expliquées par tous les critères étudiés au sein de ce même sous-groupe.

D'autres recherches récentes substituent un modèle global à une forme de modélisation dépendant du *contexte* (Kivetz, Netzer et Srinivasan 2004) ou du caractère *local* de la fonction de décision (Doksum et al. 1994), en reprenant leur vocabulaire respectif. Dans le premier cas, les auteurs désirent tenir compte des alternatives les plus proches de l'individu dans le cas d'un choix multiple afin de mesurer les effets de compromis. Dans le second, les auteurs veulent prendre en compte les variations de la force de la corrélation entre Y et X pour prédire Y à partir de X, car dans les cas d'hétérocorrélations⁸, le niveau de corrélation entre X et Y, dépend de la valeur prise par X. Ces deux études avaient deux définitions différentes du terme « local » dépendant de leur perspective de travail. En effet, il n'y a pas d'utilisation standard de l'adjectif « local » dans les précédentes études de ce type.

Ici nous considérons que l'analyse est locale si elle ne traite pas tous les cas uniformément, mais suivant leur prédictibilité, c'est-à-dire la valeur de la probabilité de choix définie plus haut.

C.1.ii. Définition de la frontière de choix

Nous appelons frontière de choix, le seuil de probabilité que l'on fixe à partir de la fonction de probabilité pour prédire de façon binaire le choix des individus. Quelle que soit la méthode de discrimination utilisée, ce seuil doit être fixé par l'utilisateur. Si les individus se répartissent en deux groupes égaux suivant leurs choix, alors ce seuil est fixé à 50%. Ainsi, lorsqu'un individu a 51% de chance de faire le choix 1, on prédit qu'il fera le choix 1. Inversement, à 49% de chances de faire le choix 1, alors on lui prédit le choix 0. 50% constitue ainsi la frontière à partir de laquelle les individus sont discriminés sur la base du choix qu'on leur prédit. Il peut arriver que cette frontière soit définie sur la base d'une autre probabilité que celle des 50%. En effet, prenons par exemple le cas du choix consistant à consommer un produit de luxe (ou très haut de gamme) versus un produit classique dans une catégorie de produits donnée. Sur la base de la population totale, la probabilité de faire le

⁸ Pour une définition de la corrélation locale et de l'hétérocorrélations, Cf. Doksum et al. (1994) à la page 572.

choix du luxe (choix noté 1) sera très faible comparée à la probabilité de faire le choix d'un produit classique (noté 0). Le seuil de prévision aura donc un niveau plus élevé que 50%. La valeur du seuil de probabilité optimal suivant laquelle baser la frontière de discrimination est celle donnant le meilleur taux de prévision (Maddala, 2001, p 327) et doit être fixée par l'utilisateur (Maddala, 1983, p 91). En effet, à chaque seuil de prévision, on peut calculer le taux de « bien classés » et de « mal classés » correspondant à ce seuil en comparant les résultats de la prévision avec les données de choix initiales. Le seuil pour lequel le taux de « bien classés » est maximal sera le seuil retenu. Une fois cette valeur fixée, elle définit la frontière de décision car elle discrimine les individus suivant les choix qui leur seront prédits.

C.1.iii. Cas marginaux et cas extrêmes

Plus un individu est proche de la frontière, en termes de probabilité de choix, plus il sera dit *marginal* ; plus un individu sera éloigné de la frontière, plus il sera dit extrême (e.g. individus avec une probabilité proche de 0 ou de 100%). Les individus ne sont pas étudiés ici individuellement mais en sous-groupes. Les sous-groupes les plus remarquables dans ce cadre sont les sous-groupes des cas marginaux et des cas extrêmes.

Les *cas marginaux* correspondent aux personnes dont les choix sont les moins prévisibles. Cette population est rencontrée dans deux cas de figure. Soit ces consommateurs sont très susceptibles de passer d'une marque à l'autre (les switchers) dans le cadre d'un choix répété, soit ils sont les moins convaincus par les deux options de choix qui leur sont présentés (les indécis). Techniquement, ces deux cas de figure se traitent de manière équivalente. Une illustration de switchers sera donnée dans l'application de la méthode au choix d'une marque de jus d'orange (section F.1.), une illustration d'indécis sera donnée dans l'application au référendum (section F.2.).

C.1.iv. Mode de représentation choisi

Depuis Hotelling (1929), il est habituel en modélisation théorique de représenter les consommateurs suivant un axe correspondant à une des caractéristiques du produit étudié. Ici l'angle de représentation choisi est assez différent, dans le sens où l'axe de caractérisation choisi est la probabilité de choix.

Dans leurs études de l'impact des coupons de réductions sur les switchers, Shaffer et Zhang (1995) choisissent de représenter sur l'axe des abscisses un ratio tenant compte des prix et des réductions de prix pour les produits A et B. En voici un extrait dans la figure ci-dessous :

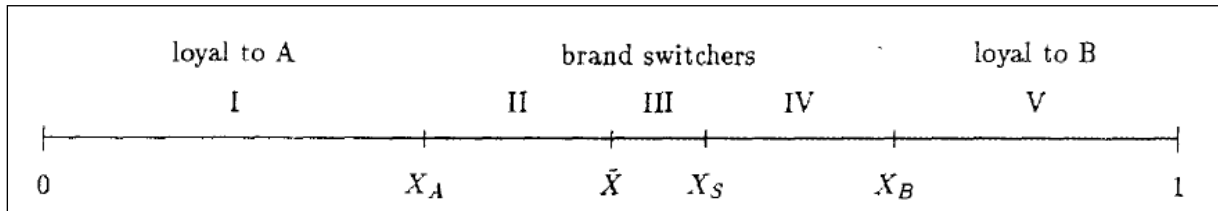


Figure C.1.1. Représentation des cas marginaux et extrêmes par Shaffer et Zhang (1995)

Il ne s'agit pas tout à fait du mode de représentation que nous avons choisi car l'axe des abscisses ne correspond pas ici à la probabilité de choix.

Nous allons voir tout de suite comment nous représentons les cas dans le cadre de notre segmentation suivant la probabilité de choix.

Etant donné notre objectif de catégorisation des individus suivant leur probabilité de choix, nous choisissons comme axes des abscisses cette probabilité de choix pour ordonner les individus depuis une probabilité de choix nulle (égale à 0%), jusqu'à une probabilité de choix certaine (égale à 100%). La frontière de décision définie plus haut nous permettra de caractériser les cas marginaux (à proximité de cette frontière), alors que les cas extrêmes se trouveront aux bornes de cet intervalle de probabilité, comme on peut le voir dans la figure C.1.2. ci dessous.

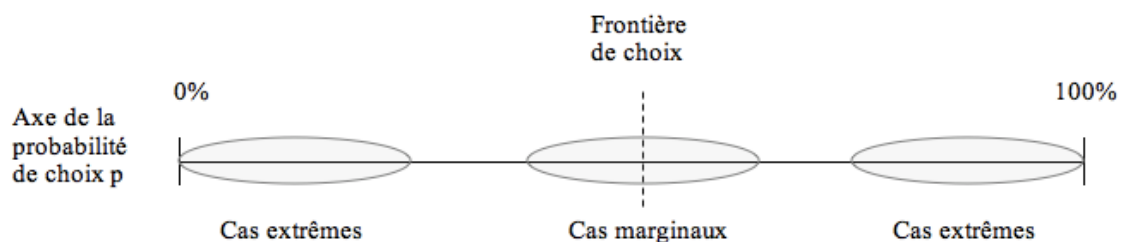


Figure C.1.2. Représentation des cas marginaux et extrêmes suivant la probabilité de choix

En ordonnée, seront placés les effets locaux des critères, au niveau de chaque sous-groupe d'individus ainsi catégorisés. La représentation finale d'une analyse locale des critères de choix sera proposée dans la section Application de cette étude (section F).

C.2. Principe de l'analyse locale

Les choix que nous étudions ici, les choix binaires, sont la plupart du temps modélisés par une régression logistique. Dans le cadre d'une régression logistique, l'impact des critères de choix sur une probabilité de choix p est modélisé via une fonction d'utilité sous-jacente u . Or une variation de u provoque une variation négligeable sur la fonction de probabilité pour les cas extrêmes, et une large variation pour les cas marginaux. Il est donc intéressant de détailler les analyses à un niveau local pour différencier cas marginaux et cas extrêmes.

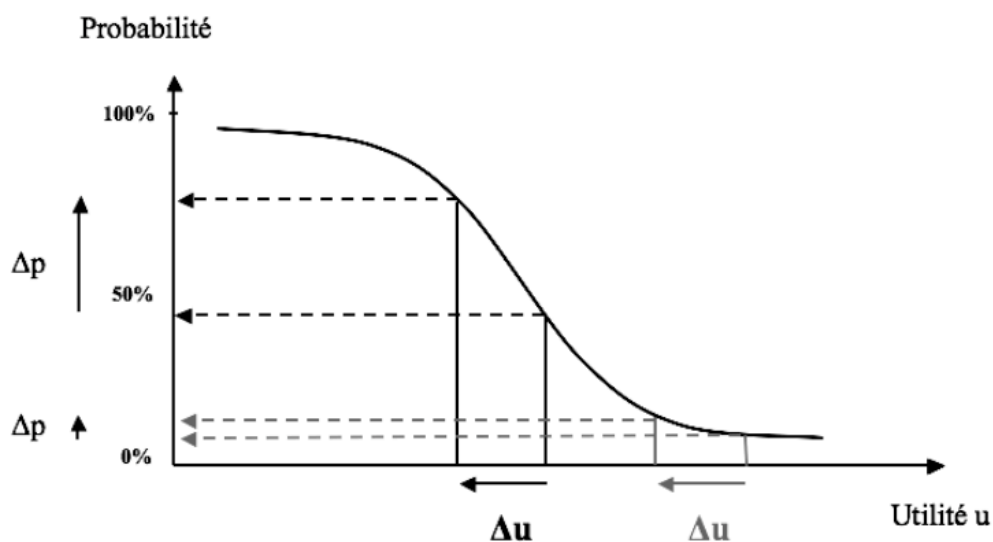


Figure C.2.1. Représentation traditionnelle des fonctions de choix binaires

Ici, nous allons donc chercher à *interpréter plus directement et plus localement le lien entre les critères de choix et la probabilité de choix*. En effet, l'influence qu'aura eu chaque critère dans la formation de p pour les individus peut varier suivant la distance qui sépare les individus de la frontière de décision (ici 50%).

C.3. Base de données utilisée

Deux bases de données seront utilisées pour appliquer notre méthode, mais nous en avons sélectionné une en particulier pour illustrer en détail notre approche dans le texte. Il s'agit de la base de données publique appelée *Orange Juice*. La base de données *Référendum* sera présentée plus loin en section F.2.

Cette base de données *Orange Juice* est constituée de données réelles proposées par Forster et al. (1998) dans un manuel de statistiques au sein d'un chapitre sur la régression logistique. Cette base est réutilisée ici avec l'aimable autorisation des auteurs. Nous avons conservé la même dénomination des variables (en anglais) par souci de cohérence avec leurs propres travaux. Cette base décrit 1070 décisions indépendantes⁹ d'achat de jus d'orange entre deux marques américaines: Minute Maid (notée MM) et son challenger Citrus Hill (notée CH et codée comme le choix 1).

Cinq critères de choix sont ici disponibles :

- le prix de vente de base (hors réduction de prix) de Minute Maid : MM Price ;
- le prix de vente de base de Citrus Hill : CH Price ;
- le taux de réduction de prix sur Minute Maid : MM Discount ;
- le taux de réduction de prix sur Citrus Hill : CH Discount ;
- la fidélité à la marque CH : Loyalty, mesurée par les préférences passées (comme Guadagni et Little, 1983).

Ces données sont intéressantes pour plusieurs raisons : la taille de l'échantillon correspond à celle communément utilisée dans les études marketing et les sondages d'opinion publique, et les variables sont des variables très souvent étudiées en marketing sur les effets desquelles nous pouvons avoir une première intuition a priori, notamment d'après la revue de littérature de la sous-section B.1.iii.

Notons ici que notre approche ne nécessite pas une méthode de recueil spécifique de données, ce qui permet de l'utiliser sur toute base de données décrivant un choix binaire. De plus,

⁹ Ces décisions d'achat ont été prises par des consommateurs tous différents, durant le même mois.

suivant la démarche adoptée pour appliquer notre méthode (Cf. section E), la nature des variables correspondant à nos critères de choix peut s'étendre à tout type de variable.

D. Première méthode : les régressions logistiques emboîtées

Cette section propose une première approche directement inspirée de la régression logistique, pour analyser plus localement l'importance des critères d'un choix binaire. Nous allons voir que cette méthode, très intuitive, pose un certain nombre de problèmes que nous devons dépasser pour obtenir une analyse AEL plus satisfaisante des critères de choix au niveau local. Cette méthode AEL sera présentée dans la section E suivante.

D.1. Régression et poupées russes

Une approche simple pour observer des effets de marge serait d'utiliser successivement plusieurs régressions logistiques, en se focalisant de plus en plus sur la population proche de la frontière de décision définie plus haut (sous-section C.1). Or a priori, nous ne connaissons que le choix final des individus, et non la probabilité de faire ce choix, il n'est donc pas possible de directement se focaliser sur les individus autour de la frontière.

Le principe de cette première approche est donc d'utiliser une première fois la régression logistique pour calculer leur probabilité de choix, puis de sélectionner les individus les plus proches de la frontière, et d'appliquer alors une seconde régression logistique sur ce sous-échantillon pour connaître les leviers influençant le plus ces individus marginaux. Ainsi nous pouvons « zoomer » une fois sur la frontière, en utilisant une première régression pour sélectionner les individus, et une seconde pour évaluer les influences des variables sur le sous-échantillon sélectionné. Ensuite nous répétons cette première étape une seconde fois, voire plusieurs autres fois, pour zoomer davantage sur la population marginale, avant d'évaluer les effets de marge. Comme nous allons le voir, le nombre d'étapes de ce type sera très vite limité par la taille de l'échantillon. Dans notre exemple, nous « zoomons » deux fois.

Nous nommons ce principe consistant à utiliser plusieurs régressions logistiques successives sur des sous-échantillons de plus en plus réduits, centrés sur la frontière, à l'image de poupées russes, les *régressions logistiques emboîtées* (que nous traduirons en anglais par *embedded logistic regressions*, à ne pas confondre avec la méthodologie traditionnelle du *nested logit*).

Dans cet exemple, afin de mettre en évidence les différences de niveau d'influence par variable suivant les sous-populations étudiées, nous calculerons aussi la statistique de Wald pour chaque variable à chaque étape afin d'étudier les différentes valeurs prises.

Donnons à présent une description formalisée du principe que nous venons de proposer.

D.2. Description formelle des régressions logistiques emboîtées

Détaillons plus précisément la technique des régressions logistiques emboîtées. Tout d'abord nous appliquons une régression logistique sur l'échantillon complet. Ensuite, nous filtrons l'échantillon de la manière suivante :

1. l'échantillon est divisé en deux sous-groupes d'individus suivant le choix prédit, avec d'un côté tous les individus pour lesquels le choix prédit est 0, de l'autre ceux pour lesquels le choix prédit est 1 ;
2. Dans chaque sous-groupe, les individus dont le choix est le plus prévisible sont éliminés. Il s'agit des individus dont la probabilité de faire le choix 1 est la plus élevée au sein du sous-groupe pour lequel le choix prédit est 1, c'est-à-dire les individus de ce sous-groupe dont la valeur individuelle de probabilité notée p_i est supérieure à la valeur médiane de p au sein de ce sous-groupe. Et nous éliminons aussi les individus dont la probabilité de faire le choix 1 est la plus faible au sein du sous-groupe pour lequel le choix prédit est 0, c'est-à-dire les individus dont la valeur p_i est inférieure à la valeur de la médiane de p au sein de ce sous-groupe.

Cette procédure est réitérée en appliquant une nouvelle régression logistique sur le nouvel échantillon obtenu à l'issue de ce filtrage afin de « zoomer » sur la frontière. Le nombre d'itérations optimal à effectuer est laissé au libre arbitre de l'utilisateur, mais il est assez naturel d'arrêter de zoomer quand les coefficients estimés ne sont plus significatifs, ce qui arrive mécaniquement du fait de la réduction de la taille de l'échantillon à chaque zoom.

Plus formellement, si S^n est l'échantillon considéré à l'étape n , et $\hat{Y}^n(x)$ est la variable indiquant le choix prédit à l'étape n pour l'individu x , et $p^n(x)$ est la probabilité estimée à

l'étape n de faire le choix 1, p_0^n la médiane des probabilités $p^n(x)$ sur le sous-ensemble pour lequel $\hat{Y}^n(x)=0$ et p_1^n la médiane des $p^n(x)$ sur le sous-groupe pour lequel $\hat{Y}^n(x)=1$, l'extraction est définie de la manière suivante :

$$S^{n+1} = \{x|x \in S^n, \hat{Y}^n(x) = 0, p^n(x) > p_0^n\} \cup \{x|x \in S^n, \hat{Y}^n(x) = 1, p^n(x) < p_1^n\} \quad [1]$$

où $\hat{Y}^n(x)$ et $p^n(x)$ ont été calculés en utilisant une régression logistique sur S^n .

Nous rappelons ici que la frontière permettant de prédire le choix des individus a été fixée durant l'estimation des paramètres de la régression logistique précédant cette extraction, et qu'elle n'est pas toujours à 50%. Comme nous l'avons indiqué précédemment, elle a été fixée pour optimiser le taux de bonnes prévisions et c'est-à-dire pour qu' \hat{Y}^n estime Y^n au plus juste.

D.3. Résultats sur les données *Orange Juice*

Les résultats¹⁰ de cette procédure sont reportés dans le tableau D.3.1 ci-dessous.

		Population totale	Première extraction	Seconde extraction
Population	Effectif	1070	535	268
	Pct	100,0%	50,0%	25,0%
Effectifs initiaux par groupe	MM	653	313	147
	CH	417	222	121
Précision	Logit	83,1%	72,5%	64,6%
	Trivial	61,0%	58,5%	54,8%
Effet des critères (Statistique de Wald et niveau de significativité)	MM Price	21,41	21,57	10,29
		0,000	0,000	0,001
	CH Price	13,91	10,58	5,78
		0,000	0,001	0,016
	Loyalty	268,32	93,08	14,99
		0,000	0,000	0,000
	MM Discount	2,42	2,96	1,47
		0,120	0,085	0,226
	CH Discount	0,96	0,76	0,93
		0,200	0,620	0,480
Effet relatif des critères	MM Price	7,0%	16,8%	31,6%
	CH Price	4,5%	8,2%	17,8%
	Loyalty	87,7%	72,6%	46,1%
	MM Discount	0,8%	2,3%	4,5%
	CH Discount	0,0%	0,1%	0,0%

Tableau D.3.1. Résultats des régressions logistiques emboîtées sur les données *Orange Juice*

¹⁰ Les calculs ont été réalisés sous XLStat 2007. Si des variations dans les résultats peuvent exister d'un logiciel à l'autre, la présentation qui suit conserve toute sa pertinence malgré les variations en question.

Les deux premières lignes décrivent la taille de l'échantillon : après chaque filtrage, la taille de l'échantillon est divisée par deux. On part ainsi de 100% des individus, pour n'en conserver plus que 50%, puis 25%. Les deux lignes suivantes décrivent la taille de chaque sous-population définie par le choix prédit par la régression logistique utilisée à chaque étape.

Pour mesurer la précision des prédictions, les deux lignes suivantes donnent le pourcentage d'individus correctement classés à chaque étape à l'issue de la régression logistique, ainsi qu'à titre de comparaison le pourcentage d'individus correctement classés en utilisant une méthode triviale consistant à prédire le même choix pour tous, à savoir le choix le plus répandu (dans notre cas MM).

L'effet de chaque critère, à chaque étape, est mesuré par sa statistique de Wald (le niveau de significativité est aussi fourni), issu de la régression logistique correspondante. Enfin, les cinq dernières lignes du tableau indiquent l'effet local relatif mesuré par le ratio entre la statistique de Wald d'un critère sur la somme de statistiques de Wald de tous les critères. Cette technique se justifie par le fait que la statistique de Wald donne la contribution individuelle d'un critère, donc la somme de ces statistiques donne l'ensemble des contributions de tous les critères, et le ratio indique ici la contribution d'un critère de manière relative par rapport à l'ensemble des contributions des critères du modèle (Cohen et West, 2003).

La ligne correspondant à la variable Loyalty indique que cette variable était la plus discriminante pour l'échantillon complet avec un effet relatif de 87.7% si l'on considère l'échantillon total. Cependant en zoomant plus près de la frontière, cette importance relative diminue jusqu'à atteindre 46.1%.

Contrairement à la variable Loyalty, la variable MM Price prend une importance croissante lorsque l'on zoome plus près de la frontière de décision. D'une importance relative de seulement 7% sur l'échantillon complet, elle atteint un effet relatif de 31.3% sur le dernier sous-échantillon. Dans une moindre mesure, il en va de même pour la variable CH Price.

Ces résultats sont concordants avec ceux de Kannan et Wright (1991), qui par une approche complémentaire de nested logit, démontrent que les cas marginaux (ou switchers) sont beaucoup plus influencés que les stayers par les variations d'éléments du marketing mix.

La variable Loyalty est plus discriminante sur la population générale que sur les cas marginaux alors que la variable MM Price est relativement plus discriminante sur les cas marginaux qu'ailleurs. Cette première méthode confirme donc notre intuition concernant l'existence de différents types de critères, car les variables n'ont pas la même influence relative suivant les différents sous-échantillons sélectionnés.

Il est intéressant de noter que les coefficients varient de manière similaire aux statistiques de Wald et que leurs signes ne changent pas entre les trois extractions : ils sont positifs pour Loyalty, MM Price et CH Disc ; et négatifs pour CH Price et MM Disc. Ceci provient du caractère symétrique de cette analyse, c'est-à-dire que l'on prend toujours autant d'individus à droite de la frontière qu'à gauche de la frontière.

L'inconvénient majeur de cette méthode est que la taille de la population est divisée par deux à chaque étape, ce qui fait diminuer la quantité d'information disponible et réduit ainsi la précision d'interprétation du modèle.

D.4. Aller plus loin que les régressions logistiques emboîtées

Le principal enjeu est de comprendre l'influence des variables à un niveau local tout en utilisant *toute l'information disponible* dans l'échantillon global. Pour concilier ces deux objectifs, notre approche est d'abord d'estimer sur l'échantillon complet les fonctions décrivant l'effet de chaque variable sur la probabilité, et dans un second temps (et un second temps seulement) d'estimer l'effet local de chaque variable sur des sous-groupes décrits précédemment. Ainsi l'estimation initiale des effets sera effectuée sur toute la population, sans perte d'information.

En utilisant la régression logistique sur l'échantillon complet, l'estimation des effets se fait sans perte d'information, mais il est impossible d'en déduire une estimation des effets locaux, par exemple une estimation de la statistique de Wald sur des sous-groupes. Et en l'utilisant sur un sous-groupe, on perd toute l'information disponible dans le reste de la population. Comme nous allons le voir dans la section suivante, cette approche nécessite l'introduction en marketing de nouvelles méthodes statistiques permettant d'élargir le cadre de la régression logistique.

E. Méthode d'analyse des effets locaux (AEL)

Dans cette section nous allons présenter une méthode complètement inédite permettant d'aller plus loin que la méthode précédemment décrite (section D) et d'atteindre les objectifs de mesure que nous nous sommes fixés (section C).

Nous allons d'abord partir de notre orientation méthodologique avant de décrire précisément les trois étapes de cette méthodologie et de faire un point sur les modes d'évaluation de la qualité des résultats pouvant être obtenus par une telle méthode.

E.1. Orientation méthodologique

Dans cette recherche, nous introduisons en marketing une nouvelle approche statistique, l'ANOVA fonctionnelle, développée par Owen (1992), permettant de travailler de manière adaptée sur la fonction de probabilité de choix. Nous utilisons aussi les Support Vectors Machines, ou Machine à Vecteurs Support, proposés par Vapnik (1995). L'usage inédit de ces méthodologies issues de la recherche en apprentissage statistique oriente clairement notre recherche dans une direction opérationnelle, liée à l'action et à la prise de décision.

Pour plus de détails sur ce type de recherche, on peut se référer à Cui et Curry (2005) qui ont commencé à ouvrir cette passerelle et qui dressent une première description des différents types de méthodes de prévision orientées vers l'action – robust analysis, analytic prediction ou structural gap analysis – en fonction de l'utilisation prévue. Parmi ces catégories, notre approche peut être classée dans le cadre des méthodes de prédiction robuste, qui consiste à :

« Accepting a trade-off between deep structural understanding and ease of interpretation in the light of the respondent effort. »

Nous tenons à insister sur l'importance et la nécessité de ce compromis (trade-off) en modélisation. En effet, comme nous le reverrons dans la discussion, il nous semble crucial de ne plus discriminer les méthodologies de modélisation, et plus particulièrement de prévision, entre des « boîtes noires » d'une côté, performantes mais ininterprétables, et des « theory testers » de l'autre, interprétables mais peu robustes ou du moins dépendants d'un très grand nombre d'hypothèses, rarement vérifiées en pratique. L'issue ne peut être qu'intermédiaire, à

l'intersection des deux cultures décrites par Breiman (2001), comme nous le proposerons dans la discussion.

E.2. Différentes étapes de la méthode AEL

Notre méthodologie comprend trois étapes successives. Nous calculons d'abord une fonction de discrimination permettant de classer les individus en deux catégories suivant leur choix prédit (0 ou 1). Ensuite une fonction de probabilité a posteriori est estimée pour donner à chaque individu une probabilité (de faire le choix 1). Cette fonction permet de classer les individus en sous-groupes (nous utiliserons en cohésion avec la littérature statistique sur ces sujets le vocable anglais « bin » pour nommer ces sous-groupes), définis par des tranches de probabilités. Enfin, dans chaque bin, sera calculé l'effet local de chaque critère, c'est-à-dire son influence au sein des individus de ce bin sur la fonction de probabilité définie dans l'étape précédente.

La première étape consiste simplement à appliquer une méthode de discrimination déjà connue (comme l'analyse discriminante ou la régression logistique). La seconde étape met en évidence la place centrale que nous accordons à la fonction de probabilité, ce qui n'est pas toujours le cas dans les méthodes de discrimination. La troisième étape fait appel à un outil d'analyse (l'ANOVA fonctionnelle ou FANOVA) qui n'avait jamais été combiné auparavant avec des outils de discrimination.

Pour chaque étape, nous avons choisi des indicateurs de performance adaptés pour valider les calculs. Ces indicateurs ont été choisis pour ne pas dépendre du type de fonction de discrimination choisi à la première étape. Ils sont décrits dans la sous-section E.6.

E.3. Première étape de la méthode AEL : Discrimination

Cette première étape consiste à produire une fonction de classification dépendant des critères de choix des individus. Notre méthodologie est assez flexible pour prendre en entrée un grand nombre des outils de discrimination actuellement disponibles. Nous tenons à présenter parmi ces outils les SVMs (Support Vector Machines) dont la qualité de prédiction constitue un atout majeur, et qui devraient avoir un grand avenir en marketing.

A partir de la fonction de discrimination choisie, nous disposons d'une fonction f de $[0,1]^d$ dans $[0,1]$ où d est le nombre de variables de prédiction, permettant de prédire une probabilité de choix à partir de d variables observées. A noter que, dans le cas le plus fréquent où les variables de prédiction ne sont pas à valeurs dans $[0,1]$, leur mise à l'échelle (par division par l'amplitude entre leur valeur maximale et minimale) est effectuée avant l'étape de classification par simple commodité.

E.3.i Flexibilité

La première étape de la méthode AEL consiste à calculer une fonction de discrimination. Cette fonction prend en entrée les valeurs des critères de choix, ou variables de prédiction pour un individu et donne en sortie une prédiction binaire du choix de cet individu (choix prédit). Une grande variété de méthodes ont été développées par les chercheurs en statistiques et en *machine learning* (dénomination anglo-saxonne de l'apprentissage statistique) pour estimer ce type de fonctions (de l'analyse discriminante de Fisher aux réseaux de neurones, pour ne donner que quelques exemples).

Notre méthode n'impose pas de choisir une de ces méthodes en particulier. Pour être plus précis, la seule condition est qu'il soit possible de calculer une fonction de probabilité de choix à partir de la fonction de discrimination. Cela écarte a priori la plupart des fonctions issues du *machine learning* (réseau de neurones, machines à vecteur support,...), parce qu'elles ne basent pas l'estimation de leur fonction de discrimination sur le calcul d'une fonction de probabilité comme cela est le cas en statistique avec par exemple la régression logistique. Cependant, comme nous le verrons dans la sous-section E.4., la plupart de ces méthodes se sont enrichies d'apports méthodologiques permettant de calculer aisément la fonction de probabilité dont nous aurons besoin dans la seconde étape. Ainsi la plupart des fonctions de classification issues de différents courants de modélisation peuvent être utilisées pour appliquer notre méthodologie.

Ceci constitue un avantage certain. En effet, le meilleur mode de discrimination *dépend souvent du problème abordé*, et la flexibilité de notre méthodologie nous permet de sélectionner durant la première étape la méthode de discrimination apportant la meilleure précision. De plus, l'AEL peut ainsi être considérée comme l'analyse complémentaire d'une

régression logistique, ou d'une autre méthode, déjà effectuée et peut ainsi permettre d'affiner l'interprétation des premiers résultats obtenus.

Par la suite, nous allons appliquer notre méthodologie à partir de deux méthodes de discrimination qu'a priori tout oppose : l'analyse discriminante et les machines à vecteur support (ou en anglais Support Vector Machines, SVM). Justifions ici ce choix. L'analyse discriminante est avec la régression logistique la méthode de discrimination la plus largement répandue en marketing. Les SVMs, issus de la recherche en machine learning, ont été récemment importés en marketing par les travaux de Cui et Curry (2005) et Evgeniou et al. (2005). L'avenir de cette méthode en marketing est prometteur étant donné leur fort pouvoir prédictif et le solide cadre théorique duquel elle est issue. Étant donné que ces deux méthodes ont des hypothèses sous-jacentes et des propriétés techniques très différentes, les utiliser toutes les deux dans notre étude empirique nous permettra d'illustrer la flexibilité de la méthode AEL. De plus, les résultats étant, au final, très similaires, ceci illustrera aussi la cohérence des résultats obtenus par la méthode AEL.

Une raison supplémentaire nous ayant amené à utiliser les SVMs est qu'ils sont encore perçus comme des boîtes noires, desquelles il serait impossible d'obtenir des résultats interprétables malgré leurs grandes qualités techniques. Nous montrerons que cette méthode peut être aussi éloquente qu'une autre si on choisit d'analyser ses résultats de manière appropriée. Ainsi, il est possible de tirer parti de leur performance technique, sans perdre en interprétabilité.

E.3.ii Présentation des Machines à Vecteurs Support (SVM)

Depuis Guadagni et Little (1983) jusqu'à West et al. (1993), toutes les grandes techniques de discrimination issues de différentes disciplines ont été adaptées en modélisation marketing. C'est dans cette tradition que se placent Cui et Curry (2005) en présentant le principe des machines à vecteurs support (SVMs) et en comparant leurs performances à celles de différentes versions de la régression logistique pour mettre en avant leur excellente capacité à prédire les choix des consommateurs. Cet article propose une présentation particulièrement éclairante et concise des SVMs, en les replaçant dans le contexte de l'apprentissage statistique. D'autres articles récents ont utilisé cette nouvelle méthodologie pour optimiser

l'analyse conjointe (Evgeniou et al., 2005), ou des arbres de décision en data mining (Bennet et al. 1998).

Du côté de la modélisation cette fois, des présentations techniques approfondies ont été proposées par Burges (1998) et Crestinaini et Shawe-Taylor (2000) à la suite des travaux fondateurs de Vapnik (1995). En complément des présentations traditionnelles, nous proposons une approche géométrique des SVMs, déjà contenue dans Vapnik (1998), développée par la suite par Bennett et Bredensteiner (2000), et que nous adaptons ici. Cette approche correspond à l'équivalent géométrique du dual du problème d'optimisation initial et propose une vue intuitive des concepts sous-jacents des SVMs, complémentaire de celle proposée par Cui et Curry (2005) et directement interprétable par des non-initiés.

Cas simple

Tout d'abord considérons le nuage bidimensionnel représenté dans la Figure E.3.1. Les individus (indexés par $i=1\dots I$, et décrits par les coordonnées des vecteurs X_i) sont divisés en deux groupes, par exemple les acheteurs ($Y_i=1$) et les non acheteurs ($Y_i=-1$) respectivement représentés par des points noirs et blancs.

Nous utilisons la distance euclidienne.

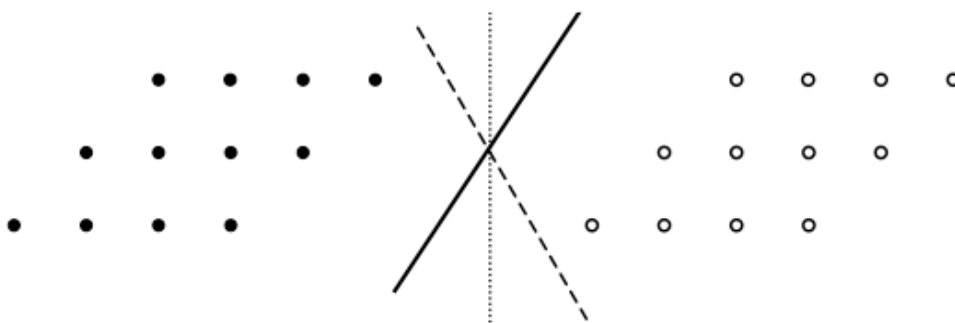


Figure E.3.1. Différentes fonctions de discrimination linéaires

Les deux groupes étant séparables, les trois droites tracées (comme toute autre droite séparant les deux groupes) peuvent être utilisées pour séparer les deux groupes et chaque droite correspond à une fonction de discrimination différente, accordant des poids respectifs différents aux deux variables de classification. La droite tracée en petits pointillés, qui est la médiatrice du segment joignant les centres de gravité des groupes, sépare les groupes suivant un seul critère qui est l'axe des abscisses (la variable de prédiction représentée sur l'axe des ordonnées n'aurait ici aucun pouvoir discriminant). La droite tracée en grands pointillés

suggère qu'une faible valeur suivant l'axe des abscisses pourrait être compensée par une valeur élevée suivant l'axe des ordonnées. Enfin la droite tracée en trait plein suggère qu'une valeur faible sur l'axe des abscisses pourrait être compensée par une valeur faible sur l'axe des ordonnées.

Quelle droite permet la meilleure classification ? Suivant quels critères choisir?

Nous allons montrer dans ce qui suit que la droite tracée en trait plein est la meilleure droite de classification car elle se base sur l'analyse des cas marginaux, et parce qu'elle induit le mode de classification le plus robuste. L'idée centrale des SVMs est de se focaliser sur les cas marginaux. Les vecteurs de supports sont d'ailleurs en réalité les cas les plus marginaux, en ce sens que ce sont les points les plus proches de cette droite, que nous appellerons par la suite droite de classification, et c'est par rapport à eux que sera déterminée la position de cette droite de classification.

Plutôt que de chercher directement la meilleure droite de classification, nous allons d'abord nous attacher à voir où ces deux groupes, acheteurs et non acheteurs, sont les plus proches l'un de l'autre. Techniquement, nous cherchons un point c appartenant à l'enveloppe convexe qui englobe tous les points du groupe des acheteurs et un point d appartenant à l'enveloppe convexe qui englobe tous les points du groupe des non-acheteurs, tels que c et d soient aussi proches que possible. Cette présentation géométrique a été formulée par Bennet et Bredensteiner (2000) par le programme quadratique suivant :

$$c = \sum_{\{i|Y_i=1\}} \alpha_i x_i$$

$$d = \sum_{\{i|Y_i=-1\}} \alpha_i x_i$$

Sous la contrainte :

[2]

$$\sum_{\{i|Y_i=1\}} \alpha_i = 1, \quad \sum_{\{i|Y_i=-1\}} \alpha_i = 1,$$

$$\alpha_i \geq 0 \quad \forall i \in 1 \dots I$$

Comme nous le voyons, c et d sont définis comme appartenant à l'enveloppe convexe de chaque groupe. Pour un optimum, les points X_i ayant des poids α_i strictement positifs soutiennent les points c et d , dans le sens où c et d sont des combinaisons linéaires de ces points et de ces points seulement (les éléments du vecteur de pondération α étant contraints séparément suivant les deux sous-groupes). Ces points ayant des poids strictement positifs sont appelés vecteurs support (et parfois vecteurs de support dans la littérature française sur le sujet, plusieurs traductions du terme anglais support vector étant admises).

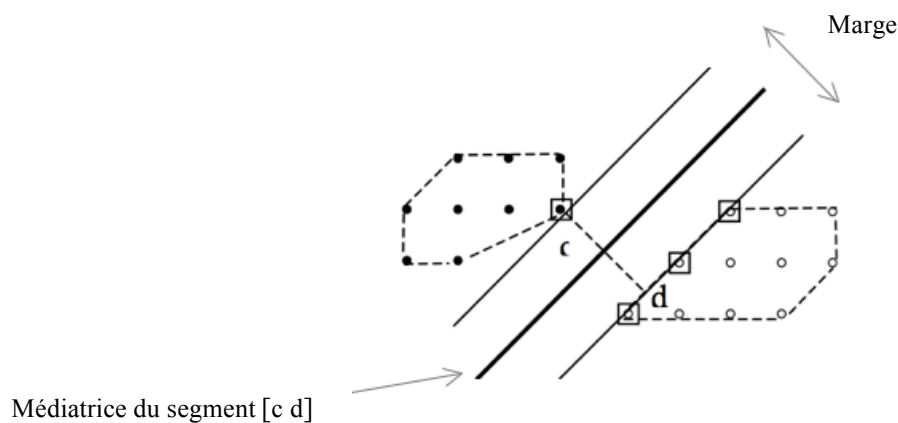


Figure E.3.2. Enveloppes convexes et détermination des vecteurs supports

La figure E.3.2. ci-dessus donne une représentation graphique de la solution de ce programme quadratique. Ici nous avons quatre vecteurs support. Notons que dans le groupe des acheteurs (en noir), il y a un seul vecteur support (c est donc précisément ce point), et que dans l'autre groupe trois vecteurs support sont nécessaires (à la définition du point d). Une fois que c et d sont identifiés, la droite de classification (à ne pas confondre avec la frontière de probabilité décrite dans la sous-section C.1.ii.) que l'on choisira sera la médiatrice du segment reliant c et d ¹¹. La marge, définie par le double de la distance minimale d'un point du nuage de points à la droite de classification, est alors maximisée. En effet, il est facile de constater que tout autre ligne mènerait à une marge plus étroite. C'est pourquoi une autre traduction française des SVMs est Séparateur à Vaste Marge.

¹¹ Dans le cas d'un espace à plus de deux dimensions (disons n) le raisonnement reste parfaitement transposable. La frontière de séparation ne sera plus alors une ligne mais un hyperplan (surface à $n-1$ dimensions).

L'avantage d'avoir une large marge est que des perturbations (petits déplacements autour de la position initiale) appliquées sur les points proches de la droite de classification impliqueront moins d'erreurs qu'avec une marge plus étroite. En effet, les points marginaux sont ceux qui risquent de passer de l'autre côté de la droite de classification et de constituer ainsi des erreurs de prédiction en cas de perturbations. Or dans le cas d'une marge large, ils ont été placés, par définition de la marge, le plus loin possible de la droite de classification. Cet avantage procure aux SVMs de grandes qualités de robustesse.

En effet, pour les lecteurs déjà initiés à l'apprentissage statistique, cette robustesse est établie par le fait qu'une large marge réduit la capacité. La capacité est le nombre de frontières de décisions possibles (Vapnik, 1995). Plus grand est ce nombre, moins grandes sont les chances d'avoir un résultat généralisable. La dimension de Vapnik-Chervonenkis (qui varie de manière opposée à la capacité) maximise la généralisabilité des résultats. Pour illustrer ce concept, prenons l'exemple où le tracé de la frontière est défini par une forme polynômiale. Plus le polynôme est de degré élevé, plus le nombre de frontières possibles est grand, donc plus la capacité est grande. Plus le polynôme est de degré faible, plus le nombre de frontières est faible et donc plus la capacité est réduite. Dans le cas d'un polynôme de degré élevé, il est possible d'avoir des frontières beaucoup plus complexes, mais ceci augmente le risque de surajustement (Vapnik, 1998).

Il est essentiel de remarquer que la droite permettant la meilleure classification est seulement déterminée par les cas les plus marginaux (les vecteurs support).

Ceci constitue une différence conceptuelle importante par rapport aux approches habituelles comme la régression logistique et l'analyse discriminante. Avec les SVMs, la droite de classification est entièrement définie par les points de chaque groupe qui sont les plus proches de l'autre groupe. La détermination de cette droite de classification n'est nullement influencée par les points de chaque groupe qui se situent loin de cette frontière, c'est-à-dire les cas extrêmes. C'est pourquoi cette approche est particulièrement appropriée dans le cadre de notre méthode qui s'attache à caractériser les leviers efficaces sur les individus marginaux, proches de la droite de classification.

Cas où les groupes ne sont pas linéairement séparables

Malheureusement, tous les problèmes de discrimination ne correspondent pas à des groupes linéairement séparables (cas où les groupes peuvent être parfaitement séparés par une droite, dans le cas bidimensionnel, un hyperplan dans les cas de dimension supérieure). Il y a deux cas potentiels dans lesquels les groupes ne sont pas linéairement séparables, d'après Gyon et Stork (1999). Celui où les deux groupes se recouvrent mutuellement et celui où la ligne optimale de classification n'est pas une droite. Même si ces deux cas apparaissent souvent simultanément, il est intéressant de les distinguer parce que leur différence de nature apporte des réponses différentes, qui seront par la suite combinées.



Figure E.3.3. Deux exemples de cas non linéairement séparables

Comme le montre la figure E.3.3.1., dans le cas d'un recouvrement des groupes, les deux enveloppes convexes se recoupent à cause de quelques cas marginaux, mais la ligne de classification optimale semble toujours être linéaire (une droite).

Dans le second cas, sur la figure E.3.3.2, la ligne de classification optimale n'est clairement pas linéaire et une frontière linéaire serait sous-optimale.

Nous allons à présent montrer brièvement comment les SVMs gèrent ces deux cas.

Considérons d'abord le cas le plus fréquent où les deux groupes se recouvrent. Les enveloppes convexes des deux groupes contiennent des points de l'autre groupe. Nous ne pouvons donc plus appliquer la méthode précédente. Cependant nous allons modifier la définition des enveloppes convexes pour qu'elles soient perméables, de manière à ce que certains points du groupe qu'elles englobent puissent se trouver en dehors de cette enveloppe, mais que l'enveloppe englobe encore l'essentiel de ce qui caractérise ce groupe. Les

enveloppes auront ainsi des tailles plus restreintes, ne se recouperont plus, et nous permettront d'appliquer la méthode précédente.

Ces enveloppes modifiées, appelées enveloppes « quasi-convexes » sont obtenues en imposant une valeur maximale D à chaque α_i . En effet, prenons le point j appartenant au groupe -1 mais trop avancé dans le groupe 1 pour faire partie de l'enveloppe quasi-convexe que l'on souhaite déterminer. Si α_j peut être égal à 1, alors l'enveloppe convexe passera nécessairement par ce point parce que s'il est loin des autres points de son groupe, il sera nécessairement sur les contours de l'enveloppe qui entoure *tous* les points du groupe (enveloppe convexe). Par contre, si α_j ne peut être égal à 1, alors l'enveloppe ne passera pas par ce point j , même si ce point influencera partiellement la détermination de l'enveloppe quasi-convexe.

Ainsi les points très éloignés du cœur de leur groupe en direction de l'autre groupe auront certes une influence sur la détermination de la forme de l'enveloppe, mais une influence limitée par cette borne D , et ils pourront se trouver en dehors de l'enveloppe. En choisissant une borne D adaptée, les enveloppes ne se recouperont plus et nous pourrons procéder comme précédemment, à savoir choisir comme droite de classification la médiatrice du segment reliant les points de chaque enveloppe les plus proches de l'autre enveloppe. Plus formellement, une des contraintes du programme d'optimisation quadratique précédent doit être modifiée : la contrainte « α_i supérieur à 0 » devient « α_i supérieur à 0 et inférieur à D . » Ainsi l'algorithme est très similaire au cas précédent, si ce n'est que certains points seront en dehors de l'enveloppe de leur groupe, et que l'enveloppe ne passera par aucun des points du groupe, même si tous les points proches de l'autre groupe auront orienté sa définition.

La borne D synthétise donc le trade-off déjà évoqué entre précision et robustesse. En effet, plus D est faible, plus les enveloppes seront resserrées autour de leurs groupes respectifs. Il y aura ainsi plus d'espace pour « faire passer la droite de classification » entre enveloppes et la marge finale (espace entre la frontière et chacune des enveloppes) sera plus grande. Ainsi plus D est faible, plus la marge est grande et par conséquent, plus la robustesse est grande.

Par contre, il y a un risque plus important de mauvais classement lorsque la valeur de D diminue. En effet, une faible valeur de D autorise un plus grand nombre de points à être en

dehors des enveloppes. Ces points hors des enveloppes risquent d'être « mal classés », puisque, rappelons-le, le choix de la droite de classification est uniquement fondé sur la séparation des enveloppes. Ainsi plus D est faible, plus le risque de commettre des erreurs de prédiction est grand, donc plus la précision est faible. En résumé, plus la valeur de D est faible, plus la discrimination sera robuste, mais moins elle sera précise.

Pour refaire le lien avec le cas précédent, lorsque que l'on a des groupes linéairement séparables, il n'est pas nécessaire de réduire la taille des enveloppes pour pouvoir séparer les groupes. Par défaut D prend donc la valeur 1 c'est-à-dire qu'il s'agit de la véritable enveloppe convexe du groupe, enveloppant tous les points du groupe sans exception. Ceci dit, il arrive que même dans le cas de deux groupes linéairement séparables, l'on veuille obtenir une plus grande marge pour améliorer la robustesse. Ainsi on fixera aussi dans ce cas une valeur de D inférieure à 1, au risque d'un peu dégrader l'ajustement.

Ce mode de contrôle du trade-off entre ajustement et précision à l'aide d'un seul paramètre est un des atouts caractéristiques des SVMs par rapport à d'autres techniques d'apprentissage statistique comme les réseaux de neurones.

Considérons à présent le second cas, pour lequel la ligne de classification optimale n'est pas une droite. Comme on peut le voir sur la figure E.3.3.2., ni les enveloppes convexes, ni les enveloppes quasi-convexes ne sont appropriées pour gérer ce cas. La technique des SVMs repose alors sur le fait de se ramener au cas linéaire en changeant d'espace vectoriel. On va se placer dans un nouvel espace dans lequel la surface de classification optimale sera linéaire, c'est-à-dire sera un hyperplan.

Pour cela, on crée de nouvelles variables à partir des variables initiales (en fait des fonctions des variables initiales) pour placer le nuage de points dans un espace de plus grande dimension. Dans cet espace de plus grande dimension, le nuage de points sera bien plus dispersé, ce qui augmente très rapidement les chances de pouvoir avoir un hyperplan comme surface de classification optimale (Vapnik, 1995). Et alors nous appliquerons simplement la procédure précédente dans cet espace de grande dimension. L'équation de l'hyperplan déterminé dans cet espace de grande dimension pourra s'exprimer à partir des variables initiales dans l'espace initial mais n'aura pas une expression linéaire (il s'agira donc d'une ligne de classification, et non plus d'une droite.). Son expression dépendra des fonctions

utilisées pour créer de nouvelles variables à partir des variables initiales (car l'expression numérique de cette ligne de classification se fera à partir des fonctions inverses des fonctions précédemment utilisées, afin de se ramener à l'espace initial).

L'ajout d'un grand nombre de dimensions permet de se ramener de manière certaine à un cas où les groupes sont linéairement séparables, mais cette manipulation comporte un fort risque du sur-ajustement. Heureusement, le critère d'optimisation de la marge et le critère de lissage par la borne D protègent de ce risque, car comme on l'a vu précédemment, ces critères privilégient la robustesse par rapport à l'ajustement.

Un dernier enjeu est l'aspect hautement calculatoire d'opérations menées dans des espaces de grande dimension. L'innovation technique proposée par les SVMs, connue sous le nom de *kernel trick*, consiste à choisir des transformations $\phi(x)$ des variables initiales pour lesquelles le produit scalaire des vecteurs transformés peut s'exprimer directement en fonction des vecteurs originaux, ce qui revient à dire qu'il existe une fonction K telle que :

$$\phi(x)^T \phi(y) = K(x,y) \quad [3]$$

Si l'on ne change pas de dimension, alors $\phi(x) = x$, et le noyau K est égal au produit scalaire :

$$K(x,y) = x^T y \quad [4]$$

Si l'on veut ajouter comme dimensions les puissances des coordonnées initiales de 0 à p, on utilise alors le noyau polynomial suivant :

$$K(x,y) = (x^T y + 1)^p \quad [5]$$

Par exemple, avec p égal à 3 on aurait pour x : $\phi(x) = (1 : x^T : x^{2T} : x^{3T})^T$

Etant donné que les dimensions ne sont pas modifiées à une constante multiplicative près disons que pour x : $\phi(x) = (a : bx^T : cx^{2T} : dx^{3T})^T$

De même pour y : $\phi(y) = (a : by^T : cy^{2T} : dy^{3T})^T$

Les constantes a, b, c et d sont les mêmes pour toutes les variables étant donné que la transformation ϕ définissant le noyau doit être la même pour toutes les variables.

Le produit scalaire est alors égal à :

$$\phi(x)^T \cdot \phi(y) = (a : bx^T : cx^{2T} : dx^{3T}) \cdot (a : by^T : cy^{2T} : dy^{3T})^T$$

$$\phi(x)^T \cdot \phi(y) = a^2 + b^2 x^T y + c^2 (x^T y)^2 + d (x^T y)^3$$

On a donc bien avec $a = 1$; $b = \sqrt{3}$; $c = \sqrt{3}$ et $d = 1$:

$$\phi(x)^T \cdot \phi(y) = (x^T y + 1)^3$$

Ce qui permet de faire tout le calcul du produit scalaire entre $\phi(x)$ et $\phi(y)$ à partir des coordonnées x et y dans l'espace initial.

Ainsi toutes les opérations sont réalisées à partir des coordonnées dans l'espace initial, ce qui réduit drastiquement le coût de calcul. Pour plus de détails sur cette astucieuse pratique, on peut se référer à nouveau à Vapnik (1995) et Cui and Curry (2005).

E.4. Deuxième étape de la méthode AEL : Allocation au sein des bins

L'étape de discrimination réalisée à l'étape précédente permet de prédire les choix de chaque individu et ainsi de les assigner dans des groupes d'individus ayant des probabilités de choix identiques.

Or dans certains cas, la méthode de classification fournit directement la probabilité a posteriori, comme dans le cas de la régression logistique (logit). Mais dans d'autres cas, comme pour les SVMs, les valeurs de sortie, à savoir les valeurs des distances des points représentant les individus à la surface de classification, ne sont pas directement interprétables en termes de probabilités.

Comme nous l'avons déjà précédemment indiqué ce cas est rarement problématique car il y a souvent plusieurs manières d'ajuster des probabilités a posteriori sur les sorties de la fonction de classification.

Pour les SVMs, la méthode la plus utilisée (Cui and Curry, 2005) est celle proposée par Platt (2000) qui consiste à redimensionner les sorties entre 0 et 1 grâce à une sigmoïde dont les paramètres sont calculés par maximum de vraisemblance. Dans l'idée, il s'agit d'une

approche similaire à celle consistant à passer du *linear probability model* au *logit* comme exposé par Maddala (2001, p.320).

Décrivons les étapes proposées par Platt (2000). Pour cela notons *dist* la fonction indiquant la distance euclidienne entre un point et la surface de classification. Le principe de la méthode de Platt est d'exprimer la probabilité de faire le choix 1 en fonction de la valeur de *dist* sous la forme d'une sigmoïde à savoir sous la forme suivante pour tout individu *i* :

$$f(x_i) = f(\text{dist}(x_i)) = 1 / \{1 + \exp (u \text{ dist}(x_i) + v)\} \quad [6]$$

Les valeurs prises par *f* seront ainsi comprises entre 0 et 1. Reste à déterminer la valeur des deux paramètres *u* et *v*. Les valeurs de ces paramètres sont ajustées en utilisant le maximum de vraisemblance sur l'ensemble des couples (*y_i*, *dist(x_i)*). Pour cela il définit au préalable un autre ensemble de couples (*t_i*, *dist(x_i)*) où les *t_i* sont définis comme suit :

$$t_i = (y_i + 1) / 2 \quad [7]$$

et prennent les valeurs 0 (si *y_i* = -1) ou 1 (si *y_i* = 1).

Les paramètres *u* et *v* sont alors fixés en minimisant l'inverse de la log vraisemblance :

$$\text{Min} - \sum_i [t_i \log(f(x_i)) + (1-t_i) \log(1 - f(x_i))] \quad [8]$$

avec $f(x_i) = 1 / [1 + \exp (u \text{ dist}(x_i) + v)]$

Cette minimisation se fait sur l'ensemble des couples (*t_i*, *dist(x_i)*). Pour les calculs, un algorithme purement technique de minimisation est proposé en annexe de l'article de Platt (1999), dans les pages 9 et 10¹².

A partir de cette fonction de probabilité, chaque individu est attribué à un bin spécifique. Chaque bin contient des individus d'une tranche donnée de probabilités. Si l'on fixe cette tranche à 20% de probabilité : le premier bin contient les individus ayant entre 0 et 20% de chances de faire le choix A, la seconde contient les individus ayant 20 à 40% de chances de faire le choix A, la dernière contient les individus ayant 80 à 100% de chances de faire le

¹² Document téléchargeable en libre accès depuis la page suivante:
<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639>

choix A. Plus la tranche choisie est étroite, plus les effets étudiés seront locaux, mais moins les estimations de ces effets seront précises. Nous reviendrons sur ce point plus tard.

Ainsi on peut s'intéresser localement aux individus en fonction de leur distance à la frontière. Rappelons que la frontière est le seuil de probabilité définie dans la sous-section C.2.ii. La plupart du temps, on étudie des données où les groupes sont équilibrés et la frontière est autour de 50%. Dans cette situation, les cas marginaux pour lesquels nous avons un intérêt particulier se situeront dans les bins du milieu, et les cas extrêmes se situeront dans les premiers et les derniers bins. Mais dans le cas, moins fréquent mais possible, dans lequel la frontière de décision est fixée à un seuil de probabilité supérieur à 50% (cas où plus d'individus font le choix 1 que le choix -1) ou inférieur à 50% (cas où moins d'individus font le choix 1 que le choix -1) alors les cas marginaux seront ceux situés dans les bins contenant ce seuil (ou autour de ce seuil). Par exemple, si très peu d'individus font le choix -1, et que dans la première étape le seuil de discrimination a été fixé à 90%, alors les cas marginaux seront situés dans les derniers bins, et les cas extrêmes dans les premiers bins.

L'avantage de l'allocation des individus dans des groupes définis par tranches de probabilités égales est que cette allocation est indépendante du seuil de probabilité choisi comme frontière, et qu'il suffit de connaître la valeur de ce seuil pour savoir dans quels bins se situent les cas marginaux (ou extrêmes).

Un mode d'allocation alternatif aurait été d'allouer les individus par quantiles de probabilités. Ce mode d'allocation est moins directement interprétable car ses bornes ne peuvent pas être fixées a priori. En effet si l'on souhaite par exemple décomposer la population en 5 sous-populations, alors le premier bin contiendra les 20% de la population dont les probabilités sont les plus faibles, et les bornes de ce premier bin seront donc par exemple entre 0% et 16,23% de chances de faire le choix 1. Pour le second bin contenant les individus ayant le moins de chances de faire le choix 1 dans la population restante (population initiale sans les individus du premier bin), alors les bornes seront par exemple de 16,23% à 34,56%, et ainsi de suite.

Les bornes des bins dépendant de la répartition de la population, elles ne peuvent pas être fixées a priori et les interprétations des effets locaux risquent d'être moins aisées. Ceci sera d'autant plus vrai lors de la représentation des effets que nous proposerons plus loin au sein

du spectre des effets locaux. Ce type d'allocation sera réservé donc aux cas où les individus seraient très mal répartis dans les bins définis par des tranches de probabilité fixes. Dans le cas contraire, il sera préférable d'allouer les individus par tranches de probabilité fixes.

E.5. Troisième étape de la méthode AEL : Mesure des effets locaux

E.5.i. ANOVA fonctionnelle (FANOVA)

Comme nous l'avons précisé auparavant, nous avons besoin de calculer l'influence de chacun des critères sur la fonction de probabilité, et ce, sur tout l'échantillon.

La plupart des méthodes de discrimination ne permettent pas de calculer facilement ces contributions – et sont habituellement qualifiées de boîtes noires à cause de ce manque d'interprétabilité. Une solution pour mesurer cette influence est pourtant de décomposer la fonction de probabilité en *fonctions* des variables initiales à l'aide d'une méthode nommée ANOVA fonctionnelle (ou FANOVA comme functional ANOVA). La méthode FANOVA a été établie par Owen (1992), sur la base de plusieurs travaux préliminaires sur ce thème, que Hooker (2004) fait remonter jusqu'à Hoeffding (1948). Owen (1992) décrit cette technique comme une décomposition ANOVA pour des fonctions continues. En effet, la fonction étudiée (ici la fonction de probabilité) est, comme on le verra par la suite, décomposée en une somme d'un terme constant et de fonctions monovariées des variables de prédiction.

Principe de FANOVA

Cette étape consiste à décomposer la fonction f sur une base de produits de fonctions orthogonales et normées. Soient (ψ_r) cette base fonctionnelle et (β_r) les coefficients de la décomposition de f , r étant l'indice générique d'indexation de cette base.

On note x un individu donné et $x^{(j)}$ sa coordonnée correspondant à la j -ième variable. On a :

$$f(x) = \sum_r \beta_r \psi_r(x) \quad [9]$$

où :

$$\psi_r(x) = \prod_{j=1}^d \phi_r(x^{(j)}) \quad [10]$$

Nous rappelons que comme indiqué page précédente, les j indexent les variables de prévision. Les coefficients de cette décomposition¹³ sont égaux à :

$$\beta_r = \int_{[0,1]^d} f(x)\psi_r(x)dx \quad [11]$$

Les suites de coefficients r_j qui interviennent dans la définition de $\psi_r(x)$ définissent le type d'effet. Quand tous les r_j sont nuls, $\psi_r(x)$ modélise un effet constant. Quand un seul coefficient r_j est égal à 1 (resp. 2), un effet du premier ordre (resp. second ordre) est modélisé. Quand deux coefficients r_j sont égaux à 1, on modélise un effet croisé.

Dans notre étude, la base (ψ_r) utilisée est celle des polynômes de Legendre ; et les suites r_j qui sont les indices des éléments sélectionnés, vérifient les propositions suivantes :

$$\begin{aligned} \|r\|_1 &= \sum r_j \leq \text{degre} \\ \|r\|_0 &= \#\{j, r_j \neq 0\} \leq \text{rang} \\ \|r\|_\infty &= \max r_j \leq \text{ordre} \end{aligned} \quad [12]$$

Le degré, le rang et l'ordre ci-dessus doivent être fixés. Les valeurs choisies pour le rang et l'ordre sont 4 ; et 8 pour le degré. Autrement dit, nous ne considérons pas des effets croisés faisant intervenir plus de 4 variables ; ou des effets au-delà du quatrième ordre.

Cette énumération peut être obtenue par une procédure récursive.

Sur un exemple simple, suivant les notations précédentes, avec une limite de 2 sur le rang et l'ordre, et 3 pour le degré, et une dimension égale à 4, la base serait constituée des fonctions de la page suivante.

¹³ On pensera à l'analogie avec la décomposition en séries de Fourier

$$\begin{aligned}
\psi_0(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= 1 \\
\psi_1(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(4)}) \\
\psi_2(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_2(x^{(4)}) \\
\psi_3(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(3)}) \\
\psi_4(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(3)})\phi_1(x^{(4)}) \\
\psi_5(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_2(x^{(3)}) \\
\psi_6(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(2)}) \\
\psi_7(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(2)})\phi_1(x^{(4)}) \\
\psi_8(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(2)})\phi_1(x^{(3)}) \\
\psi_9(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_2(x^{(2)}) \\
\psi_{10}(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(1)}) \\
\psi_{11}(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(1)})\phi_1(x^{(4)}) \\
\psi_{12}(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(1)})\phi_1(x^{(3)}) \\
\psi_{13}(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_1(x^{(1)})\phi_1(x^{(2)}) \\
\psi_{14}(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) &= \phi_2(x^{(1)})
\end{aligned} \tag{13}$$

La décomposition consiste en l'estimation des coefficients β_r . L'approche utilisée suit directement celle décrite par Jiang et Owen (2002). On observe en effet que l'intégrale permettant de calculer la valeur d'un coefficient peut s'interpréter comme l'espérance de la fonction $f(x)\psi_r(x)$, pour x distribué uniformément sur le volume $[0,1]^d$:

$$\beta_r = \int_{[0,1]^d} f(x)\psi_r(x)dx \tag{14}$$

On peut donc directement estimer empiriquement ces coefficients de la manière suivante :

$$\beta_r = \frac{1}{n} \sum_{i=1}^n f(x_i)\psi_r(x_i) \tag{15}$$

où les x_i sont une suite de variables aléatoire iid, uniformément distribuées dans $[0,1]^d$. Suivant cette hypothèse très courante pour des raisons de convergence de calcul, l'espérance de la formule 14 prend bien l'expression de la formule 15, et donc il n'y a pas lieu de changer la notation β_r dans la suite de l'exposé, pour des raisons de lisibilité.

À chaque étape, la qualité du modèle peut être mesurée par la décroissance d'un critère de « lack of fit (lof) » égal au rapport entre l'erreur quadratique et la variance intrinsèque de f :

$$lof_i = \frac{(\sum_r \beta_r \psi_r(x_i) - f(x_i))^2}{\sigma_f^2} \quad [16]$$

La procédure de shrinkage qui suit respecte aussi celle donnée et détaillée par Jiang et Owen (2002). Elle permet d'identifier au fil des itérations les fonctions de base dont les coefficients ont une forte probabilité d'être nuls. Cela permet non seulement d'accélérer les calculs, puisque la fonction $f(x_i)$ est de plus en plus rapide à évaluer à mesure que la taille de la base rétrécit ; mais cela permet surtout de réduire la variance de l'estimation de chacun des coefficients β_r identifiés comme non nuls.

Cette procédure d'intégration de Monte-Carlo a une complexité en $O(np)$; tandis qu'une estimation directe des coefficients visant à directement minimiser l'erreur quadratique aurait une complexité en $O(Np^2)$ (calcul de pseudo-inverse par décomposition en valeurs singulières), où N est le nombre total d'observations. Dans nos exemples, nous avons fixé le nombre d'itérations à $n = 10000$ en utilisant le graphe de « lack of fit » comme vérification de la convergence. Même si $N < n$, le gain en temps de calcul est tout de même de l'ordre d'un facteur 100. Lors de la présentation des résultats, les décompositions des fonctions de classification seront données, tout comme les graphiques de convergence. Seuls les termes correspondant aux 50 coefficients β_r les plus forts sont donnés.

Pour alléger les notations dans la suite de la méthode, on appelle la fonction $f_k(x)$ la composante additive pure de la variable $x_{(k)}$ défini suivant la formule suivante :

$$f_k(x) = \sum_{r, ||r||_0=1, r_k \neq 0} \beta_r \phi_r(x) \quad [17]$$

D'après les équations 9 à 15 précédentes, cette fonction de la formule 17 donne l'effet de la k -ième variable sur la fonction de probabilité étudiée pour un individu x quelconque. Elle est appelée « composante » parce que c'est à travers la somme des composantes suivant les différentes variables de prévision (qui indexent ces composantes) que l'on reconstitue la fonction de probabilité étudiée. Cette composante est dite « additive » parce qu'il faut

sommer les composantes pour reconstituer la fonction de probabilité étudiée (par opposition à des composantes multiplicatives qui auraient nécessité le produit des composantes pour reconstituer la fonction en question). Ces composantes additives sont dites « pures » parce qu'elles ne correspondent chacune qu'à une variable et une seule au sein des variables de prévision utilisées dans la modélisation (par opposition à des composantes qui seraient indexées par plusieurs variables de prévision à la fois). Cette composante permet donc d'obtenir, sous une forme fonctionnelle, une mesure de l'impact de chacune de ces variables prises séparément (terme « pure ») sur la fonction de probabilité décomposée, et de recomposer (terme « composante ») cette fonction à l'aide de la somme (terme « additive ») de ces composantes, d'où le terme de composante additive pure. Elles seront utilisées par la suite pour la mesure des effets locaux.

E.5.ii. Calcul des effets locaux

Soit B_k l'ensemble des observations appartenant au bin de probabilité centrale p et de largeur $2r$ (pour en découpage en 8 bins, le premier bin a pour valeur centrale $p_1 = 0,0625$ et pour largeur $2r = 0,125$). Nous avons :

$$B_k = \{(x)_i, |f(x_i) - p_k| < r\} \quad [18]$$

L'effet moyen absolu $E_a(j,k)$ de la variable j sur le bin k consiste en la moyenne de la valeur prise par la composante additive pur de cette variable (on élimine les effets des autres variables et les effets croisés), lorsqu'elle est évaluée sur toutes les observations assignées dans B_k :

$$E_a(j, k) = \frac{1}{|B_k|} \sum_{x \in B_k} f_j(x) \quad [19]$$

La formule de l'équation 19 ci-dessus constitue une mesure de l'effet local de la variable j au niveau du bin k , car elle est composée au numérateur de la somme des effets de la variable j pour les individus appartenant au k -ième bin, et au dénominateur de l'effectif (nombre d'individus) de cette k -ième bin. En effet, la fonction $f_j(x)$ donnant l'effet de la variable j sur la fonction de probabilité pour l'individu x , il suffit de calculer la moyenne de ces effets pour

tous les individus d'un bin pour obtenir l'effet moyen local (car correspondant à un bin et un seul) de cette variable sur ce bin. Les individus étant jugés d'égale importance (tous les individus de l'échantillon apportent une information de qualité équivalente), cette moyenne n'est pas pondérée. Si la méthode d'échantillonnage avait impliqué une pondération suivant les individus, cette pondération aurait été utilisée pour réaliser le calcul d'une moyenne pondérée en substitution d'une moyenne simple dans la formule 19 ci-dessus.

Les effets relatifs locaux signés, produits finals de l'analyse et représentés sur les graphes AEL, sont obtenus par simple normalisation des effets absolus, de manière à ce que la somme des effets relatifs de toutes les variables dans un bin soit égale à 1. Leur expression est donc :

$$E_r(j, k) = \frac{E_a(j, k)}{\sum_i |E_a(i, k)|} \quad [20]$$

Notons que ce calcul a été modifié suite à la pré-soutenance. Le jury proposait dans son rapport de prendre au numérateur la valeur réelle de l'effet moyen absolu, à la place de sa valeur absolue, ce qui a sensiblement amélioré la facilité d'interprétation des résultats.

Cette étape conclut l'application de la méthode AEL. Reste à vérifier la qualité des résultats, étape par étape.

E.6. Évaluation de la qualité des résultats

Décrivons les critères d'évaluation suivant chaque étape de la méthode AEL.

Etape 1

Pour la première étape, notre critère de qualité est bien entendu la qualité de la prévision puisque les étapes qui suivent consisteront à étudier la fonction de probabilité de choix issue de cette phase de discrimination et les impacts des critères de choix sur cette fonction de probabilité. Cette mesure nous permettra de choisir le meilleur mode de discrimination parmi plusieurs modes de discrimination testés (la régression logistique, les réseaux de neurones ou les SVMs pour ne citer qu'eux), et pour un même mode de discrimination (par exemple les

SVMs), quelle est la meilleure spécification possible (par exemple la valeur de la borne D permettant d'arbitrer entre précision et sur-ajustement).

La mesure de la qualité de prévision pourrait se faire en calculant le taux d'individus « bien classés » (pour lesquels le choix prédit sera égal au choix réel) comme nous l'avons fait pour les régressions logistiques emboîtées de la section D. Cependant cette méthode ne permettrait pas de prendre en compte un éventuel risque de sur-ajustement. C'est pour cette raison que nous calculons une moyenne d'individus bien classés par validation croisée.

Dans ce dessein, nous générons cent échantillons comprenant 90% des individus et nous calculons cent fois le taux de « bien classés » parmi les 10% d'individus à chaque fois mis à part dans un sous-échantillon de test. La moyenne de ces *taux* d'individus « bien classés » nous donne notre critère de qualité de prévision. Nous reprenons ici exactement la même méthode que celle que Cui et Curry (2005) ont utilisée pour comparer des méthodes de discrimination (notamment les SVMs).

Etape 2

Le critère qui s'impose pour ajuster la fonction de probabilité est le maximum de vraisemblance. Il est au cœur de l'estimation dans la régression logistique et de la méthode de Platt (2000) pour les SVMs.

Etape 3

La difficulté principale de l'application de cette étape consiste à choisir la bonne taille pour les bins. En effet, plus il y a de bins, plus détaillée sera la description des phénomènes locaux étudiés. Mais plus il y a de bins, plus les bins sont étroits, plus le nombre d'individus dans chaque bin décroît, et donc plus faible est la précision des estimations des effets locaux. En effet, les contributions fonctionnelles sont estimées sur l'échantillon complet pour tirer un maximum d'informations possible, et les effets relatifs locaux sont calculés ensuite sur la base du nombre d'individus dans chaque bin. Il ne peut y avoir de réponse toute faite à cette question. Par contre, nous proposons un critère de choix simple pour prendre cette décision.

Nous générons pour chaque estimation des effets locaux une estimation de la variance de ces effets en utilisant la méthode de bootstrap (suivant les mêmes paramètres qu'Evgeniou et al. 2005). L'arbitrage entre le nombre de bins et la précision des estimations des effets locaux peut donc se faire suivant le niveau d'imprécision toléré par l'utilisateur, mesuré à l'aide de la

variance des estimations des effets locaux. Si l'on s'intéresse principalement aux individus à la marge, on peut calculer la variance moyenne des estimations des effets sur les deux bins correspondant à la zone frontière et prendre cet indicateur comme un indicateur de précision des estimations. Par expérience, pour un échantillon de 1000 individus, il est difficile d'aller au-delà d'une dizaine de bins (avec en moyenne 100 individus par bin) parce que la précision serait alors vraiment faible. Pour ce type d'échantillon le nombre de bins adéquat serait plutôt situé entre 5 et 10 bins suivant le niveau de variance choisi. Encore une fois, il est laissé ici une part d'arbitraire à l'utilisateur et le meilleur moyen de se faire une idée avant d'arbitrer est de faire tourner la méthode sur différents nombres de bins et d'observer les niveaux de variance correspondants comme nous le montrerons dans la partie suivante sur les résultats.

Etape complémentaire

Il est aussi intéressant de pouvoir tester les différences entre les valeurs des effets locaux. Pour cela, il est d'abord nécessaire d'observer les distributions des estimations de ces effets relatifs locaux, car contrairement à la régression traditionnelle, nous ne connaissons pas ici a priori la distribution des estimations des effets. Il s'agit donc dans un premier temps de tester la normalité des distributions des effets bin par bin, variable par variable. A cette fin, nous utilisons le test de Jarque-Bera (1980) test habituel pour vérifier si les résidus d'une régression linéaire respectent l'hypothèse de normalité, à partir de distributions empiriques calculées par bootstrap (à nouveau suivant les mêmes paramètres qu'Evgeniou et al. 2005).

Nous verrons dans les résultats que dans la quasi-totalité des cas traités dans le cadre de cette recherche, les tests ne permettent pas de rejeter avec un risque de plus de 5% la normalité des distributions des estimations des effets. Les graphiques des distributions empiriques des effets et les résultats de tous les tests de normalité seront eux aussi fournis dans les résultats.

Ensuite nous appliquons des t-tests par variables en testant deux à deux les différences inter-bins à l'aide des estimations des variances des effets, et nous pourrions ainsi tester la variation de la valeur des effets suivant la position des sous-groupes d'individus par rapport à la frontière. Ces t-test prennent en compte les différences de variance entre les groupes non appariés traités. Au-delà du simple test de différence d'un bin à l'autre, nous proposons de tester toutes les différences entre tous les effets, sur tous les bins, comme dans l'approche Bonferroni, consistant à faire des tests deux-à-deux en utilisant un risque de première espèce.

On pourra constater par la suite que dans majorité des cas que nous allons traiter les différences entre les moyennes des estimations des effets sont significatives au seuil de 5%.

Cette approche globale permet de tester le plus objectivement possible les différences d'effets entre les populations à la marge et aux extrêmes, en ayant pris au préalable le soin de vérifier que les distributions sur lesquelles sont fondées ces tests peuvent être considérées comme gaussiennes.

D'un point de vue calculatoire, les résultats obtenus sur la mesure des effets locaux est liée à la procédure de Monte-Carlo (MC) réalisée lors de la quasi-régression (aussi notée). La procédure de Monte-Carlo est basée sur une séquence de nombres aléatoires, générée pour initier cette procédure. Comme l'estimation des effets locaux peut varier en fonction des nombres aléatoires utilisés, il est nécessaire d'itérer plusieurs fois la procédure de MC afin de mesurer la sensibilité des estimations. Cette sensibilité est ensuite mesurée par l'écart-type des estimations au sein des différentes itérations de la procédure de MC. Ici nous avons choisi de réitérer 30 fois la procédure de MC afin de mesurer l'écart-type des estimations. Durant les procédures de test visant à dire si les différences entre les effets locaux estimés sont ou non significatives, nous devons tenir compte de la taille de l'échantillon. Ici l'échantillon est composé des différentes estimations des effets locaux. La taille des échantillons correspond donc aux "nombres d'individus", c'est-à-dire ici aux nombres d'itérations de la procédure de MC, qui est de 30, et ce pour chaque bin (à ne pas confondre avec le population de consommateurs dans chaque bins). C'est pourquoi les tests d'égalité entre effets sont réalisés en prenant 30 comme taille des échantillons pour tous les tests réalisés.

Code MatLab

En complément de cette description technique, nous proposons en Annexe 7 les codes MatLab développés spécifiquement pour ces procédure de test et qui permettent à tout lecteur possédant MatLab et désireux d'appliquer cette méthodologie, d'obtenir directement :

- le tableau de l'ensemble des effets locaux, avec les signes,
- les graphiques avec la distribution des valeurs des effets selon les itérations,
- tests de normalité,
- tests de significativité des différences de moyenne des effets entre les bins.

La section concernant les tests a été placée en italique.

F. Applications de la méthode AEL

Les différentes étapes de la méthode AEL vont être appliquées sur deux bases de données afin d'en donner des illustrations concrètes. Nous commencerons par détailler les résultats sur la base *Orange Juice*, déjà utilisée dans le cadre des régressions logistiques emboîtées de la section D, et compléterons par une nouvelle illustration sur le choix de vote à un *Référendum*.

F.1. Cas du choix d'une marque de Jus d'orange

Reprenons une à une les étapes à appliquer avant d'analyser les résultats. Nous rappelons ici que la base de données utilisée a été décrite dans la sous-section C.3.

F.1.i. Discrimination

Il s'agit de prédire les choix des consommateurs entre les deux marques (MM et CH) en fonction des cinq critères disponibles à savoir :

- le prix de vente de base (hors réduction de prix) de Minute Maid : MM Price ;
- le prix de vente de base de Citrus Hill : CC Price ;
- le taux de réduction de prix sur Minute Maid : MM Discount ;
- le taux de réduction de prix sur Citrus Hill : CH Discount ;
- la fidélité à Citrus Hill: Loyalty, mesurée par les préférences passées.

Nous choisissons de prédire la préférence des consommateurs pour la marque CH.

Comme nous l'avons indiqué auparavant, nous allons tester dans cette première étape deux modes de discrimination : l'analyse discriminante de Fisher et les SVMs linéaires ou polynomiaux. En effet, nous souhaitons utiliser deux méthodes basées sur des principes très différents pour montrer à travers cet exemple la robustesse de la méthode par rapport au mode de discrimination choisi lors de cette première étape.

Pour les SVMs, nous avons testé à cette étape deux spécifications à savoir :

- la spécification linéaire pour laquelle on reste dans l'espace vectoriel initial (tout en utilisant cependant des enveloppes quasi-convexes pour éviter les risques de sur-ajustement) ;

- la spécification correspondant au noyau suivant (cf. sous-section C.3.ii) : pour tout couple de vecteur (x,y) , $K(x,y)$ est égal à $(1 + x.y)^2$, qui revient à utiliser tous les polynômes du second ordre dans les variables de prédiction initiales en plus des variables initiales pour définir la fonction de discrimination optimale (tout en utilisant là encore des enveloppes quasi-convexes pour éviter les risques de sur-ajustement).

Une fois les fonctions de discrimination évaluées¹⁴, elles nous fournissent les scores de précision suivants (critère de qualité de la première étape décrit dans la section E.6.). Les calculs ont été effectués sur l'erreur de classification traditionnelle (« bien classé » ou « mal classé » suivant la décision binaire à prévoir).

Fonction de discrimination	Erreur en validation croisée
Analyse discriminante	17.1%
SVMs linéaires	17.0%
SVMs avec noyau polynomial (degré 2)	17.3%

Tableau F.1.1. Scores de validation croisée

D'après ce tableau, tous les modes de classification ont des scores de précision assez proches (autour de 17%). Même si le meilleur score est donné aux SVMs linéaires, suivis de l'analyse discriminante et des SVMs polynomiaux, ces différences ne sont pas significatives.

L'étude de Cui et Curry (2005) aurait donné a priori un avantage aux SVMs mais étant donné le grand nombre d'individus dans la base et le petit nombre de paramètres à estimer, SVMs et analyse discriminante donnent des résultats très proches. Il en aurait été autrement avec un plus grand nombre de paramètres à estimer. Nous allons voir dans la sous-section suivante en quoi la méthode SVM a cependant un avantage dans le cas présent de l'analyse AEL.

¹⁴ Les calculs des SVMs ont été réalisés à l'aide du module Spider de MatLab, plus spécifiquement la fonction *train svm*. Ce module est très bien reconnu et diffusé en Machine Learning, notamment à travers diverses publications de haut niveau, comme par exemple: Evgeniou, Pontil, et Elisseeff (2004). Il est librement téléchargeable depuis la page: http://www.kyb.mpg.de/bs/people/spider/download_frames.html

Note sur la performance relative des fonctions de classification

Etant donné que la méthodologie AEL peut s'appliquer indépendamment de l'outil de modélisation choisi, et que donc, le choix de la fonction de discrimination n'a pas à être guidé par son pouvoir explicatif inhérent, il est en effet légitime de se demander si ce léger gain de précision sur les performances en classification est une motivation suffisante au choix des SVMs. Nous nous proposons dans cette sous-section d'apporter un éclairage nouveau sur les performances en classification des différentes approches.

La méthodologie est la suivante. Le jeu de données Orange Juice est utilisé pour élaborer trois types de fonctions de classification différentes : une analyse discriminante de Fisher ; un SVM linéaire ; un SVM polynomial de degré 2. Les résultats de la classification sont présentés de la manière originale suivante : les individus sont regroupés en 5 bins, selon la probabilité de choix estimée par la fonction de classification ; et dans chaque bin, le pourcentage des erreurs de classification des individus issus de ce bin est donné par rapport au nombre total d'erreurs effectuées par cette fonction de classification. Nous voulons savoir ici quelle est la *distribution des erreurs, et ce pour type de fonction de classification*. Les résultats sont donnés dans le tableau F.1.2. ci-après.

Rappelons que par « erreurs de classification » nous désignons toujours les individus pour lesquels le choix prédit par la méthode de modélisation considérée s'avère malheureusement être en réalité le choix contraire au choix effectivement fait par cet individu. Notons aussi qu'il est plus logique de s'intéresser au pourcentage d'erreurs selon les bins qu'au nombre absolu d'erreurs, car lors de l'allocation des individus par bin, les effectifs des bins peuvent varier d'une méthode de classification à l'autre, et les nombres absolus d'erreurs ne sont donc pas correctement interprétables. Les erreurs est donc ici calculées de la même manière que dans le tableau F.1.1 précédent.

Bin	Fischer	SVM linéaire	SVM polynomial
1	8.2%	7.8%	7.9%
2	26.8%	32.4%	36.0%
3	38.8%	30.2%	29.3%
4	19.7%	22.5%	20.1%
5	6.5%	7.1%	6.7%

Tableau F.1.2. Comparaison des erreurs en validation croisée

Ainsi, bien que les performances en classification globale soient très semblables pour les trois méthodes, la distribution des erreurs par bin est assez différente : l'analyse de Fisher peine davantage sur les exemples proches de la frontière de décision ; tandis que les SVMs ont une répartition des erreurs plus homogène suivant les bins, les erreurs de classification devenant ainsi plus courantes sur les individus moins proches de la marge. Il apparaît ainsi que le cas où la classification est vue non pas comme une fin en soi, mais comme un outil visant à identifier les individus proches de la marge, et à modéliser le comportement de la fonction de décision sur les individus proches de la frontière, le pouvoir de résolution des SVMs est plus grand. Ce résultat n'est pas surprenant d'après la présentation du mode de construction des SVMs que nous avons proposée dans la section E.3.ii.

Bien que la littérature relative au machine learning et à la classification automatique privilégie le taux d'erreur absolu comme critère d'erreur, il est assez préférable dans notre approche d'accorder plus d'importance aux individus à la marge. Cependant le praticien peut choisir de privilégier la méthode qu'il lui convient le mieux, ou qu'il lui est le plus pratique, du moment que la qualité de classification, y compris au niveau des individus à la marge lui convient. En effet, la suite de la méthodologie AEL peut s'appliquer dans tous les cas de figure où le praticien dispose d'une probabilité de choix par individu. Cette flexibilité constitue un atout important de la méthodologie AEL.

Afin d'illustrer la flexibilité et la robustesse de cette méthode par rapport à la première étape de discrimination, nous allons conserver les SVMs linéaires et l'analyse discriminante pour traiter la suite de notre exemple. Ce qui nous intéressera en particulier ne sera bien entendu pas la corrélation directe entre les valeurs des probabilités des individus suivant les différents modes de classification, qui peut recouvrir plusieurs réalités très différentes au niveau du lien entre ces probabilités et les variables de prévision, mais de comparer les impacts locaux de ces différentes variables de prévision à l'aise de la méthode FANOVA.

F.1.ii Calcul des probabilités et des effets locaux

Voici la description des deux étapes suivantes de la méthode AEL : le calcul des probabilités de choix par individu, et celui des effets locaux par critères de choix.

Calcul des probabilités de choix

La deuxième étape de calcul de la fonction de probabilité est immédiate pour l'analyse discriminante qui fournit en sortie une évaluation de la probabilité de choix pour chaque individu.

Concernant les SVMs linéaires, nous appliquons la méthode décrite dans la section E.4. Afin de donner autant de détails que possible sur cette application, nous proposons ci-dessous la figure F.1.1. qui donne la probabilité affectée à chaque individu, en fonction de la distance du point représentant cet individu à la surface de classification issue des SVMs, d'après la méthode de maximisation de la vraisemblance de Platt (2000) présentée précédemment.

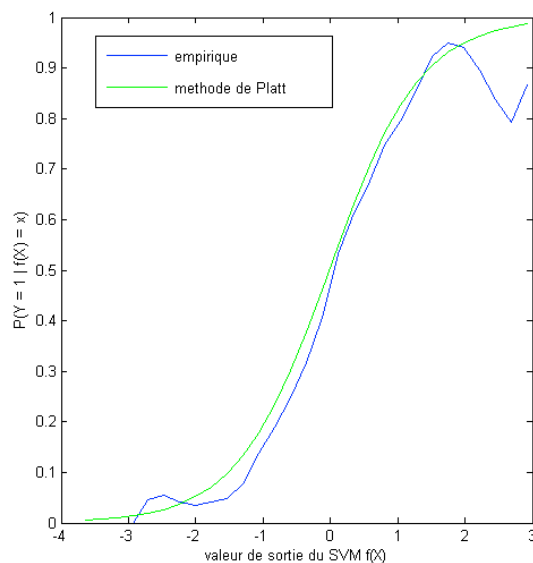


Figure F.1.1. Calcul des probabilités de choix pour les SVMs linéaires

Le graphique affiche, en fonction des valeurs de sortie des SVMs linéaires, la courbe de probabilité empirique (en bleu) et la sigmoïde obtenue par la méthode de Platt (en vert) pour estimer la probabilité de choix de chaque individu. Cette sigmoïde servira par la suite, pour les SVMs linéaires, de fonction de probabilité de choix p dans le calcul des effets locaux.

Application de FANOVA

Pour le jeu de données *Orange Juice*, nous allons noter de la manière suivante les critères de choix afin de rendre les formules aussi lisibles que possible.

$x^{(1)} = \text{MM Price}$

$x^{(2)} = \text{CH Price}$

$x^{(3)} = \text{Loyalty}$

$x^{(4)} = \text{MM Disc}$

$x^{(5)} = \text{CH Disc}$

Commençons d'abord par le cas de l'analyse discriminante. L'application de l'ANOVA fonctionnelle permet de décomposer la fonction de probabilité de choix en une somme de fonctions monovariées dépendant des critères de choix ci-dessus. Nous appellerons ces fonctions monovariées les contributions de chaque critère.

La décomposition de la fonction de décision est la suivante :

Contribution de MM price

$$+(-0.0861\phi_1(x^{(1)})) + (0.0006\phi_2(x^{(1)})) + (0.0006\phi_3(x^{(1)})) + (0.0005\phi_4(x^{(1)}))$$

Contribution de CH price

$$+(0.0652\phi_1(x^{(2)})) + (-0.0035\phi_2(x^{(2)})) + (0.0009\phi_3(x^{(2)})) + (0.0010\phi_4(x^{(2)}))$$

Contribution de Loyal

$$+(0.2848\phi_1(x^{(3)})) + (0.0013\phi_2(x^{(3)})) + (-0.0300\phi_3(x^{(3)})) + (-0.0013\phi_4(x^{(3)})) \quad [21]$$

Contribution de MM disc

$$+(0.0757\phi_1(x^{(4)})) + (-0.0032\phi_2(x^{(4)})) + (0.0054\phi_3(x^{(4)})) + (0.0053\phi_4(x^{(4)}))$$

Contribution de CH disc

$$+(-0.0809\phi_1(x^{(5)})) + (0.0047\phi_2(x^{(5)})) + (0.0012\phi_3(x^{(5)})) + (0.0023\phi_4(x^{(5)}))$$

Dans l'expression de ces contributions, la notation ϕ_i correspond au i -ième polynôme de Legendre, comme cela a été précédemment indiqué dans la sous-section E.5.i.

A partir de ces contributions, nous allons pouvoir déterminer les effets relatifs locaux (et leurs écarts-type), suivant la méthode indiquée dans la sous-section E.5.ii., consistant à calculer les moyennes (simples) des contributions ci-dessus par bin.

Ils sont présentés dans le tableau ci-dessous :

bin	MM price	CH price	Loyal	MM disc	CH disc
1	-0.10 ± 0.01	-0.01 ± 0.00	-0.58 ± 0.02	-0.18 ± 0.02	0.13 ± 0.02
2	-0.11 ± 0.01	-0.04 ± 0.01	-0.40 ± 0.02	-0.22 ± 0.02	0.23 ± 0.03
3	-0.14 ± 0.01	-0.07 ± 0.01	-0.15 ± 0.02	-0.32 ± 0.04	0.32 ± 0.04
4	-0.14 ± 0.02	-0.04 ± 0.01	-0.07 ± 0.02	-0.32 ± 0.04	0.43 ± 0.04
5	-0.08 ± 0.01	-0.10 ± 0.01	0.16 ± 0.02	-0.23 ± 0.03	0.43 ± 0.04
6	-0.05 ± 0.01	-0.06 ± 0.01	0.38 ± 0.02	-0.20 ± 0.03	0.31 ± 0.04
7	-0.05 ± 0.01	-0.05 ± 0.01	0.51 ± 0.02	-0.16 ± 0.02	0.23 ± 0.02
8	-0.06 ± 0.01	0.00 ± 0.00	0.62 ± 0.02	-0.11 ± 0.02	0.21 ± 0.02

Tableau F.1.3. Effets relatifs locaux (Fisher)

Pour nous aider à lire ces résultats de manière pratique, nous avons besoin d'une représentation graphique de l'ensemble des effets sur l'ensemble des bins. Cette représentation ne reportera pas les écarts-type qui devront être consultés dans le tableau ci-dessus.

Spectre AEL

Nous avons mis en place une méthode de représentation des effets relatifs locaux permettant d'avoir une lecture directe des résultats d'une analyse AEL. Les bins sont ordonnés de 1 à 8, et la composition des effets au sein de chaque bin est représentée par une barre verticale. Chaque barre est composée des effets locaux relatifs des critères de choix, donc la somme est toujours égale à 100%. Ainsi chaque barre représente, pour chaque bin, l'influence de chacun des critères dans le choix des individus, ou pour être plus précis une décomposition de la part prévisible du choix des consommateurs suivant les critères de choix étudiés (les variables de prédiction utilisées). Les influences négatives (jouant en défaveur de l'option de choix étudiée) sont représentées en-dessous de l'axe des abscisses, alors que les influences positives (jouant en faveur de l'option de choix étudiée) sont représentées au-dessus.

Dans la partie gauche du graphe (bins 1 à 4) on peut analyser des critères de choix des consommateurs ayant une plus grande probabilité de choisir MM que CH, de ceux dont le choix est le plus prévisible en faveur de MM (bin 1) à ceux dont le choix l'est le moins (bin4). Dans la partie droite du graphe (bins 5 à 8), une analyse des critères de choix des consommateurs ayant une plus grande chance de choisir CH, cette fois de ceux dont le choix est le moins prévisible en faveur de CH (bin 5) à ceux dont le choix l'est le plus (bin 8).

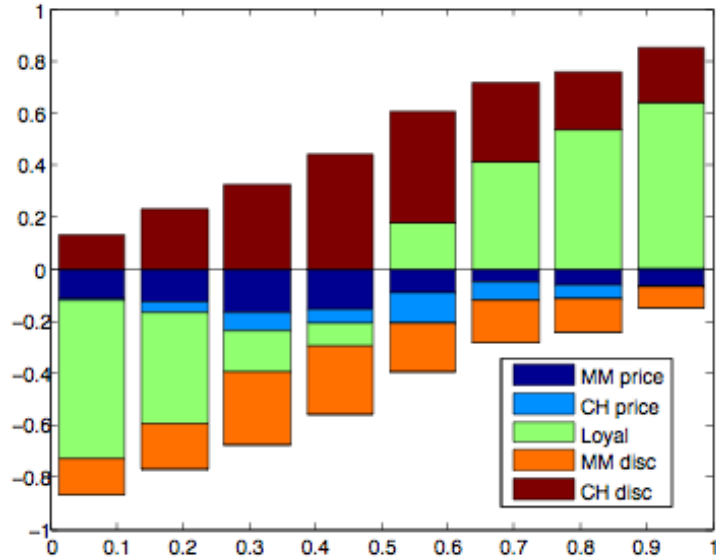


Figure F.1.2. Spectre AEL pour Orange Juice (Fisher)

En faisant de même à partir des résultats issus de SVMs linéaires, on obtient les contributions suivantes :

Contribution de MM price

$$+(-0.0793\phi_1(x^{(1)})) + (0.0021\phi_2(x^{(1)})) + (0.0004\phi_3(x^{(1)})) + (0.0037\phi_4(x^{(1)}))$$

Contribution de CH price

$$+(0.0449\phi_1(x^{(2)})) + (0.0034\phi_2(x^{(2)})) + (0.0024\phi_3(x^{(2)})) + (0.0014\phi_4(x^{(2)}))$$

Contribution de Loyal

$$+(0.2839\phi_1(x^{(3)})) + (0.0030\phi_2(x^{(3)})) + (-0.0239\phi_3(x^{(3)})) + (0.0038\phi_4(x^{(3)}))$$

[22]

Contribution de MM disc

$$+(0.0947\phi_1(x^{(4)})) + (0.0075\phi_2(x^{(4)})) + (-0.0012\phi_3(x^{(4)})) + (0.0026\phi_4(x^{(4)}))$$

Contribution de CH disc

$$+(-0.0834\phi_1(x^{(5)})) + (-0.0017\phi_2(x^{(5)})) + (0.0002\phi_3(x^{(5)})) + (0.0021\phi_4(x^{(5)}))$$

Ces contributions nous permettent de calculer les effets relatifs locaux par bin. Ils sont donnés dans le tableau F.1.4. ci-dessous :

bin	MM price	CH price	Loyal	MM disc	CH disc
1	-0.10 ± 0.00	-0.00 ± 0.00	-0.56 ± 0.01	-0.20 ± 0.02	0.14 ± 0.01
2	-0.11 ± 0.01	-0.02 ± 0.00	-0.40 ± 0.02	-0.24 ± 0.02	0.22 ± 0.02
3	-0.13 ± 0.01	-0.05 ± 0.01	-0.15 ± 0.02	-0.35 ± 0.03	0.34 ± 0.03
4	-0.10 ± 0.02	-0.06 ± 0.02	-0.00 ± 0.04	-0.37 ± 0.03	0.43 ± 0.03
5	-0.09 ± 0.02	-0.05 ± 0.01	0.17 ± 0.03	-0.28 ± 0.03	0.41 ± 0.04
6	-0.05 ± 0.02	-0.05 ± 0.01	0.34 ± 0.02	-0.23 ± 0.02	0.34 ± 0.04
7	-0.04 ± 0.01	-0.04 ± 0.01	0.49 ± 0.02	-0.17 ± 0.01	0.26 ± 0.02
8	-0.05 ± 0.00	-0.01 ± 0.00	0.61 ± 0.02	-0.13 ± 0.01	0.21 ± 0.02

Tableau F.1.4. Effets relatifs locaux (SVMs)

On peut remarquer que les valeurs des effets sont proches de celles obtenues précédemment dans le tableau F.1.3. ce qui signifie que les fonctions de probabilité issues de l'analyse discriminante et des SVMs linéaires ont des structures sous-jacentes proches et que l'analyse FANOVA met en évidence ces similarités.

En comparant la figure F.1.3. ci-dessous à la précédente, on constate que les principales différences portent sur les bins 4 et 5. Ceci s'explique par le fait que c'est pour les individus à la marge que l'on a observé le plus de différences entre ces méthodes de classification (comme indiqué dans la sous-section F.1.i.).

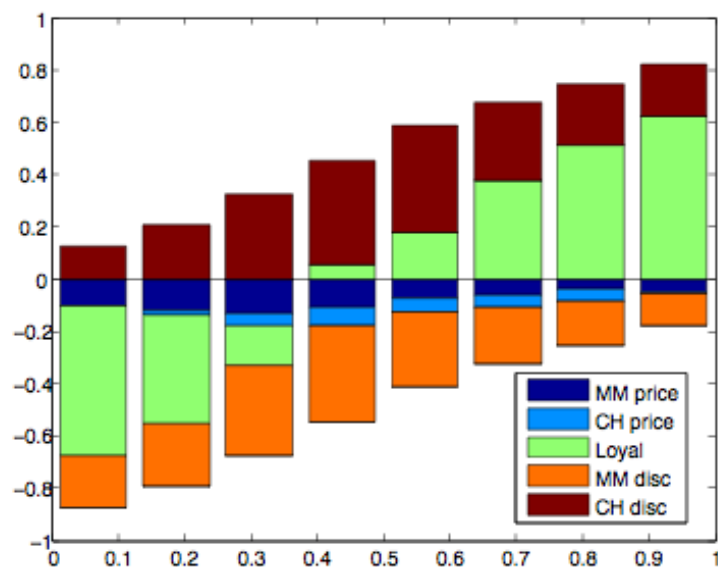


Figure F.1.3. Spectre AEL Orange Juice (SVMs)

Note sur le nombre de bins

Afin de choisir le nombre définitif de bins, nous calculons les effets locaux relatifs des variables de prédiction dans plusieurs cas de figure, de 2 à 10 bins.

Nous avons ainsi recalculé, pour différents nombres de bins, les estimations des effets locaux relatifs et leurs variances respectives, obtenues par bootstrap (suivant les mêmes paramètres qu'Evgeniou et al. 2005). Ces variances nous permettent de connaître le niveau d'incertitude dans la mesure des effets, car il s'agit de la variance des estimations de ces effets. Comme indiqué auparavant, il s'agit de faire un compromis entre le niveau de description que nous voulons obtenir (nombre de bins) et l'incertitude liée à cette description (variance des estimations des effets). Ici, nous décidons de fixer le nombre de bins à 8 pour arbitrer entre la variance moyenne de l'estimation de l'effet local relatif de Loyalty (qui décroît entre 6 et 10 bins) et celle de MM Price (qui augmente entre 6 et 10 bins), qui sont les critères de choix dont les effets sont les plus importants en moyenne.

On pourra confirmer, à partir des analyses qui suivent dans la section F.1.iii. concernant les tests d'égalité entre les effets, que ce choix de 8 bins nous donne une description assez précise des bins. En effet, la plupart des différences entre effets sont significatives, et il n'est donc pas nécessaire de passer de 8 à 6 (ou 4) bins pour avoir des résultats directement interprétables.

F.1.iii. Tests d'égalité sur les effets

Comme indiqué dans la partie E.6. il est nécessaire de vérifier que les différences de valeurs des effets relatifs entre les bins sont bien significatives. Pour cela nous allons effectuer des tests d'égalité à partir des écarts-type obtenus par bootstrap.

Cependant, il est nécessaire de vérifier préalablement que les distributions des estimations des effets relatifs locaux sont gaussiens.

Nous allons commencer par l'AEL issue de l'analyse discriminante et nous passerons ensuite à celle issue des SVMs linéaires.

Les résultats de ces tests de normalité de Jarque-Bera (1980) sont reportés dans le tableau F.1.5. ci-dessous.

Chaque case contient la valeur de la statistique de test pour la variable et le bin correspondant, puis la mention « rej. » si l'hypothèse de normalité peut être rejetée avec un risque de moins de 5% et « norm » dans le cas où l'hypothèse de normalité n'est pas rejetée.

On peut constater que l'hypothèse de la normalité n'est rejetée que 4 fois sur les 40 tests réalisés. Pour ces 4 cas, il ne sera donc pas possible de conclure sur la significativité des différences entre les valeurs des effets.

bin	MM price	CH price	Loyal	MM disc	CH disc
1	0.83 norm.	2.18 norm.	1.38 norm.	8.80 rej.	0.46 norm.
2	1.30 norm.	0.73 norm.	0.04 norm.	2.08 norm.	1.96 norm.
3	1.81 norm.	5.17 norm.	0.93 norm.	2.51 norm.	0.45 norm.
4	0.76 norm.	1.01 norm.	6.00 rej.	1.36 norm.	1.05 norm.
5	6.32 rej.	1.23 norm.	14.69 rej.	0.42 norm.	0.33 norm.
6	1.77 norm.	4.02 norm.	0.56 norm.	0.40 norm.	0.16 norm.
7	0.68 norm.	1.35 norm.	1.90 norm.	0.76 norm.	0.38 norm.
8	0.88 norm.	4.38 norm.	1.14 norm.	1.80 norm.	0.48 norm.

Tableau F.1.5. Test de normalité Orange Juice (Fisher)

En complément, les distributions empiriques des estimations des effets sont données dans la figure F.1.4. ci-après.

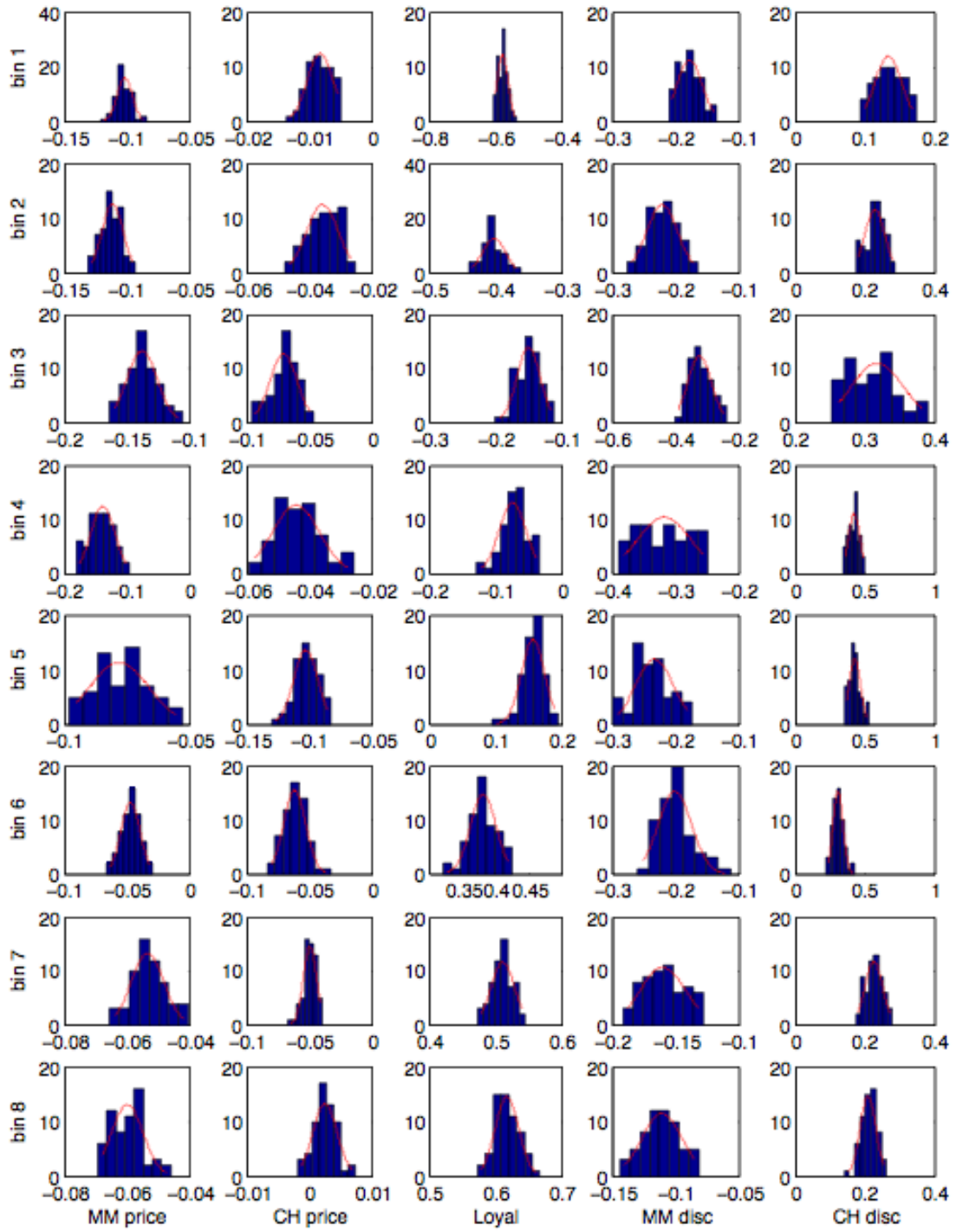


Figure F.1.4. Distributions empiriques Orange Juice (Fisher)

Les mêmes tests ont été réalisés sur les effets AEL issus des SVMs linéaires. On peut constater sur le tableau F.1.6. ci-dessous que l’hypothèse de la normalité n’est cette fois rejetée que deux fois. Ceci constitue une amélioration par rapport au cas précédent d’autant que les rejets d’hypothèses portaient dans le cas précédent sur les bins 4 et 5 qui sont particulièrement importants dans le cadre de notre étude. Il est fort probable que cette différence vienne du fait que les SVMs, même si leur score de classification global est proche de celui de l’analyse discriminante, sont plus performants à la marge, comme nous l’avons fait remarquer dans la sous-section F.1.i.

bin	MM price	CH price	Loyal	MM disc	CH disc
1	0.33 norm.	0.70 norm.	1.67 norm.	3.28 norm.	0.48 norm.
2	0.46 norm.	0.06 norm.	0.26 norm.	3.62 norm.	1.06 norm.
3	0.08 norm.	1.53 norm.	0.57 norm.	1.35 norm.	0.76 norm.
4	0.88 norm.	2.72 norm.	2.51 norm.	1.05 norm.	0.26 norm.
5	0.88 norm.	2.59 norm.	1.70 norm.	3.18 norm.	0.10 norm.
6	0.74 norm.	1.42 norm.	0.04 norm.	7.74 rej.	2.04 norm.
7	8.49 rej.	0.14 norm.	0.04 norm.	0.40 norm.	1.85 norm.
8	1.16 norm.	0.22 norm.	1.24 norm.	0.73 norm.	0.04 norm.

Tableau F.1.6. Test de normalité Orange Juice (SVMs)

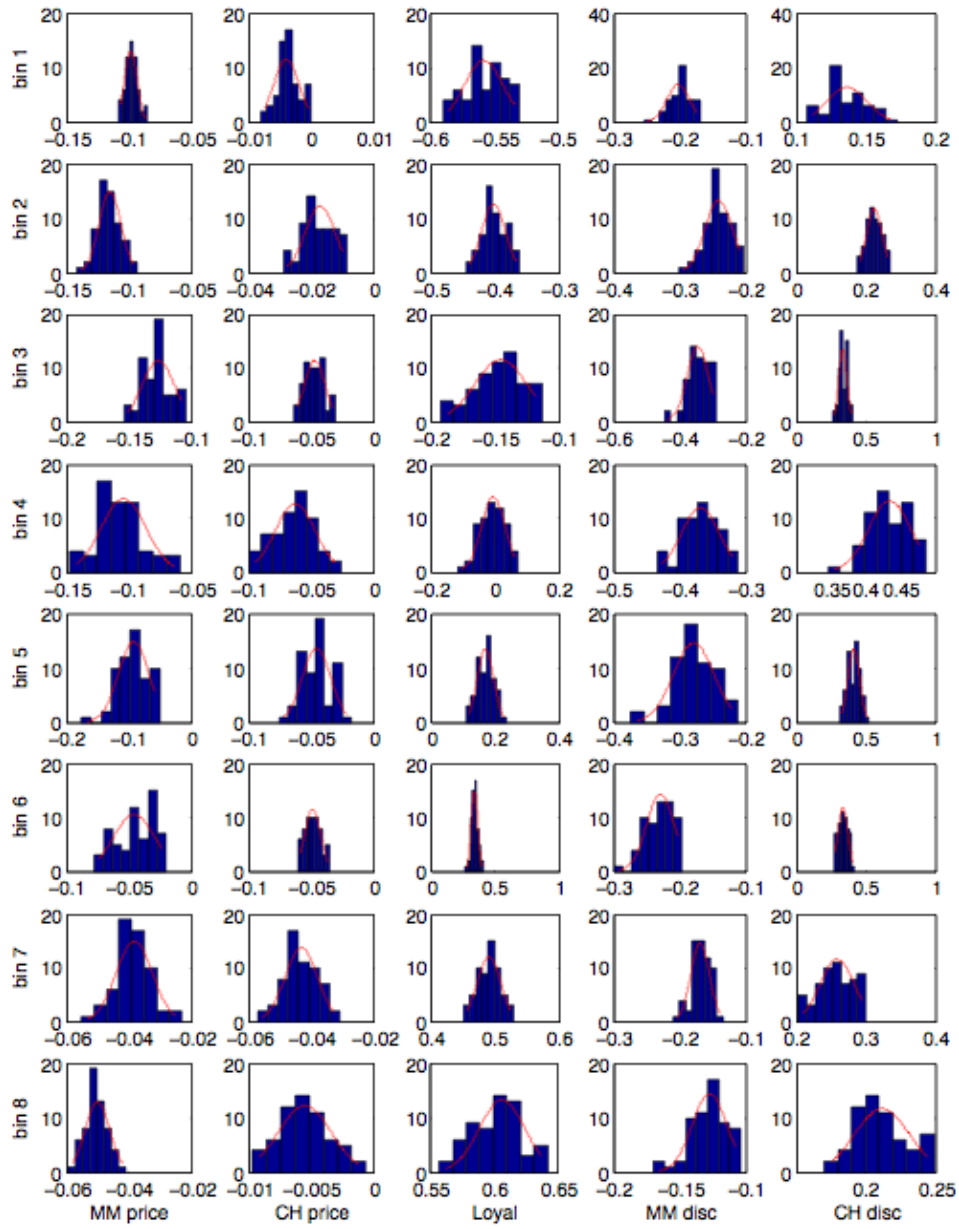


Figure F.1.5. Distributions empiriques Orange Juice (SVM)

Une fois les tests de normalité effectués, nous pouvons passer aux tests d'égalité des effets entre les différents bins, et ce, variable par variable dans les tableaux ci-dessous. Ces t-tests prennent en compte les différences de variance entre les groupes non appariés traités.

Le test t a donc été calculé en divisant les différences entre les effets proposés dans le tableau F.1.3. ci-avant, par la racine carré de la somme des carrés des deux écarts-types correspondant à ces effets, *divisés par les effectifs de chaque bin*. Prenons l'exemple de la différence entre le bin 1 et le bin 8 sur l'effet local de « loyalty » calculé à partir de l'analyse discriminante. Selon le tableau F.1.3., on obtient le résultat du tableau F.1.9. donnant la valeur 269,78 pour cette statistique de test.

Dans la présentation des résultats des pages suivantes, chaque case contient la valeur de la statistique de test pour la variable et le bin correspondant, puis la mention « diff » si l'hypothèse d'égalité entre les valeurs des moyennes peut être rejetée avec un risque de moins de 5% et « rej.» dans le cas où l'hypothèse d'égalité n'est pas rejetée et où l'on ne peut donc pas conclure à la différence entre les valeurs des moyennes.

bin	1	2	3	4	5	6	7	8
1								
2	-5.20 diff.							
3	-14.66 diff.	-9.83 diff.						
4	-10.07 diff.	-7.08 diff.	-0.09 rej.					
5	9.73 diff.	12.61 diff.	19.35 diff.	14.73 diff.				
6	30.96 diff.	31.23 diff.	34.77 diff.	24.42 diff.	12.18 diff.			
7	33.26 diff.	32.21 diff.	35.01 diff.	23.57 diff.	10.52 diff.	-3.83 diff.		
8	29.07 diff.	28.79 diff.	32.44 diff.	21.81 diff.	7.78 diff.	-7.78 diff.	-4.92 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	-27.58 diff.							
3	-30.79 diff.	-15.85 diff.						
4	-25.42 diff.	-5.16 diff.	10.87 diff.					
5	-52.61 diff.	-33.74 diff.	-12.15 diff.	-26.27 diff.				
6	-32.36 diff.	-13.97 diff.	3.45 diff.	-8.25 diff.	17.23 diff.			
7	-40.23 diff.	-10.30 diff.	9.56 diff.	-3.10 diff.	26.67 diff.	6.55 diff.		
8	20.95 diff.	38.73 diff.	36.20 diff.	33.17 diff.	58.77 diff.	39.01 diff.	51.19 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	40.56 diff.							
3	97.35 diff.	56.61 diff.						
4	105.11 diff.	67.67 diff.	15.49 diff.					
5	167.15 diff.	125.53 diff.	68.06 diff.	47.15 diff.				
6	206.69 diff.	166.90 diff.	111.96 diff.	89.03 diff.	47.38 diff.			
7	246.63 diff.	204.25 diff.	146.08 diff.	119.12 diff.	78.37 diff.	27.23 diff.		
8	269.78 diff.	227.27 diff.	169.02 diff.	140.37 diff.	101.53 diff.	49.39 diff.	23.31 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	-7.57 diff.							
3	-19.02 diff.	-12.60 diff.						
4	-16.69 diff.	-10.96 diff.	0.57 rej.					
5	-8.58 diff.	-1.72 rej.	10.44 diff.	9.09 diff.				
6	-3.89 diff.	3.11 diff.	14.82 diff.	13.06 diff.	4.50 diff.			
7	3.64 diff.	11.04 diff.	21.76 diff.	19.11 diff.	11.63 diff.	7.08 diff.		
8	14.55 diff.	20.74 diff.	28.91 diff.	25.52 diff.	20.06 diff.	16.21 diff.	11.36 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	15.27 diff.							
3	24.21 diff.	10.36 diff.						
4	35.44 diff.	21.52 diff.	10.66 diff.					
5	35.77 diff.	21.88 diff.	11.02 diff.	0.38 rej.				
6	24.06 diff.	9.65 diff.	-0.94 rej.	-11.80 diff.	-12.17 diff.			
7	16.17 diff.	-0.45 rej.	-11.23 diff.	-22.76 diff.	-23.13 diff.	-10.53 diff.		
8	14.15 diff.	-2.98 diff.	-13.69 diff.	-25.35 diff.	-25.71 diff.	-13.09 diff.	-2.73 diff.	

*Tableaux F.1.7-11. Tests d'égalité Orange Juice (Fisher)
suivant l'ordre MM Price, CH Price, Loyalty, MM Disc, CH Disc*

bin	1	2	3	4	5	6	7	8
1								
2	-9.07 diff.							
3	-11.20 diff.	-4.04 diff.						
4	-1.70 rej.	3.01 diff.	5.58 diff.					
5	1.43 rej.	4.89 diff.	6.87 diff.	2.22 diff.				
6	17.00 diff.	20.60 diff.	21.40 diff.	13.30 diff.	8.66 diff.			
7	43.29 diff.	40.23 diff.	34.47 diff.	19.31 diff.	11.75 diff.	2.35 diff.		
8	41.78 diff.	37.19 diff.	31.30 diff.	16.26 diff.	9.29 diff.	-1.45 rej.	-9.28 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	-13.89 diff.							
3	-27.78 diff.	-17.21 diff.						
4	-19.93 diff.	-14.94 diff.	-4.73 diff.					
5	-19.74 diff.	-12.66 diff.	0.55 rej.	4.74 diff.				
6	-36.73 diff.	-21.79 diff.	-0.89 rej.	4.41 diff.	-1.30 rej.			
7	-38.69 diff.	-20.30 diff.	2.19 diff.	6.34 diff.	1.08 rej.	3.73 diff.		
8	-2.94 diff.	12.17 diff.	26.71 diff.	19.39 diff.	18.97 diff.	35.32 diff.	36.96 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	36.13 diff.							
3	89.68 diff.	51.33 diff.						
4	75.69 diff.	52.47 diff.	18.16 diff.					
5	136.29 diff.	100.57 diff.	53.45 diff.	21.37 diff.				
6	168.98 diff.	131.08 diff.	82.65 diff.	42.25 diff.	26.32 diff.			
7	243.99 diff.	189.14 diff.	127.85 diff.	65.61 diff.	56.56 diff.	26.59 diff.		
8	274.49 diff.	215.94 diff.	152.49 diff.	81.16 diff.	77.42 diff.	47.31 diff.	24.75 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	-7.97 diff.							
3	-21.57 diff.	-14.94 diff.						
4	-27.46 diff.	-19.81 diff.	-2.97 diff.					
5	-11.99 diff.	-5.59 diff.	8.22 diff.	11.77 diff.				
6	-5.47 diff.	2.22 diff.	16.57 diff.	21.56 diff.	7.32 diff.			
7	8.26 diff.	15.77 diff.	27.50 diff.	34.31 diff.	17.84 diff.	13.02 diff.		
8	19.18 diff.	25.46 diff.	34.39 diff.	42.06 diff.	24.92 diff.	22.54 diff.	11.83 diff.	

bin	1	2	3	4	5	6	7	8
1								
2	17.85 diff.							
3	34.02 diff.	16.54 diff.						
4	48.72 diff.	30.15 diff.	12.85 diff.					
5	34.07 diff.	21.52 diff.	8.32 diff.	-2.15 diff.				
6	28.14 diff.	14.22 diff.	-0.02 rej.	-11.42 diff.	-7.67 diff.			
7	23.85 diff.	5.60 diff.	-11.22 diff.	-24.74 diff.	-17.41 diff.	-9.70 diff.		
8	17.12 diff.	-2.42 diff.	-19.68 diff.	-33.95 diff.	-23.86 diff.	-16.65 diff.	-8.39 diff.	

Tableau F.1.12-16. Tests d'égalité Orange Juice (SVM)
suivant l'ordre MM Price, CH Price, Loyalty, MM Disc, CH Disc

Ces tests d'égalité permettent de vérifier que dans la quasi-totalité des cas, les effets à la marge (bins 4 et 5) sont différents des effets aux extrêmes (bins 1 et 8, voire souvent 2 et 7).

De manière générale, la plupart des effets sont significativement différents d'un bin à l'autre. Les quelques fois où ces différences ne sont pas significatives correspondent à des cas proches de la diagonale c'est-à-dire des cas d'égalité entre deux bins voisins. De plus, on peut aussi constater que les différences non significatives apparaissent moins souvent dans le cas de l'AEL issue des SVMs. Ceci recoupe les remarques précédentes sur le fait que les SVMs donnent une description plus précise de la zone frontière. Ainsi sur cette zone, non seulement les tests de normalité nous indiquent que la distribution est gaussienne et de plus les tests d'égalité rendus ainsi possibles, nous indiquent que les différences d'effets inter-bins sont significatives.

Ces vérifications ayant été effectuées nous pouvons passer à l'interprétation marketing des résultats obtenus.

F.1.iv. Interprétation des résultats

En appliquant la méthodologie AEL sur ces données Orange Juice, nous pouvons confirmer plusieurs intuitions intéressantes concernant les effets locaux attendus, en cohérence avec les exemples que nous avons traités dans la section A.2, la revue de littérature de la section B.1. et avec les premiers résultats obtenus à l'aide des régressions logistiques emboîtées :

(R1) **Loyalty** est un critère de choix pour les individus extrêmes car il a un effet local relatif fort sur les cas extrêmes (avec des effets locaux relatifs de -56, -40, +49 et +61% sur les bins 1, 2, 7 et 8) et un effet local relatif sensiblement plus faible sur les cas marginaux (0 et 17% sur les bins 4 et 5). Ceci peut être attribué au fait que les consommateurs les plus fidèles feront des choix plus prévisibles sur la base de la régularité de leurs choix précédents que ne le feront les consommateurs moins réguliers pour lesquels le critère de fidélité rentre beaucoup moins en ligne de compte. À la lecture des spectres AEL des pages précédentes, on observe que la structure des effets locaux de cette variable de prédiction a une forme de « papillon » caractéristique d'un critère jouant principalement pour les cas extrêmes.

(R2) **Discounts (CH Discount et MM Discount)** sont des critères de choix pour les cas marginaux car ces critères ont des effets locaux relatifs faibles pour les cas extrêmes (bins 1, 2, 7, 8) et élevés pour les cas marginaux (bins 4 et 5). En effet, pour le critère CH Discount, les effets locaux relatifs sont respectivement de 6, 14, 22, 26 et 21% pour les bins 1, 2, 7 et 8 ; et de 43 et 41% pour les bins 4 et 5. Pour le critère MM Discount, les effets locaux relatifs sont aussi déséquilibrés entre ces deux types de bins, même si les amplitudes sont moindres avec -20, -24, -17 et -13% pour les bins 1, 2, 7 et 8 ; et -37 et -28% pour les bins 4 et 5.

Ceci s'explique aussi facilement par le fait que les consommateurs moins sensibles aux marques n'ayant pas encore de forte préférence établie attacheront plus d'importance aux promotions parce qu'ils sont plus sensibles aux variations de prix, et que la promotion est fortement susceptible d'être une raison d'achat. De plus, le rôle plus important de la variable MM Discount par rapport à CH Discount apparaît certainement parce que les campagnes de promotion de MM sont plus fréquentes et plus importantes. De plus, la marque MM étant en premium, ces réductions de prix sont censées avoir plus d'impact.

(R3) **Price (CH Price et MM Price)**: Les critères de prix ont des effets relatifs plus faibles que les autres, car il s'agit des prix corrigés des opérations promotionnelles. Cependant, on peut observer qu'ils sont légèrement plus influents sur les bins marginaux que sur les bins extrêmes. En effet, pour le cas de CH Price, les valeurs sont de -6 et -5% pour les bins 4 et 5, alors qu'ils sont quasi-nuls pour les bins 1 et 8. Concernant MM Price, ils sont de -10 et -9% pour les bins 4 et 5 contre -10 et -5% pour les bins 1 et 8.

Par contre, il est très frappant de constater l'asymétrie de l'influence de la variable MM Price à travers le spectre AEL. En effet, l'influence du prix pour les consommateurs privilégiant la marque MM (à gauche du spectre) est de -10, -11, -13 et -10% (bins 1 à 4) alors qu'elle est plus faible pour les consommateurs privilégiant la marque CH avec des valeurs -9, -5, -4 et -5% sur les bins 5 à 8. Cette asymétrie avait déjà pu être détectée par Krishnamurthi, Mazumdar, et Raj (1992) dans leur étude des switchers. Elle signifie que les consommateurs de CH ne consomment pas CH à défaut de pouvoir s'offrir la marque MM qui a un prix plus élevé. A l'inverse, les consommateurs privilégiant la marque MM la consomment en partie du fait de son prix plus élevé qu'ils interprètent comme un signal de qualité. La variable CH Price, quant à elle, a un pouvoir explicatif très limité (effets relatifs de faible amplitude), elle ne joue en défaveur de la marque CH que pour les individus des bins marginaux.

Globalement, les résultats montrent que notre méthode AEL rend possible l'identification de différents types de critères : ceux jouant sur les cas extrêmes ou sur les cas marginaux, et ceux de nature symétriques ou asymétriques.

Dans cet exemple, l'étude des variables liées au prix est particulièrement riche et instructive. Le fait que les critères de pricing de MM (prix et réductions de prix) aient globalement des effets locaux plus importants peut provenir du fait que MM avait une stratégie de pricing plus dynamique (prix relativement élevé et politique promotionnelle agressive) et que les consommateurs moins réguliers sont ainsi plus sensibles au prix et aux réductions de prix pour cette marque. De plus, il y a une différence notable entre la structure des effets du prix (asymétriques) et de la réduction de prix (équilibrés entre clients et non-clients).

F.2. Application de la méthode AEL sur la base Référendum

Appliquons à présent le même type d'analyse mais sur le cas d'un choix unique, et non répété, à savoir celui du vote à un référendum.

F.2.i. Référendum 2005

En complément de cette analyse sur de la base Orange Juice, nous avons choisi d'analyser une autre problématique concernant un choix binaire par excellence : celui d'un référendum.

Le 29 mai 2005, un référendum concernant le projet de traité constitutionnel européen a été organisé en France et s'est conclu par un refus de ce projet par 55% des électeurs contre 45% en faveur du projet. Le taux de participation a été important, atteignant près de 70% des électeurs. Cet événement a eu jusqu'à aujourd'hui un retentissement particulier parce qu'il s'agissait du refus d'un des membres fondateurs de l'Union et qu'il a été suivi d'autres refus mettant en cause tout le processus de ratification. Jean Monnet, l'un des pères de la construction européenne, avait sans doute pressenti ce type d'accident quand il disait : « l'Europe se fera dans les crises et elle sera la somme des solutions apportées à ces crises. »

Dans cette étude, notre but est de mieux comprendre les causes de ce refus en France, et pour cela, d'observer ce qui a pu influencer la décision des électeurs « à la marge » qui ont fait basculer le résultat final en faveur du non.

Il s'agit bien en effet d'un phénomène de marge, il suffirait que 5% de la population totale ait voté différemment, ce qui correspondrait à 9% de la population des électeurs ayant voté pour le « non ». En nous intéressant aux 10% des électeurs ayant voté « non » ayant eu le plus de chances de voter « oui », nous pourrions savoir quels ont été les éléments décisifs qui ont fait basculer le vote vers le « non ».

F.2.ii. Données d'enquête IPSOS

L'enquête est réalisée le jour du vote par l'institut de sondage IPSOS sur un échantillon représentatif de 2577 personnes.

Les questions que nous utilisons pour cette analyse sont les six questions suivantes, pouvant ici être considérées comme des critères de choix pour le vote. Nous en déduisons en effet sept critères: trois binaires (Vote, Opposition et Extrêmes) et quatre échelles (Projet, Construction, Confiance, Campagne).

Le choix à prédire est l'issue du vote au référendum. Cette variable est mesurée par la réponse oui ou non à la question suivante :

"Avez-vous ou allez-vous voter aujourd'hui oui au projet de constitution européenne ou non au projet de constitution européenne ?"

Le choix que nous allons analyser est la réponse « non » à cette question. Dans la lecture du spectre AEL les personnes ayant le plus de chances de voter « non » au référendum se situeront donc sur la droite du spectre.

Les variables EXTR et OPP sont mesurées par la réponse à la question suivante :

"De quel parti ou formation politique vous sentez-vous le plus proche ou le moins éloigné ?"

Les réponses possibles à cette question sont les suivantes : Extrême-Gauche, Mouvement National Républicain, Front National, Mouvement pour la France, Parti Communiste, Parti Socialiste, Les Verts. Les partis catégorisés comme extrêmes sont les suivants: Extrême-Gauche, Mouvement National Républicain, Front National, Mouvement pour la France ; et les partis d'opposition sont les suivants: Parti Communiste, Parti Socialiste, Les Verts.

Les variables EXTR et OPP représentent la proximité des personnes interrogées envers les partis extrêmes et les partis d'opposition. Si un individu est proche d'un parti dit extrême, il prendra la valeur 1 pour la variable EXTR et 0 pour la variable OPP. Si un individu est cette fois proche d'un parti d'opposition, il prendra la valeur 1 pour la variable OPP et 0 pour la variables EXTR. Si un individu est proche de la majorité au pouvoir au moment du référendum (UDF-UMP), il prendra la valeur nulle sur ces deux variables, la proximité à la majorité au pouvoir étant considérée ici comme la proximité politique par défaut.

La variable PROJET est mesurée par la réponse à la question suivante :

"Vous considérez que ce projet de constitution européenne est très bon / plutôt bon / plutôt mauvais / très mauvais ?"

Il s'agit de l'opinion sur le projet de constitution en question. La majorité des électeurs jugeait plutôt négativement ce projet, ce qui signifie que l'on doit interpréter l'impact de cette variable comme étant l'impact de l'évaluation négative des électeurs sur le projet. En bref, plus cette variable PROJET a d'influence dans l'analyse AEL, plus la désaffection des électeurs envers le projet aura eu d'influence sur l'issue de leur vote.

La variable CONSTR est mesurée par la réponse à la question suivante :

"Vous êtes favorable à la poursuite de la construction européenne : Tout à fait favorable / Plutôt favorable / Plutôt pas favorable / Pas du tout favorable?"

Il s'agit de l'opinion sur la construction européenne. La plupart des électeurs étaient favorables à la poursuite de la construction européenne, ce qui signifie que l'on doit interpréter l'impact de cette variable comme étant l'impact de l'évaluation positive des électeurs sur la construction européenne. En bref, plus cette variable CONSTR a d'influence dans l'analyse AEL, plus l'adhésion des électeurs envers la construction européenne aura eu d'influence sur l'issue du vote.

La variable CONF est mesurée par la réponse à la question suivante :

"Vous êtes confiant en ce qui concerne l'avenir de votre situation personnelle et professionnelle : Très confiant / Plutôt confiant / Plutôt inquiet / Très inquiet ?"

Il s'agit de l'opinion sur la confiance en l'avenir. La plupart des électeurs avait peu confiance en l'avenir, ce qui signifie que l'on doit interpréter l'impact de cette variable comme étant l'impact de l'opinion négative des électeurs envers leur avenir. En bref, plus cette variable CONF a d'influence dans l'analyse AEL, plus le manque des électeurs en leur avenir aura eu d'influence sur l'issue du vote.

La variable CAMP est mesurée par la réponse à la question suivante :

"Quel a été le moment de votre choix de vote : Il y a plusieurs mois / Il y a quelques semaines / Ces derniers jours / Au dernier moment ?"

Le nom que nous donnons à cette variable est CAMP, car il constitue pour nous un proxy de l'impact de la campagne dans le processus de choix. En effet, la couverture médiatique de la campagne s'est accélérée durant les semaines puis les jours précédant le jour du vote ; donc plus tard a été prise la décision de vote, plus la campagne a eu des chances d'influencer cette décision. La plupart des électeurs ont très tôt pris leur décision de vote, donc l'impact de cette variable doit être interprété comme étant l'impact de l'inefficacité de la campagne. En bref, plus cette variable CAMP a d'influence dans l'analyse AEL, plus les opinions des électeurs avant la campagne ont eu d'influence sur l'issue du vote.

F.2.iii. Hypothèses

À l'issue du scrutin, les commentaires politiques proposaient différentes interprétations :

- un vote de refus du projet en question ;
- un vote de contestation contre la politique nationale ;
- un vote de contestation contre la construction européenne en général.

Nous allons tenter de comprendre ce choix à travers les variables à notre disposition et une analyse des effets locaux adaptée.

F.2.iv. Résultats et interprétation

Dans cette première étape, nous testons plusieurs méthodes de discrimination sur la base Référendum afin de déterminer quelle est la meilleure méthode de discrimination à utiliser.

D'après le tableau F.2.1., nous choisissons les SVMs linéaires parce que cette méthode présente le plus faible taux d'erreur de classification en validation croisée.

Fonction de discrimination	Erreur en validation croisée
Analyse discriminante	15.4%
SVMs linéaires	14.2%
SVM avec noyau polynomial (degré 2)	14.9%

Tableau F.2.1. Scores de validation croisée

Définissons tout d'abord les notations formelles des variables.

$x^{(1)}$: EXTR

$x^{(2)}$: OPP

$x^{(3)}$: CAMP

$x^{(4)}$: PROJET

$x^{(5)}$: CONSTR

$x^{(6)}$: CONF

Les résultats de l'analyse FANOVA nous donnent les contributions par variable.

Contribution de extr

$$+(0.0755\phi_1(x^{(1)})) + (0.0040\phi_2(x^{(1)})) + (-0.0005\phi_3(x^{(1)})) + (-0.0005\phi_4(x^{(1)}))$$

Contribution de opp

$$+(-0.2471\phi_1(x^{(2)})) + (0.0529\phi_2(x^{(2)})) + (0.0146\phi_3(x^{(2)})) + (-0.0079\phi_4(x^{(2)}))$$

Contribution de camp

$$+(-0.1041\phi_1(x^{(3)})) + (0.0108\phi_2(x^{(3)})) + (-0.0025\phi_3(x^{(3)})) + (0.0030\phi_4(x^{(3)}))$$

Contribution de project

$$+(-0.0736\phi_1(x^{(4)})) + (-0.0017\phi_2(x^{(4)})) + (-0.0003\phi_3(x^{(4)})) + (0.0012\phi_4(x^{(4)}))$$

Contribution de constr

$$+(-0.0525\phi_1(x^{(5)})) + (0.0034\phi_2(x^{(5)})) + (0.0025\phi_3(x^{(5)})) + (0.0006\phi_4(x^{(5)}))$$

Contribution de conf

$$+(-0.0690\phi_1(x^{(6)})) + (0.0036\phi_2(x^{(6)})) + (0.0031\phi_3(x^{(6)})) + (-0.0087\phi_4(x^{(6)}))$$

[23]

Pour les mêmes raisons que précédemment, le nombre de bins le plus adapté dans ce cas pour faire un compromis entre le niveau de la description (nombre de bins) et la précision dans la mesure des effets (variance des estimateurs des effets dans chaque bin) est 8.

Nous reportons à présent dans le tableau F.2.2. ci-après la mesure de ces effets locaux relatifs au sein des huit bins.

bin	extr	opp	camp	project	constr	conf
1	-0.08 ± 0.01	-0.56 ± 0.02	-0.18 ± 0.01	-0.01 ± 0.01	-0.13 ± 0.03	0.05 ± 0.03
2	0.06 ± 0.01	-0.35 ± 0.03	-0.09 ± 0.01	0.13 ± 0.02	-0.10 ± 0.04	0.26 ± 0.05
3	0.10 ± 0.03	-0.31 ± 0.04	-0.08 ± 0.03	0.21 ± 0.02	-0.04 ± 0.05	0.25 ± 0.04
4	0.10 ± 0.03	-0.08 ± 0.07	-0.01 ± 0.03	0.23 ± 0.04	-0.14 ± 0.08	0.41 ± 0.06
5	0.12 ± 0.03	-0.17 ± 0.08	0.04 ± 0.03	0.24 ± 0.05	0.06 ± 0.06	0.36 ± 0.05
6	0.14 ± 0.02	0.18 ± 0.04	0.01 ± 0.02	0.19 ± 0.02	-0.12 ± 0.05	0.35 ± 0.05
7	0.14 ± 0.04	0.23 ± 0.03	0.07 ± 0.02	0.22 ± 0.04	-0.04 ± 0.04	0.29 ± 0.04
8	0.16 ± 0.02	0.27 ± 0.02	0.12 ± 0.01	0.17 ± 0.02	0.07 ± 0.03	0.21 ± 0.03

Tableau F.2.2. Effets locaux relatifs Référendum (SVM)

Par praticité, nous représentons ces résultats dans le spectre AEL de la figure F.2.1. ci-après.

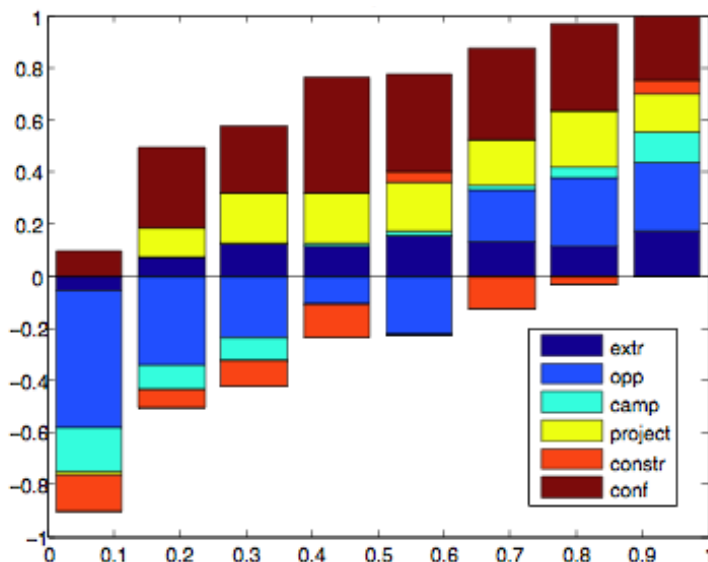


Figure F.2.1 Spectre AEL Référendum

Afin de pouvoir interpréter ces résultats de la manière la plus fiable possible, nous effectuons comme précédemment des tests de normalité et d'égalité.

Tests de normalité

Chaque case contient la valeur de la statistique de test pour la variable et le bin correspondant, puis la mention « rej. » si l'hypothèse de normalité peut être rejetée avec un risque de moins de 5% et « norm » dans le cas où l'hypothèse de normalité n'est pas rejetée.

bin	extr	opp	camp	project	constr	conf
1	0.06 norm.	1.06 norm.	1.21 norm.	1.69 norm.	2.90 norm.	1.09 norm.
2	0.26 norm.	0.63 norm.	0.92 norm.	0.61 norm.	5.41 norm.	1.27 norm.
3	1.71 norm.	0.09 norm.	2.02 norm.	0.23 norm.	0.28 norm.	1.14 norm.
4	0.89 norm.	5.51 norm.	1.12 norm.	9.36 rej.	0.82 norm.	1.36 norm.
5	0.52 norm.	3.27 norm.	57.70 rej.	0.65 norm.	0.89 norm.	2.16 norm.
6	10.96 rej.	1.33 norm.	132.79 rej.	0.50 norm.	2.00 norm.	1.00 norm.
7	1.66 norm.	0.99 norm.	1.97 norm.	1.81 norm.	0.71 norm.	0.22 norm.
8	0.28 norm.	2.85 norm.	0.02 norm.	0.68 norm.	5.16 norm.	0.06 norm.

Tableau F.2.3. Tests de normalité Référendum

Ici aussi, la plupart des distributions empiriques ne contredisent pas l'hypothèse de normalité. En effet seulement 4 hypothèses sont rejetées sur les 48 hypothèses testées. Nous pouvons donc, sereinement passer à l'analyse des résultats des t-tests.

Tests d'égalités

Ensuite nous appliquons des t-tests en prenant en compte les différences de variance entre les groupes non appariés traités. Les résultats de ces t-tests sont présentés dans les tableaux F.2.4. à F.2.9. Chaque case contient la valeur de la statistique de test pour la variable et le bin correspondant, puis la mention « diff » si l'hypothèse d'égalité entre les valeurs des moyennes peut être rejetée avec un risque de moins de 5% et « rej. » dans le cas où l'hypothèse d'égalité n'est pas rejetée et où l'on ne peut donc pas conclure à la différence entre les valeurs des moyennes.

A nouveau, la quasi-totalité des tests nous indique que les différences entre les effets sont significatives et la plupart des cas de rejet correspondent à des comparaisons entre bins voisins. Ceci confirme comme que le choix de 8 bins était ici aussi adapté.

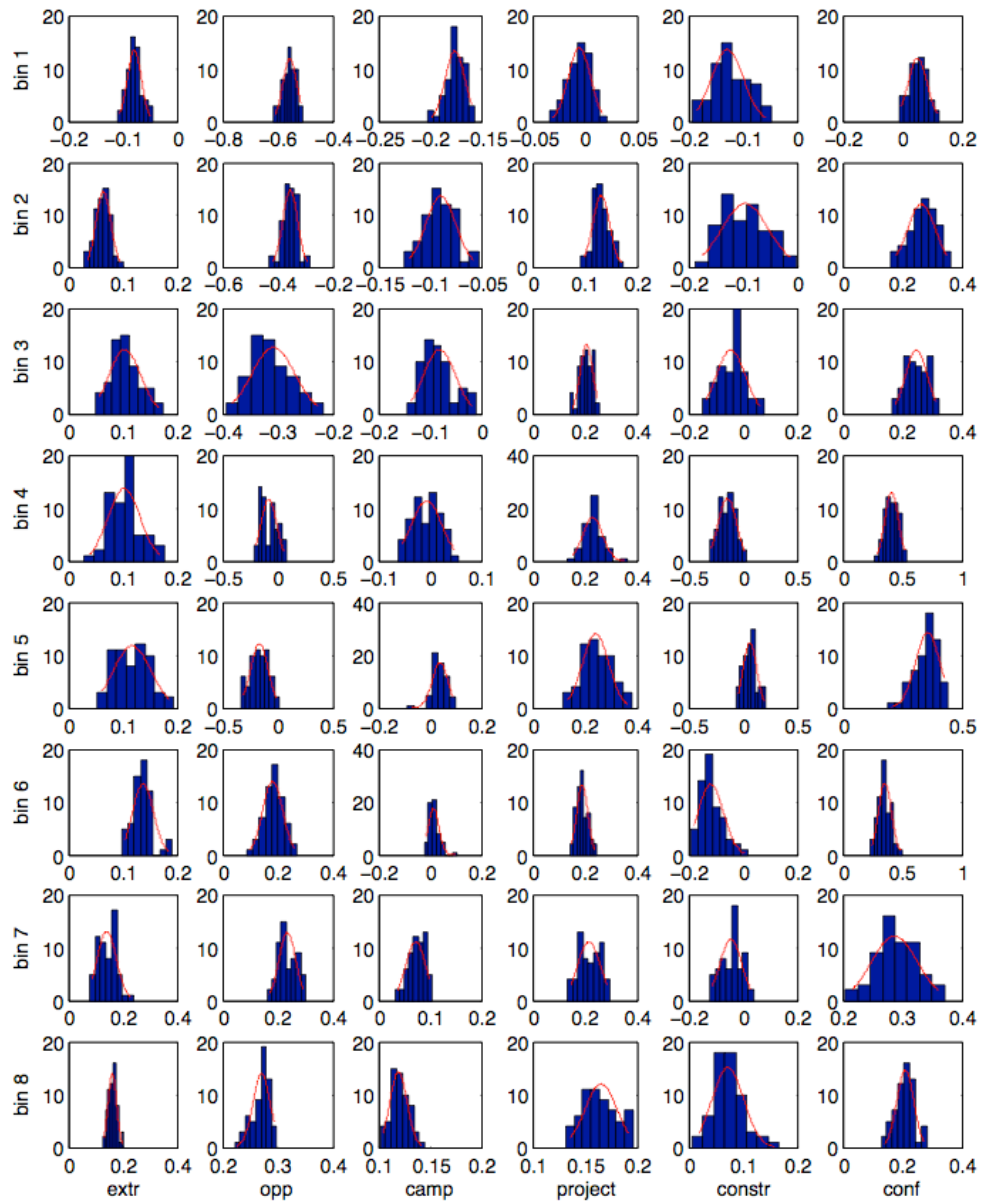


Figure F.2.2. Distributions empiriques Référendum

bin bin	1	2	3	4	5	6	7	8
1								
2	42.02 diff.							
3	31.55 diff.	6.77 diff.						
4	30.92 diff.	6.46 diff.	-0.17 rej.					
5	29.28 diff.	7.87 diff.	1.66 rej.	1.80 rej.				
6	51.30 diff.	17.11 diff.	5.29 diff.	5.44 diff.	2.77 diff.			
7	32.22 diff.	11.21 diff.	4.52 diff.	4.64 diff.	2.67 diff.	0.55 rej.		
8	65.08 diff.	25.94 diff.	9.65 diff.	9.75 diff.	6.43 diff.	5.42 diff.	2.90 diff.	

bin bin	1	2	3	4	5	6	7	8
1								
2	30.08 diff.							
3	29.37 diff.	5.42 diff.						
4	34.40 diff.	19.66 diff.	15.28 diff.					
5	25.43 diff.	12.18 diff.	8.67 diff.	-4.40 diff.				
6	92.61 diff.	65.64 diff.	50.63 diff.	17.97 diff.	21.77 diff.			
7	110.93 diff.	80.13 diff.	60.38 diff.	22.30 diff.	25.75 diff.	6.26 diff.		
8	153.69 diff.	110.14 diff.	75.37 diff.	26.30 diff.	29.39 diff.	12.50 diff.	5.77 diff.	

bin bin	1	2	3	4	5	6	7	8
1								
2	26.97 diff.							
3	16.15 diff.	1.31 rej.						
4	29.76 diff.	13.74 diff.	9.64 diff.					
5	37.04 diff.	20.92 diff.	15.34 diff.	5.89 diff.				
6	44.93 diff.	22.10 diff.	14.10 diff.	3.11 diff.	-3.70 diff.			
7	64.33 diff.	37.12 diff.	23.48 diff.	12.30 diff.	5.11 diff.	11.28 diff.		
8	127.49 diff.	67.72 diff.	34.73 diff.	22.36 diff.	13.97 diff.	25.39 diff.	12.46 diff.	

bin bin	1	2	3	4	5	6	7	8
1								
2	38.67 diff.							
3	44.07 diff.	14.23 diff.						
4	33.20 diff.	13.34 diff.	2.94 diff.					
5	26.36 diff.	11.45 diff.	3.42 diff.	0.95 rej.				
6	45.31 diff.	12.18 diff.	-2.77 diff.	-5.10 diff.	-5.13 diff.			
7	27.37 diff.	10.11 diff.	1.05 rej.	-1.39 rej.	-2.10 diff.	2.93 diff.		
8	50.30 diff.	8.47 diff.	-7.87 diff.	-8.79 diff.	-7.91 diff.	-5.20 diff.	-6.05 diff.	

bin bin	1	2	3	4	5	6	7	8
1								
2	3.48 diff.							
3	7.64 diff.	4.16 diff.						
4	-0.88 rej.	-2.76 diff.	-5.49 diff.					
5	15.47 diff.	11.70 diff.	7.34 diff.	11.01 diff.				
6	1.03 rej.	-1.95 rej.	-5.76 diff.	1.40 rej.	-12.95 diff.			
7	9.41 diff.	4.88 diff.	0.03 rej.	5.90 diff.	-8.09 diff.	6.66 diff.		
8	26.77 diff.	17.88 diff.	10.59 diff.	13.46 diff.	0.60 rej.	19.06 diff.	12.94 diff.	

bin bin	1	2	3	4	5	6	7	8
1								
2	20.74 diff.							
3	22.02 diff.	-1.59 rej.						
4	30.23 diff.	11.00 diff.	13.32 diff.					
5	27.57 diff.	7.33 diff.	9.52 diff.	-3.89 diff.				
6	26.18 diff.	6.94 diff.	8.98 diff.	-3.89 diff.	-0.13 rej.			
7	25.77 diff.	2.04 diff.	4.09 diff.	-9.98 diff.	-6.00 diff.	-5.62 diff.		
8	20.25 diff.	-5.54 diff.	-4.47 diff.	-17.44 diff.	-13.76 diff.	-12.94 diff.	-8.89 diff.	

*Tableaux F.2.4-9. Tests d'égalité Référendum
pour les variables EXTR, OPP, CAMP, PROJET, CONSTR, CONF*

Interprétation des résultats

Analysons les effets locaux, variable par variable.

La variable CONF nous indique que le manque de confiance des électeurs envers l'avenir a toujours joué quelque soit le bin en faveur du Non. Cet effet est plus important encore à la marge (impacts de 41 et 36% sur les bins 4 et 5) que pour les extrêmes (impacts de 5 et 21% pour les bins 1 et 8). On peut donc considérer que le facteur confiance a joué à la marge en défaveur du référendum et que si celui ci s'était déroulé dans un contexte socio-économique plus favorable l'issue du vote aurait pu basculer en faveur du Oui.

La variable CONSTR a un impact limité mais très souvent en faveur du Oui. Ses principaux effets peuvent être observés sur les bins 1 et 4 avec des influences de -13 et -14%. Cela signifie que l'adhésion des électeurs à la construction européenne a influencé le vote non seulement des plus convaincus (bin 1) mais aussi d'un certain nombre d'indécis (bin 4). Mais cette influence n'a pas suffi à faire basculer le vote en faveur du Oui. Ceci est particulièrement remarquable par le fait qu'elle n'a eu aucune influence sur les indécis qui ont finalement tranché contre le référendum (bin 5).

La variable PROJET nous indique que le manque d'adhésion des électeurs envers le contenu du projet de Constitution a toujours joué quelque soit le bin en faveur du Non. Cet effet n'est pas particulièrement plus marqué à la marge qu'aux extrêmes, oscillant toujours de 20%. Il est cependant intéressant de noter, que son impact a été nul pour les plus farouches partisans du Oui (bin 1). Ceci signifie que les pro-européens les plus convaincus ont à travers ce vote principalement affirmé, au-delà du contenu intrinsèque du projet, leur opinion en faveur de l'Europe.

La variable CAMP constitue par opposition à la variable CONF un exemple des critères de choix exerçant principalement son influence sur les bins extrêmes (vs les bins à la marge). En effet, ses impacts sur les 1 et 8 sont respectivement -18 et +12%, alors que ses impacts sur les bins marginaux avoisinent zéro. Ceci s'interprète par le fait que les opinions des électeurs avant la campagne ont eu plus d'influence aux extrêmes qu'à la marge. Il semble en effet logique que les électeurs aux opinions les plus tranchés soient aussi ceux qui ont pris le plus tôt leur décision vis-à-vis du choix de vote. Par contre, il est à noter que la campagne n'a

quasiment joué aucun rôle dans la prise de décision finale des électeurs indécis. La campagne n'a donc pas été déterminante dans ce scrutin et a laissé la place aux opinions a priori.

La variable OPP est la seule dont l'impact change de signe en fonction des bins. Ceci signifie que la proximité des électeurs envers les partis d'opposition a autant joué en faveur du Oui que du Non. Ceci peut s'expliquer par la diversité des consignes de vote émises par les différents partis y compris souvent au sein même de ces partis. Globalement on peut tout de même signaler que l'influence du parti joue davantage sur les bins extrêmes que sur les bins marginaux (impacts -56, -35, +23 et +27% sur les bins 1, 2, 7 et 8 ; contre -8 et -17% sur les bins 4 et 5) et que sur ces bins marginaux leur impact a plutôt joué en faveur du Oui.

Enfin, la variable EXTR a quelque soit le bin une influence positive plutôt faible, oscillant autour de 12%. On peut cependant souligner que plus on s'approche du Non, plus son impact s'amplifie (-8, 6, 10, 10, 12, 14, 14 et 16% des bins 1 à 8). Ceci signifie que les consignes de vote des partis extrêmes en défaveur de la Constitution ont porté leur fruit, et ce de manière plus cohérente que dans le cas des partis d'opposition traditionnels.

En conclusion et pour répondre à notre problématique initiale, nous pouvons dire que ce vote n'était en aucun cas un vote de sanction contre la poursuite de la construction européenne. Sur la question de savoir s'il s'agissait d'un vote sur la politique intérieure, les résultats nous indiquent certes une influence des partis politiques sur l'issue du vote mais la proximité avec un parti d'opposition semble avoir davantage influencé à la marge en faveur du Oui que du Non. D'après cette analyse, la victoire du Non tiendrait donc davantage au manque général d'adhésion des électeurs envers un projet flou desservi par une campagne peu influente et un contexte général de morosité. Ces résultats éclairent d'un nouveau jour les mécanismes ayant fait basculer à la marge l'issue du scrutin en défaveur de la Constitution européenne.

G. Discussion

En développant une approche hybride, à mi-chemin entre l'apprentissage statistique et les statistiques traditionnelles, nous avons pu proposer une méthode répondant à notre problématique sur les effets locaux. Discutons les contributions de cette recherche en modélisation et en management.

G.1. Effets de marges

Tout d'abord, la méthode AEL nous permet de mettre en évidence des effets de marge et de détecter différents types de critères de choix (marginal ou extrême, symétrique ou non). A notre connaissance, ces effets de marges ont été peu souvent étudiés en marketing, malgré leur intérêt à la fois théorique et pratique. Cette recherche constitue la première tentative pour essayer de définir, comprendre et mesurer des effets de marge. Dans la perspective de futures recherches sur ce thème, nous proposons une courte typologie des critères dans le tableau G.1.1. ci-dessous, afin de nommer les quatre types de critères de choix découverts ici.

		Proximité par rapport à la frontière de décision	
		Oui	Non
Symétrie Les effets locaux relatifs sont-ils symétriques par rapport à la frontière ?	Oui	Critère général symétrique	Critère symétrique marginal ou extrême
	Non	Critère général asymétrique	Critère asymétrique marginal ou extrême

Tableau G.1.1. Typologie des critères de choix locaux

Bien entendu cette recherche n'est qu'un premier pas vers une meilleure compréhension des effets de marge, et il serait intéressant d'avoir d'autres applications de la méthode AEL pour consolider cette approche.

G.2. Black boxes et modélisation marketing

Comme le soulignent Nalbantov, Bioch et Groenen (2006), les outils de discrimination associés à l'apprentissage statistique ont souvent été perçus par les modélisateurs en marketing comme de pure boîtes noires, performantes mais inexploitable car ininterprétables. Ce jugement est assez sévère compte tenu de la nature de nombreux travaux récents dans ce champ qui ont pour but de rendre intelligibles les boîtes noires. L'algorithme de Platt (2000) ou la décomposition FANOVA (Roosen, 1995) en sont de bons exemples. A ce propos, nous souhaiterions revenir sur le très intéressant article de Breiman (2001) qui traite de ce manque de dialogue entre différentes écoles de modélisation, principalement entre les statistiques et le machine learning, qu'il appelle « les deux cultures ».

Il y affirme que la performance en termes de prédiction peut être cruciale dans de nombreux cas réels, et ne devrait pas si souvent être reléguée au second plan derrière la significativité, comme cela est souvent le cas en statistique. En effet, meilleure est la précision de prédiction, meilleure sera l'information extraite des données. Par exemple, la régression logistique semble simple et pratique parce que nous y sommes habitués et que nous ne reconsidérons pas les hypothèses sous-jacentes à cette modélisation à chaque utilisation ; mais en réalité le principe de maximum de vraisemblance et l'ensemble des hypothèses nécessaires pour en tirer des inférences correctes sont loin d'être simples conceptuellement. Les SVMs ne sont pas forcément d'une plus grande complexité ni technique, ni conceptuelle.

D'après Breiman (2001):

« Multivariate analysis tools in statistics are frozen at discriminant analysis and logistic regression in classification even if nobody really believes that multivariate data is multivariate normal. In comparison, the one assumption made in the machine learning theory is that the data is drawn from an unknown multivariate distribution. »

Il s'agit d'un large débat qui mérite d'être ouvert en modélisation, mais qui dépasse le cadre de cette étude. Par contre, les avantages liés à l'utilisation de ces méthodes issues de l'apprentissage statistique commencent à être mis en évidence dans la recherche en modélisation marketing et nous espérons pouvoir contribuer à notre échelle à une meilleure communication entre ces « deux cultures » sachant que notre méthode s'inspire de plusieurs

courants et peut tout aussi bien intégrer les SVMs que la régression logistique. Concernant les SVMs, nous pensons qu'ils ont assez de caractéristiques statistiques intéressantes pour avoir un avenir prometteur en modélisation marketing et qu'il est essentiel de suivre la littérature en apprentissage statistique pour saisir à l'avenir des opportunités d'établir de nouvelles passerelles entre ces outils et les problématiques de modélisation marketing.

Ces questionnements renvoient aussi à la comparaison possible entre les méthodes d'apprentissage statistique et les méthodes bayésiennes traditionnelles de modélisation des choix. Le débat entre ces deux écoles de modélisation a provoqué un débat plus large sur la place de l'utilisation de lois a priori comme intermédiaires de modélisation. Pour de plus amples détails sur ce débat, nous indiquons aux lecteurs les réflexions de Vladimir Vapnik dans son essai de 2006. Il est cependant passionnant de noter de premiers rapprochements entre ces deux écoles à travers des articles combinant les deux approches méthodologiques au sein d'outils sophistiqués comme les b-SVMs de Yu et alii (2008). L'introduction d'éléments bayésiens traditionnels dans l'approche AEL constituerait un sujet d'approfondissement très pertinent pour nos futures recherches.

G.3. Utilisation pour l'optimisation

G.3.i. Quelques pistes...

Bien que les analyses aient été menées dans une optique interprétative, il serait intéressant de voir dans quelle mesure elles peuvent être utilisées à des fins d'optimisation. Reprenons le cas du jeu de données Orange Juice – SVM.

Sur la base de la fonction de discrimination obtenue dans l'étape 1, nous pouvons simuler l'impact de la variation d'un critère (par exemple une augmentation de la réduction de prix ou une baisse du prix moyen) sur les différentes sous-populations (ou bins) de consommateurs pour illustrer que certains critères sont plus efficaces pour convaincre les individus à la marge, donc plus efficaces pour gagner des parts de marché (qui ne se gagnent généralement qu'à la marge).

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Bin 8
Augmentation moyenne de la probabilité de choisir MM au sein de chaque bin (en %)	0.4	0.2	0.5	0.5	0.9	1.0	0.6	0.3
Augmentation du nombre de nouveaux consommateurs MM	0	0	0	0	6	0	0	0

*Tableau G.3.1. Simulation 1.
Augmentation de 10% du montant de la réduction de prix pour MM*

Dans le tableau G.3.1. ci-dessus nous avons simulé l'impact d'une augmentation de 10% de la valeur des rabais (10% de rabais en plus) sur la demande de MM dans chaque bin. On peut lire sur la première ligne les augmentations de probabilité de consommer MM dans chaque bin. Notons que ces augmentations sont des moyennes et qu'il peut y avoir des écarts importants entre les individus d'une même sous-population (d'un même bin). Ces augmentations sont les plus fortes pour les bins marginaux 5 et 6 (hausse respectives de la probabilité d'acheter MM de 0.9 et 1% en moyenne) alors que cette hausse est beaucoup plus faible dans les bins extrêmes (par exemple respectivement 0.4 et 0.3% pour les bins 1 et 8).

Mais il est encore plus important de constater que dans les bins marginaux, une légère augmentation de la probabilité de consommer MM est suffisante pour modifier le choix de ces individus marginaux dont la probabilité de consommer MM est proche de 50%. Nous rappelons ici que l'échantillon est composé de 1070 individus, dont 653 acheteurs de MM. Ces individus sont ceux qui se trouvaient les plus proches de la frontière des 50%, pour lesquels non seulement il y a le moins de « chemin » à parcourir pour franchir ce seuil, mais pour lesquels de surcroît l'effet de cette variation est le plus important sur la probabilité de choisir MM (pouvant aller jusqu'à 6%), ce qui explique une augmentation importante du nombre de nouveaux clients potentiels de MM. Pour le bin 6, l'augmentation de probabilité est aussi remarquable, mais comme les individus sont bien moins proches de la frontière de décision, cette augmentation ne suffit pas à potentiellement les convertir en nouveaux consommateurs.

De même, une augmentation de quelques pourcents de cette probabilité d'achat au sein des bins extrêmes n'aurait probablement pas permis d'acquérir de nouveaux consommateurs. En effet, dans les bins 1 et 2, les individus ont une trop faible propension à choisir MM pour être

convaincus par de légères variations des variables marketing disponibles et dans les bins 7 et 8, la plupart des individus sont déjà convaincus dans leur choix de MM, donc si une réduction de prix supplémentaire est proposée, ils ne feront que profiter d'un effet d'aubaine.

	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6	Bin 7	Bin 8
Augmentation moyenne de la probabilité de choisir MM au sein de chaque bin (en %)	0.7	1.3	2.2	3.1	3.5	2.6	1.8	0.7
Augmentation du nombre de nouveaux consommateurs MM	0	0	0	0	17	0	0	0

*Tableau G.3.2. Simulation 2.
Diminution de 10% du prix de base de MM.*

Une autre simulation a été menée sur un des leviers marginaux détectés dans la sous-section précédente. Il s'agit d'une baisse du prix de base de MM (MM Price) de 10%. Cette variation de prix entraîne de plus fortes variations de probabilité (allant jusqu'à 11% pour certains individus du bin 5), ce qui s'explique par le fait que le prix avait des effets locaux relatifs supérieurs à ceux des rabais dans la plupart des bins. À nouveau, on observe les plus fortes variations de probabilité dans les bins 4 et 5 (avec respectivement 3.1 et 3.5% plus de chances de choisir MM en moyenne) et les plus faibles dans les bins 1 et 8 (0.7% dans les deux cas). Ceci s'explique par le fait que le critère de prix a lui aussi plus d'effet local sur les cas marginaux. Ceci dit, nous nous attendons à davantage de conversions dans le bin 5 qui se situe à droite de la frontière de décision, parce que nous avons aussi vu que le prix de MM a des effets locaux asymétriques dans le sens où il a un effet relatif plus important à gauche (du côté du choix de CH) qu'à droite (du côté du choix de MM). Comme prévu, les seules conversions observées se font au sein du bin 5, avec 17 nouveaux consommateurs dont le choix est passé du côté de MM après cette baisse du prix de base. Rappelons à nouveau ici que la taille de la population est de 1070 individus, dont initialement 653 acheteurs de MM.

Nous souhaitons insister sur le bénéfice de bien cibler certaines campagnes sur les cas à la marge. En effet, dans un contexte de marketing direct par exemple, le fait de cibler toute la population au lieu de se focaliser sur la population marginale entraîne deux types de coûts. Tout d'abord un coût d'acquisition et de mailing inutiles pour les consommateurs convaincus de CH pour lesquels un levier marginal comme le rabais ne peut pas avoir assez d'effet (i.e. ne peut pas suffisamment augmenter leur probabilité de choisir MM) pour les convaincre de

choisir MM. Il s'agit d'un coût lié aux bins extrêmes de gauche. Mais il y a aussi un coût tout aussi important sur les bins extrêmes de droite. En effet, les consommateurs de ces bins sont déjà convaincus et auraient choisi MM même si ce rabais n'avait pas été accordé. Ils profitent donc de ce que les économistes appellent un effet d'aubaine, parce qu'ils profitent de la promotion alors qu'ils avaient déjà prévu, avant cela, de choisir MM. Ces deux coûts peuvent être grandement réduits si l'on cible la campagne sur les individus marginaux afin d'obtenir un meilleur retour sur investissement lors de ce type de campagne.

Même s'il ne s'agit pas de son premier objectif, l'AEL peut donc aussi donner des pistes d'optimisation. Ces optimisations devront par contre être par la suite réalisées avec des outils plus adaptés.

Sur le plus long terme, on peut aussi orienter ses actions suivant la nature des leviers utilisés. Ainsi, on peut s'intéresser aux critères de choix des cas extrêmes pour tenter de faire « migrer vers la droite » les individus marginaux. Dans notre cas, on peut essayer d'augmenter le niveau de Loyalty des individus à la marge (bins 5 et 6) en proposant des programmes de type « pour cinq bouteilles achetées, un cadeau vous sera offert contre vos preuves d'achat », même s'il faut rester prudent sur de telles politiques car il se peut que certains consommateurs ne maintiennent plus un tel niveau de fidélité une fois l'opération terminée. Cependant ce type d'opération pourrait entraîner chez certains une fréquence de choix de MM plus élevée et créer une habitude de consommation proche de celle observée chez les consommateurs convaincus (des bins 7 et 8).

Même s'il ne s'agit pas non plus de son objectif initial, l'AEL peut donc aussi donner des pistes de réflexion en marketing stratégique. Cependant les décisions stratégiques doivent prendre en compte un plus grand nombre d'éléments de la politique marketing et l'AEL ne peut en aucun cas suffire en tant que telle.

G.3.ii. Mais quelques limites aussi

La recherche d'interprétations du type « pour une augmentation de $x\%$ de la variable X, la probabilité d'achat augmente de $p\%$ » est une tâche périlleuse. En effet, par leur formulation même, de telles interprétations présupposent un modèle linéaire. Produire de telles interprétations exigerait donc de linéariser la fonction de décision.

Cette linéarisation peut s'effectuer globalement par le choix d'un simple modèle linéaire à l'étape de modélisation/classification comme nous l'avons fait dans la sous-section G.3.ii. précédente. Linéariser globalement constitue cependant de fait un pas en arrière et amène à renier en partie la flexibilité de la méthode qui accepte aujourd'hui n'importe quel outil de classification lors de la première étape AEL.

Cette linéarisation pourrait aussi s'effectuer localement (en considérant le gradient de la fonction de décision), auquel cas, l'interprétation prend la forme « pour un point donné (tuple de valeurs des variables utilisées), une augmentation de x% de la variable X produit une augmentation de p% de la probabilité d'achat ».

De telles interprétations, certes possibles, sont limitées grandement par deux aspects :

- Elles se font relativement à un point de départ (en lequel le gradient est évalué), ce qui en réduit l'utilité. Les résultats produits par la méthode AEL s'appliquent, eux, à l'ensemble des individus appartenant à un même bin de probabilité ;
- La linéarisation locale implique de choisir un « pas » suffisamment petit pour le calcul du gradient. Cette méthode est donc plus dans la production d'explication du type « Pour une augmentation de 0,0001% de la variable X... » que d'explication du type « Pour une augmentation de 5% de la variable X... ».

Dans les deux cas (linéarisation globale et locale), la démarche de modélisation et d'analyse se fait dans la même direction : c'est le choix du type d'explication qu'on souhaite extraire des données qui impose une forme à la fonction de décision. Or, tout l'intérêt de la méthodologie AEL est qu'elle permet une démarche inverse : rester libre du choix de la famille de fonctions de décision, et obtenir des explications a posteriori.

G.4. Implications managériales

Notre approche contribue aussi, nous l'espérons, à fournir une nouvelle grille de lecture de la demande. En effet, simplement s'intéresser à l'influence générale des leviers peut ne mener qu'à une image assez floue des effets des variables de prédiction. S'intéresser au caractère

local des effets de ces variables peut par contre apporter beaucoup d'information sur les critères de choix des cas marginaux, c'est-à-dire les meilleurs prospects. Cette information peut être utilisée dans différents contextes pour orienter les actions marketing : mise en avant de certaines caractéristiques produit, stratégie d'extension de gamme, ou communication spécifique en marketing direct. De la même manière, le caractère asymétrique de certains critères révèle comment les personnes ont pu être influencées au regard de leur choix final.

Bien entendu, nous ne négligeons absolument pas le rôle crucial des résultats obtenus par les méthodes de discrimination habituelles. Mais nous proposons, à partir de ces premières analyses, d'aller plus loin dans l'étude d'impact des variables de prédiction ; de partir de cette fonction de discrimination pour lui en faire dire plus ! Ainsi l'analyse des effets locaux peut être vue comme un complément des analyses de choix habituelles afin de s'intéresser de plus près à ce qui se passe localement. Enfin, à part quelques précautions méthodologiques, et même si l'algorithme a été basé sur des outils sophistiqués (comme la méthode de Platt ou la méthode Fanova), la méthodologie est assez facile à utiliser en pratique et donne des résultats directement interprétables pour les praticiens.

G.5. Limites et futures recherches

G.5.i. Spécification du modèle à chaque étape

La méthode AEL requiert d'utiliser le même jeu de variables, contenant à la fois des variables « historiques » ou d'attitude ; et des variables actionnables, durant l'étape d'affectation des individus aux « bins » et pour la modélisation des effets. En théorie, une approche envisageable, suggérée par les membres du jury de pré-soutenance, aurait pu consister à appliquer la méthode sur deux jeux de variables différents suivant les deux étapes LEA que sont l'allocation par bins, et l'analyse des effets. Ainsi le choix des variables aurait pu se faire en classant d'abord les individus suivant des variables décrivant leurs choix antérieurs, et d'utiliser dans un second temps des variables plus contextuelles, et plus directement actionnables.

Pour en illustrer le principe, cette approche alternative est appliquée sur les deux jeux de données étudiés dans ce chapitre.

Les variables choisies par jeu de données sont les suivantes :

- Orange Juice : Etape d'allocation : LOYALTY / Etape d'analyse : Toutes les autres
- Référendum : Etape d'allocation : CONSTR / Etape d'analyse : Toutes les autres.

En effet, la variable de fidélité *Loyalty* a été mesurée dans la base Orange Juice par les préférences passées lors des achats précédents, et la variable Construction représente l'opinion des électeurs en faveur de la construction européenne en général, c'est-à-dire de l'ensemble des choix politiques effectués par le passé dans le domaine européen, qui est celui du vote analysé.

Les résultats obtenus ainsi ne sont pas satisfaisants pour deux raisons majeures :

- A. Les effets locaux ne peuvent être mesurés sans contrôle
- B. Une démarche en deux étapes ne laisse pas s'exprimer les données

Revenons à présent sur chacun de ces deux points.

A. Les effets locaux ne peuvent être mesurés sans contrôle

Reprenons pour approcher le problème l'exemple des données Orange Juice. Dans cette nouvelle perspective, la variable *Loyalty* qui est utilisée pour la classification est exclue ensuite de la seconde étape. Ainsi la fonction de classification obtenue à l'issue de la première étape donne une probabilité de choix en fonction de la seule variable *Loyalty*. Cette fonction de classification va ensuite être décomposée en une combinaison linéaire de fonctions non linéaires univariées dépendant des autres variables du jeu de données (*Loyalty* étant exclue de cette seconde étape d'apprentissage statistique). Et c'est ici le cœur du problème.

C'est bien l'ensemble de la fonction, sur toutes les bins, qui est analysé, pour calculer la forme des effets généraux (et non locaux) des variables sur la probabilité de choix. La fonction de classification est donc analysée globalement suivant toutes les variables, sauf celle qui la constitue en premier lieu, à savoir la variable *Loyalty*. Sans cette variable essentielle à la compréhension de la fonction de décision, et donc du choix en lui-même,

l'analyse de la fonction de classification ne dispose plus de l'information nécessaire pour saisir l'utilité marginale des autres variables de discount et de prix. En effet, les variables Discount et Price n'ont un réel pouvoir prédictif qu'une fois **contrôlé** le niveau de la variable Loyalty. Sans ce contrôle, nous tombons directement dans le cadre d'un biais d'omission, amenant à surestimer et biaiser l'estimation des effets des variables restantes.

Quelques solutions pourraient être proposées pour modifier plus en profondeur la méthode initialement proposée tout en apportant des résultats techniquement valides, mais moins performants que la méthode initiale.

Etudions les ici :

- Analyser la fonction de discrimination sur chaque bin (et non dans son intégralité comme cela est actuellement le cas). En suivant cette troisième approche, le modèle disposerait effectivement de jeux de données homogènes, puisque contrôlés ou « contextualisés » au sein des bins, et les variables Price et Discount pourraient alors être passées à la loupe. Cependant cette méthode posséderait exactement les mêmes défauts que les régressions logistiques emboîtées, ou en poupées russes, présentées en début de chapitre, c'est-à-dire une diminution de la robustesse de classification et une augmentation de l'incertitude sur les paramètres estimés à l'échantillonnage ;
- Modéliser par les variables Price et Discount non pas directement la probabilité d'achat, mais les erreurs commises par une fonction de classification uniquement fondée sur la variable Loyalty. L'intuition de cette approche est qu'une première fonction de classification entièrement élaborée sur la variable Loyalty fournirait un modèle de fond, par dessus lequel on élaborerait un modèle de correction. Le modèle de correction cherche alors à expliquer le rôle des variables actionnables sur ce qui apparaît comme des erreurs de classification du premier modèle.

On notera, d'abord, que cette approche n'isole pas la variable Loyalty des autres variables – ce qui était le rôle de la deuxième approche en deux étapes. Ensuite, la méthode de quasi-régression utilisée étant une opération linéaire, *la somme des effets de deux modèles est l'effet du modèle somme*. Dans une démarche où le modèle global se décompose en une somme d'un modèle de fond et un modèle de résidu, les effets des variables Discount et Price sur le

modèle global seront les mêmes que leurs effets sur le modèle du résidu (puisque'elles ont un effet nul sur le modèle de fond). Cette piste est donc intéressante, mais équivalente à ce qui est déjà réalisé, tout en réduisant le volume d'information produit.

B. Une démarche en deux étapes ne laisse pas s'exprimer les données

La catégorisation des variables entre variables actionnables ou non introduit une dimension arbitraire dont d'après le point A, nous pouvons avantageusement nous passer. En effet, il n'est pas toujours aussi clair de faire des hypothèses sur la hiérarchie a priori des étapes. Le prix par exemple, qui dépend aussi de paramètres historiques de fond comme le positionnement ou le coût de revient, est-il réellement moins actionnable qu'une dimension de la fidélité, qui peut être modifiée par un programme promotionnel de fidélisation adapté ? Cette hypothèse sur la hiérarchie fait partie intégrante du problème que nous souhaitons résoudre en mettant en avant les différences de fond dans la nature des variables étudiées. La hiérarchie des variables est donc bien l'objectif de la méthode et non pas un de ses pré-requis. D'après le dernier paragraphe du point A ci-avant, nous pourrions imaginer aussi effectuer une analyse « forward », où à chaque étape on identifie une variable de fond, puis on étudie les effets des autres variables pour isoler une nouvelle variable de fond, mais la somme de ces modèles de résidus successifs serait identique à un modèle directement appris sur toutes les variables, et par le caractère additif de la quasi-régression, les résultats seraient identiques. Il n'y a donc aucun intérêt à suivre une telle approche forward, si ce n'est rallonger une procédure et introduire une part d'arbitraire dont nous pouvons avantageusement nous affranchir sans nuire à la qualité de la modélisation.

Par souci de rigueur, nous avons cependant tenté de mener cette approche en deux étapes en catégorisant a priori les variables Figure G.5.1 et G.5.2 ci-après.

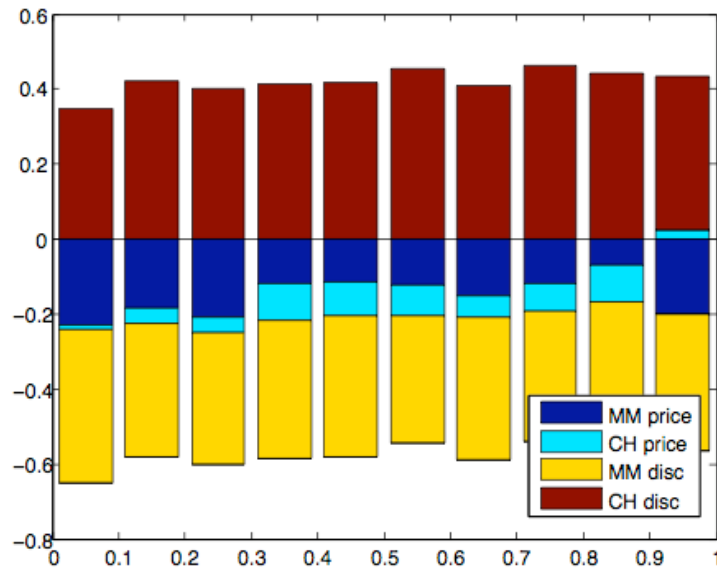


Figure G.5.1 Spectre AEL en deux étapes Orange Juice (SVM)

Dans le premier cas, on observe qu'en l'absence de contrôle par la variable Loyalty, les effets des variables Price et Discount sont dilués sur tous les bins. De plus, comme indiqué plus haut, le pouvoir explicatif de ces variables prises hors contexte est ambigu, car il est impossible de savoir quand elles captent plus ou moins du pouvoir explicatif de la variable Loyalty que nous avons ici omise.

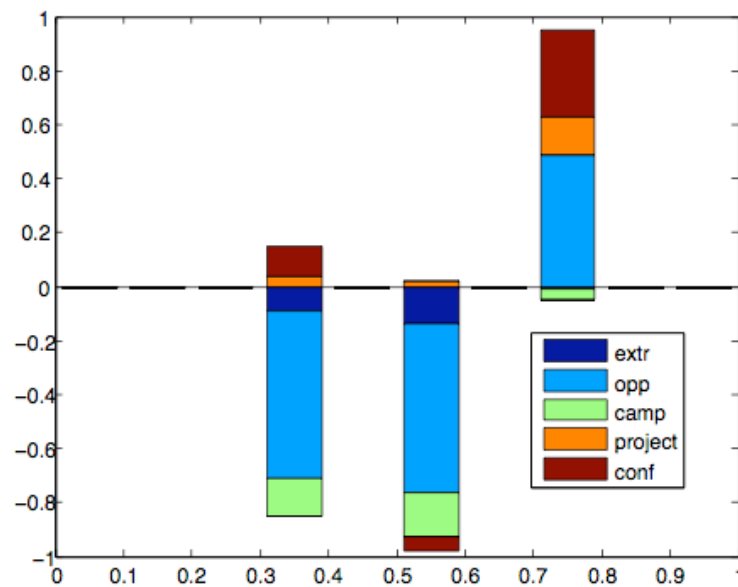


Figure G.5.2 Spectre AEL en deux étapes Référendum (SVM)

Le deuxième cas est très pathologique : la variable utilisée pour l'affectation des individus aux bins étant ternaire, la probabilité ne peut prendre que trois valeurs possibles, d'où un regroupement inégal des individus en trois groupes.

G.5.ii. Limites

Une limite intrinsèque de cette recherche est que l'approche AEL ne considère en l'état que des choix binaires, et non multinomiaux. Pourtant les choix multiples pourrait élargir le spectre d'analyse locale des critères de choix. Par exemple, en considérant le choix entre plus de deux marques (au sein d'une catégorie) ou plus de deux candidats (durant un premier tour), le choix entre l'achat d'aucun, un ou plusieurs produits d'une marque donnée. Dans leur modélisation de l'impact des promotions sur les cas marginaux pour un choix entre deux marques, Silvia-Rosso, Bucklin et Morrison (2001) considèrent cet impact non seulement en termes de choix de la marque en promotion versus l'autre marque, mais aussi en termes de stockage de produits en promotion (stock-pilling). Ceci serait possible en prenant en considération chaque couple d'options, et en ramenant ainsi l'étude d'un choix multiple à une série d'analyse locale de choix binaires.

Les deux cas d'application présentés nous ont permis d'estimer les pouvoirs explicatifs des critères à partir d'une seule occurrence de choix. Ceci constitue un atout de l'AEL qui ne nécessite pas de données de panels pour être appliquée, mais a pour conséquence que les mesures réalisées dépendent de l'occurrence en question pour un individu donné. C'est pourquoi en pratique, si des données de panels sont disponibles, il est sans doute avantageux d'appliquer plusieurs fois l'AEL pour consolider les résultats obtenus au niveau individuel suivant les différentes occurrences de choix. Une extension de cette recherche pourrait être de voir comment la mesure des effets locaux sur plusieurs choix répétés permettrait de réduire les variances des estimations des effets locaux.

G.5.iii. Futures recherches

Si nous nous sommes ici intéressés à des choix binaires et aux probabilités qui leur sont liées, c'est parce qu'il s'agit d'une problématique traditionnelle en modélisation marketing. Mais les applications des deuxième et troisième parties de notre méthodologie consistant à catégoriser les individus suivant les valeurs d'une fonction et à appliquer une FANOVA pour mesurer des effets locaux de différents critères sur cette fonction, pourraient s'appliquer à

d'autres cas de figure. Du moment que la fonction à analyser localement, prend des valeurs entre 0 et 1, tout le procédé peut être directement réutilisé. Par exemple si l'on cherche à comprendre la part de capital d'une entreprise dans une filière à l'étranger. Ou encore le taux de déchets triés par un ménage, ou bien son taux d'équipement en électroménager.

Par contre l'idée de cas marginaux ne peut être directement transposée, mais on aura tout de même une décomposition des fonctions de choix pour plusieurs sous-populations définies par rapport à la variable que l'on étudie et dont on obtiendra alors une meilleure compréhension. Ces nouvelles sous-populations constituent donc une nouvelle catégorisation qu'il serait intéressant d'analyser suivant les actions que l'on souhaite mener. Les applications de l'analyse locale sont donc nombreuses et ne se limitent pas aux choix binaires.

Il serait aussi intéressant d'appliquer cette méthodologie plusieurs fois d'affilée à l'aide de données de panels. Ainsi nous pourrions voir si l'activation de certains critères de décision à la marge permet de constater effectivement les transitions.

En réitérant encore plusieurs fois cette analyse sur une série de choix pour les mêmes individus, nous pourrions, au-delà de simples transitions, étudier d'éventuels phénomènes de migrations progressives d'individus en position de cas marginaux, vers des positions de cas extrêmes, et comprendre sous quelles conditions de telles migrations sont possibles. Ces compléments de recherche permettraient enfin de recouper nos résultats avec ceux des recherches antérieures fondées elles, la plupart du temps, sur l'observation de choix individuels répétés dans le temps, comme nous l'avons exposé dans la section B.1.

Conclusion sur la méthode AEL

Malgré l'importance potentielle de l'étude des critères de choix des switchers ou des indécis, peu d'articles avaient jusqu'à présent abordé cette question encore sous-étudiée, peut-être en partie du fait que cette population était considérée comme potentiellement moins rentable que les populations plus stables, dont on cherche à consolider la rétention. Pourtant, ces populations à la marge sont les premières cibles à viser dans une stratégie d'extension de part de marché, et aussi les premières susceptibles d'être conquises par la concurrence.

Dans un article fondateur, souvent réédité, Politz et Deming (1953) expliquent que parmi la multiplicité des causes pouvant mener à une décision de consommation, la modélisation doit se fixer comme objectif de révéler les causes les plus pertinentes en fonction de l'objectif visé:

« The function of marketing research is not to find merely causes of people behavior, nor to measure the effect of just any cause, but to discover the relevant causes, and to predict what effects on sales will result from possible actions (decisions) under certain conditions. »

Dans les conditions précises d'une stratégie d'action (offensive ou défensive) à la marge de la part de marché, nous pensons que l'analyse locale des critères de décision peut permettre de découvrir ces « relevant causes » et d'en mesurer l'impact à partir de la mesure des effets locaux. Cette motivation nous a conduit à développer cette méthode d'analyse.

L'idée sous-tendant toute cette approche est que dans beaucoup de cas, l'impact des critères de choix dépend de la prédictibilité de ce choix: on ne peut avoir le même discours pour une personne quasi-convaincue que pour une personne qui doute. Alors plutôt que de regarder l'impact global des critères de choix, nous avons proposé une méthodologie en trois temps nous permettant de mesurer les effets locaux. Par rapport aux méthodes précédentes, l'Analyse des Effets Locaux ne se focalise pas seulement sur les switchers et peut être appliquée sur un choix unique, sans nécessiter de données de panels. Enfin, à travers nos applications, nous avons illustré sa flexibilité et son efficacité, et elle nous a permis de mettre en évidence plusieurs types de critères de choix.

Du point de vue méthodologique, un certain nombre d'éléments, dont l'analyse FANOVA d'une fonction pour la mesure d'effets locaux, pourraient être utilisés dans des cadres plus larges que l'analyse de choix binaires. De même, les SVMs, présentés ici d'une manière duale par rapport aux présentations existantes en marketing, pourrait être choisis plus fréquemment, d'autant que des techniques récentes présentées dans cette étude permettent d'améliorer l'interprétabilité des résultats fournis. Nous pensons à l'avenir réutiliser ces éléments dans de nouveaux projets de recherche en modélisation.

Plus généralement, nous espérons que cette approche d'analyse locale des critères de choix, bien qu'encore naissante, ouvrira de nouvelles perspectives de recherche et aura de fructueuses implications managériales.

Transition entre la Partie 1 et la Partie 2

Dans la Partie 2 qui va suivre, nous allons mettre en évidence un autre type d'effets jusqu'alors souvent masqués en modélisation. Il s'agit des effets de cohorte, souvent masqués par les effets d'âge au cours des analyses temporelles des comportements de consommation.

A cette fin, nous avons traité le cas des comportements de fidélité à la marque et démontré que ce comportement est à la fois lié à l'âge et à la génération. En effet, si les personnes âgées sont plus fidèles aux marques automobiles lors du ré-achat d'un véhicule, ce comportement est aussi générationnel, dans le sens où les générations les plus anciennes ont été tout au long de leur vie (c'est-à-dire à tout âge) plus fidèles aux marques automobiles que les générations les plus récentes. Il co-existe donc deux raisons pour lesquelles les personnes aujourd'hui plus âgées sont plus fidèles que les jeunes : leur âge et leur génération.

Pour pouvoir mesurer simultanément les effets d'âge et de génération, tout en contrôlant les effets de période, le praticien est confronté à une multicollinéarité parfaite. Ce défi technique peut en partie expliquer pourquoi les effets de génération sont moins systématiquement pris en compte que les effets d'âge.

Cette multicollinéarité parfaite a en effet pour conséquence deux problèmes de modélisation :

- le problème de l'estimation, car les moindres carrés ne peuvent pas gérer la multicollinéarité parfaite ;
- le problème de l'intégration d'une information dite « a priori » c'est-à-dire une information exogène, car les modes d'estimation utilisés pour contourner les moindres carrés, comme la méthode PACE, ne permettent pas d'introduire d'information complémentaire sur le phénomène de consommation étudié.

Notre méthodologie, nommée APC-PLS, propose deux innovations afin d'aider le praticien à gérer ces problèmes.

La première innovation consiste à utiliser la régression PLS afin de directement pouvoir estimer les effets, sans avoir à modifier les données initiales. La seconde innovation propose un moyen d'intégrer de l'information exogène sur les effets après la phase d'estimation, par une méthode dite de « report des écarts successifs » spécialement développée et démontrée dans le cadre de cette étude.

Par son double positionnement en modélisation marketing (méthodologie APC-PLS) et en étude du comportement client (nouveaux résultats sur la composante générationnelle de la fidélité à la marque), cette étude vise à révéler le potentiel des effets de génération pour expliquer les évolutions de comportement d'achat cours du temps ; et à proposer de nouveaux moyens de mesurer ces effets de génération.

**Seconde Partie: Méthode APC-PLS d'estimation des effets d'âge-période-cohorte (APC)
appliquée à l'analyse de la fidélité à une marque automobile**

A. Introduction à l'analyse Age-Période-Cohorte en Marketing

La stratégie marketing exige de prévoir les évolutions à venir du marché sur lequel on se positionne pour sans cesse ajuster son offre et sa segmentation. Or l'âge est un des principaux critères de segmentation car il s'agit d'une des variables sociodémographiques fondamentales; et ses pendants, à savoir l'avancement dans le cycle de vie et le vieillissement, font partie des principales causes de l'évolution des comportements de consommation.

Durant la période actuelle, qui voit les enfants du baby-boom en âge d'atteindre la retraite, de nombreuses recherches en marketing portent sur les effets du vieillissement sur la consommation, et plus particulièrement sur les caractéristiques de la cible des seniors. Dans la pratique, beaucoup d'entreprises ont déjà abordé ce segment en déployant toute la panoplie des outils marketing (offre de nouveaux produits, programmes de fidélité adaptés, magazines et sites Internet spécialisés...), autant de signes faisant présager que l'offre « seniors » sera un enjeu commercial déterminant dans les années à venir.

D'autres recherches, spécialement en économie et en sociologie, abordent davantage ce thème d'actualité en termes de générations. En comparant les conditions et les styles de vie de cette génération de baby-boomers par rapports à ceux d'autres générations, ils ont pu rendre compte les acquis de cette génération par rapport aux précédentes (de la libération des femmes à l'avènement de la société de consommation de masse), et les rapports teintés de solidarité, mais aussi d'incompréhension, avec les générations suivantes. Autre angle de vue sur la même problématique, représentée par une même variable, l'âge : en effet, à une période donnée, à chaque âge correspond une et une seule cohorte.

Pour mieux éclairer le lien entre l'âge, le vieillissement, la génération et la prévision, prenons un exemple simple. Si l'on étudie les comportements d'une personne aujourd'hui âgée de 60 ans, on peut considérer que ses comportements sont conditionnés par son âge, au sens du vieillissement, c'est-à-dire les caractéristiques psychologiques et cognitives d'une personne de cet âge. Ces caractéristiques pourront avoir le même effet sur une personne aujourd'hui âgée de vingt ans quand, dans quarante ans, elle aura atteint l'âge de 60 ans. Mais ne considérer l'âge que sous l'angle du vieillissement reviendrait à omettre que si une personne

est âgée aujourd'hui de 60 ans, cela signifie aussi qu'elle est née il y a 60 ans ! Cette remarque prête d'abord à sourire, mais elle ouvre en réalité une toute nouvelle interprétation à l'âge, très différente du vieillissement : cette personne est un(e) baby-boomer et à ce titre, fait partie d'une génération ayant des spécificités propres.

Considérons les questions suivantes : quelles ont été les influences socio-économiques qu'a subies cette personne il y a quarante ans, à la fin des années soixante, quand elle avait 20 ans, qu'elle entrait à peine sur le marché de la consommation, et qu'elle connaissait ses premières expériences d'achat de biens durables ? Quelles évolutions du marché cette personne a-t-elle connues durant sa vie de consommateur ?

Quelles que soient les réponses à ces questions, elles différeront fortement des réponses à ces mêmes questions posées au sujet d'une personne ayant 20 ans aujourd'hui. Car en quarante ans, les modes de consommation et les structures des marchés de grande consommation ont énormément évolué, et ces individus auront donc connu, aux mêmes âges, des expériences de consommations très différentes. Cette dimension est la dimension générationnelle de l'impact de l'âge sur la consommation.

Pour résumer, une partie des comportements de notre individu de 60 ans sera certes expliquée par son âge au sens du vieillissement, mais son âge nous révèle aussi qu'il fait partie de la génération du baby-boom. Ceci change la donne en matière de prévision si l'on cherche à prévoir les évolutions de comportement d'une personne actuellement âgée de 20 ans quand elle aura 60 ans. En effet, une partie de ces évolutions, liées à l'âge et au vieillissement, seront similaires à celles que l'on a pu observer pour l'individu actuellement âgé de 60 ans, mais d'autres seront propres à la génération des personnes ayant aujourd'hui 20 ans (appelée génération Y), or les caractéristiques de cette génération sont déjà très différentes de celles des baby-boomers. De même, les seniors d'aujourd'hui ne seront pas tout à fait les mêmes que ceux d'hier, parce qu'ils n'ont pas vécu les mêmes expériences communes durant leur vie, et parce que notre époque diffère aussi d'hier.

Dans une perspective de prévision, on ne peut donc pas simplement interpréter l'âge comme un indicateur du niveau de vieillissement ; il faut aussi le considérer comme un indicateur de génération. A la croisée de ces points de vue, l'analyse en âge-période-cohorte (APC) permet d'appréhender le fait que l'interprétation de l'âge peut varier d'une génération à l'autre, tout

en contrôlant les effets simplement liés à la période d'observation. Mais la richesse potentielle de ce type d'analyse nécessite l'utilisation d'outils statistiques adaptés, ce qui nous conduit à considérer la complexité technique de l'analyse APC.

En effet, nous allons chercher à identifier *trois* effets : ceux de l'âge, de la période et de la cohorte, là où les données ne nous donnent que *deux* dimensions d'information. En effet, l'âge par exemple, ne *délivre* pas plus d'information que la différence entre l'année actuelle (période) et l'année de naissance (cohorte), même si elle en *contient* plus, en termes d'explications possibles des effets observés. Et l'on peut prendre le problème par les deux autres bouts, il résiste. Comme pour le ruban de Möbius, surface dont les deux faces n'en font qu'une, ou encore le vase de Klein, volume n'ayant qu'une seule surface, à la fois intérieure et extérieure, il nous manque ici... une dimension.



Figure A.1. : Représentations du nœud de Möbius (a) et du vase de Klein (b)

Nous reconnaissons sans détour qu'il existe, même dans le champ plutôt rude de la modélisation statistique, des problématiques a priori plus engageantes. Pour illustrer l'intérêt de notre étude, prenons une autre métaphore, issue cette fois de la chimie analytique. Ces effets sont comparables à des substances que l'on aurait mélangées au sein d'une solution aqueuse dont on ne connaîtrait plus la composition a posteriori. Pour déterminer la concentration des diverses substances, pour les démêler, le chimiste utilisera un catalyseur, provoquant une réaction entre ces substances pour en produire d'autres, cette fois mesurables (e.g. sous forme d'un précipité). Le catalyseur n'apparaît ni dans les réactifs, ni dans les produits, mais il permet, pour ainsi dire, de démêler la situation. Dans notre cas, ce catalyseur, permettant de choisir une option parmi toutes les combinaisons d'effets ayant pu mener à la situation observée, sera appelé *information a priori*. Il s'agit d'une information postulée par le chercheur pour démêler les effets étudiés. Cette information ne peut, par nature, être

vérifiée, mais l'on cherche à utiliser l'information la plus proche possible de la vérité. En effet, la qualité de la réponse apportée par cette opération dépendra de la qualité, en termes de véracité, de l'information apportée. Dans une terminologie analytique, le problème APC n'est pas totalement solvable, car il n'est pas identifiable. Notre but n'est donc pas de prétendre réussir à résoudre l'insolvable, mais de proposer une méthode apportant la solution la plus juste possible compte tenu de l'information a priori disponible.

Comme cela sera présenté dans la revue de littérature et la comparaison des méthodologies actuellement disponibles, la multicollinéarité parfaite entre les variables APC soulèvent deux enjeux techniques :

- Problème n°1 : Le problème de l'estimation, car les moindres carrés ne peuvent pas être appliqués si les variables explicatives sont parfaitement colinéaires ;
- Problème n°2 : Le problème d'introduction de l'information a priori, qui est une information exogène, que le praticien doit pouvoir intégrer dans sa modélisation.

Le choix de la régression PLS, qui n'a encore jamais été utilisée pour modéliser des phénomènes APC, permet de résoudre le premier problème. L'estimation est ainsi rendue possible sans avoir à modifier les données initiales, ni à fixer des conditions a priori sur les effets à estimer.

Le problème n°2 est résolu quant à lui par la technique que nous avons nommée « technique de report des écarts successifs ». Cette technique, spécifiquement développée dans le cadre de cette thèse, permet au praticien d'introduire l'information souhaitée, et ce après la phase d'estimation, tout en maintenant exactement la même qualité de prévision. Il s'agit à ce jour de la seule technique permettant une telle opération.

Ces deux innovations constituent le cœur d'APC-PLS, et sont très largement détaillées au cours de la section D et certaines des annexes de la présente thèse.

En ce qui concerne le plan de cette Partie 2, elle présente d'abord le double positionnement de notre recherche, à la fois théorique et technique (section B). Théorique d'abord, en analysant les influences potentielles de l'âge et de la cohorte sur la fidélité à la marque (section C).

Technique ensuite, en proposant une approche différente de la modélisation des effets à mesurer (section D). Cette méthodologie sera appliquée à deux exemples d'étude de fidélité à la marque : les marchés automobiles en France et en Allemagne (section E). Ainsi nous pourrons répondre aux hypothèses formulées sur les influences de l'âge et de la cohorte à travers l'exemple du marché automobile au sein des deux pays étudiés (section F).

Enfin nous discuterons des implications de notre recherche (section G) sans oublier d'en signaler les limites et pistes de futures recherches (section H).

B. Double positionnement

Le but initial de ce projet de recherche est de faire le lien entre la littérature sur l'impact du vieillissement sur les modes de consommation et la littérature sur la composante générationnelle du comportement du consommateur.

Cependant l'analyse des données sur ce thème nous a aussi conduit à travailler sur les méthodes de modélisation en elles-mêmes et à proposer d'affiner ces méthodes vis-à-vis d'une caractéristique particulière de ces données : la multicollinéarité parfaite entre les variables d'âge, de période et de cohorte. Cette section vise à clarifier et expliquer ce double positionnement.

B.1. Influences de l'âge et de la cohorte sur la fidélité à la marque

B.1.i. Des points de vue complémentaires

Au carrefour des études sur le vieillissement du consommateur et de celles sur les générations de consommateurs, cette recherche a pour but de mettre en évidence comment ces deux composantes sont indissociables. En effet, d'un point de vue technique, seule une prise en compte simultanée de ces deux phénomènes permet de mesurer le plus justement possible les impacts de ces deux composantes du comportement, et de les comparer. Et du point de vue de la théorie marketing, leur étude combinée permet de comprendre comment ces deux notions s'éclairent l'une l'autre. Or trop souvent ces deux notions sont étudiées séparément. Peut-être qu'elles sont considérées comme faisant partie de deux champs différents : le domaine de la psychologie concernant le vieillissement, et celui de la sociologie, concernant les générations. Sans doute aussi du fait de la difficulté technique de les analyser en même temps, comme nous le verrons dans la sous-section suivante B.2.

Il en ressort que l'âge est bien plus souvent utilisé pour expliquer la dynamique temporelle des évolutions de consommation que ne le sont les cohortes, et l'intérêt pour ce sujet a connu récemment un très net regain, comme en témoigne l'initiation d'un groupe de travail

international sur ce thème par le Choice Symposium (Yoon et al., 2005)¹⁵. Beaucoup de références existent sur ce thème de recherche et dans la section C, nous donnons un aperçu des résultats obtenus suivant trois différents modes d'opérationnalisation de l'âge à savoir l'âge chronologique, l'âge subjectif et le cycle de vie.

Cependant nous tenons à mettre l'accent sur les opportunités liées à l'interprétation du changement par le facteur générationnel, qui permet de prendre en compte des facteurs sociologiques importants, et qui a été trop longtemps négligé dans la recherche en marketing. Au-delà de l'arrivée massive des baby-boomers dans la catégorie des seniors, les sujets liés aux caractéristiques des nouvelles générations n'ont de cesse d'irriguer l'actualité. Entre les très rentables nouveaux seniors qui se sentent toujours plus jeunes, les jeunes adultes qui aménagent leur transition vers la maturité en usant de l'adulthood, à l'image du Tanguy d'Etienne Chatiliez (2001), ou encore les jeunes adolescentes imitant les jeunes femmes dans un syndrome Lolita généralisé (visible à travers les succès de la chanteuse Alyzée ou des poupées Bratz, et auquel Arte a consacré en 2003 une soirée Théma¹⁶), ces phénomènes de génération apparaissent souvent au premier plan des changements de comportement et de consommation, sans que l'on ait toujours en main les bonnes clefs pour les interpréter.

De plus, cette approche est déjà souvent prise en compte dans la littérature en marketing opérationnel: Dann (2007), ou encore Smith et Clurman (2007), étudient les problématiques qui se posent aux responsables marketing face à la génération des baby-boomers arrivant sur le marché des seniors, alors que d'autres s'intéressent déjà aux générations suivantes, i.e. les générations X et Y (Bailey et Ulman, 2005 ; Johnson et Hanson, 2006), jusqu'à faire de l'analyse par génération un véritable fonds de commerce (Marconi, 2000). Pour finir, Noble et Schewe (2003) proposent plusieurs outils de segmentation par cohorte, et certains, comme Kritz et Arsenault (2006) plaident même pour une entrée de l'analyse des cohortes dans les cours et les manuels de marketing.

Cette approche est aussi bien entendu très répandue en économie et en sociologie. Bodier (1999) montre que l'effet de l'âge sur le niveau et la structure de la consommation n'est plus le même quand on en isole l'effet de la génération, Recours et Hebel (2007) étudient également ainsi les comportements alimentaires afin de mieux prédire ceux des générations

¹⁵ Un compte-rendu de ces travaux a été publié dans la revue *Marketing Letters* en 2008.

¹⁶ Les petites-filles de Lolita, Soirée Théma du jeudi 13 mars 2003 sur Arte.

futures. Chauvel (1998) démontre clairement comment les générations constituent un facteur structurant au sein de la société et pourquoi de trop grandes inégalités générationnelles mettent aujourd'hui en péril notre modèle méritocratique. Ces études révèlent des effets jusqu'alors sous-étudiés et nuancent l'effet de l'âge par rapport à celui des cohortes.

Autant sur le plan théorique que managérial, le marketing a d'après nous aujourd'hui pris un retard certain en matière d'étude des effets de cohorte. Pourtant de nombreuses opportunités de recherche vont apparaître en cette période de papy-boom et de mutations des cycles de vie. Il semble donc particulièrement important de compléter les études actuelles sur l'influence de l'âge en les complétant par l'étude de l'influence de la cohorte.

B.1.ii. Choix d'un champ d'application : la fidélité à la marque sur le marché automobile

Pour aborder cette problématique, nous avons besoin d'un champ d'application précis et nous avons choisi celui de la fidélité à la marque sur le marché automobile. Ce thème correspond en effet parfaitement aux phénomènes que l'on cherche à étudier pour plusieurs raisons :

- l'étude des générations et du vieillissement se fait généralement dans une optique de long terme, dans laquelle les changements liés à ces variables sont effectivement observables. Or la fidélité à la marque pour des biens de consommation durable est un indicateur marketing crucial dans une optique de long terme ;
- le choix d'un bien de consommation durable est un choix impliquant, dans le processus duquel rentrent en compte plusieurs aspects, y compris les aspects économiques et sociaux, ce qui permet d'envisager une influence de la génération dans l'acte d'achat ;
- il s'agit d'un marché pour lequel, dans la très grande majorité des cas, le consommateur est lui-même acheteur et prescripteur, donc pour lequel les variables individuelles, dont l'âge (au sens du vieillissement), peuvent avoir des effets détectables dans le processus de décision ;
- le marché de l'automobile est lui aussi particulièrement adapté pour cette recherche car il a fortement évolué au cours du siècle, notamment en termes de marques disponibles, et que les consommateurs ont donc été exposés à des offres très différentes au cours du temps.

Malgré cette proximité entre les thèmes de la fidélité et des effets combinés de l'âge et de la cohorte, aucune étude ne tente, à notre connaissance, d'établir de liens entre la fidélité à la marque et ces deux effets combinés d'âge et de cohorte. Certains travaux, dont Lambert-Pandraud, Laurent et Lapersonne (2005), exposent comment les personnes plus âgées ont un ensemble de considération plus restreint et privilégient les marques établies de longue date, et l'on pourrait en déduire une plus grande tendance à la fidélité avec l'âge.

Cependant, les effets de cohortes semblent pouvoir partiellement expliquer la fidélité à la marque. En effet, certaines cohortes sont potentiellement plus attachées à un groupe restreint de marques, et différentes cohortes ont pu être exposées à des marchés très différents d'une période à l'autre. Ainsi dans le secteur automobile, les consommateurs âgés n'ont été exposés qu'à des marques nationales durant leur jeunesse et lors de leurs premiers achats, alors que les jeunes aujourd'hui sont exposés à un ensemble de marques beaucoup plus large, en provenance de différents pays, et ce depuis qu'ils sont en âge d'acheter un véhicule. Aux mêmes âges, ces deux groupes de consommateurs ont donc été exposés à des structures de marché très différentes, ce qui peut influencer leur comportement d'achat sur le long terme, et donner lieu à de fortes différences dans leurs comportements d'achat à l'âge senior.

Le choix de la fidélité à la marque comme champ d'application de l'étude de la combinaison des effets d'âge et de cohorte a ainsi le double avantage : celui d'être adapté aux caractéristiques du phénomène étudié, et celui de fournir un thème de recherche encore sous étudié malgré son fort potentiel.

B.2. Modélisation des effets d'âge et de cohorte

Pour analyser l'évolution des comportements au cours du temps au sein d'une population, nous avons décidé de prendre en compte deux effets principaux: l'impact du vieillissement, mesuré par l'âge A , et l'impact de la cohorte à laquelle appartient la personne observée, mesurée par son année de naissance C . Cependant dès que nous observons les comportements d'une population sur plusieurs périodes successives afin d'en comprendre l'évolution, nous avons besoin de contrôler l'effet potentiel de la période d'observation P , mesurée par l'année à laquelle est observé ce comportement. C'est trois variables d'âge, de période d'observation

et de cohorte, presque toujours étudiées simultanément dans les études prenant en compte l'âge et la cohorte, sont souvent regroupées sous le sigle APC¹⁷.

B.1.i. Multicolinéarité parfaite entre les variables étudiées

Chacun de ces trois *effets* possède des interprétations spécifiques qui s'avèrent pertinentes pour la segmentation d'un marché et la prévision de ses tendances. Cependant les *valeurs* prises par ces trois variables ont un désavantage majeur, car ces trois variables sont parfaitement colinéaires. En effet, l'âge A est exactement égal à la période d'observation P moins l'année de naissance C . Relation que l'on peut résumer par la formule [1] ci-dessous :

$$A = P - C \quad [1]$$

Comme nous l'avons indiqué dans l'introduction de cette partie, ce problème ne peut être résolu mathématiquement qu'en introduisant de l'information dite a priori, c'est-à-dire en faisant une (ou des) hypothèse(s) a priori sur la valeur des effets pour certaines modalités des variables APC.

Cette information est obtenue de manière exogène et n'est pas testable, mais il s'agit du seul moyen d'obtenir une estimation des effets APC, quelque soit le mode d'estimation choisi et la manière d'introduire cette information.

Sa qualité et donc celle des estimations qui en découlent, reposent donc simplement sur la qualité de la connaissance ou de l'expertise que l'on possède a priori au sujet du phénomène étudié, d'où le nom d'information a priori. Meilleure est la qualité de cette information a priori, meilleure sera la qualité d'estimation des effets APC étudiés. Pour modéliser ces effets, il est donc crucial d'avoir une bonne idée a priori de l'information à utiliser.

¹⁷ Ce sigle est identique en anglais et en français (Age – Period – Cohort en anglais).

B.2.ii. Méthode de modélisation des effets

Une partie de ce travail de recherche a consisté à récapituler les méthodes disponibles pour introduire cette information a priori et dans la section D nous les présentons sous suivant deux catégories :

- les méthodes de modélisation APC fondées sur les moindres carrés,
- les méthodes de modélisation APC non fondées sur les moindres carrés.

Dans le premier cas, l'information a priori est introduite avant estimation, au stade de la spécification du modèle, afin de permettre l'estimation des effets par la méthode des moindres carrés. En effet, la multicolinéarité parfaite ne permet pas de lancer l'estimation par les moindres carrés et il est donc nécessaire de fixer des contraintes d'égalité sur certains des effets à estimer. Par exemple, si on pense a priori que le fait d'avoir 20 ans n'a pas une grande différence d'influence sur la variable à expliquer par rapport à l'influence liée au fait d'avoir 21 ans, alors on dira que ces deux modalités de la variable âge ont le même effet. Ceci se traduit par le fait qu'au sein de la régression de la variable à expliquer sur l'ensemble des modalités des trois variables APC, les coefficients des modalités A_{20} et A_{21} de la variable A seront fixés comme égaux¹⁸. Le choix de ces contraintes constitue l'information a priori introduite dans le modèle estimé. On passe alors d'un cas de multicolinéarité parfaite, à un cas de multicolinéarité forte entre les modalités restantes. L'estimation est alors possible, même si la multicolinéarité forte entre les prédicteurs (les modalités restantes des variables APC) endommage la qualité d'estimation par les moindres carrés.

Dans le second cas, les méthodes n'utilisent pas les moindres carrés pour estimer le modèle et n'ont donc pas à introduire l'information a priori dès le stade de la spécification. Par contre, il est toujours nécessaire d'introduire l'information a priori pour résoudre le problème évoqué. Dans ce cas l'information est introduite, mais implicitement, et sans contrôle de l'utilisateur, au moment de l'estimation. Ces méthodes permettent donc d'estimer un modèle contenant toutes les modalités de toutes les variables APC, mais ne laissent pas à l'utilisateur la possibilité d'introduire l'information a priori qui lui semble la plus vraisemblable.

¹⁸ Les notations utilisées dans cette recherche seront toutes explicités au début de la section D traitant des options de modélisation des phénomènes APC.

En résumé :

- les méthodes fondées sur les moindres carrés permettent d'introduire l'information a priori désirée, mais la qualité de l'estimation est réduite par la forte multicollinéarité du modèle ainsi spécifié ;
- les méthodes non fondées sur les moindres carrés permettent d'estimer le modèle malgré la multicollinéarité, forte ou parfaite, entre les variables du modèle, mais ne permettent pas de choisir l'information a priori introduite.

Enfin, il est important de noter qu'aucune de ces méthodes ne permettent de vérifier la sensibilité de l'estimation des effets par rapport à l'information a priori introduite. Or même si une bonne expertise du phénomène étudié permet d'avoir en sa possession une information a priori fiable, cette information est la plupart du temps imparfaite, et rien ne permet de tester sa validité. Il serait donc souhaitable de pouvoir observer les variations des résultats obtenus en fonction de légères perturbations de l'information initialement introduite, ce qui n'est possible avec aucune des méthodes disponibles à l'heure actuelle.

La méthode APC-PLS que nous proposons dans cette recherche contribue à la recherche dans le domaine en combinant les avantages des deux catégories définies ci-dessus en permettant :

- de maintenir une bonne qualité d'estimation (via PLS) ;
- de garder le contrôle de l'information a priori introduite.

De plus, cette méthode comprend une étape finale de contrôle de la sensibilité des résultats obtenus en fonction de la variation de l'information a priori autour de l'information a priori initialement introduite.

Cette seconde contribution, plus méthodologique, nous a permis in fine de tester les hypothèses de notre contribution théorique par divers moyens et d'ainsi consolider les résultats obtenus à partir de différentes méthodes de modélisation.

Par la suite, seront d'abord établies nos hypothèses sur les influences de l'âge et de la période dans la section C. Puis la section D présentera les aspects méthodologiques de la recherche, permettant de valider les hypothèses sur les effets APC dans les sections E pour les résultats et F pour leurs interprétations.

C. Influences de l'âge et de la cohorte sur la fidélité à la marque

Les caractéristiques d'âge et de cohorte ont potentiellement de fortes influences sur les comportements de consommation, et parmi eux, sur notre fidélité à la marque. Dans les sous-sections suivantes, seront abordées séparément ces influences pour l'âge et pour la cohorte, afin de mettre en évidence les différentes causes respectives à l'origine de ces influences pour chacune des deux variables. Dans chacune de ces sous-sections, nous allons d'abord présenter les différentes définitions de ces variables, avant d'établir les liens entre ces variables et la fidélité à la marque, pour aboutir aux hypothèses concernant l'influence de ces variables sur la fidélité à la marque. Enfin, une dernière sous-section récapitulera ces informations et ces hypothèses - qui seront analysées dans la section F.

C.1. Influence de l'âge sur la fidélité à la marque

Dans la pratique marketing, l'âge constitue, avec le revenu, l'un des principaux critères de segmentation. Il est donc naturel de constater que la littérature académique en marketing a beaucoup étudié l'impact de l'âge des individus sur leurs comportements de consommation. Pour une synthèse sur ce thème, l'on peut se référer aux travaux de Yoon et al. (2005), Cole et al. (2008) et de Lambert-Pandraud (2000).

C.1.i. Définitions de l'âge

Avant d'analyser plus précisément l'influence potentielle de l'âge sur la fidélité à la marque, nous passons en revue les trois principales opérationnalisations de la variable âge dans la recherche en marketing, c'est-à-dire l'âge chronologique, l'âge subjectif et le cycle de vie ; à la suite de quoi nous donnons la définition de l'âge retenue dans le cadre de cette recherche.

Age chronologique

La première définition de l'âge retenue en recherche marketing est évidemment l'âge réel de l'individu, également appelé l'âge chronologique (Guiot, 2006). De nombreuses recherches

montrent que les capacités cognitives, et par conséquent le processus de décision, varient très fortement avec l'âge, particulièrement pour les âges avancés.

Parallèlement, plusieurs recherches portent sur les variations des attitudes en fonction de l'âge. Citons par exemple les études sur les attitudes vis-à-vis :

- de la publicité (Dawson et Spangenberg, 1987; Perrien et al, 1997 ; Drolet et al, 2007),
- des consommations culturelles (Holbrook et Schindler, 1994),
- des préférences pour les marques (Bass et Talarzyk, 1972),
- du matérialisme (Nguyen et Roedder John, 2007),
- et du marketing en lui-même (Barksdale et Darden, 1972).

Notons dès à présent qu'un handicap commun à la plupart des études traitant de l'âge est qu'elles travaillent sur des données en coupe transversale, c'est-à-dire sur une seule période. Dans ce cas, on n'observe chaque cohorte qu'à un seul âge des âges de cette cohorte, et il est donc impossible de s'intéresser aux effets de cohorte. Ainsi, les catégories utilisées sont figées dans le temps (jeunes vs seniors) au lieu d'être étudiées dans une dynamique temporelle (les jeunes dans les années 2000), ce qui risque de masquer certains phénomènes. Les adolescents des années 2000 diffèrent de ceux des années 1990 par de nombreux aspects. Pour ne citer que quelques éléments différenciants en termes de consommation, leur exposition à l'internet et à la téléphonie mobile a été bien plus précoce, ils ont principalement connu l'euro, et considèrent qu'Oasis ou Nirvana sont des groupes passés de mode. Il existe donc toujours un risque latent d'attribuer à l'âge des effets de cohorte, comme nous l'évoquions précédemment. C'est pourquoi par la suite nous étudierons l'âge et la cohorte comme des facteurs concomitants.

Age subjectif

L'âge subjectif peut être défini comme l'âge ressenti. Guiot (2001) le définit comme l'« âge de référence auquel un individu s'identifie en fonction des rôles sociaux qu'il lui attribue. » C'est l'âge que les personnes citent quand on leur demande l'âge des personnes ayant les mêmes activités ou les mêmes centres d'intérêt qu'elles. Il est donc davantage fondé sur un mode de vie (et de consommation) que sur un âge chronologique réel.

De nombreuses recherches utilisent cette notion pour étudier l'effet de l'âge (Kastenbaum et al. 1972 ; Stephens 1991 ; Wilkes 1992). En publicité, Guiot (2001) compare les liens entre les âges chronologiques et subjectifs avec l'implication envers la mode, et observe que l'âge subjectif est le plus corrélé des deux. De même, en marketing du tourisme, Marthur, Sherman et Shiffman (1998) trouvent que le comportement des personnes ayant un âge subjectif de dix ans inférieur à leur âge réel ont tendance à se comporter comme les personnes ayant cet âge. Bien entendu, pour que cela soit possible, il faut supposer que seulement une minorité de personnes ont un âge subjectif différent de leur âge réel par rapport aux comportements majoritairement observés chez des personnes de leur âge, sinon par itération, le principe ne fonctionne plus.

Du point de vue managérial, l'usage de l'âge subjectif permet d'étudier plus efficacement les réactions des personnes âgées aux stratégies marketing. Par exemple, Moschis et Mathur (2006) pèsent le pour et le contre des campagnes de promotion ciblées sur cette tranche d'âge au vu du risque d'une impression de stigmatisation, spécialement ressentie par les individus ayant des âges subjectifs peu avancés.

Cycle de vie

Enfin, l'âge de l'individu est souvent ramené à une étape spécifique du cycle de vie comme l'accès à la majorité, le départ du foyer familial, l'entrée dans la vie professionnelle, le mariage ou le concubinage, l'arrivée ou le départ d'enfants au sein du foyer, les déménagements ou encore la retraite. Altman (2003) étudie par exemple les comportements optimaux d'investissement et d'épargne au cours des cycles de vie, thème récurrent en marketing depuis les années 50. Une étape du cycle de vie très utilisée dans l'étude des effets de l'âge est la retraite, qui a un fort impact sur les comportements de consommation (Neugarten, 1979 ; Lepisto, 1997). Burnett (1989) prétend même que, dans certains cas, l'étape de la retraite peut servir de meilleure variable de prévision et de segmentation que l'âge chronologique lui-même, notamment en ce qui concerne les comportements de visite et de fidélité aux magasins.

Modélisation de l'âge

Il y a trois raisons principales pour lesquelles nous avons choisi de modéliser l'âge par l'âge chronologique : l'une pratique, deux théoriques.

La raison pratique provient du fait que l'âge chronologique est le seul de ces trois aspects de l'âge qui soit disponible dans la plupart des bases de données marketing et des études de consommation. En effet, l'âge perçu et le cycle de vie exigent d'avoir recueilli des informations complémentaires, soit à travers des questions très spécifiques (âge perçu), soit par recoupement de diverses autres questions (cycle de vie). Ainsi, l'application de notre méthodologie se verrait cantonnée à un cadre très restreint. Cependant cette raison pratique ne peut à elle seule justifier d'un tel choix.

La première raison théorique provient du fait qu'il s'agit de l'information la moins biaisée des trois. L'âge perçu peut en effet être largement influencé par la sensibilité du répondant à son niveau de désirabilité sociale. On mesure moins ainsi la perception de son âge par un individu, que sa tendance à céder, durant le temps de la passation du questionnaire, au conformisme d'une société prônant les valeurs de la jeunesse. De même le cycle de vie laisse une grande part à l'interprétation de l'interviewé: un professeur émérite se considèrera-t-il comme retraité? Face à ces confusions potentielles, l'âge chronologique apporte une mesure objective et homogène du phénomène de vieillissement étudié.

La seconde raison concerne la qualité d'intégration de l'âge chronologique. En effet, cette implémentation peut être vue comme un proxy (une approximation) des deux autres. Ce qui n'est le cas ni du cycle de vie, ni de l'âge perçu. Elle regroupe donc la plupart des impacts que l'ont peut envisager concernant le vieillissement. Certes elle n'identifie pas toutes les causes par elles-mêmes, mais permet d'en mesurer les effets, évoqués par la suite sous le nom d'effets d'âge.

Pour ces raisons, par la suite, quand nous parlerons d'âge, il s'agira de l'âge chronologique.

C.1.ii. Influences de l'âge sur la fidélité à la marque

Les effets d'âge couvrent une large gamme d'impacts menant à une recherche d'information limitée. Parmi eux on peut compter l'âge biologique et le déclin physique, une limitation de l'espace géographique de recherche, mais aussi un déclin cognitif limitant le nombre d'options traitées, une plus grande prudence dans la prise de décision et une plus grande

sélectivité socio-émotionnelle (Lambert-Pandraud, Laurent et Lapersonne, 2005). Deux facteurs affectifs liés à l'âge pourraient aussi jouer en faveur d'une plus grande fidélité à la marque : l'attachement (Kleine et Baker, 2004) et la nostalgie (Holbrook, 1989).

Cependant, nous soutiendrons ici que, même si le facteur de nostalgie est davantage visible avec l'âge, il s'agit en réalité d'un facteur lié à la cohorte. En effet, la nostalgie porte sur des objets, ici les marques, qui dépendent entièrement d'une expérience vécue à un âge et une période donnés. Typiquement, les marques que l'on a connues durant sa jeunesse ou à un moment clef de son cycle de vie. Nous les retraiterons donc plus en détail dans la section C.2. sur l'impact de la cohorte sur la consommation.

Les quatre principaux liens entre l'âge et la fidélité à la marque restants sont donc :

- le vieillissement biologique,
- le déclin cognitif,
- l'aversion au changement,
- la sélectivité socio-émotionnelle,
- l'attachement.

Nous allons ici reprendre chacun d'entre eux pour en détailler le contenu théorique.

Le vieillissement biologique

Le vieillissement, phénomène continu décrit comme un « déclin des capacités physiques » (Barrère, 1992), a été avancé pour expliquer la réduction du nombre de distributeurs visités (ici les concessionnaires) par les personnes plus âgées. L'argument consiste simplement à dire qu'il est plus difficile pour ces personnes de se déplacer. Cependant, parmi les personnes âgées de 60 à 74 ans¹⁹, plus de 80% sont tout à fait mobiles, alors qu'il en va de même pour seulement 34% des personnes âgées de 80 ans et plus (David et Sarzec, 1996). La contrainte physique pourrait ainsi réduire le nombre de distributeurs visités pour les personnes très âgées (75 ans et plus), mais cette explication semble moins pertinente pour les jeunes seniors (entre 60 et 74 ans). Dans le cadre de la fidélité à la marque, la réduction du nombre de distributeurs potentiels joue en défaveur du nombre de marques incluses dans l'ensemble de considération,

¹⁹ Les tranches d'âge utilisées ici sont celles arbitrairement choisies dans la recherche citée ici.

et donc en faveur de la fidélité aux marques prises en compte. Spécialement sur le marché automobile européen, dans lequel la plupart des concessionnaires sont encore mono-marque.

Le déclin cognitif

Lié à la mémoire de travail (Park et Gutchess, 2004) et l'intelligence fluide (Chasseigne, Mullet et Stewart, 1997 ; Chasseigne et Mullet, 2004), le déclin cognitif a été évoqué comme un antécédent potentiel de la réduction de la recherche d'information par les consommateurs plus âgés, car la mémoire de travail et l'intelligence fluide sont jugées nécessaires pour prendre en considération et comparer de nouveaux modèles dans le cadre de la consommation de biens durables aussi complexes que l'automobile.

Plusieurs tests psychométriques convergent pour mettre en évidence ce déclin cognitif (Salthouse, 1991). Ces conclusions sont renforcées et étayées dans des études récentes mettant en œuvre des techniques d'imagerie neurologique (Heden et Gabrieli, 2004) pour montrer que le déclin lié à l'âge prend place dans le cortex préfrontal dorsolatéral, qui est en charge de fonctions comme la mémoire de travail (MacPherson, Philips et Della Sala, 2002). La mémoire de travail opère l'encodage de l'information en direction de la mémoire de long terme, mais aussi le traitement conscient de l'information concernant les événements récents (Park et Gutchess, 2004). Les données neurologiques mettent en évidence des difficultés particulières des personnes âgées dans les tâches de maintenance et de manipulation de l'information dans la mémoire de travail (Yoon et Cole 2006, p13).

À partir de l'âge de 60 ans, les personnes peuvent aussi subir une réduction de leur mémoire explicite, forme de mémoire qui rend possible le traitement conscient de l'information et de ses sources (e.g., se souvenir à quel constructeur correspond une publicité et quand elle a été vue). Par exemple, le rappel d'une série de mots ou d'un texte décline significativement avec l'âge (Zelinsky et Burnight, 1997). La mémoire de travail permet aussi aux gens de manipuler plusieurs éléments d'information simultanément et de les comparer (Mather, 2003). Du fait de ce déclin de la mémoire de travail, les personnes plus âgées peuvent éviter certains efforts cognitifs, tels que la comparaison de trop d'alternatives de choix, et plutôt privilégier des heuristiques simples. Par exemple, Cole et Blasubramanian (1993) mesurent la préférence pour la première option satisfaisante lorsqu'un problème se complexifie.

Ces aspects du déclin cognitif pourraient avoir un impact sur les variables que nous analysons. Dans le cas du choix complexe d'un bien durable, les consommateurs pourraient décider de se restreindre aux choix entre des alternatives faciles d'accès, faisant appel à des connaissances déjà intégrées depuis longtemps.

De surcroît, l'âge est négativement corrélé avec l'intelligence fluide (intelligence requise pour traiter de nouveaux problèmes ou de nouvelles situations), mais l'intelligence cristallisée (intelligence basée sur l'expérience) reste intacte (Salthouse, 1991). Le déclin de l'intelligence fluide est continu, mais il franchit des seuils de performance. Ces seuils d'âge ne sont pas universels. Ils dépendent des individus et de leur vécu. Chasseigne, Mullet et al. (1997, 2004) font l'hypothèse d'une « diminution graduelle de la flexibilité cognitive chez les adultes âgés », particulièrement dans des cas de relations inversées entre des indicateurs et leurs conséquences. Sur le marché de l'automobile, si les prix sont gonflés pour permettre de plus gros rabais, les consommateurs âgés seraient moins capables d'utiliser cette relation inverse, même avec une aide visuelle. Par contre, leur expérience (intelligence cristallisée qui reste elle intacte) peut leur permettre de conserver des réflexes et des raisonnements acquis de longue date.

Le déclin de la mémoire de travail et de l'intelligence fluide devraient entraîner une réduction du nombre de marques, de modèles et de concessionnaires pris en considération. Racheter un véhicule d'une même marque peut alors devenir un choix heuristique (Jacoby et al. , 1978).

L'aversion au changement

L'aversion aux risques liés au changement, même si la situation présente est loin d'être idéale, est un phénomène bien documenté en gérontologie. Wallach et Kogan (1961) et Botwinick (1966) demandent à des sujets de choisir entre deux options : rester à un poste sûr bien que médiocre, avec des perspectives d'augmentation et d'évolution très limitées, ou choisir une position qui pourrait mener avec une probabilité p à une augmentation de salaire et avec une probabilité $1-p$ à de mauvaises conditions financières. Les sujets âgés ont une probabilité plus marquée de choisir de ne pas changer, quelle que soit la valeur de p (Botwinick ,1978). Botwinick (1978) suggère deux hypothèses pour expliquer cette résistance au changement. Premièrement, étant donné leur déclin intellectuel, les personnes âgées pourraient vouloir éviter de prendre des décisions. Les résultats précédents corroborent cette hypothèse. Ensuite,

les personnes plus âgées pourraient avoir tendance à vouloir minimiser le risque lié aux décisions qu'ils prennent, spécialement quand ce risque est financier. Cependant Botwinick (1978) remarque aussi que lorsque l'option de statu quo n'est pas disponible ou que l'option complexe est plus rémunératrice (Okun et Elias, 1977), les sujets âgés ont une fonction d'utilité semblable à celle des plus jeunes. Des études de laboratoire récentes dans lesquelles les sujets âgés ont à choisir des cartes suivant leur ratio niveau de rémunération / niveau de risque établissent qu'ils ont un comportement proche de celui des sujets plus jeunes (MacPherson, Phillips, et Dalla Sala, 2002).

Ainsi le comportement d'achat des personnes plus âgées peut être influencé par cette aversion au changement (plus qu'au risque), qui les mènerait à répéter leurs achats précédents du moment qu'ils en ont été satisfaits. Rester engagé avec la même marque ou le même concessionnaire est une manière d'éviter la complexité d'une nouvelle décision, tout comme racheter une nouvelle version d'un modèle que l'on connaît.

Sélectivité socio-émotionnelle

La théorie de la sélectivité socio-émotionnelle avance que les personnes âgées qui perçoivent leur horizon temporel comme plus limité attachent plus d'importance à leurs sentiments et à leurs émotions et ont moins tendance à s'intéresser à la nouveauté. Ils donnent la priorité à des contacts émotionnels proches et familiers, plutôt qu'à des contacts plus récents, plus informatifs (Carstensen, Isaacowitz, et Charles, 1999 ; Isaacowitz, Charles et Carstensen, 2000). Quand on leur demande de choisir pour une activité sociale entre plusieurs compagnies potentielles suivant trois degrés de familiarité à savoir un membre de leur famille immédiate, une connaissance récente ou l'auteur d'un livre qu'ils viennent de lire, 65% des sujets âgés choisissent la personne la plus familière, contre seulement 30% chez les jeunes sujets. Des résultats similaires sont obtenus auprès de jeunes sujets dont l'horizon temporel a été artificiellement raccourci à cause, par exemple, d'un déménagement imprévu à effectuer dans la semaine ou d'une maladie soudaine diagnostiquée comme fatale (Fredrickson et Cartensen, 1990). On peut analyser d'une manière similaire la relation des personnes âgées avec leur concessionnaire traditionnel, avec lequel elles ont établi un rapport de confiance sur des années, particulièrement dans le cas d'un achat aussi important que celui d'un véhicule.

La sélectivité socio-émotionnelle peut aussi biaiser la mémoire des personnes au sujet de leurs choix précédents. Dans une expérimentation, Mather et Johnson (2000) trouvent que les sujets âgés ont plus tendance que les jeunes sujets à préférer mettre en avant les aspects positifs de l'option qu'ils ont choisie plutôt que les aspects négatifs de l'option qu'ils ont rejetée. De même, dans Lambert-Pandraud et al. (2005), les consommateurs plus âgés ont plus tendance à être satisfaits de leurs précédents choix de voitures. Sur une échelle de 1 à 10, les scores élevés (8 et plus) étaient donnés par 56% des jeunes, 67% des personnes d'âge moyen, 78% des jeunes seniors et 86% des seniors âgés. De plus, pour un niveau de satisfaction donné, les personnes âgées ont davantage tendance à considérer la marque précédente, et elle seulement, dans leur achat suivant. Cette tendance à défendre leurs précédents choix reflète leur fidélité plus élevée à la marque et aux concessionnaires.

L'attachement

Un consommateur peut développer un attachement à une voiture, identique à l'attachement que l'on peut porter à toute autre possession matérielle. Cette relation conduirait à racheter cette voiture plutôt qu'à en essayer d'autres. Bien que l'attachement ait initialement été introduit à propos du lien entre les parents et enfants (Bowlby, 1979), il a été appliqué aux consommateurs de tout âge. Dans leur revue de littérature, Kleine et Baker (2004, p.4) définissent l'attachement à une possession matérielle comme « une propriété multi-facettes de la relation entre un individu... et un objet matériel spécifique que l'individu s'est psychologiquement approprié, débanalisé et singularisé au sein d'une interaction personne-objet. » La plupart des caractéristiques de l'attachement définies par Kleine et Baker (2004) peuvent s'appliquer à l'automobile. La voiture peut être considérée comme une possession qui est une part du soi étendu (Belk, 1998). Dans Belk (1988), « l'habillement..., le logement..., et les automobiles sont tous considérés comme une « seconde peau » dans laquelle les autres peuvent nous voir » (p.151), « un individu peut utiliser ses possessions personnelles comme les bijoux, l'automobile, le maquillage, et les vêtements peuvent l'aider à se définir en tant que tel » (p.152). De la même manière, la définition formelle de l'attachement, proposée par Ball et Tasaki (1992, p.158) s'applique aussi à la voiture : « la mesure dans laquelle un objet qu'un individu a possédé, souhaite posséder, ou a précédemment possédé, est utilisé par cet individu pour contribuer à sa construction personnelle ». D'après Thomson, McInnis et Park (2005), « des attachements forts se développent au cours du temps et sont souvent basés sur des interactions entre un individu et

un objet d'attachement... Ces interactions motivent la construction de sens et lient de fortes émotions à l'objet d'attachement. » (p78).

Ainsi, à des âges plus avancés, les personnes ont davantage d'attentes ayant une signification émotionnelle car leur horizon temporel est vu comme limité (Fung et Carstensen, 2003) et leurs attachements pourraient amener les utilisateurs à racheter un véhicule de même marque.

C.1.iii. Hypothèse sur les effets d'âge sur la fidélité à la marque dans le cas de l'automobile

Ces différents effets liés à l'âge (le vieillissement physique et cognitif, l'aversion au changement, la sélectivité socio-émotionnelle, et l'attachement) convergent pour prévoir une préférence, avec l'âge, pour les choix plus familiers en matière de consommation.

Ceci nous permet d'élaborer l'hypothèse suivante :

(H1) En vieillissant, les consommateurs ont une plus grande tendance, quand ils achètent une voiture, à ré-acheter un véhicule de la même marque que leur précédent véhicule.

Cette hypothèse sera testée sur la base des deux applications de la section E sur les marchés français et allemand, et discutée au sein de la section F. A présent, nous allons étudier de manière similaire les sources d'influence potentielle du facteur cohorte sur la fidélité à la marque. Bien que la littérature marketing sur les cohortes soit moins riche que celle sur l'âge, des bases théoriques complémentaires vont nous permettre d'associer à cette hypothèse H1 sur les effets du vieillissement, une hypothèse H2 sur les effets de cohorte.

C.2. Influence de la cohorte sur la fidélité à la marque

Nous avons vu dans la section C.1. en quoi les effets du temps peuvent souvent être interprétés en termes de vieillissement (à travers la variable âge). Cependant, il est fréquent de voir attribuer à l'âge des effets s'avérant dus à la cohorte. En effet, nombreux sont les exemples, dans les informations quotidiennes, d'analyses qui tiennent uniquement compte de l'effet d'âge et sous-estiment la composante générationnelle.

Prenons quelques cas récents :

- Cas sociologique : les personnes âgées et l'usage du téléphone portable,
- Cas marketing : les jeunes et les musiques contemporaines, telles que le métal ou l'électro,
- Cas culturel : les personnes âgées parlant moins souvent anglais.

Dans ces trois cas, les raisons d'adoption des pratiques citées sont souvent davantage liées à la génération qu'à l'âge. Les jeunes consommateurs d'aujourd'hui qui utilisent un téléphone portable, écoutent des musiques dites contemporaines et parlent anglais, conserveront ces pratiques au cours de leur vie et ce jusqu'à être à leur tour des seniors.

Afin de bien marquer la différence avec les effets d'âge étudiés dans la section précédente, illustrons un effet de cohorte par un petit cas fictif. Dans les données ci-dessous, les valeurs possibles de l'âge sont indiquées sur la première colonne, et les périodes d'observation sur la première ligne. Le contenu de chaque cellule donne le taux de fidélité correspondant à chaque tranche d'âge pour la période correspondante.

Période	2000	2001	2002
Age de l'acheteur	Taux		
20	<u>40%</u>	35%	30%
21	45%	<u>40%</u>	35%
22	50%	45%	<u>40%</u>
23	<u>60%</u>	50%	45%
24	65%	<u>60%</u>	50%
25	70%	65%	<u>60%</u>

Tableau C.2.1. Exemple fictif d'effet de cohorte

Ici, chaque cohorte conserve le même taux de fidélité au cours du temps et il n'y a en réalité aucun effet du vieillissement, même si à une période donnée, plus une personne est âgée plus son taux de fidélité est élevé. Ainsi, dans chacune des trois enquêtes (en 2000, 2001 et 2002), la fidélité semble plus élevée pour les personnes plus âgées.

Les différences de taux de fidélité en termes d'âge reflètent ici des différences en termes de cohorte. Ces effets de cohortes peuvent être notamment expliqués par le fait que l'offre a pu évoluer au cours des années (e.g. plus de variété) et donc que les différentes cohortes n'ont pas été exposées aux mêmes offres aux mêmes âges.

Après avoir donnée une définition de la cohorte, nous allons exposer les différents effets qu'elle peut avoir sur le comportement (conservatisme, nostalgie, self-consistency) afin d'en déduire une hypothèse H2 sur son impact sur la fidélité à la marque.

C.2.i. Définitions des cohortes

L'impact de la cohorte est parfois moins directement observable que l'effet de l'âge. En effet, la cohorte résume l'ensemble des expériences vécues aux mêmes âges par un groupe de personnes, nées la même année ou à peu de distance.

Le choix de l'unité de temps par année constitue en démographie la norme pour définir une cohorte, nous réutilisons ici cette norme.

Définition théorique

Repartons de la définition de Ryder (1965, p.745), proposée par Rentz, Reynolds et Stout (1983) : « *the aggregate of individuals (within some population definition) who experienced the same event within the same time interval.* » Ici l'événement définissant la cohorte sera la naissance. Dans d'autres recherches en marketing, notamment en Customer Relationship Management (CRM), ce peut être l'année d'adhésion à un service (Schweidel, Fader et Bradlow, 2008). Dans l'éducation supérieure, le terme est synonyme d'une promotion d'entrants. Dans la vie personnelle, il peut s'agir du mariage ou du premier enfant. Dans la consommation, ce peut être l'accès à l'autonomie financière, ou une acquisition très impliquante, comme l'achat d'un logement.

Pour conserver toute la généralité de notre approche, nous considérerons ici l'année de définition de la cohorte comme l'année de naissance des individus étudiés.

Cohorte, génération et cohorte générationnelle

Notons dès à présent que la cohorte ne doit pas être confondue avec la notion de génération qui est plus vague et qui regroupe plusieurs cohortes. Citant Mannheim (1953), Jaworski et Sauer (1985, p.32) indiquent que la notion de génération « devrait être restreinte à la descendance familiale », ce qui représente des cycles de trente ans²⁰. Par exemple, Smith et Clurman (1997) définissent trois générations : la génération « mature » des personnes nées entre 1909 et 1945, celle des baby-boomers nés entre 1946 et 1964 et la génération X pour ceux nés entre 1965 et le début des années quatre-vingts. Pour les générations suivantes, les termes ne sont pas encore clairement définis. On parle aussi bien de la génération Y (succession des Xs) que de la génération N (pour Network).

Meredith et Schewe (2002, p16) définissent quant à eux sept cohortes générationnelles basées sur les événements marquants ayant eu lieu quand les personnes étaient dans la tranche d'âge de 17 à 23 ans. Les personnes de 80 ans et plus ont connu durant leur prime jeunesse la grande dépression puis la seconde guerre mondiale. La génération d'après-guerre, correspondant aux personnes âgées de 64 à 80 ans, a connu la guerre froide. Les « premiers boomers » âgés de 54 à 62 ans (nés entre 1946 et 1954) ont grandi avec la guerre du Vietnam, tandis que les « boomers tardifs » âgés de 44 à 53 ans (1955-65) ont connu le scandale du Watergate et la crise pétrolière. La génération X âgée de 32 à 42 ans (née entre 1965 et 1975) était exposée au crash des marchés financiers de 1987 et à la chute du mur de Berlin de 1989, alors que les personnes âgées de moins de 30 ans ont connu l'avènement d'Internet et les attentats du 11 septembre (Meredith et Schewe, 2002).

Concernant les générations, les risques de confusion lexicale sont assez importants, car il n'existe pas de définition officielle de chaque génération. Et les risques de confusion ne s'arrêtent pas là. Par exemple, Mannheim (1952) introduit la notion intermédiaire de cohorte générationnelle, définie sur une tranche de cinq à dix ans. Pour cette raison, toute la suite de

²⁰ Mais Rindfleisch (1994) estime que l'appartenance à une cohorte peut ainsi avoir plus d'impact sur le comportement de consommation que l'appartenance familiale.

l'analyse sera fondée sur des cohortes annuelles définies par l'année de naissance, et si le terme « générationnel(le) » est utilisé, il se réfère lui aussi à cette définition de la cohorte.

Mise en oeuvre de la cohorte

Il nous arrivera dans le texte d'utiliser alternativement les termes de cohorte et de génération mais nous basons notre analyse sur les cohortes annuelles définies par l'année de naissance.

Ainsi définie par l'année de naissance des individus, la cohorte a une signification très particulière. Ces groupes ont été exposés aux mêmes événements socio-économiques aux mêmes âges, évoluent en parallèle, suivant les mêmes normes sociales; ce qui implique que leurs expériences ont été constituées en grande partie des mêmes ingrédients, et ce dans une chronologie identique.

Les personnes d'une même cohorte ont une histoire sociale commune (Schewe et Noble, 2000) et certains événements, spécialement ceux vécus pendant l'adolescence et la jeunesse (Schuman et Scott, 1989), peuvent avoir un impact déterminant sur toute la durée de la vie (Bodier, 1999), conditionnant leur système de valeurs et d'interprétation. Cette expérience psychosociale similaire a un impact sur le comportement social, mais aussi sur le comportement de consommation, comme indiqué par Schewe et Meredith (1994, p.51) : « Ces moments clés influencent les valeurs, les préférences, les attitudes et le comportement d'achat, et ce tout au long de la vie. »

C.2.ii. Influence de la cohorte sur la fidélité à la marque

Parallèlement à ce qui a été présenté pour l'âge, nous allons ici recenser les principales causes pouvant expliquer l'influence de la cohorte sur la fidélité à la marque. Les trois principales causes que nous avons pu repérer sont les suivantes :

- le conservatisme,
- la nostalgie,
- la self-consistency.

Elles nous permettront d'établir dans la prochaine sous-section une hypothèse sur l'influence globale de la cohorte sur la fidélité à la marque.

Conservatisme

Une résistance en termes d'attitudes et de comportements peut être davantage révélatrice de la cohorte que de l'âge. En effet, il serait hasardeux de penser que les personnes aujourd'hui âgées ont un comportement plus conservateur à cause de leur âge, car elles peuvent être bien moins conservatrices que ce à quoi les prédisposait leur conditionnement générationnel. Par exemple, Danigelis et Cutler (2007) étudient les stéréotypes attribués aux personnes âgées concernant l'intolérance et le conservatisme en modérant le lien entre l'âge et ces attitudes par les effets de cohortes. Ils montrent que l'évolution au sein des cohortes peut aller vers une plus grande tolérance et un moindre conservatisme au cours du vieillissement, mais que cette évolution est masquée par le fait que leur point de départ est très différent. Ainsi les différences qui subsistent avec les générations suivantes, et donc avec les personnes plus jeunes au moment de l'observation, demeurent importantes.

Transposé au marketing, ce conservatisme relatif des cohortes permet d'anticiper une certaine constance des comportements de consommation au sein d'une même cohorte. Ainsi certaines attitudes comme le patriotisme économique peuvent avoir un grand impact sur le mode de consommation, particulièrement pour le choix de marques pour un bien durable, tel qu'un véhicule automobile.

Nostalgie

Holbrook et Schindler (1991,p.330) définissent la nostalgie comme "une préférence ... envers des objets (personnes, lieux ou choses) qui étaient plus communs (populaires, très répandus, à la mode) quand on était plus jeune (la jeunesse, l'adolescence, l'enfance, ou même avant la naissance)." Les psychologues expliquent qu'une personne traverse une période critique durant le "début de sa vie" en termes de développement psychologique (Bornstein, 1989, cité dans Schindler et Holbrook, 2003, p277). Ce concept a d'abord été appliqué dans le cadre de la théorie de la *gestalt* qui considère que l'empreinte appliquée au début du développement animal conditionne le comportement (Lorenz, 1951; cité dans Holbrook et Schindler 1989

p.550 et dans Holbrook et Schindler 1991, p.331) et a été mise en évidence dans de nombreuses études dont la plus citée reste sans doute celle de Schuman et Scott (1989).

En marketing, Divard et Robert-Demontrond (1997) font une revue étendue de la littérature sur la nostalgie. Ils font référence à Baker et Kennedy (1994) qui proposent de distinguer trois types de nostalgie : la nostalgie réelle, la nostalgie simulée et la nostalgie collective. La nostalgie réelle correspond à des remémorations d'événements précis vécus par l'individu (par exemple : la photo de la classe de lycée). La nostalgie simulée se réfère à des expériences indirectes et transmises par les médias ou par des proches, pouvant avoir eu lieu avant même la naissance de l'individu (exemple de la nostalgie des jeunes d'aujourd'hui pour les événements de Woodstock). La nostalgie collective est celle qu'« il est possible d'éprouver pour des dimensions représentatives ou symboliques d'une culture, une génération ou une nation ». Baker et Kennedy (1994) précise que l'émotion est dans ce cas ressentie de manière relativement homogène par des individus ayant le même champ d'expériences ou le même arrière-fond culturel. Il s'agit donc bien ici d'un effet qui s'exprime davantage à travers la cohorte qu'à travers l'âge, car ce groupe n'est pas défini par la suite, au cours du temps, par son âge, mais par l'âge commun qu'avaient les individus à l'époque de cette expérience commune, c'est-à-dire la cohorte.

Les travaux les plus populaires sur ce thème en matière de comportement du consommateur sont ceux d'Holbrook et Schindler (1994, p.414). De la même manière, ils expliquent que les “consommateurs se forgent leurs goûts durant des périodes sensibles” spécialement dans “l'adolescence tardive ou la prime jeunesse” (Holbrook et Schindler 1989, p.119)²¹ et suggèrent que les consommateurs conservent cette empreinte de préférences tout au long de leur vie. Holbrook (1989,1994) démontre ainsi que les individus conservent dans le temps leurs préférences pour les consommations culturelles qu'ils ont connues durant leur jeunesse²².

Dans une autre catégorie de produits, Rentz, Reynolds et Stout (1983,1991) montrent que l'analyse des cohortes peut sensiblement modifier l'évaluation des consommations de boissons par âge. Une analyse en coupe semble indiquer que les consommateurs consomment

²¹ Période aussi décrite sous la forme “late teens and early twenties” (Schindler et Holbrook, 1993, p.551),

²² Il est intéressant de remarquer que ces préférences de consommation pourraient potentiellement s'étendre au support de consommation lui-même. Cf. Annexe 1.

de moins en moins avec l'âge. En réalité, une analyse longitudinale montre qu'au sein de chaque cohorte, le niveau de consommation est constant suivant l'âge (on ne consomme pas moins avec le temps), mais que le niveau de consommation de chaque cohorte augmente par rapport aux précédentes.

Malgré le peu d'analyses marketing sur le thème des cohortes, on observe à travers ces deux séries d'études des phénomènes cohérents en termes de formations des goûts (culturels ou gustatifs). Les consommations auxquelles a été exposé l'individu à une certaine période sont des repères conditionnant ces consommations futures, tels des marqueurs sociaux.

Dans les paragraphes suivants, nous montrons que le consommateur ne fait pas que subir ces influences. En effet, il a aussi tendance à les cultiver activement pour se renforcer une identité (par rapport à son histoire et à son passé) : il s'agit du concept de self-consistency.

Self-consistency

Sur le plan psychologique, Rosenberg (1979) expose l'effet du besoin de self-consistency par laquelle les individus ont tendance à conserver les points de vue et attitudes développés dans leur jeunesse dans un effort pour conserver la conception qu'ils ont d'eux-mêmes et qu'ils renvoient aux autres. Ce besoin mène à une certaine constance face aux évolutions sociales, parfois malgré le jugement que les autres peuvent porter sur eux (Swann, Griffin et Gaines, 1987). Du point de vue de la consommation, ce besoin de self-consistency s'exprime aussi d'après Belk (1990) à travers certains objets dont le rôle est en partie d'assurer une certaine stabilité dans nos vies et d'enraciner une partie de notre identité dans le passé (notion d'extended self). Au-delà de l'empreinte de la nostalgie, des liens avec le passé sont donc activement maintenus par la consommation et l'usage de certains objets pour consolider la construction personnelle.

Dans la perspective de Belk, nous proposons que ces objets constituent autant des « vecteurs » vers un ensemble de souvenirs et d'émotions passées que de simples « marqueurs » du passé. Ce point de vue permet d'enrichir notre explication du lien existant entre consommations actuelles et expériences passées en détaillant la dynamique.

Pour résumer, la figure C.2.1. ci-après illustre le rôle des choix de consommation comme aller-retour entre le présent et une période de référence à travers les phénomènes de nostalgie (comme marqueur identitaire, influence depuis le passé vers le présent) et de self-consistency (comme vecteur identitaire, consolidation depuis le présent vers le passé). Comme à chaque cohorte correspond une période de référence différente, les constantes liées à cet aller-retour varient d'une cohorte à l'autre, et les caractérisent donc différemment.

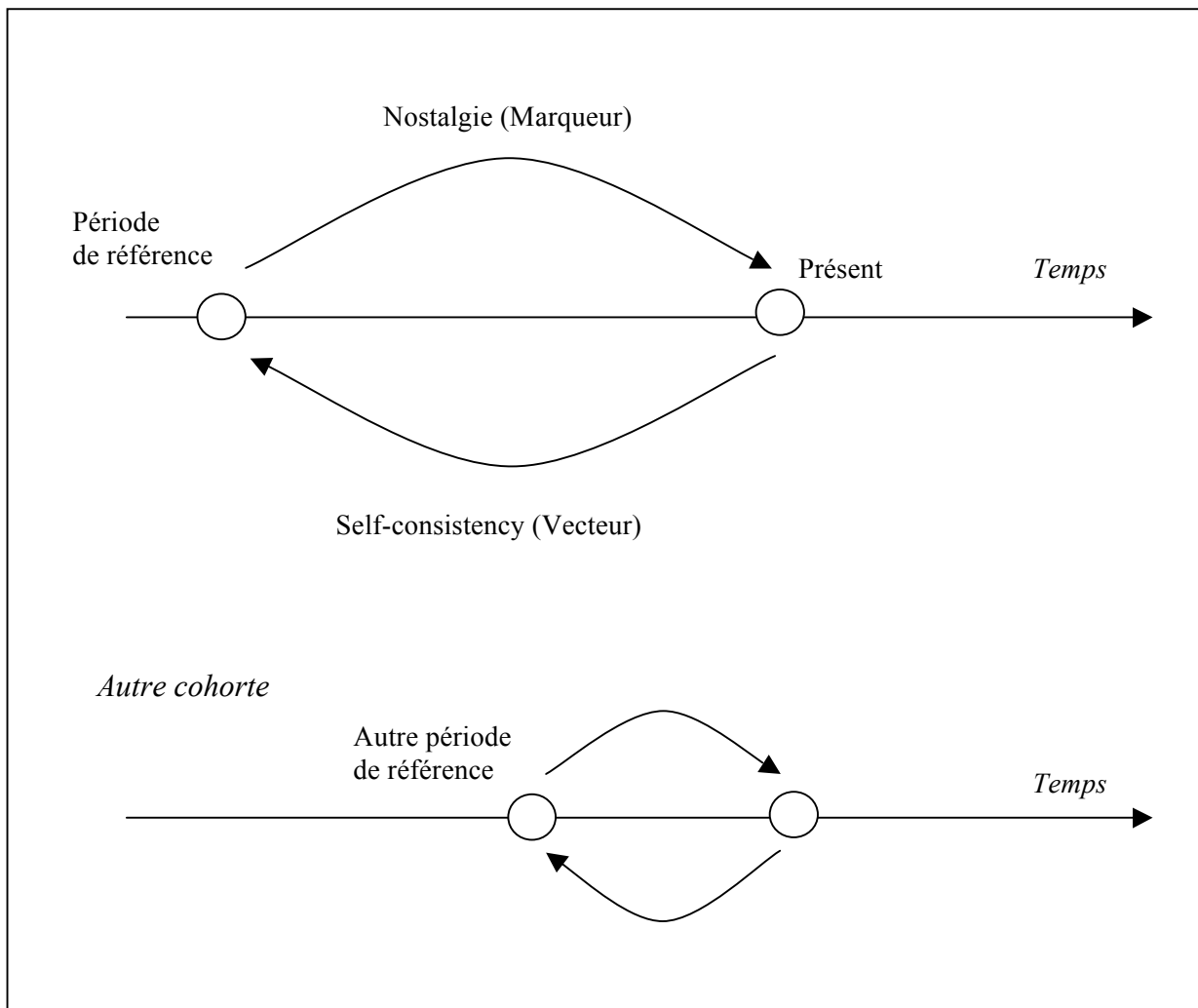


Figure C.2.1. Nostalgie et self-consistency appliquées à deux cohortes différentes

Ces deux concepts, associés au conservatisme, confèrent aux cohortes des éléments communs qui leur sont propres et qui vont dans le sens d'une certaine constance de leurs comportements en fonction des expériences vécues durant leurs périodes de référence respectives.

C.2.iii. Hypothèse sur les effets de cohortes dans le cas de l'automobile

Nous étudions la fidélité à la marque, et ce dans le cadre de l'achat d'un nouveau véhicule. Même si nous avons mis en évidence certaines constantes générationnelles liées aux expériences vécues durant une période de référence, pourquoi penser que les cohortes plus anciennes aient conservé des niveaux de fidélité plus élevés ?

Ceci est lié à la structure du marché en France au moment où chaque cohorte a été pour la première fois en âge d'acheter une voiture ou de s'intéresser aux marques automobiles dans une optique d'achat. Cette histoire commune crée une relation particulière entre les marques sur le marché durant cette période de référence et les consommateurs de cette cohorte. Et ce lien va perdurer au fil des années pour les raisons évoquées précédemment.

Or le nombre de marques automobiles a sensiblement augmenté au cours du siècle dernier. Donc plus une cohorte est récente, plus grand est le nombre de marques automobiles auxquelles elle a été exposée à ces âges clefs, plus large l'ouverture qu'elle aura à différents constructeurs, modèles, modes d'achat, et plus faibles seront ses chances de rester fidèle à une même marque.

Pour aller plus en détail, reprenons l'approche de Schewe et Meredith (2002) présentée en section C.2.ii. consistant à repérer les caractéristiques du contexte socio-économique de chaque génération durant sa période de référence. Dans notre contexte, cela consiste à étudier la structure de l'offre, en termes de nombre de marques (et de part de marché de ces marques), à différentes périodes de référence.

Pour le marché français de l'automobile cette information est résumée dans le tableau C.2.2. ci-après. La première colonne décrit les différentes périodes de référence potentielles, la seconde les dates d'introduction et la nationalité des nouvelles marques disponibles sur ce marché, la troisième la part de marché des marques françaises et enfin dans la dernière colonne figurent les générations correspondant aux différentes périodes de référence potentielles décrites dans la première colonne²³.

²³ Idéalement, nous aurions pu raisonner année par année et, donc cohorte par cohorte, mais les données de part de marché n'étaient pas disponibles à une telle échelle. De même nous n'avons pas les parts de marché par marque mais le cumul des parts de marché des marques nationales. Cependant, cette description nous donne une image assez fine de l'évolution du marché pour chaque grande période pour l'utiliser dans cette recherche.

Période de référence	Date d'introduction des marques par pays d'origine	Part de marché des marques nationales	Cohorte arrivant sur le marché
Avant-guerre et seconde guerre mondiale	France: Panhard (1891), Peugeot (1896), Renault (1899), Salmson (1908), Citroën (1918), Simca (1934) Allemagne: Daimler-Benz (1926) EU: General Motors (Pontiac, Cadillac, 1939)	Très élevée, mais non disponible	Dépression 1912-21 Seconde Guerre mondiale 1922-27
Post-war 1945-1960	GB: Jaguar (1951) Allemagne: Porsche (1951), Opel (1959) Italie: Fiat (1958)	Idem	Génération d'après-guerre 1927-43
1960-69	EU: Ford (1960) Allemagne: Volkswagen (1960), Mercedes (1960), BMW (1961), Autobianchi (1965) Italie: Alfa Romeo (1961), Lancia (1966) GB: Austin (1962), Triumph (1963) Allemagne: Opel (1963), Mercedes (1964), BMW (1964), Audi (1968) République tchèque: Skoda (1965) Suède: Volvo (1967), Saab (1968) Japon: Honda (1967), Suzuki (1968)	89% en 1962	Boomers 1946-54
1970-79	Allemagne: Audi (1970) EU: Chrysler (1970) Japon: Toyota (1972), Datsun (1972), Mazda (1975), Nissan (1978) ²⁴ , Mitsubishi (1979) Espagne: Seat (1972) Russie: Lada (1976) GB: Rover (1978)	79% en 1972	Boomers tardifs 1955-65
1980-90	Japon: Toyota (1983) Suède: Saab (1984)	69% en 1982	Génération X 1965-75
1990-2000	Japon: Hyundai (1992), Suzuki (1995), Subaru (1997), Daihatsu (2000), Corée: Kia (1994), Daewoo (1996) Suisse: Smart (1999) Japon-EU: Lexus (1999)	59,8% en 1992 57,2% en 1998 55% en 2007	Génération Y 1975-1985

*Ventes - Broustail et Greggio (2000, p. 212) et avec l'aimable autorisation d'une entreprise du secteur automobile.

Tableau C.2.2. Structure du marché automobile français au cours du temps.

²⁴ En remplacement de Datsun, marque japonaise elle aussi et lancée six ans plus tôt.

Les personnes de 80 ans et plus regroupées dans la génération « Dépression » ou « Seconde Guerre mondiale », ainsi que les consommateurs de la génération d'après-guerre âgés de 64 à 80 ans ont connu au moment de leurs 20 ans un marché de l'automobile réservé à une petite minorité, et dominé par les trois marques nationales historiques (Peugeot, Renault, Citroën).

D'après Thierry Morin, PDG de Valeo²⁵, chaque constructeur proposait un nombre très limité de modèles. Les consommateurs avaient donc un choix limité de modèles et de marques. Par exemple, avant la seconde guerre mondiale, Citroën avait une gamme de trois modèles en moyenne : il s'agissait en 1939 par exemple de la traction et du TUB²⁶.



Figure C.2.2. Reproduction du TUB de Citroën, modèle gris de 1939.

Cette gamme s'est progressivement élargie pour atteindre huit modèles dans les années 1970 (Broustail et Greggio, 2000). La Citroën DS a presque été l'unique voiture haut de gamme sur le marché français entre 1958 et 1970. Ainsi les personnes à la recherche d'une voiture haut de gamme se contentaient de renouveler leur DS, et en prenaient la version la plus récente. Il y avait aussi quelques autres marques de luxe, mais elles ne perdurèrent jamais longtemps, à l'instar de Facel-Vega, qui malgré sa réputation internationale, ne produisit qu'entre 1954 et 1964. De plus, les consommateurs étaient bien plus habitués aux pannes (par exemple, une panne de voiture était un classique par jour de gel).

Les « Boomers » âgés entre 54 et 62 ans (nés entre 1946 et 1954) ont grandi exposés principalement aux marques nationales bien établies et à quelques marques importées principalement européennes (Fiat, Opel, Volkswagen, Volvo, ainsi que l'exception de Ford).

²⁵ Conversation privée datant du 16 mars 2008.

²⁶ Acronyme : Traction Utilitaire de type B.

Alors que les « Boomers tardifs » âgés entre 44 et 53 ans (1955-1965) pouvaient étendre leur choix à des marques japonaises (Honda, Toyota Suzuki). La génération X, âgée de 32 à 42 ans (née entre 1965 et 1975) a vu l'arrivée de nouveaux entrants comme les marques coréennes (Daewoo...) et d'Europe de l'Est (Lada, Skoda).

Thierry Morin ajoute que le marché automobile offre à présent un nombre beaucoup plus grand de modèles par segment. En effet, les principales marques proposent entre 10 et 15 modèles différents dans différentes versions. Un consommateur est donc davantage susceptible d'être intéressé par un modèle en particulier, quitte à changer de marque. De même, les jeunes consommateurs acceptent beaucoup moins facilement les pannes que leurs prédécesseurs et ils sont donc beaucoup plus enclins qu'eux à changer de marque au premier incident (Thierry Morin).

Ainsi les cohortes plus âgées auraient plus tendance à être fidèles à une marque parce qu'elles ont été exposées à beaucoup moins de modèles et de marques lors de leur premier achat. Par contraste, les cohortes plus jeunes ont naturellement plus tendance à changer de marque parce qu'elles ont toujours été dans un univers comprenant beaucoup de marques et de modèles et qu'elles tolèrent assez mal d'avoir à subir des pannes.

Par ailleurs, la cohorte de baby-boomers nés entre 1946 et 1964 pourrait avoir encore d'autres raisons de changer de marque (Schewe and Noble, 2000). En regardant de plus près cette cohorte, Lipke (2000) cite Nat Puccio, Vice-président exécutif et Directeur du planning stratégique chez McCann Erickson à New York en indiquant que :

« les baby-boomers vont réécrire nos théories sur la fidélité à la marque chez les consommateurs âgés. »

En effet, chacune des perspectives théoriques sur le vieillissement du consommateur, décrites ci-avant, pourraient aider à prédire qu'avec l'âge, les baby-boomers évolueront différemment des cohortes précédentes. On peut d'abord s'attendre à les voir moins soumis aux effets du vieillissement biologique décrit précédemment. L'argument étant que ces cohortes sont en meilleure santé, ont une plus grande espérance de vie et bénéficient d'une meilleure condition physique. De même, les baby-boomers bénéficient d'un niveau d'éducation moyen plus élevé que les générations précédentes, ce qui devrait mener à de meilleurs scores aux tests

psychométriques, même à des âges avancés. Leur meilleure condition physique devrait aussi améliorer leurs capacités cognitives dans leurs décisions quotidiennes (Stuart Hamilton, 2004). En effet, Flynn (1987), cité par Zelinsky et Kennison (2007, p.547), a observé une augmentation des scores de logique entre des individus de 19 ans de 1950, et des individus du même âge testés en 1980. Des résultats similaires sont disponibles pour les personnes âgées. Dans Zelinsky et Kennison (2007), des personnes âgées de 74 ans ont obtenu des mesures d'intelligence fluide aussi bonnes que celles obtenues avec des personnes de 60 ans testées 16 ans plus tôt (p.554).

Les baby-boomers sont aussi plus ouverts au changement et à la flexibilité, dans le sens où ils ont toujours été plus mobiles, et ont plus souvent divorcé : les divorces représentent 33% des mariages en France en 2000 comparé à 24% en 1980 (Insee 2005). Aux Etats-Unis, les divorces ont doublé entre 1965 et 1975 (Almanach mondial 1991, cité par Rindfleisch, 1994 ; p.473), ce qui signifie que les baby busters nés entre 1965 et 1970 sont moins enclins que les baby boomers de 1950-1955 à suivre l'influence parentale, mais plus enclins à suivre l'influence de leurs congénères. A travers ces exemples, on peut s'attendre à ce que la voiture du père ne soit pas aussi populaire auprès des jeunes cohortes que ce qu'elle le fut pour la cohorte précédente. En effet, Coleman (2006, P. 192) établit que « les baby boomers ont été des pionniers et des acteurs du changement durant toute leur vie » et en citant Peterson (2005) « qu'ils avaient le désir d'essayer de nouveaux produits. » Pour résumer, les cohortes récentes étant plus ouvertes aux changements et moins attachées aux relations de longue durée, le facteur de sélectivité socio-émotionnelle pourrait ne pas jouer autant que pour les cohortes précédentes dans leur relation à la marque, et donc ne pas favoriser autant que pour les cohortes antérieures la fidélité à la marque.

Parce qu'elles ont été exposées à des marchés de plus en plus concurrentiels au cours du temps, et d'après les effets de conservatisme, de nostalgie et de self-consistency exposés plus haut, les cohortes plus récentes devraient avoir moins tendance que les générations antérieures à ré-acheter une voiture de la même marque.

(H2) Les cohortes récentes ont une moins grande tendance, quand ils achètent une voiture, à ré-acheter un véhicule de la même marque que leur précédent véhicule.

Cette hypothèse est rarement testée en pratique, parce qu'il est difficile d'observer une même cohorte à différents âges. En effet, cela nécessite d'avoir plusieurs périodes d'observation, alors que l'on travaille souvent sur des données transversales (et non longitudinales). Dans ce cas, on observe une cohorte à un seul âge, celui de l'âge qu'elle a durant la période d'observation, et tout l'effet potentiel de la cohorte est attribué à l'effet d'âge. Alors qu'ici les cohortes devraient pouvoir être observées à différents âges de cette cohorte. Et c'est seulement en observant des effets persistant à tout âge au sein des cohortes (et donc différents des effets du vieillissement) que l'on peut détecter des effets de cohorte.

Notons qu'H1 et H2 ne s'excluent pas l'une l'autre et qu'il est possible d'avoir simultanément des effets d'âge et de cohorte. D'ailleurs le but de cette recherche n'est en aucun cas de contester l'importance des effets du vieillissement sur la fidélité à la marque, mais plutôt d'en compléter l'analyse par une prise en compte des effets de cohortes qui peuvent lui être potentiellement associés. L'importance relative de ces deux effets peut par contre différer d'une population à l'autre, comme nous le verrons dans la comparaison des résultats en France et en Allemagne.

C.3. Résumé et rappel des hypothèses

Avant d'aborder les aspects plus méthodologiques de cette recherche, nous résumons ici les apports théoriques issus de la revue de littérature sur les effets d'âge et de cohorte nous ayant permis d'établir nos hypothèses sur les liens entre âge, cohorte et fidélité à la marque.

C.3.i. Résumé de la revue de littérature

Pour reprendre les propos de Rentz, Reynolds et Stout (1991), nous pensons que les prévisions du comportement du consommateur doivent intégrer les trois dimensions de l'analyse APC :

« The basic thesis of our article is that the response to the changing age distribution should be based on an understanding of the dynamics of aging, cohort succession, and environmental influences that produce consumers's responses to a given product or set of products, and not just the changing size of the age distribution. »

L'intégration des effets de cohorte est donc cruciale pour avoir une vision complète des dynamiques temporelles au sein de la population. Mais nous insistons aussi sur le fait qu'elle est essentielle pour avoir une mesure précise des effets d'âge, qui auraient sans cela tendance à être surestimés, comme nous le verrons par la suite.

Les effets de période sont quant à eux principalement intégrés par souci de contrôle. Ils interviennent quand on remarque une variation similaire et simultanée des comportements dans toutes les tranches d'âge à une période donnée. Un tel effet peut être dû soit à une cause externe qui peut être un événement économique ou politique (e.g. une nouvelle régulation du marché, une récession), ou soit à un événement sur le marché étudié (e.g. l'introduction d'une nouvelle marque, une rupture technologique comme le diesel et demain peut-être les véhicules électriques).

Notons pour finir que dans cette étude, les effets seront étudiés sans prendre en compte d'éventuels effets d'interaction. En effet, pour Knoke et Hout (1974), les interactions sont déjà implicitement intégrées dans l'analyse APC. Un effet d'interaction entre l'âge et la période se résume à l'impact d'avoir un âge donné à une période donnée, ce qui est un effet de cohorte. Quant aux interactions entre âge et cohorte, ou période et cohorte, elles semblent difficilement justifiables ou tout du moins interprétables.

C.3.ii. Résumé des hypothèses

A l'issue de cette section théorique, voici les deux hypothèses que nous testerons pas la suite :

- l'influence de l'âge sur la fidélité à une marque automobile :

(H1) En vieillissant, les consommateurs ont une plus grande tendance, quand ils achètent une voiture neuve, à ré-acheter un véhicule de la même marque que leur précédent véhicule.

- l'influence de la cohorte sur la fidélité à une marque automobile :

(H2) A âge égal, les cohortes récentes ont une moins grande tendance, quand ils achètent une voiture neuve, à ré-acheter un véhicule de la même marque que leur précédent véhicule.

La section suivante présente plus précisément le champ d'étude des effets APC, les différentes méthodes existantes sur le sujet, et l'innovation méthodologique que nous proposons afin de tester et commenter ces hypothèses H1 et H2 dans les sections suivantes.

D. Méthodes de modélisation des effets d'âge et de cohorte

Dans une première sous-section, nous définissons les concepts de multicollinéarité ainsi que les effets étudiés avant d'indiquer les notations qui seront utilisées par la suite. Puis les deux sous-sections suivantes présentent respectivement les méthodes fondées ou non sur les moindres carrés. Enfin la méthode APC-PLS élaborée dans cette recherche est décrite en détail dans une dernière sous-section.

D.1. Définitions et notations

D.1.i. Définitions de la multicollinéarité

Avant toute chose, nous souhaitons distinguer dans cette section la notion usuelle de multicollinéarité, qui désigne une forte corrélation entre les prédicteurs, de la multicollinéarité parfaite, qui désigne l'existence d'une relation linéaire parfaite entre une partie ou la totalité des prédicteurs.

Multicollinéarité usuelle

Dans le cas de la multicollinéarité usuelle, les prédicteurs sont très corrélés, de ce fait il existe une relation linéaire forte (mais non parfaite) entre les prédicteurs. Une relation linéaire forte se traduit notamment par le fait qu'il existe au moins un prédicteur, noté x_k qui peut être régressé par la méthode des moindres carrés ordinaires sur l'ensemble des autres prédicteurs X_k avec un R^2 très élevé, que l'on note R_k^2 .

Or rappelons que la matrice de variance-covariance du vecteur b des estimations des coefficients par les moindres carrés ordinaires s'exprime, en notant σ^2 la variance des termes aléatoires:

$$\text{Var}[b|X] = (X'X)^{-1} \sigma^2 \quad [2]$$

Et que la distribution de ce vecteur b , en notant β la vraie valeur de ce vecteur, est la suivante :

$$b | X \sim N(\beta, (X'X)^{-1} \sigma^2) \quad [3]$$

Comme rappelé par exemple dans Greene (2003, p.56), si l'on s'intéresse à l'estimation du coefficient d'un prédicteur k quelconque, on obtient:

$$\text{Var}[b_k|X] = \sigma^2 / [(1 - R_k^2) * \sum_{i=1 \dots n} (x_{ik} - \bar{x}_k)^2] \quad [4]$$

$$\text{et } b_k | X \sim N(\beta_k, \text{Var}[b_k|X]) \quad [5]$$

Ainsi, plus le lien de multicollinéarité entre un prédicteur et les autres sera fort, plus la variance de l'estimation de son coefficient sera grande. Par conséquent, le niveau de significativité du coefficient sera d'autant plus faible. Cet effet s'oppose à l'autre caractéristique des estimations de coefficients, à savoir que plus la variance de la variable qui lui est associée est grande, plus la variance de l'estimation de ce coefficient est réduite.

Ces deux caractéristiques ont une traduction intuitive simple :

- une variable à forte variance fournit une information claire sur la variable en question: même si cette information ne permet pas de forcément prédire les variations de la variable dépendante (pouvoir prédictif), l'effet du prédicteur peut être estimé avec plus de précision ;
- une variable fortement corrélée avec d'autres variables partage de l'information avec ces autres variables, et il est par conséquent difficile de déterminer l'origine de cette information commune et donc d'estimer l'effet de chaque prédicteur.

A contrario, un modèle de régression idéal serait d'après cette même formule, un modèle pour lequel le pouvoir prédictif des prédicteurs est fort (σ^2 faible), avec des prédicteurs à forte variance et orthogonaux entre eux (R_k^2 nuls). Un exemple typique de ce cas de figure est réalisé lors de la collecte d'information par les plans d'expérimentation.

Pour améliorer la qualité d'estimation des paramètres du modèle testé, on choisit souvent d'éliminer une ou plusieurs des variables colinéaires, afin de réduire la multicollinéarité du modèle. On utilise alors de l'information a priori, inspirée par l'expérience ou par la théorie, afin de choisir quelle(s) variable(s) éliminer. Dans le cas de l'analyse APC, il ne s'agit pas d'une bonne méthodologie. Par la suite, nous reviendrons plus amplement sur cette technique dans le cadre de son application au cas de la multicollinéarité parfaite.

Multicollinéarité parfaite

Dans le cas de la multicollinéarité parfaite, il existe une relation linéaire parfaite (sans résidus) entre les prédicteurs (ou un sous-groupe de prédicteurs). Ceci implique qu'il existe au moins un prédicteur k tel que R_k^2 soit égal à 1 et que son coefficient ne soit pas estimable (variance infinie d'après la formule 1). La matrice $X'X$, n'est plus de plein rang, donc n'est plus inversible, et il est impossible de calculer la valeur du vecteur b . La différence avec le cas précédent est qu'ici la régression linéaire n'est plus applicable du tout.

Greene (2002, p.74) résume bien la gravité de cette situation en notant que:

“The case of an exact linear relationship among the regressors is a serious failure of the assumptions of the model, not the data.”

Même si leurs libellés sont proches, les problèmes de forte multicollinéarité et de multicollinéarité parfaite (exact multicollinearity en anglais) diffèrent dans leur nature.

Notons ici que, malgré la différence fondamentale entre ces deux problématiques, il est aisé de passer techniquement de la multicollinéarité parfaite à une multicollinéarité usuelle en perturbant légèrement le lien linéaire entre les prédicteurs pour que cette relation, bien que souvent encore forte, ne soit plus parfaitement linéaire. Ce passage « de pis en mal » est à l'origine de plusieurs méthodes de résolution du problème APC.

D.1.ii. Multicolinéarité en APC

Il existe deux raisons principales pour lesquelles les effets d'âge, de période et de cohorte sont rarement étudiés simultanément :

- D'abord, ce type d'analyse requiert d'avoir une base de données longitudinales. C'est-à-dire qu'il faut pouvoir observer, de manière répétée, plusieurs tranches d'âges sur plusieurs années. En effet, des données transversales (cross-sectional data en anglais), qui sont les plus souvent disponibles pour étudier les effets d'âge n'apportent de l'information que sur une seule année, et n'apportent donc qu'une année d'observation par cohorte. Certaines données de panel suivent une même cohorte sur plusieurs années, mais ceci n'apporte de l'information que sur une seule cohorte. Pour mener une analyse APC, il est nécessaire de collecter l'information sur toutes les tranches d'âge pendant plusieurs années ;

- Ensuite, la contrainte majeure provient du fait que les trois variables étudiées : âge a , période p et cohorte c , sont parfaitement colinéaires. Cette dépendance linéaire résulte de la relation suivante évidente entre l'âge, la période (mesurée par l'année d'observation) et la cohorte (mesurée par l'année de naissance) : $\text{Age} = \text{Période} - \text{Cohorte}$, ou $A = P - C$ [Rappel de 1].

Comme proposé par Vandescrick (2001), partons d'une représentation traditionnelle des trois variables sur trois axes.

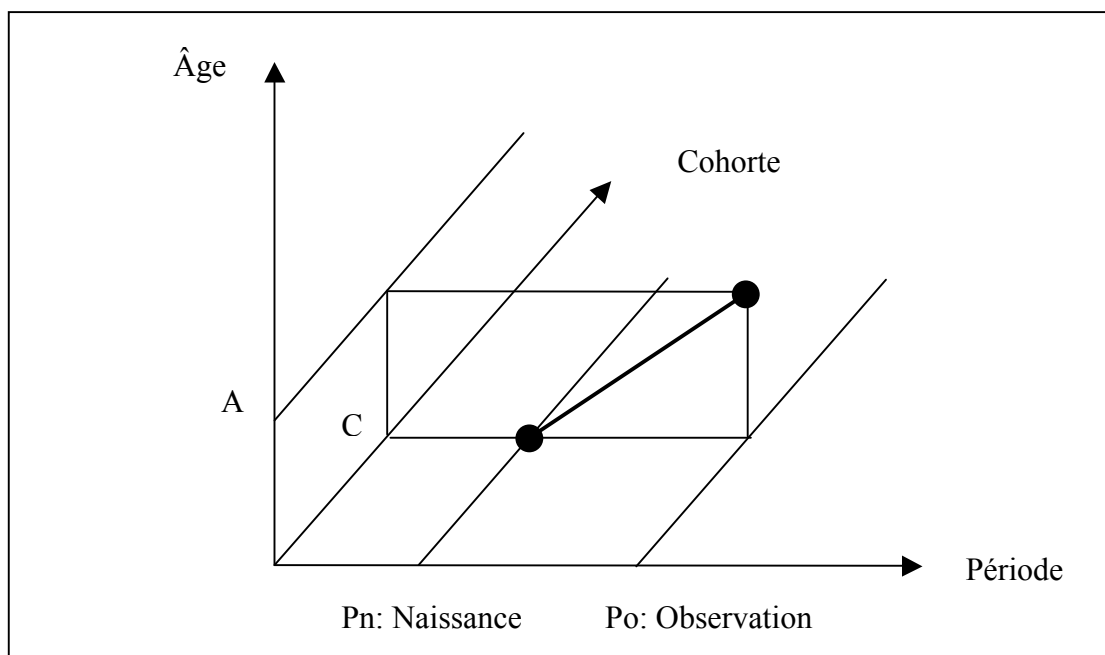


Figure D.1.1. Parcours d'un individu dans l'espace (Âge, Période, Cohorte)²⁷

²⁷ D'après Vandescrick (2001).

Sur la figure D.1.1, l'axe des abscisses représente la période, l'axe des ordonnées l'âge et l'axe de profondeur la cohorte.

Prenons un individu i quelconque qui naît à la période P_n , période de naissance qui correspond à la cohorte C , et qui est observé à la période d'observation P_o , à laquelle il a atteint l'âge A . Durant toute sa durée de vie, son âge varie en même temps que la période, suivant un axe parallèle à la première bissectrice du plan (Âge x Période). Le segment décrivant l'évolution de l'individu se situe dans le plan de la cohorte C .

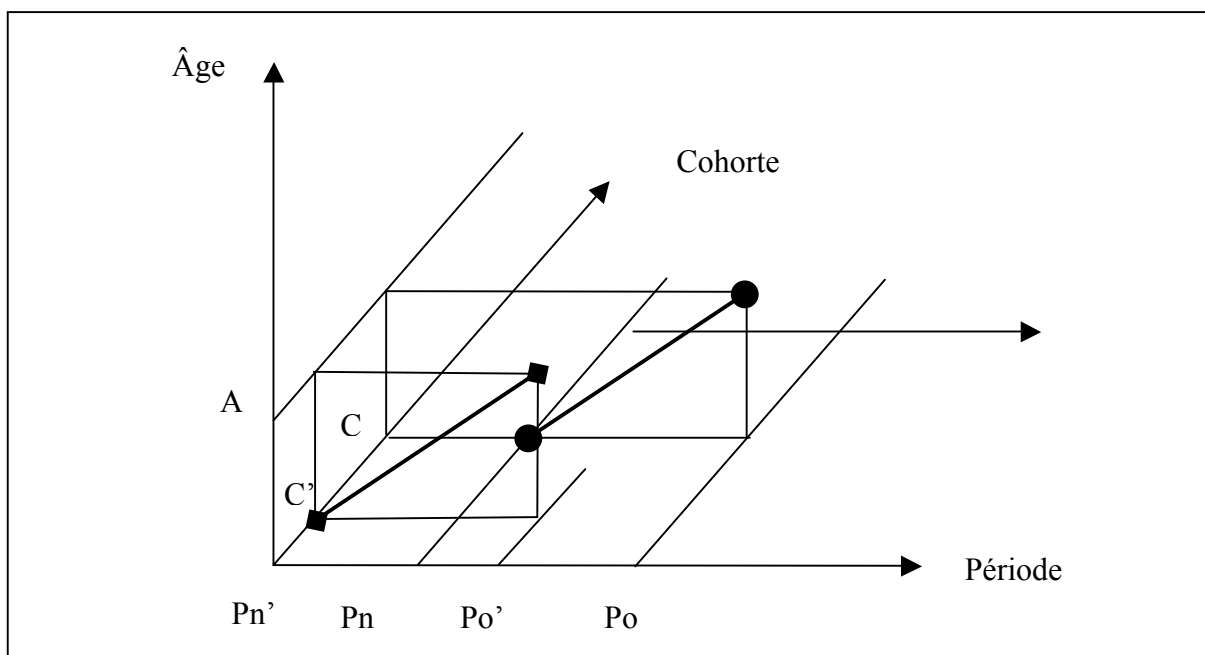


Figure D.1.2. Parcours de deux individus dans l'espace (Âge, Période, Cohorte)

Prenons à présent un second individu i' , d'une génération C' différente de C que l'on va observer au même âge A à la période P_o' . Dans cet exemple, l'âge est choisi identique, égal à A , par souci de clarté. La même représentation graphique étant réalisable avec des individus ayant des âges différents au moment de l'observation (respectivement P_o et P_o').

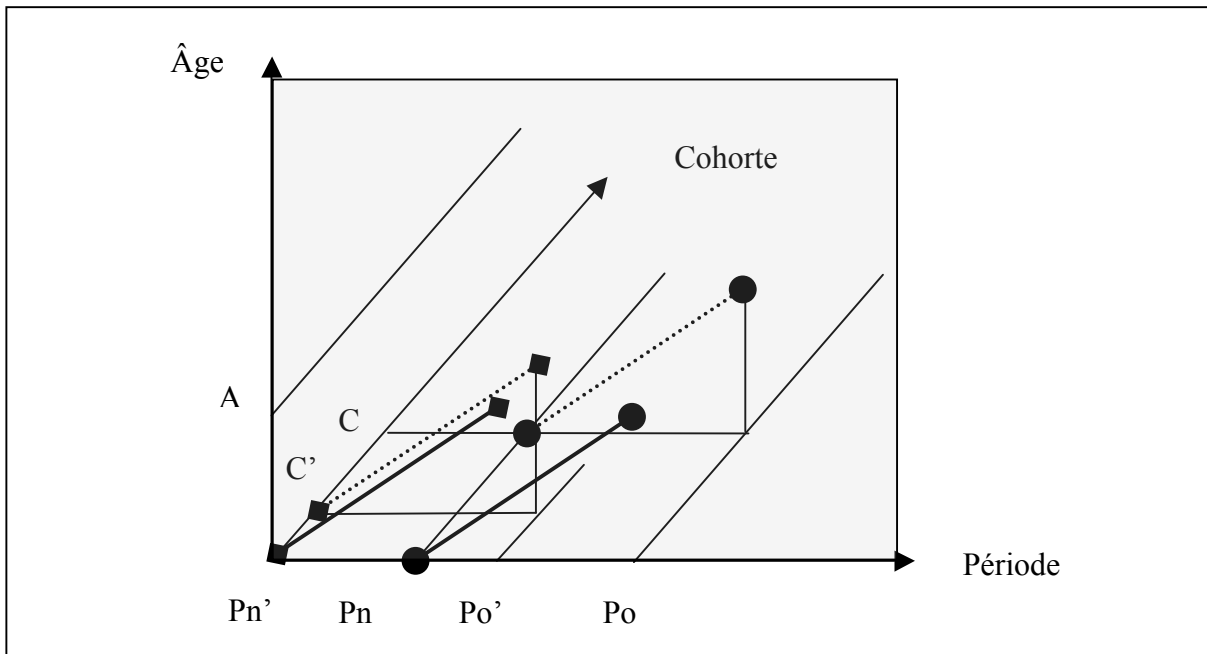


Figure D.1.3. Projection de deux parcours sur le plan ($\hat{\text{Age}} \times \text{Période}$)

Les droites correspondant aux deux parcours appartiennent à des plans parallèles (C et C') et sont toutes deux parallèles à la première bissectrice du plan ($\hat{\text{Age}} \times \text{Période}$). Elles sont donc parallèles entre elles et peuvent être projetées sur le plan ($\hat{\text{Age}} \times \text{Période}$).

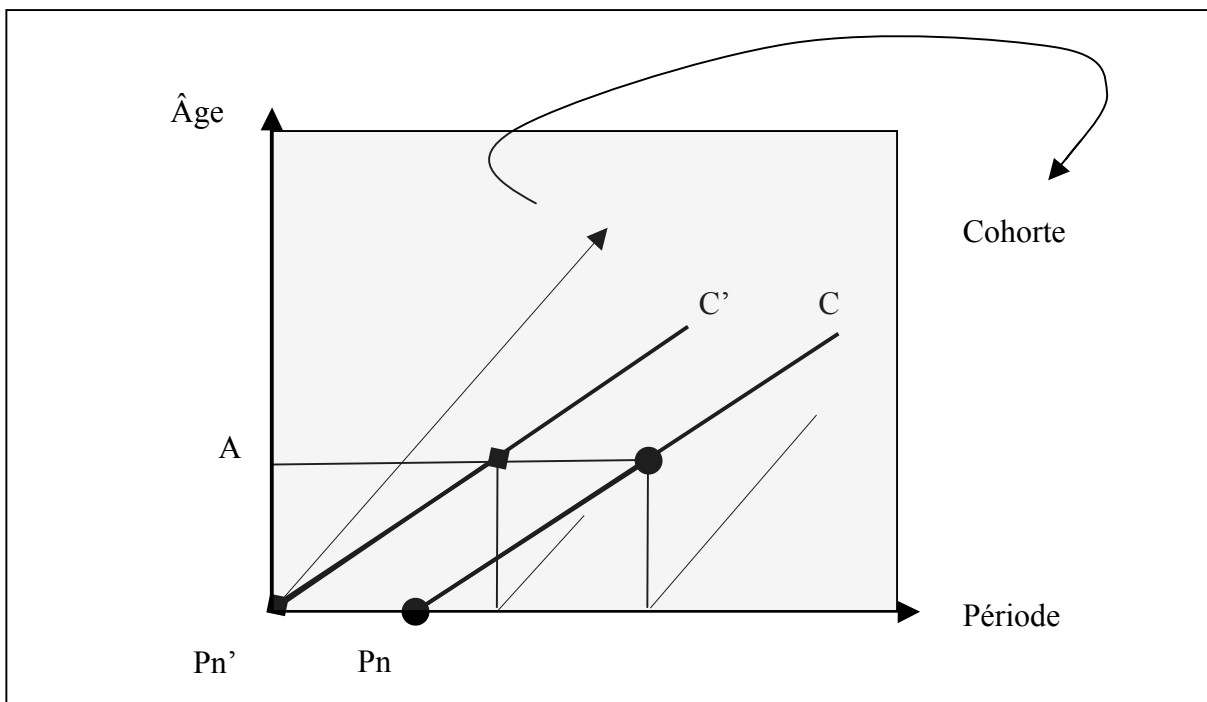


Figure D.1.4. Deux parcours projetés sur le plan ($\hat{\text{Age}} \times \text{Période}$)

L'axe de profondeur a ainsi été éliminé sans perte d'information. En effet, l'âge et la période ont des coordonnées inchangées, et la cohorte est lisible sur les parallèles de la première bissectrice. En effet, de par la parfaite multicolinéarité entre les variables, C' regroupe tous les individus, et seulement les individus, nés à la première période Pn', et C ceux nés à Pn. Les cohortes se lisent donc « en diagonale » et les trois variables sont dans un seul plan.

Cette opération peut être effectuée en projetant sur chacun des trois plans. Les démographes, axés sur la dynamique des populations, ont tendance à placer les cohortes en abscisse. En effet ils peuvent ainsi analyser à la fois les cohortes (de gauche à droite) et les âges (de bas en haut) en reléguant l'analyse des périodes à un second plan.

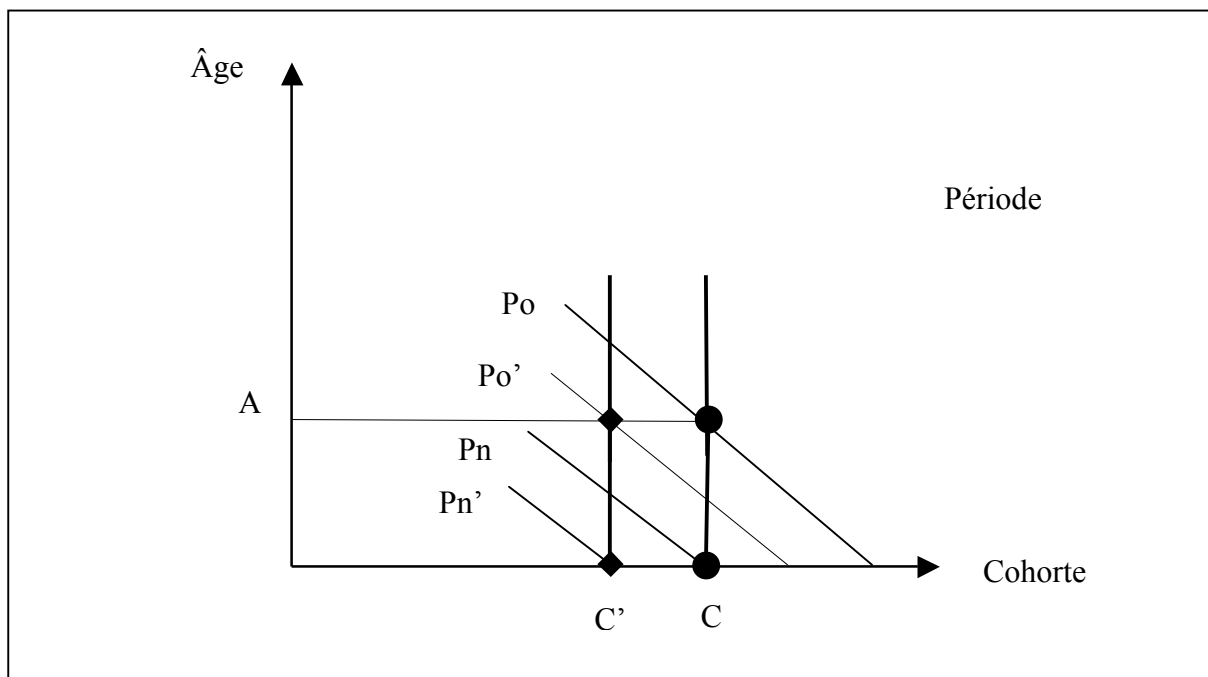


Figure D.1.5. Projections sur le plan (Âge x Cohorte)

En effet si l'on s'intéresse aux deux parcours précédents, on observe toujours l'âge A sur une ligne horizontale, mais la lecture de l'évolution au sein d'une cohorte (C' puis C sur la figure C.3.5.) se fait à présent verticalement de bas en haut. Les périodes se lisent cette fois en diagonale.

Par contre, ces diagonales ne forment plus un angle de 45 degrés avec l'axe des ordonnées mais avec un angle de -45 degrés, les périodes les plus à droite étant les périodes les plus récentes. Ainsi, en partant de la droite du graphique, on peut repérer tout d'abord les périodes

de naissance des cohortes C' et C qui sont respectivement Pn' et Pn. Puis les périodes d'observations respectives des deux cohortes à l'âge A, c'est-à-dire Po' et Po.

Ce type de figure s'avère très pratique pour comparer les profils des cohortes. Une fois les taux (ici le taux de fidélité lors de l'achat) indiqués dans chaque cellule, comme dans le tableau D.1.1. ci-dessous, les taux des cohortes se comparent cohorte par cohorte à âge égal en lisant le graphique horizontalement.²⁸

Le marketing est plus axé sur les évolutions par période et par âge, peut-être par méconnaissance des phénomènes de cohortes, il privilégie donc l'affichage sur le premier plan que nous avons présenté à savoir le plan (Age x Période). Nous prendrons donc ce parti par la suite. Il est en effet pratique de savoir lire les tableaux traditionnels « en diagonale » pour éventuellement y relever des effets de cohorte. Comme vu précédemment, les données chiffrées affichées de cette manière forment un diagramme de Lexis et ce type de tableau se lit en ligne pour les âges, en colonne pour les périodes et en diagonale pour les cohortes.

50	62%	53%	58%	58%	55%	54%	55%	53%	57%	53%	53%	1954
49	58%	59%	55%	57%	53%	55%	53%	56%	51%	53%	54%	1955
48	56%	51%	57%	55%	53%	54%	54%	56%	49%	50%	54%	1956
47	56%	54%	51%	53%	56%	58%	53%	53%	52%	51%	48%	1957
46	57%	44%	52%	53%	52%	50%	51%	52%	56%	56%	52%	1958
45	57%	52%	47%	49%	52%	53%	54%	49%	50%	52%	53%	1959
44	54%	53%	47%	60%	50%	45%	52%	55%	48%	51%	49%	1960
43	50%	51%	45%	54%	50%	43%	50%	44%	52%	52%	45%	1961
42	51%	50%	46%	52%	43%	46%	52%	45%	48%	45%	46%	1962
41	48%	52%	49%	42%	50%	48%	47%	47%	48%	45%	45%	1963
40	50%	49%	45%	49%	51%	50%	45%	48%	49%	45%	43%	1964
Age	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Cohorte
	Période											

Tableau D.1.1. Extrait du diagramme de Lexis de la base de données Fidélité Automobile

²⁸ Il existe une troisième disposition consistant à placer les périodes en abscisse et les cohortes en ordonnée, mais il s'agit de la manière la moins usitée car les effets d'âge, qui sont très souvent d'un intérêt direct, ne sont pas directement lisibles dans cette disposition.

Dans le tableau D.1.1. ci-avant, on présente un extrait de la base de données sur la fidélité automobile. Les cellules contiennent les taux de fidélité à la marque lors de l'achat d'un véhicule de remplacement, pour des âges allant de 40 à 50 ans, et des cohortes de 1954 à 1964, pour toutes les périodes d'observation, à savoir entre 1994 et 2004.

La diagonale sélectionnée (gris clair) représente les données disponibles pour la cohorte née en 1956 pour cette tranche d'âge (nous disposons aussi de données pour les âges 38 et 39 ans, observées en 1995 et 1994, non présentées dans ce tableau).

On observe que les personnes nées en 1956 ont un taux de fidélité à la marque de 45% à l'âge de 40 ans. En remontant la colonne de la période 1996, on remarque que sur cette période, le taux de fidélité augmente avec l'âge pour atteindre 57% à 48 ans et 58% à 50 ans. Cet effet peut être dû au vieillissement, indépendamment des cohortes, mais aussi au fait que les cohortes précédentes, qui sont plus âgées en 1996, ont des taux de fidélité supérieurs. Si l'on veut savoir comment le taux de fidélité évolue au sein de la cohorte des personnes nées en 1956, on doit se référer à la diagonale grisée partant des observations de 1996 (elles avaient alors 40 ans) jusqu'à celles de 2004 (elles avaient 48 ans). Au sein de cette cohorte, le taux de fidélité augmente pour atteindre 54% à 48 ans. Cet effet peut être dû au vieillissement de cette cohorte, mais aussi à l'évolution du taux de fidélité suivant les périodes, indépendamment des cohortes observées.

En suivant la ligne des personnes observées à l'âge de 40 ans, on s'aperçoit qu'aucune tendance claire n'apparaît sur l'évolution du taux, ce qui peut être lié à l'absence d'effet de période sur cette tranche d'âge.

Notons ici que le fait que la fidélité soit plus faible pour les personnes de 48 ans nées en 1956 (et observées à l'âge de 48 ans en 2004) que pour les personnes de 48 ans observées en 1996, nous rappelle les dangers de l'interprétation uniquement basée sur des données transversales car l'on voit ici que les variations par tranches d'âge pour une même année sont en partie dues au vieillissement (car au sein de la cohorte 1956 les taux de fidélité augmentent avec l'âge) mais qu'elles sont aussi potentiellement partiellement liées à la cohorte (car sinon leur taux de fidélité aurait été de 57% à 48 ans). Si l'on ne tient pas compte de ce potentiel effet de cohorte, l'ensemble de la variation du taux de fidélité est affecté par l'effet d'âge, ce qui peut mener à une surestimation. Ce phénomène est encore plus marqué pour les personnes de 40

ans observées en 1994, car les personnes de 50 ans de la même année ont un taux de fidélité de 62%. Dans une analyse transversale, on serait tenté d'affecter l'ensemble de la variation (12%) à un effet de vieillissement, mais lorsque l'on regarde ce qu'est devenu cette cohorte 1954, qui atteint 50 ans en 2004, on observe que leur taux de fidélité n'a varié que de 3% entre 40 et 50 ans, ce qui signifie que d'autres effets sont en jeu.

D.1.iii. Notations utilisées

Afin de présenter les différents modèles de manière cohérente, nous allons définir ici les notations utilisées dans le reste de la recherche.

Dans ces définitions, δ représente le delta de Kronecker.

Y : Variable à expliquer. Ici taux de fidélité à la marque lors du ré-achat.

X : Vecteur générique de variables indépendantes quand leur nature n'est pas indiquée.

A : Variable d'âge chronologique, prenant ici les valeurs de 20 à 80. Cette variable a été décomposée en une série d'indicatrices formellement indexées par i , allant de 1 à I .

$A_i = \delta(A, i)$: pour $i=1 \dots I$, variable indicatrice prenant la valeur 1 si la variable A prend la valeur i , et la valeur 0 sinon.

P : Variable de période d'observation, prenant ici les valeurs de 1994 à 2004. Cette variable a été décomposée en une série d'indicatrices formellement indexées par j , allant de 1 à J .

$P_j = \delta(P, j)$: pour $j=1 \dots J$, variable indicatrice prenant la valeur 1 si la variable P prend la valeur j , et la valeur 0 sinon.

C : Variable de cohorte, ou année de naissance, prenant ici les valeurs de 1914 à 1984. Cette variable a été décomposée en indicatrices formellement indexées par k , allant de 1 à K .

$C_k = \delta(C, k)$: pour $k=1 \dots K$, variable indicatrice prenant la valeur 1 si la variable C prend la valeur k , et la valeur 0 sinon.

A présent, nous allons terminer cette sous-section par une définition des effets mesurés avant de passer à la suite de cette section sur la modélisation.

D.1.iv. Définitions des effets APC

Afin de clarifier la suite de notre exposé, nous proposons ici deux définitions des deux types d'effets mesurés dans les analyses APC, les effets nets et bruts. La notation des effets nets est apportée en complément des notations de la page précédente.

Effet brut

Les effets bruts sont les effets directs d'une des variables APC sans contrôler ces effets par les deux autres variables APC. Prenons par exemple la variable A. L'effet brut de cette variable sera obtenu en n'expliquant Y que par cette variable. On parlera des effets bruts de la variable d'âge pour l'ensemble des coefficients des A_i , i allant de 1 à I (lorsque l'on explique Y par l'ensemble des A_i , i allant de 1 à I).

Ces effets bruts ne sont que l'écart moyen (suivant une modalité d'une des variables APC) de la variable à expliquer Y par rapport à sa valeur moyenne. Ces effets, comme nous l'avons vu précédemment, sont d'un intérêt très réduit. En effet, ils sont mesurés sans contrôler les influences des autres variables APC et sont donc souvent surévalués, comme nous le verrons en détail par la suite. Même en les analysant séparément, nous ne pouvons pas comparer les effets des différentes variables APC. Pour cela, il faut mesurer les effets nets de ces variables.

Effet net

Les effets nets sont eux les coefficients d'un modèle de prévision prenant en compte simultanément les trois variables APC. Prenons à nouveau l'exemple de la variable A. L'effet net de cette variable sera obtenu en expliquant Y par les trois variables APC, c'est-à-dire l'ensemble des indicatrices A_i , P_j et C_k , i allant de 1 à I, j de 1 à J et k de 1 à K ; et on parlera des effets nets de la variable d'âge pour l'ensemble des coefficients des A_i , i allant de 1 à I au sein de ce modèle.

Contrairement aux effets bruts, ils ne peuvent être obtenus directement à partir des données. Ce sont à ces effets que l'on s'intéressera par la suite, et qui sont difficiles à obtenir car il existe peu de méthodologies de prévision prenant en charge la multicollinéarité parfaite.

Notations complémentaires pour les effets nets

Les effets des modalités $A_{i, i=1\dots I}$, de A seront notés $a_{i, i=1\dots I}$

Les effets des modalités $P_{j, j=1\dots J}$, de A seront notés $p_{j, j=1\dots J}$

Les effets des modalités $C_{k, k=1\dots K}$, de A seront notés $c_{k, k=1\dots K}$

D.2. Illustration de la multicollinéarité APC sur la base de données France

Nous allons ici présenter la base de données utilisée pour le premier exemple d'application de notre méthode et de test de nos hypothèses, à savoir la base de données d'achat de véhicules en France. La base de données sur le marché allemand a exactement la même structure. De plus, nous l'utilisons pour illustrer les principes de multicollinéarité parfaite et d'effets bruts.

D.2.i. Nos données : des Courants Commerciaux au pseudo-panel

Le questionnaire utilisé est extrait d'une étude européenne menée consécutivement à l'achat d'un véhicule automobile par un particulier. Cette étude, appelée « Courants Commerciaux », est menée à la demande d'un consortium de constructeurs pour comprendre les comportements d'achat des consommateurs. Par extraction ont été éliminés les véhicules professionnels.

Nous analysons ici les données concernant les consommateurs français et allemands sur une période de 11 ans allant de 1994 à 2004. Le nombre de personnes sondées varie entre 13537 et 18316 d'individus pour les données françaises, comme indiqué dans le tableau D.2.1. ci-dessous, ce qui correspond à un taux d'échantillonnage très élevé lors de l'enquête.

Dans toutes les études de modélisation sur le thème de l'analyse âge-période-cohorte, l'erreur de mesure potentielle liée à la qualité du recueil des données est considérée comme négligeable. Pourtant il faudra prendre quelques précautions concernant les premières et les dernières cohortes pour lesquelles, comme nous le verrons plus loin, seul un nombre limité de points d'observation est disponible.

<i>Période</i>	<i>Effectifs</i>
1 994	15 589
1 995	13 527
1 996	14 291
1 997	13 862
1 998	14 750
1 999	14 659
2 000	16 560
2 001	16 749
2 002	18 316
2 003	17 278
2 004	17 169
Total	172 750

Tableau D.2.1. Effectifs de l'enquête Courants Commerciaux par année (France)

Cette enquête recueille la marque et le type de la voiture précédente et de la voiture nouvellement acquise. Ces données nous permettent de déduire pour chaque individu s'il a ré-acheté un véhicule de la même marque ou s'il en a changé lors de son achat. Nous en déduisons ensuite les taux de ré-achat pour chaque triplet de valeurs des variables âge - période (année d'observation) - cohorte (année de naissance).

Cette opération nous fournit 671 observations. En effet, l'âge varie entre 20 et 80 ans (61 âges différents). Cette tranche d'âge recense la plupart des rachats d'automobiles neuves, les rachats étant très peu fréquents avant 20 ans ou après 80 ans. Pour chacune des 11 périodes, de 1994 à 2004, on obtient donc ainsi 61 observations (nombre d'âges différents). Or, comme on le verra en détail par la suite, pour chaque couple âge - période, ne correspond par multicollinéarité parfaite qu'une et unique cohorte. Par conséquent, la table contient 671 observations correspondant à chacun des triplets de valeurs des variables (âge, période, cohorte). Les cohortes varient elles, de 1914 (1994-80) à 1984 (2004-20).

Cette agrégation permet de constituer ce que l'on nomme un pseudo-panel. Cette technique consiste à suivre un comportement moyen au travers d'enquêtes successives, plutôt que d'analyser un vrai panel dans lequel les données auraient été recueillies auprès d'un même groupe de personnes ne variant pas au cours du temps. Comme indiqué par Bonier (1999), « suivre le comportement moyen d'une cohorte revient à suivre le comportement d'un ménage représentatif de la cohorte : en utilisant les données moyennes pour chaque cohorte, (...) on a alors des données temporelles qui permettent de suivre des cohortes comme on

suivrait des ménages dans un véritable panel. » Dans notre cas, il aurait été difficile d'utiliser de réelles données de panel, étant donné que l'achat automobile n'est pas communément renouvelé chaque année. Cependant, les données ne sont pas forcément moins bonnes que celles issues d'un panel car elles permettent d'éviter deux inconvénients des panels : l'attrition du panel et les phénomènes d'apprentissage des individus du panel.

Les démographes reportent traditionnellement des données de pseudo-panel dans un diagramme âge-période aussi appelé « diagramme de Lexis ». La multicolinéarité parfaite entre ces trois variables permet en effet de les représenter simultanément dans un même plan. Ce n'est pas d'ailleurs la seule façon de représenter ces trois axes dans un seul plan. Vandescrik (2001) expose deux autres modes de représentation possibles et explique par ce biais l'origine d'une mauvaise dénomination qui a attribué à Lexis la paternité d'un diagramme qu'il avait en réalité emprunté à Zeuner (1869).

50	62%	53%	58%	58%	55%	54%	55%	53%	57%	53%	53%	1954
49	58%	59%	55%	57%	53%	55%	53%	56%	51%	53%	54%	1955
48	56%	51%	57%	55%	53%	54%	54%	56%	49%	50%	54%	<u>1956</u>
47	56%	54%	51%	53%	56%	58%	53%	53%	52%	51%	48%	1957
46	57%	44%	52%	53%	52%	50%	51%	52%	56%	56%	52%	1958
45	57%	52%	47%	49%	52%	53%	54%	49%	50%	52%	53%	1959
44	54%	53%	47%	60%	50%	45%	52%	55%	48%	51%	49%	1960
43	50%	51%	45%	54%	50%	43%	50%	44%	52%	52%	45%	1961
42	51%	50%	46%	52%	43%	46%	52%	45%	48%	45%	46%	1962
41	48%	52%	49%	42%	50%	48%	47%	47%	48%	45%	45%	1963
40	50%	49%	45%	49%	51%	50%	45%	48%	49%	45%	43%	1964
Age	<u>1994</u>	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Cohorte
	Période											

Tableau D.2.2. Extrait du diagramme de Lexis de la base de données Fidélité Automobile

Dans le diagramme du tableau D.2.2. ci-dessus, les âges se lisent en ligne et les périodes en colonne, et qui permet de lire les données de cohorte « en diagonale à 45° ». L'extrait ci-dessus du diagramme de Lexis de nos données nous permet par exemple de lire les taux de fidélité durant l'année 1994 pour les âges de 40 à 50 ans dans la première colonne (colonne

grisée en foncé sur l'extrait) pour voir que pour cette année les taux de fidélité entre les personnes de 40 et 50 ans diffèrent en valeur absolue de 14% (62% à 50 ans et 48% à 41 ans).

S'agissant des cohortes, la lecture est un peu plus inhabituelle. Si l'on considère la case concernant les personnes de 40 ans en 1996 (dernière ligne, troisième colonne), on peut déduire rapidement que ces personnes sont de la cohorte 1956 (en calculant $1996-40$). Si l'on cherche à suivre l'évolution de cette cohorte dans le temps, il suffit de prendre en compte le fait que l'année suivante, en 1997, ces personnes auront 41 ans. Sur le diagramme, on s'est donc déplacé à la fois d'une case vers la droite et vers le haut, pour progresser « en diagonale ». En suivant cette diagonale jusqu'aux personnes ayant 46 ans en 2002, on peut suivre cette cohorte 1956 sur toute notre fenêtre d'observation (diagonale grisée en clair sur l'extrait) pour voir qu'au sein de cette cohorte, le taux de fidélité a progressé de 9% en valeur absolue durant les neuf années d'observation (de 45% à 54%).

Le nombre de cohortes (entre 1914 et 1984) est donc égal au nombre de diagonales d'une telle table à savoir 71 ($11+61-1$). Ainsi, on obtient au maximum 11 observations (nombre maximum de périodes) et ce pour les cohortes allant de 1924 ($1914 + 11-1$) à 1974 ($1984 - 11 + 1$) alors que les autres cohortes, à savoir les dix plus anciennes et les dix plus récentes, ne sont observées que sur certaines périodes. Ainsi pour les cohortes nées de 1914 à 1924, le nombre d'observations s'incrémente de 1 à 11 observations, et il en est de même en redescendant pour les cohortes nées de l'année 1974 à l'année 1984. Si le nombre d'observations est en-dessous de cinq observations au sein d'une cohorte, les résultats sont à manipuler avec précaution. En effet, ces cohortes ne sont observées que pour un nombre restreint d'âges différents ce qui les expose à un risque de biais. Par la suite, les résultats pour les cohortes allant de 1914 à 1918 et de 1980 à 1984 seront inclus dans les analyses mais pour éviter tout risque de biais, ils seront à interpréter avec précaution, et nous nous concentrons donc sur une interprétation sur la période allant des cohortes 1919 à 1979.

D.2.ii. Description des effets bruts

Cette base de données est explorée dans une première approche en analysant les effets bruts de chacune des trois variables APC.

Représentation graphique des effets bruts

Pour donner une première description des données, les taux de fidélité par âge, période et cohorte sont représentés dans les figures D.2.1 à D.2.3 ci-dessous.

Comme définis dans la sous-section D.1.iv., les effets de chacune de ces trois variables sur la fidélité sont ici dits « bruts » car il n'est pas tenu compte de l'effet simultané des deux autres variables (e.g. P et C) lorsque l'on observe l'impact d'une des variables (e.g. A) sur le taux de fidélité Y. Cette absence de contrôle biaise les effets observés tels quels et tout le problème sera donc d'attribuer les effets respectifs à chacune des variables.

Une autre précaution importante, et rarement évoquée dans les études sur ce sujet, est l'interprétation de la moyenne sur les graphiques de représentation des effets bruts. En effet, il s'agit de la valeur moyenne de la fidélité, à travers les âges, périodes et cohortes, sans tenir compte des pondérations des effectifs suivant chacune des observations, c'est-à-dire des triplets (âge, période, cohorte). Cette moyenne constitue pourtant le meilleur référent dans le cadre de l'analyse APC car elle indique la moyenne des effets et permet donc de comparer l'effet d'une variable par rapport à l'effet moyen de toutes les variables. Si on veut obtenir une estimation de la moyenne générale sur une population, il faudra utiliser le modèle prédictif pour obtenir une prédiction du taux de fidélité par catégorie APC et réaliser ensuite une moyenne pondérée de ces prévisions suivant la taille de chaque catégorie.

Notons enfin que dans les graphiques suivants, les courbes en gras indiquent l'effet brut des modalités de la variable considérée (âge, période ou cohorte), et les segments verticaux indiquent l'intervalle de confiance à 95% pour chaque cas (les bornes de l'intervalle de confiance de cette proportion de ré-achat p ayant été calculées à partir de l'estimation de la variance empirique de la proportion par $p(1-p)/(\text{nombre d'observations})$). Cet intervalle est lui aussi proposé à titre indicatif car l'estimation des effets d'une variable est ici réalisé sans le nécessaire contrôle par les deux autres variables.

La ligne horizontale en pointillés indique le taux de fidélité moyen.

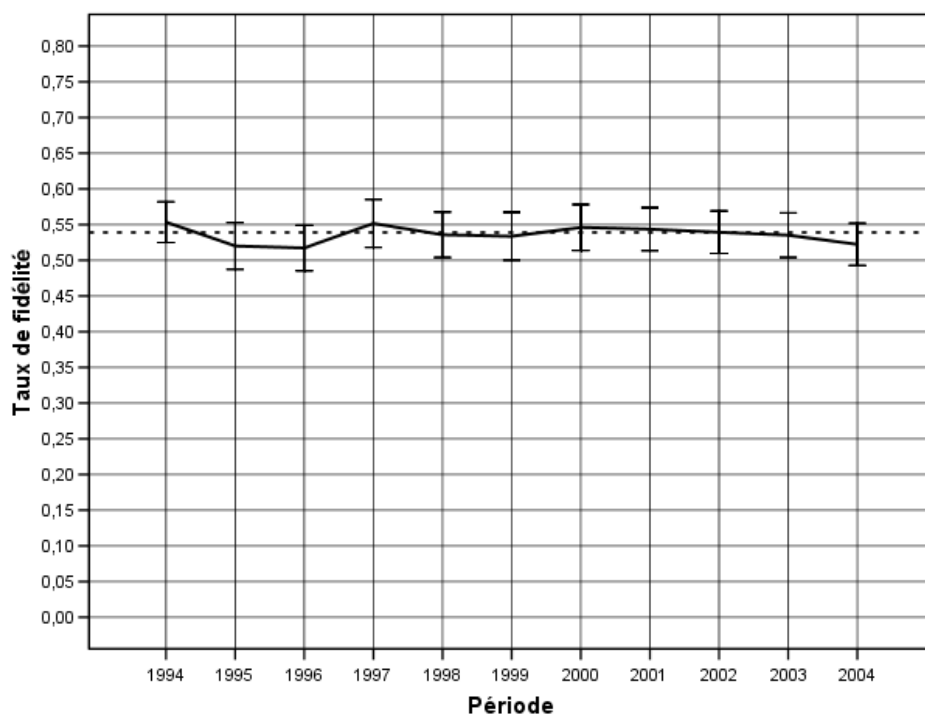


Figure D.2.1. Effets bruts de la période sur le taux de fidélité²⁹

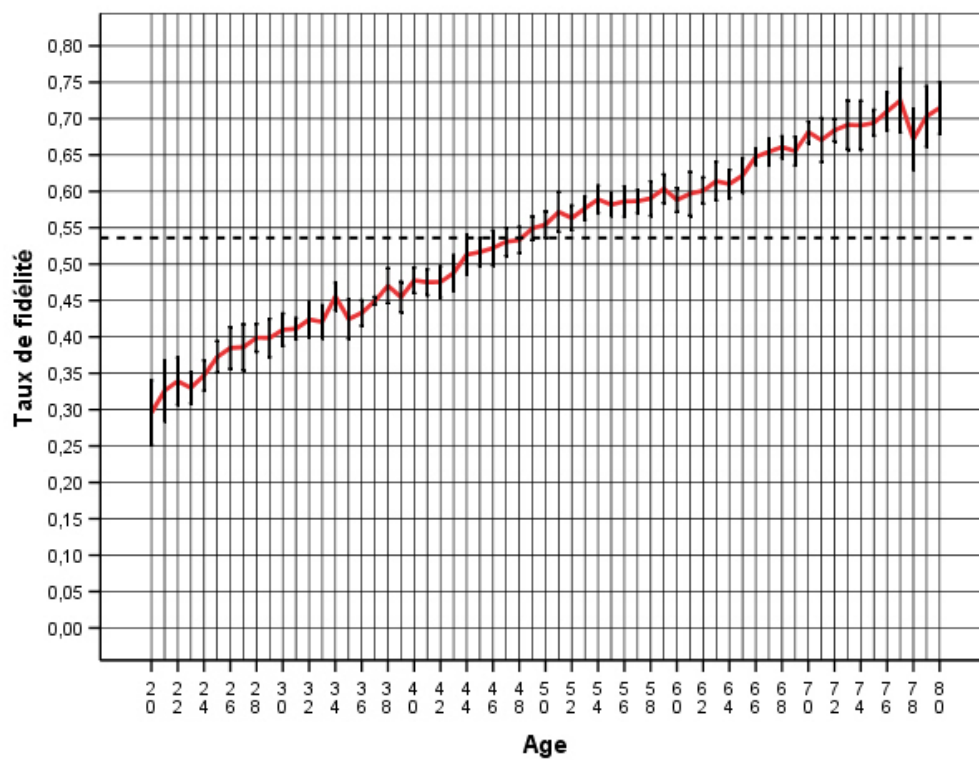


Figure D.2.2. Effets bruts de l'âge sur le taux de fidélité

²⁹ L'échelle a été ici choisie pour être identique suivant les trois variables APC.

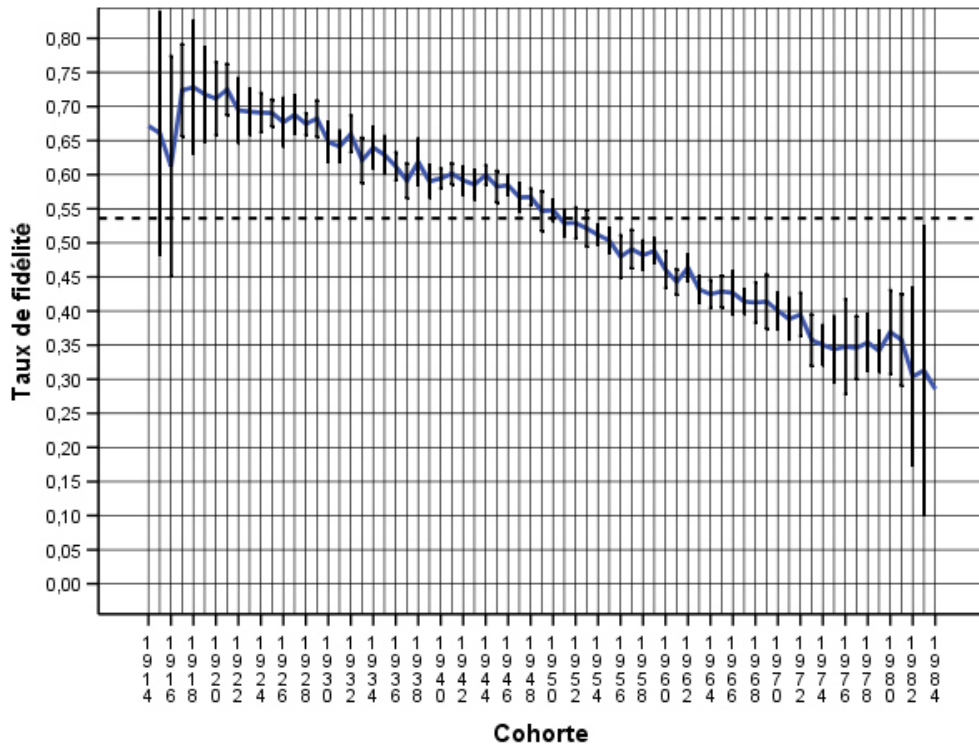


Figure D.2.3. Effets bruts de la cohorte sur le taux de fidélité

A priori l'âge est la variable suivant laquelle on observe les plus grandes variations, sans pouvoir dire pour autant a priori si cette variable aura le plus grand pouvoir explicatif une fois ses effets corrigés des effets de période et de cohorte. Dans cette première approche, sans contrôle par les autres variables (période et cohorte), on observe une variation de 30% à 72 % du taux de fidélité en fonction de l'âge ce qui donne un effet brut global d'environ 42% pour l'âge. Regardons à présent les effets de la cohorte et de la période.

S'agissant de la cohorte, l'effet brut est lui aussi apparemment important, car même en tenant compte des fourchettes basses de ces premiers intervalles de confiance, qui sont plus grands que pour les âges, et en laissant de côté les années à moins de 5 observations, on observe une variation du taux, décroissant cette fois, allant de 65% à 38% de ré-achat, ce qui fournit un effet brut d'au moins 27%. Les effets d'âge et de cohorte vont dans le même sens : les populations les plus âgées, qui sont aussi celles appartenant aux plus anciennes cohortes, sont les plus fidèles.

Les effets bruts de la période varient quant à eux très peu, entre 52% et 55%. Les premiers intervalles de confiance indiquent que ces effets ne sont pas significatifs mais il faut prendre

garde au fait qu'ils sont calculés sur la base d'une seule variable, alors qu'il faudrait avoir contrôlé les effets des autres variables préalablement au calcul des intervalles de confiance des effets considérés.

D.2.iii. Description de combinaisons d'effets APC

Examinons à présent comment l'on peut représenter simultanément deux des effets bruts APC sur un même graphique.

Il serait possible de placer en abscisse les cohortes et d'avoir pour chaque âge une courbe séparée. Si l'âge a un effet constant quelle que soit la cohorte (ligne horizontale), alors l'effet observé est dû à l'âge (et visible par l'écart vertical entre ces lignes horizontales). Cependant, étant donné les 61 âges différents, il est difficilement envisageable de tous les représenter simultanément.

Nous avons regroupé les âges par tranches d'âge de dix ans, rassemblant les âges de 20 à 30 ans, puis de 31 à 40 ans,... jusqu'à la tranche de 71 à 80 ans³⁰. Les six courbes représentant l'évolution du taux de fidélité par cohorte en regroupant les observations par tranches d'âge sont exposées dans la figure D.2.4. ci-après.

³⁰ La première classe regroupe 11 âges (de 20 à 30 ans) alors que les autres en contiennent 10 (de 31 à 41 ans...) mais ce léger déséquilibre est inévitable sachant que le nombre d'âges n'est pas divisible par 10.



Figure D.2.4. Effets combinés de l'âge et de la cohorte sur le taux de fidélité

La tendance générale est décroissante, car la fidélité décroît avec la cohorte et croît avec l'âge. Cependant, notons que les courbes des tranches d'âges sont définies sur des intervalles de cohortes qui se chevauchent, ce qui permet d'observer, au sein d'une même cohorte (sur une même ligne verticale), l'effet de l'âge séparément de l'effet de cohorte (car celle-ci est identique sur un axe vertical). Par une lecture du graphique, on peut en conclure que cet effet d'âge est bien présent parce que les courbes sont bien les unes en dessous des autres quand on passe d'une tranche d'âge à la suivante (sachant que de gauche à droite, les tranches d'âge regroupent des âges de moins en moins élevés).

On peut illustrer ce phénomène sur les deux dernières tranches d'âge : de 20 à 30 ans (dernière courbe) et de 31 à 40 ans (avant-dernière courbe). Les lignes horizontales ajoutées sur le graphe indiquent les niveaux médians de chaque courbe sur leur zone de chevauchement (de 1964 à 1973). On en déduit bien que les individus observés quand ils sont plus âgés sont plus fidèles que *les individus appartenant aux mêmes cohortes, observés quand*

ils sont moins âgés. Autrement dit, des individus appartenant à la même cohorte sont plus fidèles quand on les observe à un âge plus avancé.

De même l'on peut aussi observer l'effet de cohorte car les courbes par tranche d'âges ne sont pas horizontales : si elles l'étaient, alors cela signifierait que le taux de fidélité est constant par rapport aux cohortes (placées en abscisse) au sein d'une même tranche d'âge. Dans ce cas, il n'y aurait pas d'effet de cohorte mais simplement des effets d'âge car la décroissance serait due simplement au fait de passer d'un segment à l'autre, c'est-à-dire d'une tranche d'âges à l'autre. Ce raisonnement pourrait être mis en défaut par le fait que chaque courbe ne représente pas un et un seul âge, car après regroupement, une partie de la variation à l'intérieur de chaque groupe pourrait être due à de simples variations d'âge. Cependant, pour chaque cohorte, le point d'une courbe est une *moyenne sur tous les âges de la tranche d'âges que représente cette portion de courbe* donc la décroissance partielle au sein d'une courbe ne peut être attribuée à l'âge. Et ceci est retranscrit par le graphique. En effet, si l'on observe la courbe de la tranche d'âge 20-30 ans, elle a une amplitude intra-groupe de 18% (entre 28 et 46%) et celle de la tranche d'âge 31-40 ans en a une de 11% (entre 51% et 40%). Or si l'on se réfère à nouveau à la figure 2 on peut voir que les variations dues à l'âge sur ces tranches d'âges sont respectivement de 10% (de 30% à 40%) et 6% de (41% à 47%) au plus. La décroissance de chacune de ces courbes est donc bien due aux effets de cohorte, qui ont eux aussi une composante propre, une fois l'effet d'âge défalqué de leur effet brut.

De la même manière, nous aurions pu placer en abscisse les âges et observer des groupes de cohortes. Par souci de clarté, nous avons choisi la disposition de la figure D.2.4. car il y a moins d'effets d'âge que d'effets de cohortes, donc (en maintenant des tranches de 10 ans) moins de courbes à afficher sur une même figure en plaçant les cohortes en abscisse.

Une analyse graphique plus approfondie permet donc, à défaut de les mesurer précisément, de mettre en évidence la présence des effets propres de l'âge *et* de la cohorte. La difficulté que nous rencontrons pour interpréter ces graphiques plus avant, par rapport à un jeu de variables plus traditionnel, provient du fait que les variables sont « imbriquées », car liées par une relation de dépendance linéaire, l'âge étant obtenu en soustrayant l'année de naissance (cohorte) de l'année d'observation (période).

Nous présentons dans les sections D.3. et D.4. plusieurs méthodes visant à séparer ces effets imbriqués dans le but d'avoir une mesure des effets nets respectifs de chacune des variables d'âge, de période et de cohorte, mais avant cela justifions la nécessité de passer des effets bruts aux effets nets dans une analyse APC.

D.2.iv. Critique de l'analyse des effets bruts

Envisageons le cas où l'analyse des données ne se ferait qu'en considérant les variables usuelles d'âge et de période. Dans ce cas, l'analyse préliminaire des graphes suggèrerait un effet principal de l'âge d'une amplitude d'environ 40%, et un effet quasi-nul de la période, mais cette analyse est faussée par l'omission des effets de cohortes repérés dans la sous-section D.2.ii. ci-avant.

L'omission de la prise en compte des effets de cohorte est fréquente, même lorsque des données longitudinales sont disponibles, car il est habituel d'attribuer l'effet du temps au vieillissement et à la période, sans tenir compte de la dynamique sous-jacente des cohortes. Théoriquement, ceci revient à privilégier le point de vue conjoncturel par rapport au point de vue structurel. En effet, l'avancement de la période et le vieillissement sont des événements contemporains de l'observation: ils se produisent au moment où l'on observe le phénomène étudié. Par contre, les cohortes capitalisent une somme d'événements passés, qui ont conditionné partiellement les attitudes et les perceptions des consommateurs et n'auront donc un impact qu'en différé sur des comportements à venir : la structure du marché auquel les individus sont exposés dans leur jeunesse explique en partie la manière dont ces individus perçoivent encore ce marché, et ce tout au long de leur vie.

Cependant, si l'on n'omet pas d'analyser le graphique des cohortes, on détecte un effet concomitant de l'âge et de la cohorte (Cf. figure D.2.4.). Si l'on souhaite mesurer l'effet d'âge net, par opposition à l'effet brut décrit dans les graphiques précédents, il faut déduire de cet effet d'âge l'effet de cohorte (ainsi que l'effet de période, moins souvent oublié).

Considérons l'exemple simplifié ci-dessous que nous ré-utiliserons par la suite³¹. Imaginons un cas théorique simplifié avec deux périodes, les années 3 et 4, deux âges 1 et 2 ans, et donc

³¹ Dans cet exemple, seuls les taux de fidélité sont manipulés et les effectifs par classe n'ont pas d'importance.

trois cohortes (2+2-1), à savoir les cohortes 1, 2 et 3. Nous étudions un taux quelconque, avec les observations suivantes :

		Période	
		P ₃	P ₄
Age	A ₁	C ₂ Y=10%	C ₃ Y=30%
	A ₂	C ₁ Y=20%	C ₂ Y=40%

Tableau D.2.3. Exemple simplifié d'effets APC

En calculant les effets bruts moyens notés liés à A₁ et A₂, et P₃ et P₄, on obtient les résultats de la première colonne du tableau ci-après.

Effets bruts		Hypothèse sur les effets nets excluant les cohortes		Hypothèse sur les effets nets incluant les cohortes	
Indicatrice	Effet	Indicatrice	Effet	Indicatrice	Effet
A ₁	20%	A ₁	10%	A ₁	0%
A ₂	30%	A ₂	20%	A ₂	0%
P ₃	15%	P ₃	0%	P ₃	0%
P ₄	35%	P ₄	20%	P ₄	30%
				C ₁	20%
				C ₂	10%
				C ₃	0%

Tableau D.2.4. Interprétations de l'exemple simplifié

Bien entendu, ces effets bruts ne permettent pas de retrouver les données du tableau D.2.3. En prenant par exemple la catégorie des individus ayant 1 an durant l'année 3, si l'on somme les effets bruts de l'âge 1 (20%) et ceux de la période 3 (15%), on ne retrouve pas le taux observé (mais 35% au lieu de Y=10%).

Cherchons à retrouver les chiffres du tableau D.2.3. en partant sur la base des effets bruts liés de l'âge et de la période uniquement. Pour cela, fixons pour commencer l'effet le plus faible à zéro par commodité, à savoir celui de la période 3 (initialement de 15%). Conserver le même écart entre les deux effets de période nous conduit à fixer l'effet net de P_4 à 20% (35%-15%). On en déduit, pour retrouver les chiffres du tableau D.2.3, les effets des 1 et 2 respectivement à 10 et 20%, et l'on obtient la solution indiquée dans la deuxième colonne du tableau D.2.4. A partir de cette solution, on retrouve bien les valeurs de Y du tableau D.2.3.

Mais cette hypothèse sur les effets nets est en réalité basée sur l'hypothèse implicite qu'il n'y a pas d'effets de cohorte. On pourrait en effet faire des hypothèses très différentes sur la structure des effets APC, comme proposé dans la troisième colonne du tableau D.2.4., dans laquelle les phénomènes observés sont attribués exclusivement à la cohorte et nullement à l'âge. Nous verrons plus loin comment passer de la deuxième à la troisième colonne.

Dans ce cas théorique, impossible de trancher entre ces diverses hypothèses potentielles. Cela illustre comment, sans complément d'information, les phénomènes observés d'âge, pourraient tout aussi bien être expliqués par une combinaison d'effets de cohortes et de périodes.

Ce cas révèle comment l'intégration des effets de cohortes peut remettre en cause l'analyse des phénomènes par les seules variables d'âge et de période puisque techniquement, il est tout à fait possible de se passer des effets d'âge pour tout aussi bien décrire ces phénomènes. Sans atteindre ce cas extrême qui consiste à annuler les effets d'âge, nous verrons par la suite que ces effets seront potentiellement modifiés par l'introduction des effets de cohorte.

Dans les deux sections suivantes, nous passons à présent aux méthodes de modélisation par elles-mêmes, que nous illustrerons à l'aide de la base de données France.

D.3. Méthodes de modélisation APC fondées sur les moindres carrés

D.3.i. Omission de variable

Une méthode immédiate, mais assez simpliste, pour résoudre le problème de la multicollinéarité parfaite consiste à éliminer une des trois variables considérées et d'estimer les paramètres pour les variables restantes suivant la méthode des moindres carrés. Cette méthode est appelée Omitted Variable (OV)³² Estimation par Maddala (2001, p. 287). Pour obtenir un panorama complet de cette méthode sur nos données, l'effet de chaque variable APC a été estimé seul.

Par exemple pour l'âge, nous estimons d'abord les effets a_i par le modèle suivant:

$$Y = b_0 + \sum_{i=21...80} a_i A_i + \varepsilon \quad [6]$$

Nous rappelons que dans cette recherche, Y représente le taux de fidélité et $A_i, i=1...I$ les indicatrices des différentes modalités de la variable d'âge A . La constante s'interprète alors comme la fidélité moyenne, une fois contrôlés les effets d'âge (et eux seuls). C'est par rapport à cette constante que sont modulées les prévisions à l'aide des modalités de la variable Age.

Puis l'effet de chaque variable APC a été estimé associé à l'effet de chacun des deux autres prédicteurs. Pour l'âge toujours, on estimera les effets a_i associés aux modèles suivants :

$$Y = b_0 + \sum_{i=21...80} a_i A_i + \sum_{j=1995...2004} p_j P_j + \varepsilon \quad [7]$$

$$Y = b_0 + \sum_{i=21...80} a_i A_i + \sum_{k=1915...1984} c_k C_k + \varepsilon \quad [8]$$

Les résultats de ces trois séries de régressions sont présentés dans les tableaux D.3.1, D.3.2. et D.3.3. ci-après. Les coefficients fournis sont les a_i modalités A_i de la variable Age dans le tableau D.3.1., p_j modalités P_j de la Période dans D.3.2., et c_k modalités C_k de la Cohorte dans D.3.3.

³² Traduit ici par Omission de variable.

Indicatrices	Modèle avec Age				Modèle avec Age et Période ³³				Modèle avec Age et Cohorte			
	<i>a_i</i>	<i>ET</i>	<i>T</i>	<i>Sig.</i>	<i>a_i</i>	<i>ET</i>	<i>t</i>	<i>Sig.</i>	<i>a_i</i>	<i>ET</i>	<i>t</i>	<i>Sig.</i>
(Constante)	0,296	0,011	26,831	< 0,01	0,313	0,011	27,717	< 0,01	0,273	0,047	5,826	< 0,01
Age = 21	0,031	0,016	1,961	0,05	0,031	0,015	2,065	0,04	0,032	0,016	2,041	0,04
Age = 22	0,043	0,016	2,782	0,01	0,043	0,015	2,930	< 0,01	0,045	0,016	2,801	0,01
Age = 23	0,034	0,016	2,201	0,03	0,034	0,015	2,318	0,02	0,035	0,016	2,170	0,03
Age = 24	0,051	0,016	3,290	< 0,01	0,051	0,015	3,465	< 0,01	0,056	0,016	3,433	< 0,01
Age = 25	0,077	0,016	4,918	< 0,01	0,077	0,015	5,180	< 0,01	0,086	0,017	5,160	< 0,01
Age = 26	0,089	0,016	5,711	< 0,01	0,089	0,015	6,015	< 0,01	0,100	0,017	5,904	< 0,01
Age = 27	0,090	0,016	5,784	< 0,01	0,090	0,015	6,092	< 0,01	0,103	0,017	6,010	< 0,01
Age = 28	0,103	0,016	6,616	< 0,01	0,103	0,015	6,968	< 0,01	0,117	0,018	6,681	< 0,01
Age = 29	0,103	0,016	6,606	< 0,01	0,103	0,015	6,957	< 0,01	0,117	0,018	6,569	< 0,01
Age = 30	0,114	0,016	7,318	< 0,01	0,114	0,015	7,708	< 0,01	0,129	0,018	7,082	< 0,01
Age = 31	0,115	0,016	7,406	< 0,01	0,115	0,015	7,800	< 0,01	0,130	0,019	7,037	< 0,01
Age = 32	0,128	0,016	8,215	< 0,01	0,128	0,015	8,652	< 0,01	0,140	0,019	7,456	< 0,01
Age = 33	0,125	0,016	8,027	< 0,01	0,125	0,015	8,454	< 0,01	0,140	0,019	7,303	< 0,01
Age = 34	0,160	0,016	10,260	< 0,01	0,160	0,015	10,806	< 0,01	0,174	0,019	8,973	< 0,01
Age = 35	0,129	0,016	8,264	< 0,01	0,129	0,015	8,704	< 0,01	0,141	0,020	7,138	< 0,01
Age = 36	0,137	0,016	8,799	< 0,01	0,137	0,015	9,268	< 0,01	0,148	0,020	7,417	< 0,01
Age = 37	0,154	0,016	9,861	< 0,01	0,154	0,015	10,386	< 0,01	0,163	0,020	8,052	< 0,01
Age = 38	0,174	0,016	11,189	< 0,01	0,174	0,015	11,785	< 0,01	0,184	0,021	8,926	< 0,01
Age = 39	0,159	0,016	10,210	< 0,01	0,159	0,015	10,753	< 0,01	0,167	0,021	7,992	< 0,01
Age = 40	0,182	0,016	11,669	< 0,01	0,182	0,015	12,290	< 0,01	0,187	0,021	8,866	< 0,01
Age = 41	0,179	0,016	11,514	< 0,01	0,179	0,015	12,127	< 0,01	0,182	0,021	8,482	< 0,01
Age = 42	0,180	0,016	11,521	< 0,01	0,180	0,015	12,134	< 0,01	0,178	0,022	8,222	< 0,01
Age = 43	0,192	0,016	12,328	< 0,01	0,192	0,015	12,984	< 0,01	0,190	0,022	8,679	< 0,01
Age = 44	0,217	0,016	13,951	< 0,01	0,217	0,015	14,694	< 0,01	0,212	0,022	9,544	< 0,01
Age = 45	0,221	0,016	14,169	< 0,01	0,221	0,015	14,923	< 0,01	0,213	0,022	9,479	< 0,01
Age = 46	0,226	0,016	14,529	< 0,01	0,226	0,015	15,303	< 0,01	0,217	0,023	9,543	< 0,01
Age = 47	0,235	0,016	15,087	< 0,01	0,235	0,015	15,890	< 0,01	0,223	0,023	9,705	< 0,01
Age = 48	0,237	0,016	15,235	< 0,01	0,237	0,015	16,046	< 0,01	0,222	0,023	9,540	< 0,01
Age = 49	0,254	0,016	16,263	< 0,01	0,254	0,015	17,128	< 0,01	0,233	0,023	9,925	< 0,01
Age = 50	0,259	0,016	16,604	< 0,01	0,259	0,015	17,488	< 0,01	0,234	0,024	9,870	< 0,01
Age = 51	0,276	0,016	17,698	< 0,01	0,276	0,015	18,640	< 0,01	0,249	0,024	10,378	< 0,01
Age = 52	0,268	0,016	17,171	< 0,01	0,268	0,015	18,085	< 0,01	0,238	0,024	9,834	< 0,01
Age = 53	0,281	0,016	18,030	< 0,01	0,281	0,015	18,990	< 0,01	0,249	0,024	10,176	< 0,01
Age = 54	0,293	0,016	18,826	< 0,01	0,293	0,015	19,829	< 0,01	0,259	0,025	10,485	< 0,01
Age = 55	0,286	0,016	18,341	< 0,01	0,286	0,015	19,318	< 0,01	0,251	0,025	10,062	< 0,01
Age = 56	0,290	0,016	18,635	< 0,01	0,290	0,015	19,627	< 0,01	0,252	0,025	10,036	< 0,01
Age = 57	0,291	0,016	18,658	< 0,01	0,291	0,015	19,652	< 0,01	0,254	0,025	10,027	< 0,01
Age = 58	0,295	0,016	18,896	< 0,01	0,295	0,015	19,902	< 0,01	0,258	0,026	10,088	< 0,01
Age = 59	0,308	0,016	19,760	< 0,01	0,308	0,015	20,812	< 0,01	0,272	0,026	10,539	< 0,01
Age = 60	0,293	0,016	18,768	< 0,01	0,293	0,015	19,767	< 0,01	0,256	0,026	9,839	< 0,01
Age = 61	0,301	0,016	19,310	< 0,01	0,301	0,015	20,338	< 0,01	0,268	0,026	10,207	< 0,01
Age = 62	0,305	0,016	19,583	< 0,01	0,305	0,015	20,626	< 0,01	0,271	0,026	10,247	< 0,01
Age = 63	0,318	0,016	20,419	< 0,01	0,318	0,015	21,506	< 0,01	0,286	0,027	10,721	< 0,01
Age = 64	0,315	0,016	20,179	< 0,01	0,315	0,015	21,253	< 0,01	0,285	0,027	10,583	< 0,01

³³ La proximité de la valeur de ces effets avec ceux du premier modèle de ce tableau peut s'expliquer par le fait qu'aucun des effets de période n'est significatif en lui-même, comme l'indique la première partie du tableau suivant numéroté D.3.2. (Tous les calculs effectués sous SPSS V7 et vérifiés sous XLStat 2008 V1).

Indicatrices	Modèle avec Période				Modèle avec Période + Age				Modèle Cohorte et Age			
	a_i	ET	T	Sig.	a_i	ET	t	Sig.	a_i	ET	t	Sig.
Age = 65	0,326	0,016	20,925	< 0,01	0,326	0,015	22,038	< 0,01	0,295	0,027	10,889	< 0,01
Age = 66	0,352	0,016	22,582	< 0,01	0,352	0,015	23,784	< 0,01	0,321	0,027	11,736	< 0,01
Age = 67	0,359	0,016	23,008	< 0,01	0,359	0,015	24,233	< 0,01	0,328	0,028	11,922	< 0,01
Age = 68	0,365	0,016	23,426	< 0,01	0,365	0,015	24,673	< 0,01	0,334	0,028	12,031	< 0,01
Age = 69	0,360	0,016	23,067	< 0,01	0,360	0,015	24,295	< 0,01	0,328	0,028	11,721	< 0,01
Age = 70	0,386	0,016	24,741	< 0,01	0,386	0,015	26,058	< 0,01	0,355	0,028	12,594	< 0,01
Age = 71	0,375	0,016	24,051	< 0,01	0,375	0,015	25,331	< 0,01	0,345	0,028	12,166	< 0,01
Age = 72	0,388	0,016	24,914	< 0,01	0,388	0,015	26,241	< 0,01	0,358	0,029	12,509	< 0,01
Age = 73	0,396	0,016	25,391	< 0,01	0,396	0,015	26,742	< 0,01	0,364	0,029	12,642	< 0,01
Age = 74	0,395	0,016	25,343	< 0,01	0,395	0,015	26,692	< 0,01	0,361	0,029	12,453	< 0,01
Age = 75	0,398	0,016	25,562	< 0,01	0,398	0,015	26,923	< 0,01	0,362	0,029	12,398	< 0,01
Age = 76	0,414	0,016	26,569	< 0,01	0,414	0,015	27,983	< 0,01	0,377	0,029	12,819	< 0,01
Age = 77	0,429	0,016	27,535	< 0,01	0,429	0,015	29,001	< 0,01	0,390	0,030	13,181	< 0,01
Age = 78	0,376	0,016	24,141	< 0,01	0,376	0,015	25,427	< 0,01	0,346	0,030	11,603	< 0,01
Age = 79	0,407	0,016	26,134	< 0,01	0,407	0,015	27,525	< 0,01	0,382	0,030	12,689	< 0,01
Age = 80	0,419	0,016	26,880	< 0,01	0,419	0,015	28,311	< 0,01	0,398	0,030	13,096	< 0,01
Qualité	R² ajusté = 91,04				R² ajusté = 91,92				R² ajusté = 91,47			

Tableau D.3.1. Régressions pour la mesure des effets d'âge avec omission de variable

Indicatrices	Modèle avec Période				Modèle avec Période + Age				Période + Cohorte			
	p_i	ET	T	Sig.	p_i	ET	t	Sig.	p_i	ET	t	Sig.
(Constante)	0,553	0,016	35,288	< 0,01	0,313	0,011	27,717	< 0,01	0,672	0,034	19,479	< 0,01
Période = 1995	-0,033	0,022	-1,504	0,13	-0,033	0,016	-5,308	< 0,01	-0,028	0,006	-4,424	< 0,01
Période = 1996	-0,036	0,022	-1,630	0,10	-0,036	0,006	-5,753	< 0,01	-0,025	0,006	-3,954	< 0,01
Période = 1997	-0,002	0,022	-0,081	0,94	-0,002	0,006	-0,284	0,78	0,015	0,006	2,311	0,02
Période = 1998	-0,018	0,022	-0,796	0,43	-0,018	0,006	-2,810	0,01	0,005	0,006	0,846	0,40
Période = 1999	-0,020	0,022	-0,888	0,37	-0,020	0,006	-3,133	0,00	0,010	0,006	1,621	0,11
Période = 2000	-0,007	0,022	-0,336	0,74	-0,007	0,006	-1,184	0,24	0,029	0,006	4,552	< 0,01
Période = 2001	-0,010	0,022	-0,442	0,66	-0,010	0,006	-1,558	0,12	0,033	0,006	5,175	< 0,01
Période = 2002	-0,014	0,022	-0,634	0,53	-0,014	0,006	-2,236	0,03	0,036	0,006	5,659	< 0,01
Période = 2003	-0,018	0,022	-0,822	0,41	-0,018	0,006	-2,901	< 0,01	0,038	0,006	6,039	< 0,01
Période = 2004	-0,031	0,022	-1,402	0,16	-0,031	0,006	-4,947	< 0,01	0,033	0,006	5,098	< 0,01
Qualité	R² ajusté = -0,58				R² ajusté = 91,92				R² ajusté = 92,03			

Tableau D.3.2. Régressions pour la mesure des effets de période avec omission de variable

Indicatrices	Modèle avec Cohorte				Modèle avec Cohorte et Période				Modèle Cohorte et Age			
	c_k	ET	T	Sig.	c_k	ET	t	Sig.	c_k	ET	t	Sig.
(Constante)	0,672	0,041	16,322	< 0,01	0,672	0,034	19,479	< 0,01	0,273	0,047	5,826	< 0,01
Cohorte=1915	-0,011	0,050	-0,211	0,83	0,013	0,042	0,077	0,94	-0,002	0,044	-0,050	0,96
Cohorte=1916	-0,059	0,048	-1,252	0,21	-0,042	0,040	-1,049	0,29	-0,036	0,042	-0,863	0,39
Cohorte=1917	0,052	0,046	1,134	0,26	0,062	0,039	1,592	0,11	0,072	0,041	1,741	0,08
Cohorte=1918	0,057	0,045	1,257	0,21	0,063	0,038	1,664	0,10	0,076	0,040	1,892	0,06

Indicatrices	c_k	ET	T	Sig.	c_k	ET	t	Sig.	c_k	ET	t	Sig.
Cohorte=1919	0,046	0,044	1,041	0,30	0,050	0,037	1,335	0,18	0,069	0,040	1,723	0,09
Cohorte=1920	0,040	0,044	0,909	0,36	0,039	0,037	1,053	0,29	0,065	0,040	1,631	0,10
Cohorte=1921	0,053	0,044	1,223	0,22	0,048	0,037	1,317	0,19	0,079	0,039	2,011	0,04
Cohorte=1922	0,023	0,043	0,523	0,60	0,014	0,037	0,392	0,70	0,050	0,039	1,280	0,20
Cohorte=1923	0,021	0,043	0,478	0,63	0,009	0,036	0,255	0,80	0,051	0,039	1,297	0,20
Cohorte=1924	0,019	0,043	0,452	0,65	0,006	0,036	0,168	0,87	0,051	0,039	1,300	0,19
Cohorte=1925	0,019	0,043	0,442	0,66	0,006	0,036	0,157	0,88	0,057	0,040	1,438	0,15
Cohorte=1926	0,005	0,043	0,128	0,90	-0,008	0,036	-0,215	0,83	0,048	0,040	1,202	0,23
Cohorte=1927	0,016	0,043	0,380	0,70	0,003	0,036	0,083	0,93	0,060	0,040	1,511	0,13
Cohorte=1928	0,003	0,043	0,068	0,95	-0,010	0,036	-0,287	0,77	0,053	0,040	1,327	0,18
Cohorte=1929	0,011	0,043	0,247	0,80	-0,003	0,036	-0,074	0,94	0,068	0,040	1,701	0,09
Cohorte=1930	-0,023	0,043	-0,544	0,59	-0,037	0,036	-1,012	0,31	0,041	0,040	1,022	0,31
Cohorte=1931	-0,030	0,043	-0,709	0,48	-0,044	0,036	-1,207	0,23	0,041	0,040	1,012	0,31
Cohorte=1932	-0,012	0,043	-0,276	0,78	-0,025	0,036	-0,693	0,49	0,068	0,040	1,677	0,09
Cohorte=1933	-0,051	0,043	-1,175	0,24	-0,064	0,036	-1,760	0,08	0,037	0,041	0,919	0,36
Cohorte=1934	-0,032	0,043	-0,738	0,46	-0,045	0,036	-1,242	0,21	0,064	0,041	1,575	0,12
Cohorte=1935	-0,043	0,043	-0,989	0,32	-0,056	0,036	-1,539	0,12	0,061	0,041	1,490	0,14
Cohorte=1936	-0,060	0,043	-1,387	0,17	-0,073	0,036	-2,010	0,04	0,050	0,041	1,223	0,22
Cohorte=1937	-0,080	0,043	-1,869	0,06	-0,094	0,036	-2,582	0,01	0,037	0,041	0,891	0,37
Cohorte=1938	-0,053	0,043	-1,237	0,22	-0,066	0,036	-1,832	0,07	0,071	0,041	1,713	0,09
Cohorte=1939	-0,081	0,043	-1,895	0,06	-0,095	0,036	-2,612	0,01	0,049	0,041	1,180	0,24
Cohorte=1940	-0,077	0,043	-1,796	0,07	-0,091	0,036	-2,495	0,01	0,056	0,042	1,357	0,18
Cohorte=1941	-0,070	0,043	-1,632	0,10	-0,083	0,036	-2,301	0,02	0,067	0,042	1,601	0,11
Cohorte=1942	-0,080	0,043	-1,860	0,06	-0,093	0,036	-2,571	0,01	0,061	0,042	1,466	0,14
Cohorte=1943	-0,086	0,043	-2,006	0,05	-0,100	0,036	-2,743	0,01	0,057	0,042	1,362	0,17
Cohorte=1944	-0,072	0,043	-1,673	0,09	-0,085	0,036	-2,349	0,02	0,075	0,042	1,770	0,08
Cohorte=1945	-0,089	0,043	-2,072	0,04	-0,102	0,036	-2,822	0,00	0,060	0,042	1,408	0,16
Cohorte=1946	-0,087	0,043	-2,030	0,04	-0,101	0,036	-2,772	0,01	0,066	0,042	1,555	0,12
Cohorte=1947	-0,105	0,043	-2,441	0,01	-0,118	0,036	-3,260	0,00	0,051	0,043	1,210	0,23
Cohorte=1948	-0,104	0,043	-2,426	0,02	-0,118	0,036	-3,242	0,00	0,056	0,043	1,301	0,19
Cohorte=1949	-0,125	0,043	-2,903	< 0,01	-0,138	0,036	-3,807	0,00	0,039	0,043	0,902	0,37
Cohorte=1950	-0,124	0,043	-2,881	< 0,01	-0,137	0,036	-3,780	0,00	0,043	0,043	1,003	0,32
Cohorte=1951	-0,143	0,043	-3,322	< 0,01	-0,156	0,036	-4,303	0,00	0,030	0,043	0,704	0,48
Cohorte=1952	-0,142	0,043	-3,314	< 0,01	-0,156	0,036	-4,294	0,00	0,037	0,043	0,858	0,39
Cohorte=1953	-0,150	0,043	-3,502	< 0,01	-0,164	0,036	-4,516	< 0,01	0,034	0,043	0,787	0,43
Cohorte=1954	-0,160	0,043	-3,725	< 0,01	-0,173	0,036	-4,781	< 0,01	0,030	0,044	0,693	0,49
Cohorte=1955	-0,168	0,043	-3,909	< 0,01	-0,181	0,036	-4,999	< 0,01	0,028	0,044	0,649	0,52
Cohorte=1956	-0,192	0,043	-4,465	< 0,01	-0,205	0,036	-5,658	< 0,01	< 0,019	0,044	0,204	0,84
Cohorte=1957	-0,181	0,043	-4,209	< 0,01	-0,194	0,036	-5,354	< 0,01	0,025	0,044	0,574	0,57
Cohorte=1958	-0,190	0,043	-4,413	< 0,01	-0,203	0,036	-5,597	< 0,01	0,023	0,044	0,527	0,60
Cohorte=1959	-0,183	0,043	-4,264	< 0,01	-0,197	0,036	-5,420	< 0,01	0,037	0,044	0,827	0,41
Cohorte=1960	-0,211	0,043	-4,907	< 0,01	-0,224	0,036	-6,182	< 0,01	0,012	0,044	0,280	0,78
Cohorte=1961	-0,229	0,043	-5,324	< 0,01	-0,242	0,036	-6,676	< 0,01	< 0,011	0,044	0,023	0,98
Cohorte=1962	-0,208	0,043	-4,838	< 0,01	-0,221	0,036	-6,100	< 0,01	0,026	0,045	0,594	0,55
Cohorte=1963	-0,240	0,043	-5,583	< 0,01	-0,253	0,036	-6,983	< 0,01	< 0,011	0,045	-0,027	0,98
Cohorte=1964	-0,247	0,043	-5,743	< 0,01	-0,260	0,036	-7,172	< 0,01	< 0,013	0,045	-0,072	0,94
Cohorte=1965	-0,243	0,043	-5,645	< 0,01	-0,256	0,036	-7,056	< 0,01	< 0,017	0,045	0,163	0,87

Indicatrices	c_k	ET	T	Sig.	c_k	ET	t	Sig.	c_k	ET	t	Sig.
Cohorte=1966	-0,245	0,043	-5,704	< 0,01	-0,258	0,036	-7,126	< 0,01	0,009	0,045	0,207	0,84
Cohorte=1967	-0,258	0,043	-5,992	< 0,01	-0,271	0,036	-7,468	< 0,01	0,004	0,045	0,093	0,93
Cohorte=1968	-0,259	0,043	-6,029	< 0,01	-0,272	0,036	-7,511	< 0,01	0,008	0,045	0,186	0,85
Cohorte=1969	-0,258	0,043	-5,993	< 0,01	-0,271	0,036	-7,468	< 0,01	0,016	0,045	0,345	0,73
Cohorte=1970	-0,271	0,043	-6,310	< 0,01	-0,284	0,036	-7,844	< 0,01	0,010	0,046	0,213	0,83
Cohorte=1971	-0,283	0,043	-6,594	< 0,01	-0,297	0,036	-8,181	< 0,01	0,010	0,046	0,222	0,82
Cohorte=1972	-0,277	0,043	-6,440	< 0,01	-0,290	0,036	-7,998	< 0,01	0,025	0,046	0,555	0,58
Cohorte=1973	-0,314	0,043	-7,310	< 0,01	-0,327	0,036	-9,030	< 0,01	-< 0,012	0,046	-0,046	0,96
Cohorte=1974	-0,321	0,043	-7,476	< 0,01	-0,335	0,036	-9,226	< 0,01	0,003	0,046	0,057	0,95
Cohorte=1975	-0,328	0,043	-7,594	< 0,01	-0,342	0,036	-9,388	< 0,01	0,002	0,046	0,034	0,97
Cohorte=1976	-0,324	0,043	-7,467	< 0,01	-0,343	0,037	-9,364	< 0,01	0,011	0,047	0,232	0,82
Cohorte=1977	-0,326	0,044	-7,459	< 0,01	-0,350	0,037	-9,500	< 0,01	0,016	0,047	0,337	0,74
Cohorte=1978	-0,318	0,044	-7,225	< 0,01	-0,344	0,037	-9,258	< 0,01	0,030	0,047	0,638	0,52
Cohorte=1979	-0,330	0,044	-7,424	< 0,01	-0,360	0,038	-9,579	< 0,01	0,026	0,048	0,550	0,58
Cohorte=1980	-0,302	0,045	-6,707	< 0,01	-0,336	0,038	-8,822	< 0,01	0,063	0,048	1,300	0,19
Cohorte=1981	-0,314	0,046	-6,818	< 0,01	-0,349	0,039	-8,967	< 0,01	0,057	0,049	1,163	0,25
Cohorte=1982	-0,368	0,048	-7,736	< 0,01	-0,403	0,040	-10,042	< 0,01	0,005	0,050	0,108	0,91
Cohorte=1983	-0,359	0,050	-7,118	< 0,01	-0,394	0,043	-9,257	< 0,01	0,024	0,053	0,453	0,65
Cohorte=1984	-0,386	0,058	-6,632	< 0,01	-0,419	0,049	-8,511	< 0,01	0,013	0,059	0,213	0,83
Qualité	R² ajusté = 88,65				R² ajusté = 92,03				R² ajusté = 91,47			

D.3.3. Régressions pour la mesure des effets de cohorte avec omission de variable

Signalons qu'au sein de chaque groupe d'indicatrices, une des variables doit être choisie comme niveau de référence pour éviter la multicollinéarité parfaite entre les indicatrices. Ce choix étant arbitraire et ne modifiant pas les résultats obtenus, nous avons choisi la première indicatrice de chaque groupe (Age=20, Cohorte=1914 et Période=1994) comme niveau de référence par commodité. Il n'y a aucune méthode pour fixer a priori ces indicatrices de référence. Chaque coefficient d'indicatrice s'interprète par rapport à l'indicatrice de référence de la variable correspondante.

Concernant l'âge, que l'on considère l'effet simple ou associé avec l'effet des périodes ou des cohortes, toutes les indicatrices ont des coefficients positifs et significatifs, par rapport à l'âge de référence de 20 ans, et ce dans tous les cas avec des R² ajustés variant de 91,04% à 91,92%. On peut constater que les écart-types des coefficients (coefficients des indicateurs qui constituent notre mesure des effets APC) diminuent légèrement par l'ajout des variables de période et augmentent avec l'ajout des variables de cohortes (à cause de la colinéarité entre l'âge et la cohorte).

Les effets simples des indicatrices de période ne fournissent pas un modèle acceptable (F d'une valeur de 0,616 et un R^2 ajusté négatif) et les effets associés avec la prise en compte de l'âge (resp. des cohortes) ont eux des coefficients significatifs (7 sur 10, et resp. 8 sur 10), mais le signe de ces coefficients change en fonction de la variable ajoutée. En effet, les modalités de la variable âge augmentent dans le même sens que celles de la variable période et dans le sens opposé à celles de la variable cohorte (plus l'âge augmente, plus l'année de naissance s'éloigne).

Le résultat le plus intéressant est sans doute celui concernant les cohortes. En effet, considérées seules, toutes les indicatrices ont des coefficients négatifs par rapport à la cohorte de référence 1914 et pour la plupart significatifs (41 sur 70). L'ajout des indicatrices de période permet d'atteindre le meilleur R^2 ajusté de tous les modèles ici considérés (92,03%) et d'augmenter le nombre de coefficients significatifs (48 au lieu de 41). Par contre, associé à l'effet d'âge, l'effet de cohorte disparaît avec seulement une indicatrice de cohorte significative, et un R^2 ajusté tout aussi élevé d'une valeur de 91,47 %. On peut en déduire ici que la colinéarité entre les variables d'âge et de cohorte dégrade davantage l'estimation des effets de cohorte. Or ces effets étaient significatifs d'après la régression n'intégrant que les indicatrices de cohorte. Une interprétation de la régression intégrant les effets d'âge et de cohorte ne tenant pas compte de la multicollinéarité entre ces variables amènerait à conclure à un simple effet d'âge alors qu'un effet de cohorte pourrait intervenir dans le phénomène étudié.

Ce cas illustre bien le danger de la régression multiple dans les cas de forte colinéarité, mais aussi le risque d'entreprendre une résolution de l'analyse âge-période-cohorte en utilisant la méthode d'omission d'une variable. En effet, à en juger par la non-significativité du modèle n'incluant que les variables de période, la meilleure variable à omettre serait à priori la période. S'ensuit une régression multiple sur les deux variables restantes, à savoir l'âge et la cohorte, qui mènerait à conclure au seul effet de l'âge.

Par ailleurs, nous tenons à signaler que nous sommes conscients, comme le démontrent Sengupta et Bhimasankaram (1997), que la colinéarité peut être due dans certains cas à un petit nombre d'observations perturbant le modèle. Cependant, dans notre cas, le schéma de relation linéaire entre les prédicteurs s'applique identiquement à toutes les observations, et le

problème de la colinéarité entre les indicatrices de cohorte et d'âge dans la régression qui les associe ne pourrait pas être diminué par la suppression de certaines observations.

L'application de cette méthode sur notre base de données et l'examen détaillé des résultats ainsi obtenus, nous amène à conclure sur deux points importants :

- On constate un fort biais potentiel dans l'estimation des effets d'âge lié à l'omission de la variable cohorte, biais traditionnellement considéré comme une erreur de spécification et appelé biais d'omission ;
- La forte colinéarité entre les jeux d'indicatrices d'âge et de cohorte provoque une augmentation de la variance des coefficients des indicatrices non seulement pour l'âge (Tableau D.3.1.), mais surtout pour les coefficients des cohortes (Tableau D.3.3.), qui se trouvent ainsi jugés non significatifs.

En résumé, cette méthode a deux principaux défauts étant d'amener à un biais de spécification (choix de l'omission d'une des variables) et à une forte variance sur les effets mesurés (choix des moindres carrés comme méthode d'estimation). Pour ces deux raisons, cette méthode sera écartée de la suite de cette recherche sur la mesure des effets APC.

D.3.ii. Contraintes d'égalité

Dans la même ligne que la méthode précédente, la plupart des articles de marketing (Rentz, Reynolds et Stout, 1983 ; Rentz et Reynolds, 1991) proposent d'imposer une contrainte plus fine que l'omission d'une variable complète, qui consiste à égaliser les coefficients de deux indicatrices au sein d'un même effet. Dans ce cas aussi, il s'agit de modifier la spécification du modèle initial pour pouvoir utiliser les moindres carrés comme méthode d'estimation.

Par exemple, l'on peut décider a priori que les effets de période de deux périodes successives sont égaux. On passe alors d'un cas de multicollinéarité parfaite à un cas de forte multicollinéarité, ce qui revient à aller de « pis en mal ». En effet, grâce à l'introduction de ce type d'information a priori, le modèle devient estimable par les moindres carrés. Cependant la forte colinéarité restante entre les jeux d'indicatrices restant donne lieu à de très fortes variances sur les effets estimés. De plus, comme nous allons le constater, les résultats obtenus sont très sensibles à la contrainte

choisie, c'est-à-dire à l'information a priori introduite dans le modèle. D'autant que cette forte sensibilité n'est pas simplement constatée sur les coefficients des indicatrices de la variable sur laquelle porte la contrainte, mais sur tous les coefficients des indicatrices des trois variables APC.

Signalons dès à présent un cas particulier de cette méthode de contrainte d'égalité évoquée par Bodin (1999). Elle consiste à regrouper des sous-populations par tranches d'âges et de générations afin de là aussi contourner le problème d'estimation par un modèle linéaire. Cette pratique est équivalente à une combinaison de contraintes d'égalité, car on impose qu'au sein de chacun des sous-groupes ainsi constitués, certains effets d'âge (pour une classe d'âge), de période (pour un groupe de périodes) ou de cohorte (pour un groupe de cohortes) soient fixés égaux entre eux a priori. Il s'agit d'une combinaison de contraintes simples (que nous allons décrire ici) au sein d'une contrainte multiple. En effet au lieu de poser $a_{20} = a_{21}$, par exemple, on regroupera les âges par tranche de cinq ans et on aura pour les premiers âges traités, quatre contraintes d'égalité au lieu d'une à savoir : $a_{20} = a_{21} = a_{22} = a_{23} = a_{24}$. Nous allons voir qu'une contrainte d'égalité simple endommage déjà sévèrement l'estimation des effets. A fortiori, l'usage de contraintes d'égalité multiples ne pourrait que l'endommager un peu plus, et ne sera donc pas traité ici.

Observons ce qui se produit sur notre jeu de données si nous imposons différentes contraintes. Les contraintes les plus plausibles, et qui sont pratiquement les seules utilisées dans la littérature, consistent à fixer l'égalité de deux indicatrices successives. A priori, toutes ces contraintes semblent aussi pertinentes les unes que les autres. En effet, on peut se dire a priori qu'on ne fait pas une grande concession en imposant l'égalité de deux effets correspondants à des coefficients successifs parce que le phénomène étudié varie en pratique assez peu d'une année sur l'autre. Le problème, comme nous allons le voir est qu'une telle contrainte a des répercussions sur l'estimation de tous les effets du modèle et que ces répercussions dépendent fortement de la contrainte imposée.

Nous allons prendre deux égalités qui sont a priori aussi anodines l'une que l'autre :

- une égalité sur les deux premières périodes, à savoir 1994 et 1995
- une égalité sur les deux premières générations, à savoir 1914 et 1915

Ces deux contraintes sont retranscrites dans les deux modèles suivants :

$$Y = b_0 + \sum_{i=21 \dots 80} a_i A_i + \sum_{j=1996 \dots 2004} p_j P_j + \sum_{k=1915 \dots 1984} c_k C_k + \varepsilon \quad [9]$$

$$Y = b_0 + \sum_{i=21 \dots 80} a_i A_i + \sum_{j=1995 \dots 2004} p_j P_j + \sum_{k=1916 \dots 1984} c_k C_k + \varepsilon \quad [10]$$

Les critères d'ajustement des modèles estimés à partir de ces deux modèles sont très satisfaisants avec des R^2 de 94%, des R^2 ajustés de 92,5%. Par contre, les faiblesses de la méthode apparaissent très vite lorsque l'on s'intéresse aux estimations des effets.

La figure D.3.1. représente les coefficients des indicatrices d'âge estimés avec la contrainte d'égalité sur les périodes (trait plein) avec un intervalle de confiance à 95% sur leurs valeurs (traits en pointillés). Ceci nous montre dans le cas de l'égalité sur les périodes que les effets d'âges sont estimés avec des variances croissantes en fonction de l'âge ce qui montre que la qualité d'estimation des effets d'une variable (ici l'âge) peut être réduite par la fixation d'égalités a priori sur une autre variable (ici la période). Ces coefficients ne sont pas biaisés si l'information a priori introduite par cette contrainte est parfaitement correcte, cependant les variances des coefficients estimés n'en demeurent pas moins très élevées, et croissent avec la valeur de la modalité d'âge concernée.

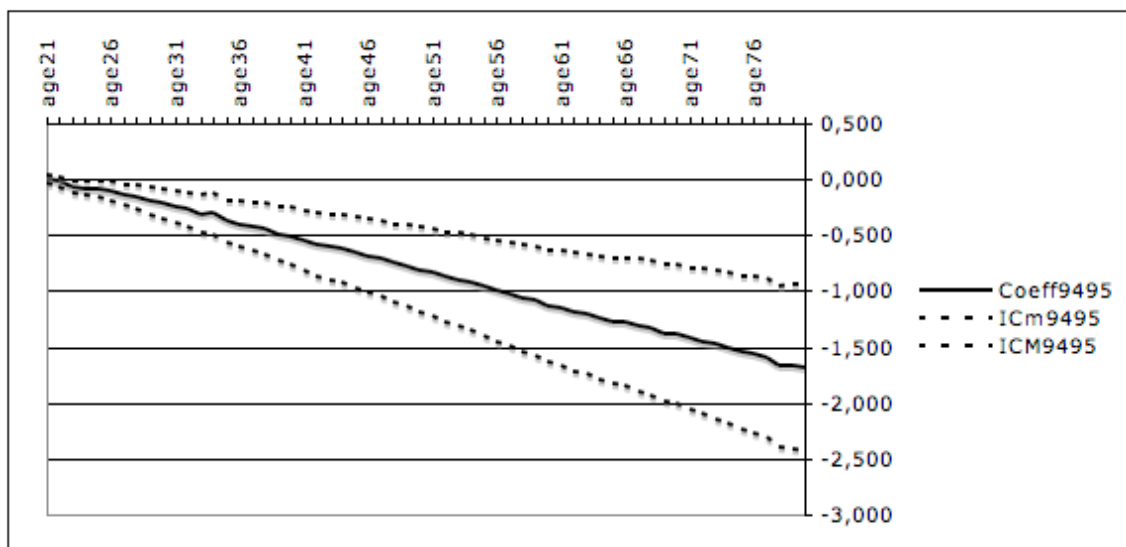


Figure D.3.1. Effets d'âge estimés par une régression multiple (OLS) avec la contrainte d'égalité des effets des périodes 1994 et 1995.

La figure D.3.2. ci-après compare les évaluations des effets d'âge suivant les deux contraintes d'égalité et l'on peut observer que les résultats varient fortement d'une contrainte à l'autre, non seulement en termes de direction mais aussi d'amplitude générale des effets. Là encore, si une des deux contraintes est valide (au sens où elle correspond à une information a priori correspondant à la réalité), alors une des deux solutions est la bonne. Cependant, ces résultats mettent en évidence la très forte sensibilité des résultats à la contrainte appliquée sur les données. Ceci est en partie dû au fait que ces contraintes d'égalité sont en réalité très fortes (égalité stricte), et que le mode d'estimation par les moindres carrés y est très sensible (car on se situe dans un contexte de multicollinéarité très forte entre les jeux d'indicatrices).

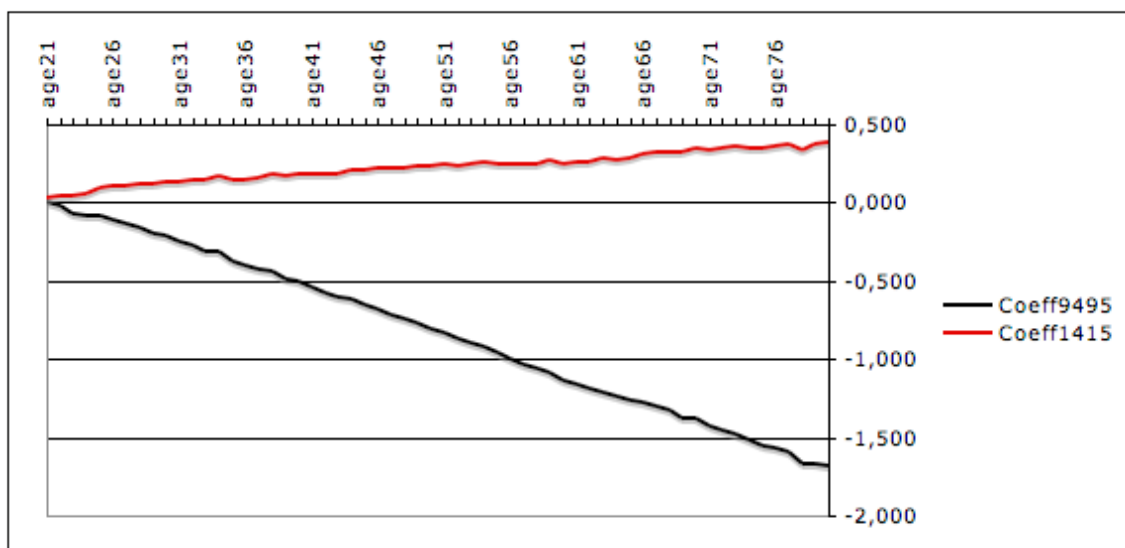


Figure D.3.2. Comparaison des effets d'âge estimés à partir d'une contrainte sur deux périodes successives (1994 et 1995) ou sur deux cohortes successives (1914 et 1915).

Et le diagnostic va même s'aggraver lorsque l'on s'intéresse à l'évaluation des effets de cohorte. En effet, la figure D.3.3. ci-dessous illustre une divergence encore plus grande entre les deux évaluations.

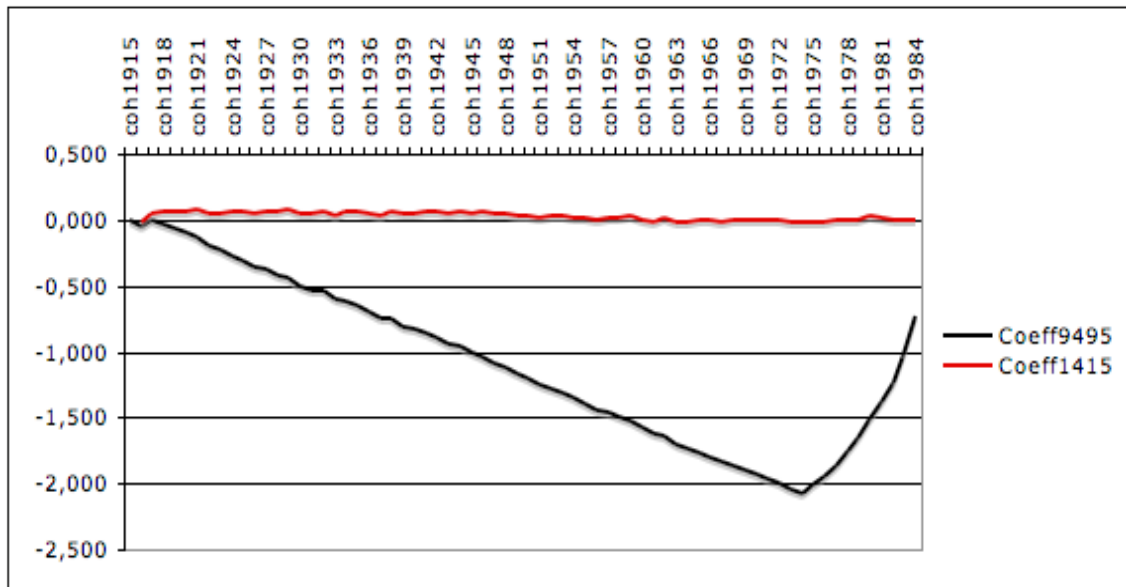


Figure D.3.3. Comparaison des effets de cohorte estimés à partir d'une contrainte sur deux périodes successives (1994 et 1995) ou sur deux cohortes successives (1914 et 1915).

L'application de cette méthode de contrainte d'égalité en utilisant des contraintes a priori très banales met en évidence deux inconvénients majeurs de cette approche :

- Les variances de toutes les indicatrices sont fortes parce que l'on se place ainsi dans un cas de multicollinéarité certes non plus parfaite, mais qui reste très forte et que les moindres carrés supportent très mal;
- Les résultats obtenus sont très sensibles aux contraintes imposées, et ce non seulement sur la variable concernée par la contrainte, mais aussi sur les autres variables du modèle via la forte multicollinéarité entre les jeux d'indicatrices. Cette forte sensibilité va jusqu'à inverser la direction des effets mesurés.

Malgré ces inconvénients, cette approche reste la plus utilisée dans les modélisations APC en marketing et constitue une des meilleures à disposition à ce jour.

Cependant, nous pensons qu'un des moyens de contourner les problèmes engendrés par la forte multicollinéarité entre les jeux d'indicatrices restantes consiste à ne pas estimer le modèle par les moindres carrés ordinaires, mais par une méthode capable d'estimer des données multicollinéaires. Ceci ne résoudra pas le problème du biais latent dans les

modélisation APC, mais permettra potentiellement d'obtenir des estimations des effets de plus faible variance.

Ensuite, il nous semble important de pouvoir examiner le niveau de sensibilité des résultats à l'information a priori introduite. En effet, cette information étant invérifiable et difficilement sujette à une formulation très précise (de type égalité stricte), il serait préférable de pouvoir vérifier de quelle manière les résultats obtenus varieraient si l'on introduisait une information légèrement différente de celle initialement introduite.

Par la suite, nous allons présenter dans la section D.4. deux autres méthodes de modélisation APC, cette fois indirectement fondées sur les moindres carrés. Puis la section D.5. détaillera notre méthodologie, qui a aussi l'avantage de ne pas reposer sur les moindres carrés, mais qui permet de surcroît d'examiner la sensibilité des résultats à l'information a priori choisie.

D.4. Méthodes de modélisation APC non fondées sur les moindres carrés

D.4.i. Régression sur composantes principales (PCR)

La méthode de régression sur composantes principales (PCR) n'est plus qu'indirectement basée sur les moindres carrés car elle consiste à régresser la variable à expliquer, sur les facteurs d'une analyse en composantes principales des prédicteurs, et non plus directement sur les prédicteurs eux-mêmes. Pour une bonne description de la méthode, illustrée de quelques applications, le lecteur pourra consulter Massy (1965).

Elle pourrait sembler a priori avantageusement adaptée à la problématique APC parce qu'elle permet de contourner le problème de la multicollinéarité parfaite des variables, sans avoir à détériorer les données en amont, comme dans le cas des méthodes précédentes (omission de variable ou contrainte d'égalité).

En effet, cette multicollinéarité, même parfaite, n'empêche pas de mener une ACP sur toutes les indicatrices de toutes les variables APC. Au final il y aura simplement quatre facteurs de moins que le nombre d'indicatrices initialement introduites, car la matrice de variance-covariance aura quatre valeurs propres nulles (une par groupe d'indicatrices pour A, P et C ; plus la relation de multicollinéarité parfaite entre les variables). Ensuite, nous régresserons la variable à expliquer Y sur les facteurs, qui sont orthogonaux, avant de désagréger les facteurs en combinaisons linéaires des prédicteurs initiaux en appliquant le changement de base inverse à celui permettant de passer des variables initiales aux facteurs.

En appliquant cette méthode sur nos données, le modèle est significatif (F-test de 60,05 et p-value correspondante inférieure à 0,1 %) et les critères d'ajustement sont bien entendu très satisfaisants (R^2 égal à 95,5% et R^2 ajusté de 94%).

Les résultats complets par effet sont fournis dans les premières colonnes du tableau récapitulatif de l'Annexe 5. On remarque que la multicollinéarité entre les variables initiales réapparaît inévitablement lors de la désagrégation des facteurs. Ainsi la variance des estimations est assez importante. En prenant un seuil de significativité à 1%: 41% des effets d'âge estimés ne sont pas significatifs, tout comme 55,5% des effets de période et 21,1% des

effets de cohorte. Ce problème constitue un des défauts majeurs de la PCR utilisée dans le cadre de la multicollinéarité et des recherches visent à l'atténuer, comme celles de Hwang et Nettleton (2002) en sélectionnant les composantes principales retenues en fonction des colinéarités potentielles entre variables initiales. Mais en faisant ainsi, on introduit implicitement de l'information exogène au moment de l'estimation. Cette information constitue une information a priori dont on ne connaît pas la nature, donc qu'on ne contrôle pas, mais qui aura pourtant une influence sur les résultats.

Cette méthode a donc deux inconvénients imbriqués :

- La régression linéaire sur les facteurs (vs. les jeux d'indicateurs) permet l'estimation mais fournit de estimations des effets nets à très forte variance, ce qui rend la plupart d'entre eux non significatifs ;
- La suppression de certains facteurs avant la régression du taux sur les facteurs permettrait de résoudre ce premier inconvénient, mais en impliquerait un autre, plus gênant encore dans notre contexte, celui d'introduire de l'information a priori dont on ne connaît absolument pas la nature.

Mais comme nous le verrons plus loin, nous pensons qu'il serait préférable d'éviter d'utiliser l'estimation par les moindres carrés, même dans un second temps. Pour cela, nous privilégierons la méthode des moindres carrés partiels, ou régression PLS, pour procéder à l'estimation de notre modèle APC. En effet, comme souligné par Cazes (1992) dans sa comparaison des deux méthodes d'estimation, la régression PLS génère des facteurs expliquant non seulement les liens entre les prédicteurs comme dans la PCR (Abdi, 2003), mais aussi les liens entre les prédicteurs et la variable à expliquer.

D.4.ii. Parsimonious Cohort Estimation (PACE)

La méthode PACE étant plus innovante que la précédente, elle sera d'abord présentée plus en détails avant d'être appliquée et commentée.

Principe de la méthode PACE

Cette méthode d'optimisation quadratique est proposée par Rust et Yeung (1995). Elle se nomme Parsimonious Cohort Estimation (PACE) car elle part du principe que, parmi les

différents modèles disponibles du fait de la sous-identification du problème APC, le principe de parcimonie (ou rasoir d'Ockham) nous permet de choisir un modèle unique.

Ce modèle est le plus parcimonieux dans le sens où il fait intervenir un minimum de variables. Pour les variables restantes, les coefficients doivent avoir des valeurs minimales, toujours, d'après les auteurs, suivant le même critère de parcimonie. Le critère d'ajustement reste les moindres carrés, certes, mais les moindres carrés sont une *contrainte* à satisfaire au sein d'un problème d'optimisation et non la méthode d'estimation utilisée.

Pour résumer, voici les trois critères hiérarchiques que la méthode PACE vise à satisfaire :

1. Minimiser la somme des carrés des erreurs (MCO),
2. Réduire au minimum le nombre de variables incluses dans le modèle,
3. Minimiser les valeurs absolues des coefficients estimés pour les variables retenues.

La traduction algébrique de cette procédure consiste en la formule ci-dessous, avec la notation SCR représentant la somme des carrés des résidus :

Minimiser

$$\begin{aligned}
 h = & M_1 \times [\sum_i a_i^2 + \sum_j p_j^2 + \sum_k c_k^2] \\
 & + M_2 \times [(\sum_i a_i^2)(\sum_j p_j^2) + (\sum_i a_i^2)(\sum_k c_k^2) + (\sum_j p_j^2)(\sum_k c_k^2)] \\
 & + M_3 \times [(\sum_i a_i^2)(\sum_j p_j^2)(\sum_k c_k^2)] \\
 & + M_4 \times \text{SCR}
 \end{aligned}
 \tag{11}$$

Avec $0 \ll M_1 \ll M_2 \ll M_3 \ll M_4$

Reprenons en détail les critères d'optimisation choisis :

- le terme en M4 sera minimal si la somme des carrés des résidus est minimale, ce qui consiste à intégrer les MCO dans la fonction à minimiser et à lui donner un poids plus important que tout le reste;
- le terme en M3 est minimal si tous les coefficients d'une des trois variables, âge, période ou cohorte, sont égaux à zéro (c'est-à-dire qu'on élimine une de ces trois variables), ce qui consiste à privilégier la parcimonie « par variable » ;
- le terme M2 sera minimal si deux des trois variables sont écartées, ce qui consiste à privilégier, dans un second temps, la parcimonie « par couple de variables » ;

- le terme M1 sera minimal si les valeurs des coefficients estimés pour les variables finalement retenues (après les étapes M3 et M2) sont de norme minimale, ce qui consiste à privilégier, dans un dernier temps, la parcimonie sur la valeur des effets.

Pour ordonner ces critères, les valeurs des constantes M4 à M1 ont des valeurs décroissantes et des ordres de grandeur différents. D'un point de vue théorique il est intéressant de constater qu'ici l'information a priori choisie (la parcimonie) est introduite *durant* l'estimation, et non *avant* estimation comme cela était le cas dans les méthodes utilisant directement les moindres carrés. Nous reviendrons dans la discussion sur le fait que les méthodes présentées dans cette recherche diffèrent aussi suivant l'étape durant laquelle est introduite l'information a priori.

Application de la méthode PACE

Dans leur recherche, Rust and Yeung (1997) travaillaient sur 3 âges, observés sur trois périodes, correspondant à 5 cohortes. Il y avait donc 11 effets à estimer. Dans notre cas, le nombre d'effets s'élève à 143. Nous avons donc dû adapter la méthode d'optimisation, sur la base de la fonction `fminsearch` de Matlab (2007) afin d'estimer les coefficients dans des temps de calcul raisonnables. Le code est fourni dans l'annexe 4. Nous avons ajouté une constante au modèle, ce qui nous fournit des coefficients comparables à ceux obtenus par les autres méthodes qui permettent elles aussi d'avoir un terme constant.

Ceci dit, en pratique cette méthode nous a posé deux importants problèmes de mise en oeuvre. D'abord, même en profitant des qualités de Matlab et d'un processeur à 2Ghz, l'estimation n'a été obtenue qu'après plus de 10h de calcul. Ce premier problème en soulève un second : dans ce cadre, il devient impossible d'imaginer pouvoir obtenir des intervalles de confiance sur les estimations obtenues. En effet, cette méthode exigerait d'appliquer une technique de bootstrap pour estimer les variances des effets, ce qui demande de réitérer des estimations des centaines de fois, ce qui n'est ici pas envisageable.

Les résultats sont présentés dans la deuxième partie du tableau de l'annexe 5. Les figures D.4.1 à D.4.3. ci-après donnent la représentation graphique de ces coefficients. Il est possible de mesurer la qualité d'ajustement du modèle par une mesure du R^2 , qui devait être optimisé par M4. Il est ici de 97.44%.

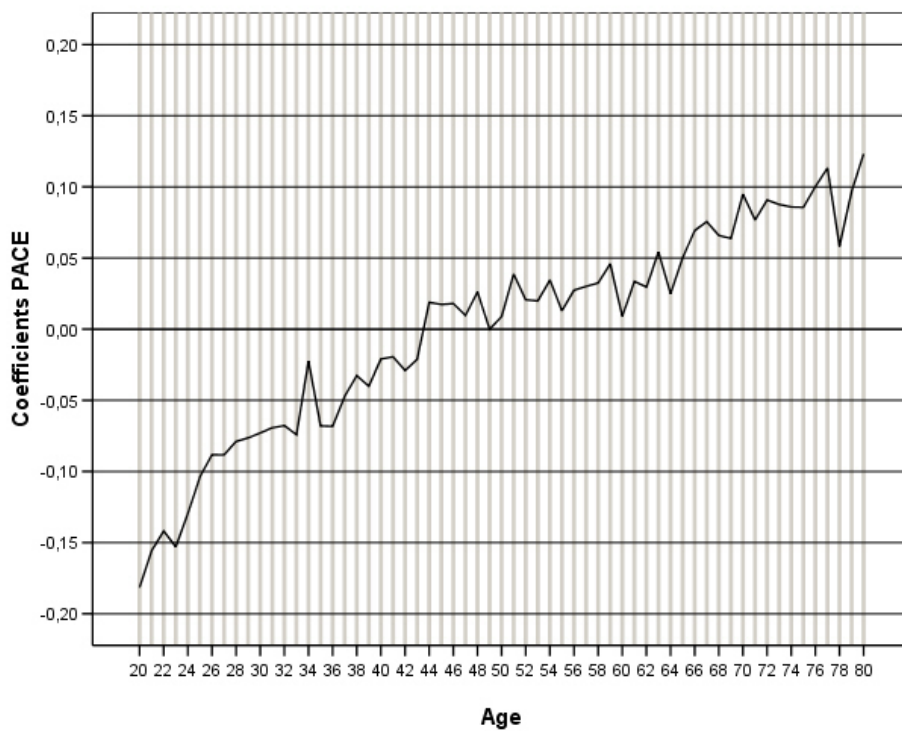


Figure D.4.1. Effets d'âge obtenus par la méthode PACE

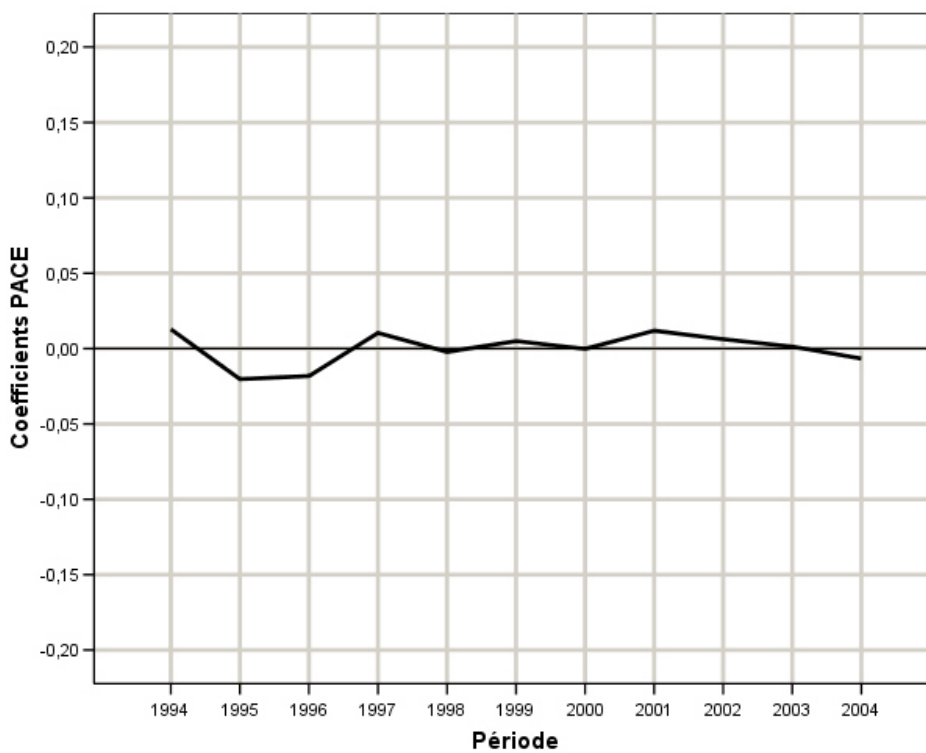


Figure D.4.2. Effets de période obtenus par la méthode PACE

facile à éliminer est celle comportant le moins de modalités. Sur des données réelles, il s'agit souvent de la variable période, car on observe les individus sur tous les âges de 20 à 60 ans, c'est-à-dire sur au moins 40 ans. On a donc de ce fait au moins autant de cohortes. Par contre, les phénomènes sont en général observés sur une plus courte durée (donc sur un moins grand nombre de périodes), donc c'est quasi-systématiquement la période qui sera éliminée lors du processus d'estimation. Le critère de parcimonie reviendra donc en pratique à prendre des effets de période nuls, ce qui est équivalent à prendre comme information a priori que les effets de période sont nuls, sans avoir pu pour autant choisir cette information a priori.

Si ce n'est pas toujours le cas dans l'exemple théorique utilisé (simulation) dans cet article, c'est qu'il s'agit d'un cas avec 3 périodes et 3 âges et donc que les deux variables ont le même nombre de modalités. Mais ce cas n'apparaît que très peu fréquemment en pratique.

Le second inconvénient majeur que nous avons déjà évoqué auparavant concerne le mode d'optimisation et le calcul des intervalles de confiance. L'optimisation quadratique sous contrainte est assez périlleuse par rapport à d'autres méthodes disponibles en statistiques, car elle est très coûteuse en puissance de calcul (convergence lente), et présente le risque d'optima locaux. Dans notre cas, il faut plus de dix heures de calcul pour obtenir les résultats présentés. Par la suite, nous avons exprimé la formule à minimiser à l'aide d'un gradient, et cette solution a permis de réduire le temps de calcul d'un peu plus d'une heure. Cependant, si l'on souhaite ré-échantillonner les données quelques centaines de fois pour obtenir un intervalle de confiance sur les estimations des coefficients, on est de nouveau face à une barrière.

Encore une fois cette barrière technique n'est pas visible sur les données simulées utilisées dans l'article d'origine parce que le cas comporte peu d'effets à estimer (trois périodes, trois âges et cinq cohortes), mais elle apparaît handicapante sur des données réelles. Et, concernant les intervalles de confiance, les données simulées étaient peu bruitées, donc leur calcul ne s'imposait pas. Mais dans un cas réel, le modélisateur ne connaît pas le modèle sous-jacent et doit avoir un moyen de contrôler l'incertitude liée à l'estimation.

Suite à cet examen détaillé des différentes approches de modélisation APC disponibles, la section D.5. est consacrée à la présentation et l'application de la méthode que nous avons développée sur ce sujet, nommée APC-PLS.

D.5. Méthodologie APC-PLS

Nous présentons dans cette partie notre approche de modélisation APC nommée APC-PLS. Suivront les applications de cette méthode sur les deux exemples que nous traitons dans cette recherche, à savoir la fidélité à une marque automobile sur les marchés français et allemand.

D.5.i. Principe et étapes de la méthode APC-PLS

Principe

D'après Maddala (2001, p.65),

“Most of the discussions of the multicollinearity problem and its solution are based on criteria based on the intercorrelations between the explanatory variables. However this is an incorrect approach.”

Ainsi, la clef de la résolution des problèmes de type âge-période-cohorte est d'après nous d'accepter dès le début la relation linéaire entre les prédicteurs et, par conséquent, d'écarter les méthodes utilisant les moindres carrés ordinaires (MCO), incompatibles avec ce type de données.

En effet comme nous l'avons vu, les méthodes précédentes utilisent directement ou indirectement les moindres carrés ordinaires. Comme ils ne sont pas applicables directement, des étapes préliminaires sont nécessaires pour « se ramener » à des données sur lesquelles l'on peut appliquer les MCO. Au lieu d'être contraints de modifier les données pour éviter que la multicollinéarité parfaite présentée dans la section D.1. ne bloque l'estimation, nous préférons nous placer dans un cadre où la multicollinéarité n'est pas un problème d'estimation, c'est-à-dire celui de la méthodologie PLS. Ainsi nous traiterons le problème de multicollinéarité parfaite pour ce qu'il est. À savoir non pas un problème d'estimation à proprement parlé, mais un problème théorique d'usage de l'information a priori.

La régression PLS, malgré ses nombreux avantages techniques (Tenenhaus, 1999) et son développement dans plusieurs champs comme la chimie (Wold et al. 1996) n'est pas encore devenue très populaire dans la recherche académique en marketing depuis la présentation de

l'approche par Fornell et Cha (1992, 1994) au sein de la communauté. Les premières applications ont été proposées en France par Valette-Florence (1988,1990) et Tenenhaus (1998). Un récent exemple d'application est fourni par Tenenhaus et al. (2005). Dans le cas de l'analyse âge-période-cohorte, l'approche PLS apporte pourtant une supériorité technique décisive qui nous a amené à l'utiliser comme méthode d'estimation.

La méthodologie que nous proposons se compose des trois étapes suivantes :

- Après une discrétisation et une normalisation des variables APC, nous appliquons la régression PLS de Y, taux de fidélité à expliquer, sur les indicatrices APC à savoir l'ensemble des A_i , P_j et C_k disponible, sans supprimer aucune des indicatrices, ni imposer de contrainte a priori sur les valeurs des effets a_i , p_j et c_k de ces indicatrices ;
- A partir du premier jeu d'effets ainsi obtenu, nous appliquons la méthode de « report des écarts consécutifs » pour introduire l'information a priori désirée ;
- Enfin, nous examinons la robustesse des résultats par rapport à des variations de cette information a priori.

Ces trois étapes sont présentées dans les sous-sections suivantes, de D.5.ii. à D.5.iv.

D.5.ii. Etape 1 : Régression PLS

Préparation des données

Cette sous-étape consiste à discrétiser les variables d'âge, période et cohorte (resp. A,P,C) en autant de variables binaires (resp. A_i , P_j et C_k) qu'il existe de modalités différentes au sein de ces variables. Elle permet d'obtenir une estimation année par année des effets, alors que conserver les variables entières imposerait implicitement un lien linéaire entre les effets individuels des âges (resp. des cohortes, des périodes). Notons qu'il n'est pas nécessaire, comme dans le cas de la discrétisation dans le cadre d'une régression linéaire OLS, d'éliminer une des modalités qui serviraient de modalité de référence.

Afin d'améliorer l'interprétation des résultats, les indicatrices sont centrées-réduites afin d'obtenir des coefficients standardisés plus facilement comparables et interprétables.

Régression PLS

Pour exposer l'intérêt de la régression PLS dans le cadre du problème âge-période-cohorte, nous utiliserons l'interprétation des moindres carrés partiels de Garthwaite (1994) car elle propose une décomposition intéressante des différentes étapes de cette méthode.

Le principe de la régression PLS est de ne pas directement régresser la variable dépendante sur les prédicteurs, mais sur des combinaisons linéaires des prédicteurs, appelées composantes, définies non seulement pour résumer l'information contenue dans les prédicteurs mais aussi celle concernant le lien entre les prédicteurs et la variable dépendante Y .

L'équation de régression s'exprimera donc ainsi :

$$Y = \tau_0 + \tau_1 T_1 + \tau_2 T_2 + \dots + \tau_K T_K + \varepsilon \quad [12]$$

Dans laquelle chaque composante T_k , $k=1, \dots, K$, est une combinaison linéaire des prédicteurs et la corrélation entre une paire quelconque de composantes est égale à 0, c'est-à-dire :

$$\forall j, k \quad \rho(T_j, T_k) = 0. \quad [13]$$

Prenons ici notre notation générique Y de la variable à expliquer et X du vecteur de m prédicteurs X_n ($n=1, \dots, N$). Parallèlement, notons U_1 et v_{1n} ($n=1, \dots, N$) les versions centrées-réduites correspondantes. On régresse tout d'abord la variable dépendante U_1 sur chaque prédicteur v_{1n} , pris isolément, et l'on obtient N prédictions de U_1 notées $U_{1(n)}$ sous la forme:

$$U_{1(n)} = t_{1n} v_{1n} \quad [14]$$

avec $t_{1n} = v'_{1n} U_1 / (v'_{1n} v_{1n})$

Ces estimations des t_{1n} sont elles aussi potentiellement biaisées par la multicolinéarité, mais comme nous l'avons vu, ce problème n'est pas évitable. Par contre, nous pouvons ici conserver tous les prédicteurs initiaux et n'avons pas à fixer de contrainte a priori pour lancer l'estimation. Dans notre perspective d'utilisation de la méthode PLS, seules les valeurs prédites nous intéressent, et non pas le poids précis à donner à chaque variable, étant donné que ces poids seront par la suite retravaillés à l'occasion d'une nouvelle répartition.

La manière la plus naturelle, dans le cadre d'un modèle linéaire, de combiner ces différentes prévisions de U_1 est d'en faire une moyenne pondérée :

$$T_1 = \sum_{n=1 \dots N} w_{1n} (t_{1n} v_{1n}) \quad [15]$$

Reste à spécifier la valeur des pondérations w_{1n} . Dans l'approche PLS traditionnelle, les poids sont définis par $v'_{1n} v_{1n}$, donc ces pondérations sont proportionnelles à la variance de v_{1n} qui est estimée par $v'_{1n} v_{1n} / (n-1)$. Dans le cas de la régression sur les indicatrices APC, cette pondération est particulièrement pertinente. En effet, la variance d'une indicatrice est d'autant plus grande que l'âge ou la cohorte en question est présent(e) dans la base de données, car la variance d'une proportion est croissante sur l'intervalle $[0,0.5]$ et aucun âge, ni aucune cohorte ne regroupe plus de 50% des individus d'une table APC dès que le nombre d'âge considéré dans la table est supérieur à 2. Concernant les variables de périodes, le nombre d'observations étant également réparti suivant les différentes périodes (toujours le même nombre d'âges et de cohortes par période) la variance est constante. Mais comme les indicatrices ont été préalablement centrées-réduites, le problème est ici résolu à la source.

Une fois la première composante déterminée à partir des pouvoirs prédictifs des prédicteurs initiaux, les composantes suivantes sont déterminées de manière récurrente grâce au principe suivant. T_1 en tant que combinaison linéaire de prévisions de U_1 est un prédicteur de Y mais il reste potentiellement de l'information à extraire des prédicteurs pour prédire Y : il s'agit de l'information contenue dans les prédicteurs non contenue dans T_1 . Cette information peut être estimée par les résidus de la régression univariée de chaque prédicteur sur T_1 . De même l'information qui reste à prévoir après utilisation de T_1 peut être estimée par les résidus de la régression de U_1 sur T_1 . On note respectivement ces résidus v_{2n} ($j=1, \dots, N$) et U_2 . La composante suivante, T_2 est alors estimée à partir de v_{2n} ($n=1, \dots, N$) et U_2 , comme T_1 l'avait été à partir de v_{1n} ($j=1, \dots, N$) et U_1 .

Il est aisé de comprendre dans ce cas comment les composantes sont orthogonales entre elles car T_1 est composée des prédicteurs initiaux centrés-réduits v_{1n} , or, d'après les propriétés de la régression linéaire, les v_{1n} sont orthogonaux aux résidus v_{2n} des régressions des v_{1n} sur T_1 , donc T_1 est orthogonal à l'ensemble des v_{2n} et par conséquent à T_2 qui en est une combinaison linéaire. L'ensemble des propriétés évoquées sur ces deux premières composantes est itérativement transférable aux autres composantes T_3, \dots, T_K construites par une méthode identique à celle utilisée par T_2 .

Le choix du nombre de composantes est réalisé par validation croisée comme dans beaucoup de méthodologies de réduction de dimensions. Les critères pouvant être retenus sont la PRESS, Prediction Error Sum of Squares (Tenenhaus, 1998), et l'AIC, Akaike (1973) Information Criterion (Akaike, 1973 ; Naik, Hagerty et Tsai, 2000). Le critère du Q^2 que nous avons choisi est aussi celui utilisé par Tenenhaus et al. (2005) et il est implémenté dans le logiciel XLStat (XLStat, 2008, p307).³⁴

La mesure générique de qualité en PLS notée $Q^2_{cum}(h)$, définie dans Tenenhaus (1998) et utilisée notamment par Tenenhaus et al. (2005), inclut la statistique PRESS obtenue par validation croisée. C'est une mesure de la contribution des h premières composantes à la qualité prédictive du modèle (et des sous-modèles quand nous aurons plusieurs variables dépendantes). Il peut être interprété comme un « R^2 cross-validé ». Le critère de sélection d'ajout d'une nouvelle composante sera qu'après ajout de la composante considérée, Q^2_{cum} augmente significativement (d'après l'application d'un F-test). La valeur finale du Q^2_{cum} donnera une mesure de la qualité générale du modèle, tenant compte de la qualité prédictive, et complémentaire du R^2 .

En appliquant la régression PLS sur ces données ainsi préparées, nous obtenons un premier jeu d'effets nets des variables APC. En effet, le lien entre facteurs et variables initiales, ou prédicteurs, se fait directement à travers la formule [12] ci-avant.

Pour vérifier la significativité des effets ainsi obtenus nous avons besoin d'intervalles de confiance sur ces effets. L'algorithme PLS est ainsi souvent couplé avec une estimation en Bootstrap des intervalles de confiance. Cette méthode consiste à ré-échantillonner un grand

³⁴ Nous tenons ici à remercier Thierry Fahmy, fondateur et dirigeant d'AddinSoft, pour ses précisions détaillées sur l'algorithme implémenté dans le module XLStat PLS-PM.

nombre de fois l'échantillon initial et à recalculer sur chacun de ces nouveaux sous-échantillons les coefficients pour pouvoir estimer leur dispersion. Et c'est à partir de l'estimation de leur dispersion que l'on peut estimer un intervalle de confiance pour les coefficients en question.

En utilisant les notations de la sous-section D.1., un premier jeu d'effets APC est donc obtenu sans introduire d'information a priori de la manière suivante :

$$Y =_{PLS} b_0 + \sum_{i=1...I} a_i A_i + \sum_{j=1...J} p_j P_j + \sum_{k=1...K} c_k C_k + \varepsilon \quad [16]$$

Ce qui dans le cas de notre base de données se traduit de la manière suivante :

$$Y =_{PLS} b_0 + \sum_{i=20...80} a_i A_i + \sum_{j=1994...2004} p_j P_j + \sum_{k=1914...1984} c_k C_k + \varepsilon \quad [17]$$

Ce premier jeu d'effets (a_i , p_j , c_k) pourra dans une deuxième étape être modifié par la méthode de permutation linéaire de « report des écarts consécutifs » afin de tenir compte de l'information a priori disponible et d'examiner la sensibilité du nouveau jeu d'effets ainsi obtenu par rapport aux variations de l'information a priori introduite.

D.5.iii. Etape 2 : Introduction de l'information a priori

Dans cette section, nous présentons d'abord la technique de report des écarts consécutifs avant de l'illustrer sur quelques exemples.

Principe de la technique

À cause de la relation de multicollinéarité parfaite entre les variables APC, dont les effets sur le taux étudié sont modélisés à l'aide de leurs effets nets, il est possible d'obtenir les mêmes prévisions (taux estimés) à partir de plusieurs jeux d'effets nets différents. Comme nous l'avons souvent évoqué auparavant, la multicollinéarité parfaite implique qu'il existe plusieurs solutions (ici des jeux d'effets nets) de qualité identique (car fournissant exactement les mêmes taux estimés) au problème traité (estimer le taux à partir d'un jeu d'effets nets).

Quelle que soit la méthode utilisée lors de l'estimation, il sera toujours possible de trouver un ensemble de jeux d'effets nets fournissant exactement les mêmes prévisions du taux étudié.

Ce problème est intrinsèque à la multicolinéarité. Cependant, cette particularité nous conduit à développer une méthodologie nous permettant de déduire d'un jeu d'effets nets tous les autres jeux d'effets nets amenant à des taux estimés identiques. Ceci ne signifie pas que tout jeu d'effet peut être utilisé pour effectuer des prévisions identiques, mais qu'il en existe plusieurs et qu'il faudra ensuite choisir entre ces jeux d'effets nets en fonction de l'information a priori disponible.

Algorithme de la méthode de report des écarts consécutifs

Étant donné une explication, c'est-à-dire un jeu d'effets, quel algorithme peut engendrer des explications équivalentes, c'est-à-dire fournissant les mêmes estimations des prévisions, en redistribuant les effets sur les modalités des variables APC ?

Voici les notations utilisées pour présenter cet algorithme. Ce sont les mêmes que celles utilisées jusqu'à présent :

Les a_i , $i=1 \dots I$, sont les effets des modalités A_1, \dots, A_I de la variable A, A_1 étant l'âge le plus faible ; p_j , $j=1 \dots J$ les effets des modalités P_1, \dots, P_J de la variable P, P_J étant la période la plus récente et c_k , $k=1 \dots K$ les effets des modalités C_1, \dots, C_K de la variable C, C_K étant la cohorte la plus récente. Nous notons ici de plus y la prévision de Y liée à un jeu d'effets nets.

Soit m une constante.

Considérons la liste de $I+J+K+1$ éléments, ou tuple θ suivant :

$$\theta = ((a_i), (p_j), (c_k), m) \quad [18]$$

qui définit la liste des effets d'un modèle.

Il permet de prédire, à partir d'une paire âge et période (A_i, P_j) observée, la variable Y à modéliser en fournissant l'estimation y de la manière suivante :

$$y(\theta(A_i, P_j)) = a_i + p_j + c_{j-i} + m \quad [19]$$

Notre problématique peut donc se retraduire à présent en la recherche d'une transformation φ de θ , voire une famille entière de transformations fournissant un nouveau jeu d'effets donnant exactement les mêmes prédictions de Y , c'est-à-dire, pour tout $A_i, i=1\dots I$ et $P_j, j=1\dots J$ on doit avoir :

$$y(\varphi(\theta(A_i, P_j))) = y(\theta(A_i, P_j)) \quad [20]$$

Il s'agit de l'objectif de cette méthode : pouvoir faire varier les valeurs des effets A, P et C sans modifier les valeurs prédites pour y . Reste à trouver une telle famille de transformation respectons cette contrainte.

Montrons qu'il existe effectivement une famille de transformations respectant cette contrainte. Cette famille sera paramétrée par les paramètres : α, π, γ et v , et les éléments de cette famille de transformations sont notés $\varphi_{\alpha, \pi, \gamma, v}$.

Les transformations de cette famille n'ont pas d'interprétation directe. Chaque transformation représente chacune un *ensembles de permutations linéaires* permettant de passer d'un jeu d'effets à l'autre sans modifier les estimations de Y . Nous verrons plus loin le sens de v .

Etant donné $\theta = ((a_i), (p_j), (c_k), m)$ et des paramètres de transformation α, π, γ et v ayant des valeurs réelles données, définissons les nouveaux effets issus de θ de la manière suivante :

$$\varphi_{\alpha, \pi, \gamma, v}(\theta) = \theta' = ((a'_i), (p'_j), (c'_k), m') \quad [21]$$

Construits par récurrence :

$$a'_1 = \alpha$$

$$a'_{i+1} = a'_i + ((a_{i+1} - a_i) + v)$$

$$p'_J = \pi$$

$$p'_{j-1} = p'_j - ((p_j - p_{j-1}) - v)$$

$$c'_K = \gamma$$

$$c'_{k-1} = c'_k - ((c_k - c_{k-1}) + v)$$

$$m' = m + (a_1 - \alpha) + (p_J - \pi) + (c_K - \gamma)$$

Notons ici que par souci de cohérence, les notations employées ici seront les même dans tout le cours de l'étude.

Démontrons à présent que le θ' ainsi défini vérifie bien l'équation [20] ci-avant. Si cette démonstration est réalisée, nous aurons atteint l'objectif fixé par cette équation [20].

Par construction de m' , on vérifie que $y(\theta'(A_1, P_j)) = y(\theta(A_1, P_j))$.

Pour une valeur de j fixée et en utilisant la définition de a'_{i+1} et de c'_{k-1} , on montre par récurrence ascendante sur i que cette propriété est vérifiée pour tout A_i .

Pour une valeur de i fixée et en utilisant les définitions de p'_{j-1} et de c'_{k-1} , on montre par récurrence descendante sur j que cette propriété est vérifiée pour tout P_j .

Nous avons donc ainsi défini comment passer d'un jeu d'effets à l'autre à partir d'une fonction indexée par quatre paramètres :

- α, π, γ pour l'étalonnage des bornes ;
- v pour la réallocation.

Cette méthode permet de passer d'un jeu d'effet à l'autre sans modifier les prévisions, ce qui est l'objectif fixé par l'équation [20].

Elle est fondée sur la colinéarité parfaite :

$$A = P - C$$

entre les variables complètes A, P et C (par rapports aux indicatrices A_i, P_j, C_k).

Nous reformulons cette colinéarité parfaite au niveau des indicatrices A_i, P_j, C_k suivant le principe de vases communicants, comme exposé dans la démonstration ci-avant à travers les trois expressions suivantes:

$$\begin{aligned} a'_{i+1} &= a'_i + ((a_{i+1} - a_i) + v) \\ p'_{j-1} &= p'_j - ((p_j - p_{j-1}) - v) \\ c'_{k-1} &= c'_k - ((c_k - c_{k-1}) + v) \end{aligned}$$

Cette méthode est validée par la démonstration mathématique de la page précédente. Une rédaction « pas à pas » des quatre récurrences simples contenues dans cette démonstration est proposée aux lecteurs n'étant pas familier des démonstrations par récurrence dans l'annexe 8.

Elle a été élaborée spécifiquement dans le cadre de cette recherche. Il existe peut-être d'autres méthodes permettant d'atteindre l'objectif de l'équation [20], mais à notre connaissance, aucune autre n'a été publiée à ce jour. Nous ne démontrons pas qu'il s'agit de la seule manière de répondre à cet objectif, mais démontrons simplement qu'elle satisfait cet objectif.

En effet, elle satisfait en tous points aux objectifs fixés, avec simplicité (car fondée sur lien analytique entre colinéarité des variables et colinéarité des indicatrices) et efficacité (car n'utilisant que quatre paramètres). Elle sera donc utilisée par la suite.

Le paramètre v indique de combien les écarts consécutifs, définis comme les différences entre les effets nets de deux modalités consécutives d'une même variable vont être modifiées. Par exemple pour la variable Age dans le premier jeu d'effets nets, les écarts consécutifs seront les $(a_{j+1} - a_j)$ pour les effets nets des deux modalités consécutives A_j et A_{j+1} . D'un jeu d'effets nets à l'autre, ces écarts consécutifs varient. Le principe de notre méthode appelé technique de report des écarts consécutifs, consiste à dire qu'en appliquant une même variation v à tous ces écarts consécutifs des variables âge et cohorte, et la variation inverse à tous les écarts consécutifs de la variable période, nous obtiendrons des jeux fournissant la même estimation Y . Notre paramètre de report des écarts consécutifs v est donc cette variation, d'un jeu d'effets nets à un autre, entre les écarts consécutifs des modalités des variables APC. Ce paramètre v est le même pour toutes les modalités de toutes les variables.

Malgré l'interprétation donnée ci-dessus du paramètre v , soulignons le fait que la valeur optimale de v , tout comme celles des trois autres paramètres α , π , γ , n'est pas choisie par l'analyste, mais dépendra de l'information a priori que l'utilisateur souhaitera intégrer. Ainsi l'utilisateur n'a par exemple pas à choisir l'indicatrice dont le coefficient subit la plus petite modification v . Ces paramètres ne permettent que d'optimiser la spécification du modèle à partir de l'information a priori, et n'ont pas à être manipulés directement par l'utilisateur de la méthode.

N.B. Il est aussi crucial de comprendre ici que la technique s'applique sur les *prédictions du taux*, et non sur le taux Y lui-même. Du passage d'un jeu d'effets nets à l'autre, la qualité du modèle reste la même puisque les deux jeux d'effets fournissent exactement les mêmes prédictions du taux. Cette technique est donc applicable a posteriori et permet de modifier les valeurs des effets nets à résultat constant.

Nous allons expliquer par la suite pourquoi dans notre cas de figure, cette technique est particulièrement importante dans l'optique d'intégrer dans le modèle de l'information a priori sur les effets nets étudiés.

Exemple d'application de la technique de report des écarts consécutifs

Reprenons l'exemple théorique simplifié qui nous a servi dans la section D.2. pour la mise en garde sur l'interprétation abusive de données APC dans la partie de l'étude concernant l'interprétation des effets bruts.

		Période	
		P ₃	P ₄
Age	A ₁	C ₂ Y=10%	C ₃ Y=30%
	A ₂	C ₁ Y=20%	C ₂ Y=40%

Tableau D.5.1. Exemple simplifié d'effets APC (Rappel du tableau D.2.3.)

En calculant les effets bruts moyens notés liés à A₁ et A₂, et P₃ et P₄, on obtient les résultats de la première colonne du tableau ci-après.

Jeu d'effets nets initial		Nouveau jeu d'effets nets obtenu avec $v = -10\%$	
Indicatrice	Effet	Indicatrice	Effet
A ₁	a ₁ = 10%	A ₁	a ₁ '= 0%
A ₂	a ₂ = 20%	A ₂	a ₂ '= 0%
P ₃	p ₃ = 0%	P ₃	p ₃ '= 0%
P ₄	p ₄ = 20%	P ₄	p ₄ '= 30%
C ₁	c ₁ = 0%	C ₁	c ₁ '= 20%
C ₂	c ₂ = 0%	C ₂	c ₂ '= 10%
C ₃	c ₃ = 0%	C ₃	c ₃ '= 0%

Tableau D.5.2. Première application de la technique de report des écarts consécutifs (adapté à partir du tableau D.2.4.)

Dans la section D.2. nous avons indiqué qu'il était possible passer d'un jeu d'effet à un autre sans préciser comment cela avait été réalisé. En réalité, ce nouveau jeu d'effets a été obtenu à l'aide de la technique de report des écarts consécutifs avec le paramètre $v = -10\%$, et dans ce cas, les valeurs des paramètres des bornes sont : $\alpha = 0\%$, $\pi = 30\%$ et $\gamma = 0\%$.

En effet :

$$a_1' = \alpha = 0\%$$

$$a_2' = a_1' + (a_2 - a_1) + v = 0\% + (20\% - 10\%) - 10\% = 0\%$$

$$p_4' = \pi = 30\%$$

$$p_3' = p_4' - (p_4 - p_3 - v) = 30\% - ((20\% - 0\%) - (-10\%)) = 0\%$$

$$c_3' = \gamma = 0\%$$

$$c_2' = c_3' - (c_3 - c_2 + v) = 0\% - (0\% - 0\% - 10\%) = 10\%$$

$$c_1' = c_2' - (c_2 - c_1 + v) = 10\% - (0\% - 0\% - 10\%) = 20\%$$

Pour conclure, revenons sur la nouvelle valeur de la constante, qui était nulle. Comme les nouvelles prédictions issues de ce second jeu correspondaient déjà aux précédentes dans notre exemple, cette nouvelle valeur devrait aussi être nulle.

Rappel : $m' = m + (a_1 - \alpha) + (p_J - \pi) + (c_K - \gamma)$

Ici : $m' = 0 + (10\% - 0\%) + (20\% - 30\%) + (0\% - 0\%) = 0$.

Comme nous allons le voir, cette technique permettant d'explorer les différents jeux d'effets disponibles peut être utilisée pour introduire de l'information a priori au sein des premiers résultats issus de notre première étape (régression PLS sur l'ensemble de indicatrices) durant laquelle nous n'avons pas encore introduit cette information nécessaire à la bonne évaluation des effets APC.

Introduction de l'information a priori à partir de l'exemple d'application

Pour introduire une information a priori, il faut lui donner l'expression d'une contrainte analytique. Cette contrainte analytique sera imposée dans le programme d'optimisation dépendant du paramètre d'optimisation v . On teste toutes les valeurs possibles et on en déduit les valeurs qui correspondent le mieux à la contrainte. Dans une approche similaire à la détermination de ρ dans les modèles de correction d'auto-corrélation, c'est par la procédure qu'on déduit v , qui n'est en aucun cas choisi directement par l'utilisateur.

Pour illustrer cette démarche, appliquons en le principe sur l'exemple précédent. D'autres illustrations plus poussées seront données dans la suite de l'étude : l'application de la méthode dans la section E.1.ii. et une section dédiée à cet effet dans la section de discussion G.4.

Supposons que l'on sache de manière exogène que dans le phénomène modélisé par cet exemple fictif, les effets de période soient nuls. Cette information n'est pas vérifiable mais sera jugée comme assez fiable pour être utilisée afin de choisir le nouveau jeu d'effets $((a'_i), (p'_j), (c'_k), m')$ le plus apte à expliquer ce phénomène à partir du jeu d'effets initial $((a_i), (p_j), (c_k), m)$ issu de la régression PLS de Y sur toutes les indicatrices $((A_i), (P_j), (C_k))$.

Introduire cette information revient à trouver a posteriori parmi les jeux d'effets disponibles le (ou les) jeu(x) tel(s) que :

$$\text{Min}_{\alpha, \pi, \gamma, v} \quad \sum_j p'_j{}^2$$

avec : $((a'_i), (p'_j), (c'_k), m') = \varphi_{\alpha, \pi, \gamma, v}((a_i), (p_j), (c_k), m)$ [22]

et la définition de $\varphi_{\alpha, \pi, \gamma, v}$ donné dans la formule [21].

Dans le cas, notre programme se résume donc à minimiser $(p_3^2 + p_4^2)$. Il n'y a pas de solution algébrique à ce type de programme, mais la solution peut être trouvée par des méthodes d'optimisation telle le simplex. Etant la nature du programme à résoudre, l'application du simplex peut mener à un minimum local. Il est donc nécessaire de réaliser au préalable un quadrillage discret de l'espace des valeurs possibles pour les paramètres à optimiser et de réappliquer la méthode du simplex avec comme point de départ chacune des intersections de ce quadrillage afin d'éviter les minimums locaux et d'aboutir au minimum minimorum, qui est le minimum recherché.

D'après le processus itératif décrit dans la formule [21], nous avons besoin pour cela de fixer $\pi = 0\%$ et $v = +20\%$. Par choix d'étalonnage, les deux autres paramètres de la technique sont fixés tels que $\alpha = 10\%$, $\pi = 0\%$ et $\gamma = 20\%$.

Ainsi :

$$a_1'' = \alpha = 10\%$$

$$a_2'' = a_1'' + (a_2 - a_1) + v = 10\% + (20\% - 10\%) + 20\% = 40\%$$

$$p_4'' = \pi = 0\%$$

$$p_3'' = p_4'' - (p_4 - p_3 - v) = 0\% - ((20\% - 0\%) - 20\%) = 0\%$$

$$c_3'' = \gamma = 20\%$$

$$c_2'' = c_3'' - (c_3 - c_2 + v) = 20\% - (0\% - 0\% + 20\%) = 0\%$$

$$c_1'' = c_2'' - (c_2 - c_1 + v) = 0\% - (0\% - 0\% + 20\%) = -20\%$$

A nouveau pour le calcul de la constante, nous nous attendons à une valeur nulle.

$$\text{Ici : } m' = 0 + (10\% - 10\%) + (20\% - 0\%) + (0\% - 20\%) = 0.$$

Ainsi nous obtenons un nouveau jeu d'effets satisfaisant notre objectif d'effets de période nuls dans le tableau D.5.3. ci-après. Cette utilisation de la méthode de report des écarts pour introduire de l'information a priori va être à présent détaillée.

Nouveau jeu d'effets nets obtenu avec $v = + 20\%$	
Indicatrice	Effet
A1	$a_1 = 10\%$
A2	$a_2 = 40\%$
P3	$p_3 = 0\%$
P4	$p_4 = 0\%$
C1	$c_1 = -20\%$
C2	$c_2 = 0\%$
C3	$c_3 = 20\%$

Tableau D.5.3. Deuxième application de la technique de report des écarts consécutifs

Introduction de l'information a priori par la technique de report des écarts consécutifs

Maddala (2001) propose de définir l'information a priori (prior information en anglais) par « toute information que l'on pourrait avoir sur des paramètres inconnus du modèle, cette information venant de la théorie économique ou d'études empiriques antérieures. » Ici il s'agit de la valeur que l'on va donner a priori à certains des effets APC que l'on cherche à estimer. Cette information exogène doit être choisie par le modélisateur en fonction des sources d'informations externes auxquelles il a pu avoir eu accès.

Nous définissons l'*information a priori initiale (IAPI)* l'information a priori choisie pour obtenir ce premier jeu d'effets par notre technique de report d'écarts consécutifs à partir de l'estimation d'un premier jeu d'effets par la régression PLS.

Cette information doit être exprimée sous la forme d'une fonction F des nouveaux effets à trouver. Elle est utilisée comme critère d'optimisation dans le programme suivant :

$$\text{Min}_{\alpha, \pi, \gamma, v} F((a'_i), (p'_j), (c'_k), m')$$

avec : $((a'_i), (p'_j), (c'_k), m') = \varphi_{\alpha, \pi, \gamma, v}((a_i), (p_j), (c_k), m)$ [23]

et la définition de $\varphi_{\alpha, \pi, \gamma, v}$ donnée dans la formule [21].

Dans le cas précédent pour l'information a priori était que les effets de période sont faibles ou nuls. La fonction F exprime cette information en prenant la forme de la somme des carrés des nouveaux effets de période dans la formule [22].

Si l'information a priori consiste à donner une valeur précise à un effet la fonction F prendra alors la forme de la différence entre la nouvelle valeur de cet effet et la valeur en question.

Si l'information a priori consiste à égaliser deux effets la fonction F prendra alors la forme de la différence entre les nouvelles valeurs de ces deux effets.

Une fois F choisie par le modélisateur en fonction de l'information a priori qu'il souhaite intégrer dans la modélisation des effets APC, il lui suffit de déduire du programme de minimisation de la formule [23] les valeurs des paramètres de la technique de report des écarts correspondants. Ces paramètres lui permettront à l'aide de la technique de report des écarts de la formule [21] de déduire le nouveau jeu d'effets $((a'_i), (p'_j), (c'_k), m')$ le plus apte à expliquer ce phénomène à partir du jeu d'effets initial $((a_i), (p_j), (c_k), m)$ issu de la régression PLS de Y sur toutes les indicatrices $((A_i), (P_j), (C_k))$.

Ainsi toute information a priori exprimable comme d'une fonction (des nouveaux effets nets) à minimiser peut être introduite dans la méthode de modélisation APC-PLS. L'information a priori n'est donc introduite ni au stade de la spécification du modèle (comme dans les méthodes de la section D.3.), ni au stade de l'estimation (resp. D.4.), mais après l'estimation, donc a posteriori.

Ceci constitue une seconde différence majeure (en plus du mode d'estimation) par rapport aux méthodes précédentes sur laquelle nous reviendrons plus en détails dans la section de discussion

G.

D.5.iv. Etape 3 : Etude de sensibilité à l'information a priori

Grâce aux deux premières étapes de la méthode APC-PLS, un premier jeu d'effet a pu être estimé et l'information a priori a pu être introduite à l'issue de cette estimation pour détecter le jeu d'effets nets correspondant le mieux à cette information a priori initiale (IAPI).

Cependant, il reste toujours un certain degré d'incertitude concernant cette information car elle provient soit de la théorie, soit d'études empiriques externes, mais ne peut, quelle que soit la méthode choisie, ni être déduite des données APC, du fait de la multicollinéarité parfaite des données, ni être établie comme parfaitement exacte.

Il serait difficile de demander au modélisateur, même s'il est expert dans le domaine étudié, de formuler une hypothèse stricte, de type « les effets de telle variable sont nuls » ou « l'amplitude des effets de cette variable est de 5% sur la période considérée » ne serait-ce que par prudence. Par contre, il est envisageable qu'il trouve plausible des versions plus prudentes de ces hypothèses comme : « les effets de telle variable sont plutôt proches de zéro », ou : « l'amplitude des effets de cette variable doit être de l'ordre de 5% à quelques points près ». Dans une modélisation APC, l'information a priori utilisée concerne souvent la période. En effet, les effets de période concernent l'ensemble des événements exogènes (de type inflation, nouvelle législation, nouveaux entrants sur le marché considéré) sur lequel un expert du secteur économique concerné peut avoir une idée a priori, ne serait ce qu'en termes de tendance. Et l'on souhaite estimer les autres effets (le plus souvent sur la cohorte et l'âge) de manière assez robuste par rapport à l'incertitude liée à l'information a priori.

C'est pourquoi nous souhaitons pouvoir vérifier que les résultats obtenus sur les effets ne seront pas trop modifiés par de légères variations de cette information a priori introduite (IAPI), c'est-à-dire, si les résultats ne sont pas trop sensibles à la formulation de l'IAPI³⁵.

En effet, dans le cas où l'IAPI est parfaitement exacte, les effets trouvés grâce à la méthode APC-PLS sont égaux aux vrais effets du phénomène étudié.

³⁵ Nous tenons ici à remercier le Professeur Hubert Gatignon pour sa remarque durant la prés-soutenance de thèse du 20 mai 2009 concernant la distinction importante entre robustesse et sensibilité en modélisation.

Par contre, il est bien plus probable que l'I-API ne soit pas parfaitement exacte et que la bonne information a priori soit légèrement différente de la version formulée dans l'I-API. Dans ce cas, qui est le plus courant, l'examen des variations des résultats de la méthode APC-PLS par rapport à de légères variations de l'I-API devient nécessaire pour s'assurer de la stabilité des résultats trouvés. Ainsi l'incertitude vis-à-vis de l'exactitude de l'I-API peut être mieux appréhendée.

Cet examen va être réalisé par le calcul de ce que l'on nomme l'intervalle de sensibilité (IS) des résultats par rapport à des variations de l'I-API. Partant de l'I-API, il faut vérifier que les résultats ne varient pas trop si l'on reformule légèrement cette hypothèse en la remplaçant par une hypothèse voisine. Pour cela nous utilisons à nouveau la technique de report des écarts consécutifs, en appliquant des variations de v à partir du jeu d'effets obtenu grâce à l'I-API. En effet, alors que les trois autres paramètres de cette technique permettent d'étalonner les seuils de valeur des trois variables APC, v modifie l'allocation des effets nets suivant les différentes variables. De légères modifications de v impliquent donc de légères modifications de l'information a priori autour de l'I-API. Ainsi, nous définissons les intervalles de sensibilité IS_h , pour les effets $((a'_i), (p'_j), (c'_k), m')$ issus de l'I-API, comme les intervalles comprenant les valeurs prises par ces effets nets lors d'une variation de l'I-API d'une amplitude de h autour de sa valeur initiale.

Les bornes inférieures $((a^-_i), (p^-_j), (c^-_k), m^-)$ et supérieures $((a^+_i), (p^+_j), (c^+_k), m^+)$ pour chacun des effets nets sont définies de la manière suivante :

$$\begin{aligned} \text{Bornes inférieures des } IS_h : ((a^-_i), (p^-_j), (c^-_k), m^-) &= \varphi_{\alpha, \pi, \gamma, v=-h}((a'_i), (p'_j), (c'_k), m') \\ \text{Bornes supérieures des } IS_h : ((a^+_i), (p^+_j), (c^+_k), m^+) &= \varphi_{\alpha, \pi, \gamma, v=h}((a'_i), (p'_j), (c'_k), m') \end{aligned} \quad [25]$$

avec la définition de $\varphi_{\alpha, \pi, \gamma, v}$ donnée dans la formule [21].

Cet intervalle représente l'ensemble des valeurs prises par les effets nets pour une variation de h (souvent exprimée en pourcentage) de l'I-API. Un intervalle de robustesse à 1% signifie que l'on a fait varier les jeux d'effets nets initialement déduits de l'I-API à l'aide de reports des écarts consécutifs de plus ou moins 1% comme exprimé dans la formule [25]. L'intervalle de robustesse que nous voulons définir dépend de la grandeur h et donc de la distance que l'on veut autoriser par rapport à l'hypothèse initiale. Plus l'information a priori initiale (I-API) sera jugée plausible, plus faible sera la valeur du paramètre h .

A l'issue de cette troisième étape, la méthode APC-PLS nous a permis :

- d'estimer les effets en évitant les moindres carrés ;
- d'introduire une information a priori choisie (IAPI) ;
- d'examiner la sensibilité des résultats par rapport à l'IAPI.

Appliquons à présent cette méthode sur nos deux cas d'application concrets.

E. Application de la méthode APC-PLS sur deux exemples

E.1. Application au cas automobile en France

E.1.i. Première étape de la méthode APC-PLS

Avant toute chose nous avons discrétisé les variables d'âge, période et cohorte en autant de variables binaires qu'il existe de modalités différentes au sein de ces variables. Il n'est pas nécessaire d'éliminer des modalités comme dans les méthodes OLS parce que la régression PLS supporte la multicollinéarité parfaite. Nous avons ainsi des indicatrices d'âge variant de 20 à 80 ans (noté A_i à A_{80} pour les graphiques), des indicatrices de cohorte de 1914 à 1984 (resp. C_{1914} à C_{1984}), et des indicatrices de période d'observation de 1994 à 2004 (resp. P_{1994} à P_{2004}). Ces indicatrices conservent les notations présentées dans la sous-section D.1. à savoir A_i , P_j et C_k , i allant de 20 à 80, j de 1994 à 2004 et k de 1914 à 1984. Leurs effets nets seront notés en minuscules : a_i , p_j et c_k , i allant de 20 à 80, j de 1994 à 2004 et k de 1914 à 1984. Le taux de fidélité est quant à lui centré pour améliorer la lisibilité des résultats. Sa valeur moyenne est de 54.2%.

Nous appliquons ensuite la régression PLS sur les données ainsi préparées pour obtenir un premier jeu d'effets nets d'après la formule [17] de la sous-section D.5.ii :

$$Y =_{PLS} m + \sum_{i=20...80} a_i A_i + \sum_{j=1994...2004} p_j P_j + \sum_{k=1914...1984} c_k C_k + \varepsilon \quad [26]$$

Les indices de qualité de l'estimation sont très satisfaisants, avec un R^2 très élevé de 92,7%, ce qui est normal étant donné le nombre de prédicteurs employés. Le risque est plutôt ici de subir un sur-ajustement (over-fitting). Dans notre cas, l'utilisation du critère de PRESS dans le choix du nombre de composantes limite ce risque. Pour plus de précision sur ce mode

courant de gestion de l'over-fitting, se référer à Tenenhaus (1999). Ici l'on se restreint à une composante, car elle a un Q^2 de 86,7% et la seconde composante n'améliore pas le Q^2 cum de plus de 5%.

Enfin, pour vérifier la significativité des effets ainsi obtenus nous avons besoin d'intervalles de confiance sur ces effets calculés par bootstrapping avec 100 itérations en éliminant aléatoirement pour chaque itération 10% des points d'observation via XLStat (2008).

E.1.ii. Deuxième étape de la méthode APC-PLS

Pour obtenir un premier jeu d'effets nous devons formuler une information a priori initiale (IAPI) sur certains des effets APC. L'IAPI choisie est que les effets de périodes sont faibles sur l'espace de temps considéré. En effet, le niveau de fidélité dans un secteur de biens de consommation durables est un phénomène qui a une certaine inertie au cours du temps, et aucun événement économique majeur de type nouvelle régulation du marché ou arrivée massive de nouveaux entrants n'a pu être observé sur ce marché durant cette période. Il s'agit de l'information exogène la plus fiable que nous ayons sur les phénomènes de type APC concernant ce marché et c'est celle que nous souhaitons intégrer dans notre modélisation.

D'après la seconde étape de notre méthodologie APC-PLS, nous devons trouver une manière d'exprimer cette information a priori sous la forme d'une fonction F à optimiser. Dans notre cas, nous allons exprimer F sous la forme de la somme des carrés des nouveaux effets nets des indicatrices de périodes p_j , j de 1994 à 2004.

$$F = \sum_{j=1994 \dots 2004} p_j^2 \quad [27]$$

Sur laquelle nous appliquons le programme d'optimisation de la formule [23] avec le critère d'optimisation proposé dans la formule [22], basé sur la technique de report des écarts décrites dans la sous-section D.5.iii. Cette deuxième étape nous permet d'obtenir le jeu d'effets présenté dans les figures E.1.1. à E.1.3. suivantes.

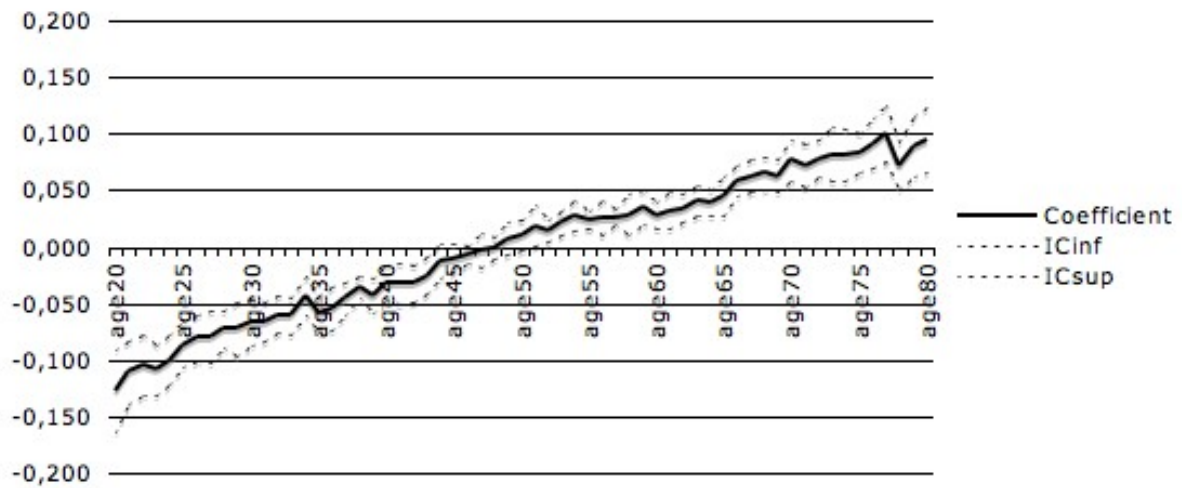


Figure E.1.1. Estimation des effets d'âge par la méthode APC-PLS

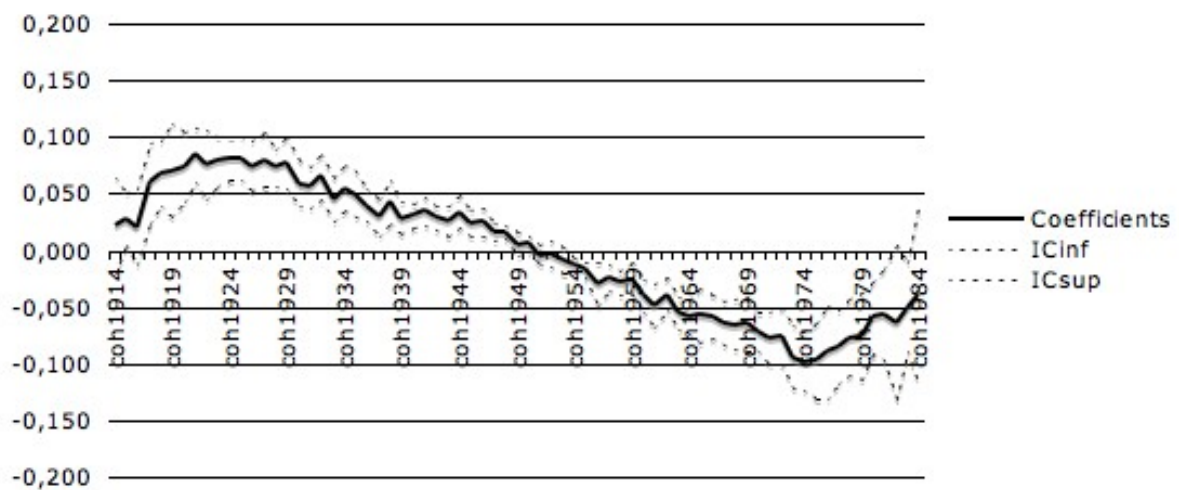


Figure E.1.2. Estimation des effets de cohorte par la méthode APC-PLS

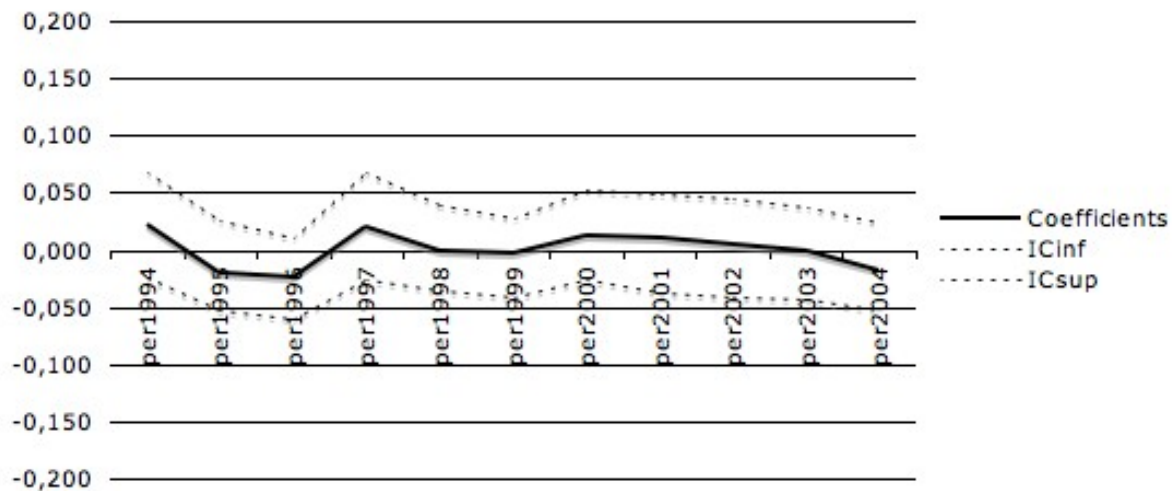


Figure E.1.3. Estimation des effets de période par la méthode APC-PLS³⁶

Notons ici que, contrairement à la plupart des autres méthodes précédemment citées, nous ne pouvons pas introduire l'information a priori selon laquelle tous les effets de période seraient nuls. Nous partons d'un jeu d'effets nets obtenu à l'issue de l'estimation PLS et ensuite, nous appliquons la méthode de report des écarts consécutifs qui nous permet de trouver un nouveau jeu d'effets nets correspondant au mieux à notre IAPI.

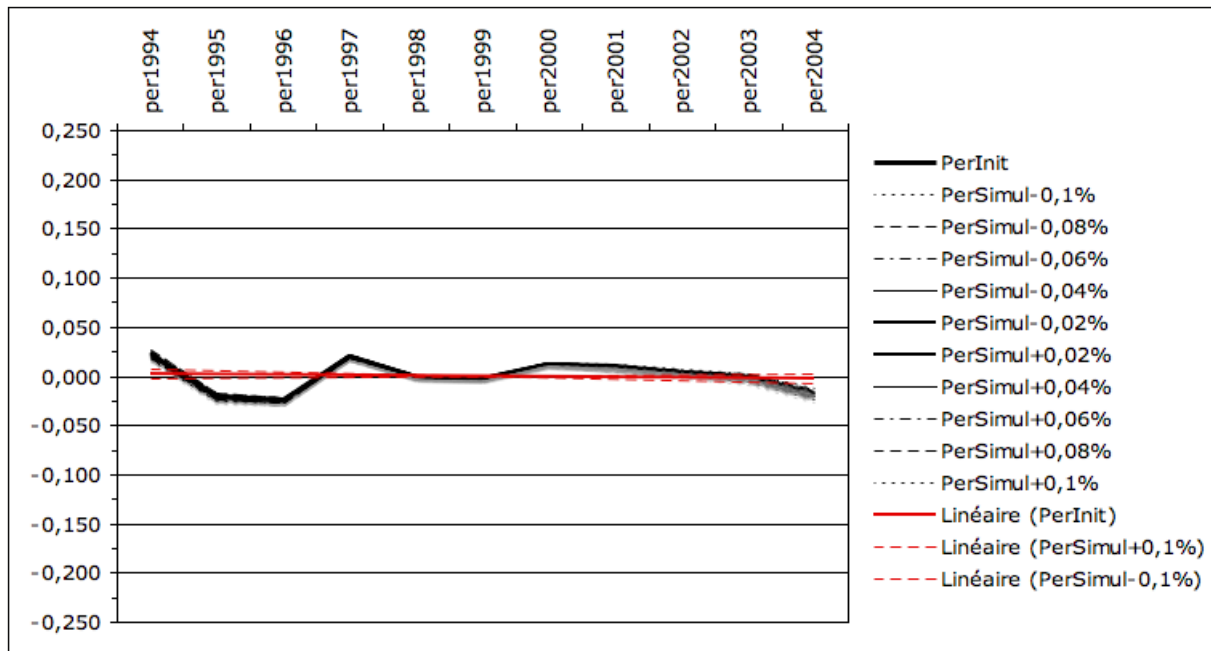
Dans ce cas précis, aucun des effets de période n'est au final significatif, et leur amplitude est très faible comparée aux effets d'âge et de cohorte trouvés à l'issue de cette seconde étape.

³⁶ Nous tenons ici à remercier le Professeur Sandrine Macé pour sa remarque durant la pré-soutenance de thèse du 20 mai 2008 concernant une erreur de calcul sur des intervalles de confiance pour ce graphique.

E.1.iii. Troisième étape de la méthode APC-PLS

A partir de ce nouveau jeu d'effets nets obtenu par l'introduction de l'IAPI, nous cherchons à vérifier si les résultats obtenus sont sensibles à la formulation de cette IAPI, c'est-à-dire si de légères variations autour de cette information initiale n'entraînent pas de larges modifications des estimations des autres effets, à savoir l'âge et la cohorte. En effet, comme nous l'avons vu dans l'étude des méthodes antérieures traitant l'approche APC, une variation, même minime, de l'information a priori peut avoir de lourdes conséquences sur l'estimation des effets de toutes les variables. Nous allons donc imposer de légères variations aux effets de période (notre IAPI étant qu'ils sont quasi-nuls) à travers le paramètre h de la technique de report des écarts et observer les conséquences de ces variations sur les valeurs des autres effets à l'aide des intervalles de sensibilité IS_h décrits dans la section D.5.iv.

Nous allons calculer les intervalles de sensibilité avec des niveaux de variation h égaux à 0,02% ; 0,04% ; 0,06% ; 0,08% et 0,1%. Les nouveaux jeux d'estimateurs sont décrits, par variable, sur les figures E.1.4. à E.1.6. suivantes.



³⁷ L'échelle a été choisie pour être identique pour les trois variables APC.

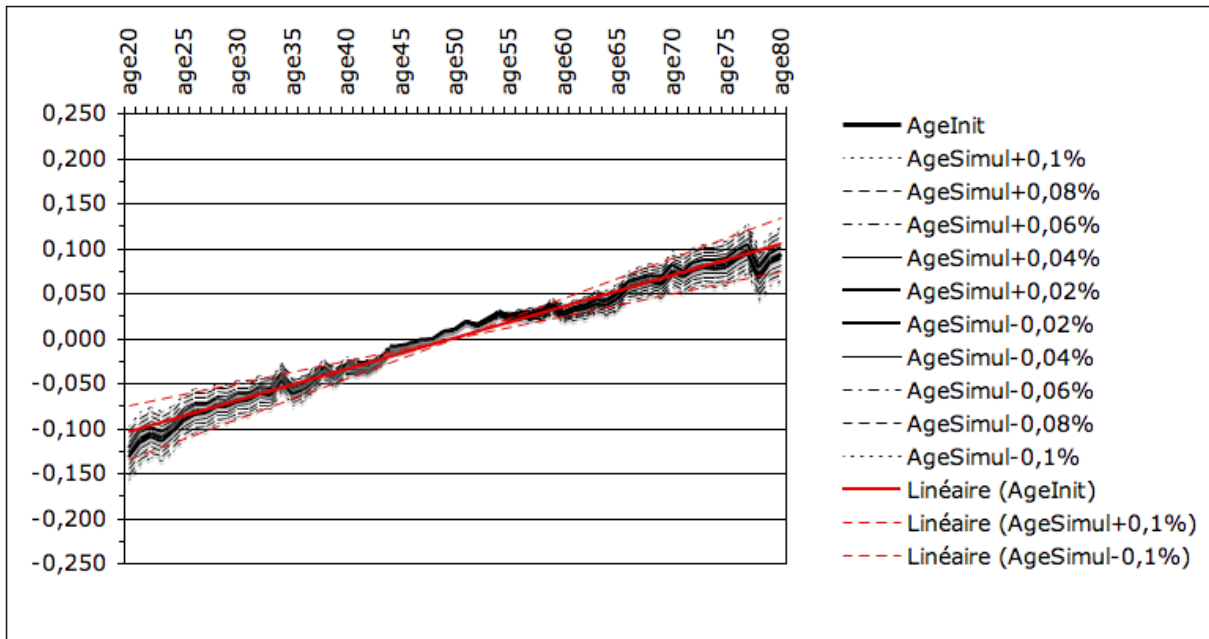


Figure E.1.5. Intervalle de sensibilité des effets d'âge
à 0,02% ; 0,04% ; 0,06% ; 0,08% et 0,1%

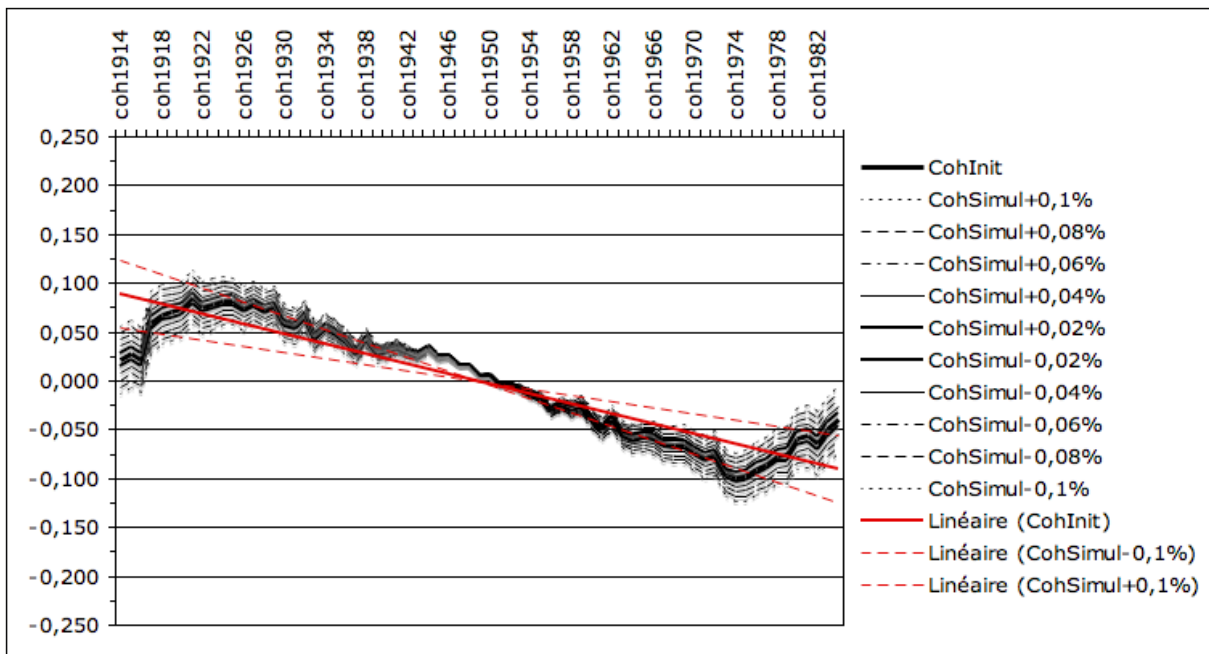


Figure E.1.6. Intervalle de sensibilité des effets de cohorte
à 0,02% ; 0,04% ; 0,06% ; 0,08% et 0,1%

Les figures E.1.4 et E.1.6. indiquent tout d'abord que les variations appliquées aux effets de période ne perturbent pas trop les tendances observées sur les effets d'âge et de cohorte (du moins en termes d'amplitude et de direction). Cependant ces figures témoignent du fait qu'une tendance plus marquée à la baisse dans les effets de période entraîne une diminution

de l'amplitude des effets d'âge et une augmentation de celle des effets de cohorte. Inversement, une tendance plus marquée des effets de période à la hausse entraîne une diminution de l'amplitude des effets de cohorte et une augmentation de celle des effets d'âge. Ainsi on constate une légère modification des amplitudes relatives des effets de ces variables: lorsqu'on prend des effets de période croissants, l'effet de l'âge aura un effet accru aux dépens de l'effet de cohorte ; et réciproquement des effets de période décroissants auront tendance à favoriser des effets de cohorte aux dépens des effets d'âge.

Des intervalles de sensibilité complémentaires sont proposés en annexe 6. Les résultats de cette analyse APC-PLS seront analysés dans la section F de cette recherche, après une nouvelle application de la méthode sur le cas allemand, dans la sous-section E.2. suivante.

E.2. Application au cas automobile en Allemagne

Nous souhaitons pouvoir valider notre méthodologie et nos résultats sur une autre base de données. Le marché allemand de l'automobile a le mérite d'avoir des caractéristiques légèrement différentes du marché français tout en nous permettant d'émettre les mêmes hypothèses et de les vérifier à nouveau sur un autre marché.

Mais le marché de l'automobile allemand a aussi ses caractéristiques propres. Les marques nationales leaders ne sont pas les mêmes et sont encore plus puissantes qu'en France. De plus, le marché allemand est resté fermé pendant plus longtemps après-guerre. Nous pensons donc que le taux de fidélité sera cette fois, au moins autant, si ce n'est plus influencé par la cohorte que par l'âge.

Concernant les effets bruts, nous avons des niveaux de fidélité supérieurs suivant les trois variables APC à ceux constatés sur le marché français. Les tendances sont similaires : augmentation du taux de fidélité avec l'âge, diminution du taux avec la cohorte, ainsi qu'une légère diminution avec le temps. Mais l'étude de ce nouveau cas APC nécessite aussi l'application d'une méthodologie adaptée capable de nous fournir les effets nets, et non bruts.

Pour les mêmes raisons que pour le marché français, nous choisissons comme IAPI des effets de période quasi-nuls. Du point de vue technique, les critères de qualité sont du même ordre que pour l'analyse précédente (R^2 de 0,935 et Q^2_{cum} de 0,874) et nous avons aussi

sélectionné une seule composante dans la régression PLS ($Q^2_{cum}(2)$ de 0,843). Enfin, les intervalles de sensibilité sont calculés pour des valeurs de h allant jusqu'à 0,1%. Les résultats issus de cette analyse sont présentés dans les figures E.2.1. à E.2.3. par type de variable APC.

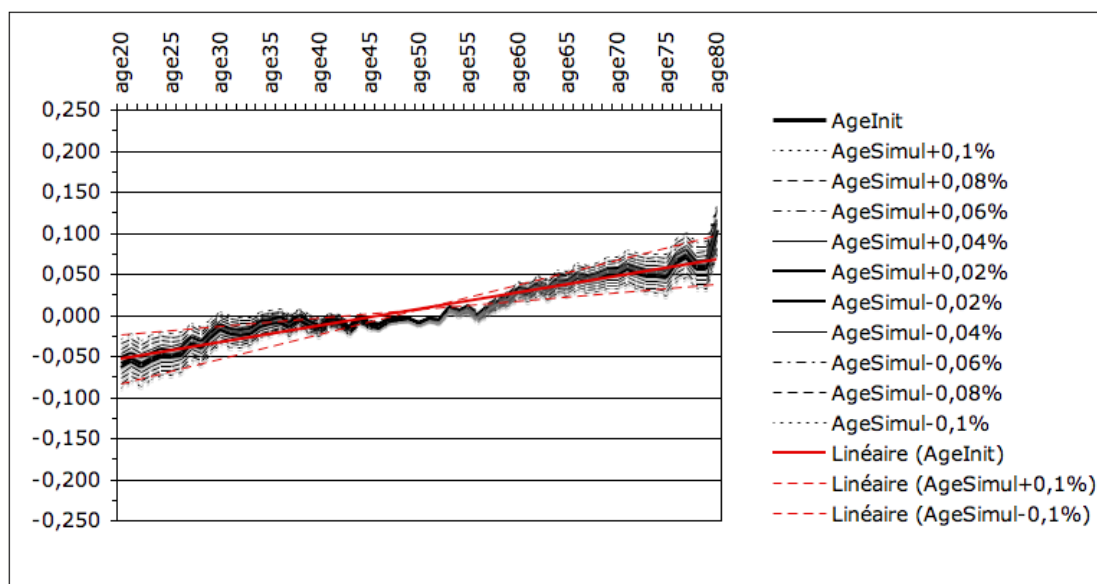


Figure E.2.1. Intervalle de sensibilité des effets d'âge
à 0,02% ; 0,04% ; 0,06% ; 0,08% et 0,1%

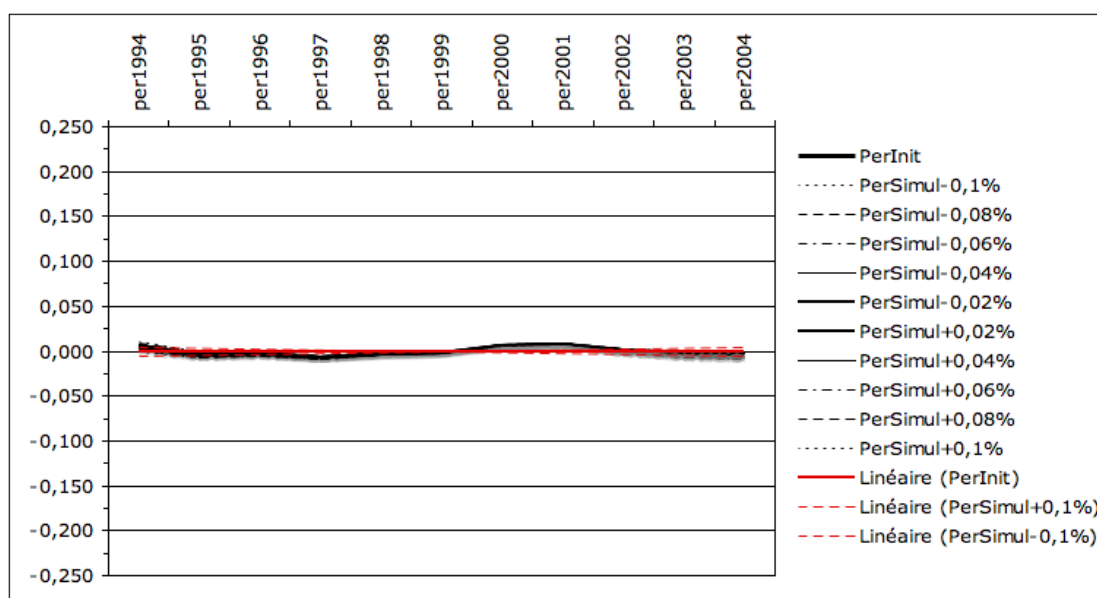


Figure E.2.2. Intervalle de sensibilité des effets de période
à 0,02% ; 0,04% ; 0,06% ; 0,08% et 0,1%³⁸

³⁸ L'échelle a été choisie pour être identique pour les trois variables APC.

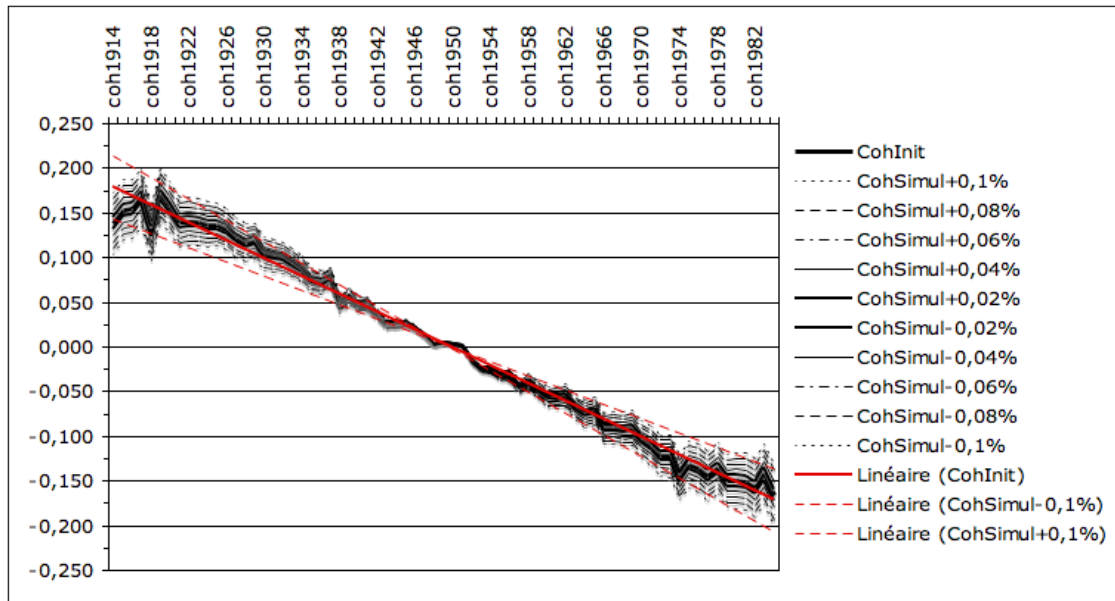


Figure E.2.3. Intervalle de sensibilité des effets de cohorte
à 0,02% ; 0,04% ; 0,06% ; 0,08% et 0,1%

Les effets d'âge sont intéressants dans le sens où ils évoluent de manière moins linéaire que dans les cas français. En effet, le niveau de fidélité est à peu près constant entre 35 et 55 ans (effets quasi nuls) et il augmente rapidement à partir de 55 ans. Comme prévu, une autre différence avec le cas français est que les effets de cohortes sont de plus grande amplitude (30%) que les effets d'âges (15%). Si l'on s'intéresse à la sensibilité des résultats, on peut constater que les tendances restent très similaires sur un intervalle de robustesse de 0,1%.

A l'issue de l'application de la méthode APC-PLS sur ces deux marchés qui clôture la partie méthodologique de cette recherche (avec la discussion proposée en section G sur ce sujet), nous allons à présent analyser les résultats ainsi obtenus au regard des hypothèses que nous avons élaborées sur les influences de l'âge (H1 de C.1.iii) et de la cohorte (H2 de C.2.iii.) dans la section F ci-après.

F. Retour sur les hypothèses

F.1. Rappel des hypothèses

Rappelons les hypothèses formulées au début de cette recherche.

(H1) En vieillissant, les consommateurs ont une plus grande tendance, quand ils achètent une voiture neuve, à ré-acheter un véhicule de la même marque que leur précédent véhicule.

(H2) A âge égal, les cohortes récentes ont une moins grande tendance, quand ils achètent une voiture neuve, à ré-acheter un véhicule de la même marque que leur précédent véhicule.

F.2. Validation des hypothèses dans le cas automobile en France

À partir des résultats de la section précédente, nous pouvons affirmer qu'en prenant comme information a priori le fait que les effets de période sont d'amplitude très réduite, les deux hypothèses (H1) et (H2) sont validées³⁹. En effet, on peut observer sur la figure E.1.5. que les effets d'âge ont un impact croissant sur la fidélité à la marque quand l'âge augmente, et que les effets de cohorte ont un impact décroissant sur la fidélité en direction des cohortes les plus récentes. Si l'on cherche à comparer les amplitudes des effets d'âge et de cohorte, l'âge atteint une amplitude maximale de 22% contre 18% pour la cohorte. Les deux effets ont donc des amplitudes comparables avec un avantage pour l'âge.

Ces résultats recourent les études antérieures sur le fait que l'interprétation de l'âge en termes de vieillissement ait un effet très important sur le comportement (ici sur la fidélité à une marque automobile), mais cet effet aurait été largement surestimé (presque doublé au regard des effets brut de la figure D.2.2.), si nous n'avions pas tenu compte des effets de cohorte.

Même si l'on ne s'intéresse qu'à l'impact de l'âge sur le comportement, il est donc essentiel, si les données le permettent, de contrôler l'impact des cohortes afin de ne pas surestimer l'effet de l'âge en lui attribuant tous les effets des cohortes.

Si l'on cherche à éprouver la sensibilité de ce résultat vis-à-vis des variations de l'information a priori choisie, on observe sur les graphiques de sensibilité des sections E.1. et E.2. que

³⁹ Nous rappelons ici que les effets de période sont considérés comme très réduits suivant les différentes période d'observation des achats, c'est-à-dire entre 1994 et 2004.

même d'assez fortes variations autour de l'IAPI ne changent pas la direction des résultats. Ceci signifie que même en postulant des effets croissants (ou décroissants) de la variable période sur la fidélité, les hypothèses concernant l'impact positif de l'âge et négatif de la cohorte sur la fidélité restent vérifiées.

F.3. Validation des hypothèses dans le cas automobile en Allemagne

Comme dans l'analyse du marché français, nous avons utilisé comme information a priori que les effets de période sur la décennie étudiée était de très petite amplitude. Et comme précédemment, nous avons envisagé d'éventuelles petites perturbations autour de cette information a priori.

Nous validons les deux hypothèses (H1) et (H2) car sur ce marché aussi :

- en vieillissant, les consommateurs ont davantage tendance à ré-acheter la même marque, ce qui était le résultat le plus prévisible puisque les effets de l'âge liés au vieillissement doivent être les mêmes quel que soit le marché considéré. Par rapport à la France, l'amplitude des effets d'âge est de 18% comme on peut le voir sur la figure E.2.1. ce qui est un peu plus faible que les 22% constaté sur la France. Ceci est bien entendu lié à l'importance relative des effets de cohorte en Allemagne ;
- les cohortes les plus récentes sont moins fidèles que les cohortes précédentes, et contrairement au cas de la France, l'amplitude des effets de cohorte observée sur la figure E.2.3. est supérieure à celle des effets d'âge à un niveau de 33% (contre 18% en France). Ceci peut s'expliquer par les récentes difficultés des marques nationales (depuis la fin des années 70) qui autrefois étaient prédominantes sur un marché allemand très protégé, combiné avec le fait que le marché allemand était particulièrement protectionniste à la sortie de la guerre (années 50, 60), ce qui a créé un très fort attachement potentiel chez les personnes issues des plus anciennes générations.

Au final, nos hypothèses sont vérifiées pour les deux marchés. Les effets sont équilibrés en France entre âge et cohorte, alors qu'en Allemagne la fidélité à la marque s'explique pour

deux tiers par la cohorte et un tiers par l'âge. Ces différences mettent en avant l'importance de la prise en compte de l'évolution historique des structures des marchés pour comprendre les comportements actuels et prévoir leurs évolutions.

Les sections suivantes visent à discuter cette recherche (section G) et à en montrer certaines limites (section H), avant de conclure cette étude.

G. Discussion

G.1. Comparaison des méthodes PCR, PACE, et APC-PLS

Nous proposons de récapituler les avantages et les défauts des méthodes PCR, PACE et APC-PLS pour mettre en évidence la contribution de notre méthode dans la modélisation des effets âge-période-cohore.

	PCR	PACE	APC-PLS
Vitesse d'estimation	Quasi-instantanée	Lente	Quasi-instantanée
Variance sur l'estimation des effets	Forte	Non disponible	Faible
Critère d'ajustement	OLS	OLS	PLS
Moment d'introduction de l'information a priori	Pendant l'estimation	Pendant l'estimation	Après l'estimation
Choix de l'information a priori utilisée	Impossible (information inconnue)	Impossible (critère de parcimonie)	Possible
Contrôle de la robustesse	Non disponible	Non disponible	Intervalle de sensibilité

Tableau G.1.1. Comparaison théorique des méthodes PCR, PACE et APC- PLS.

Notre méthode est donc la seule non directement fondée sur les moindres carrés permettant de garder le contrôle de l'information a priori que l'on souhaite introduire et de contrôler la sensibilité des résultats aux variations autour de cette information a priori à l'aide des intervalles de sensibilité. Ce critère est décisif dans le choix de la méthode car elle permet à l'utilisateur d'ajouter de l'information qu'il a pu recueillir par ailleurs au sujet des phénomènes observés. L'information a priori, qui est indispensable à un ajustement des résultats d'une modélisation APC peut ainsi être fixée par l'utilisateur et non plus imposée par la méthode d'estimation choisie. Nous revenons sur ce point dans la section G.2.

Par ailleurs, elle permet d'avoir aussi un meilleur contrôle, par rapport aux autres méthodes APC disponibles, sur l'erreur d'estimation en elle-même (erreur liée à l'information a priori mise à part), parce qu'elle propose des estimations des intervalles de confiance permettant de juger de la significativité des effets mesurés (contrairement à la méthode PACE) et que les variances des estimations des effets sont moins grandes que pour les méthodes dont l'estimation est fondée sur le moindres carrés (comme la méthode des contraintes d'égalité). De plus, l'information a priori est contrôlée (v.s. PCR ou PACE) et ne se limite pas à des contraintes d'égalité (comme dans les méthodes basées sur la régression linéaire) mais couvre toute information pouvant s'exprimer comme une fonction (des effets nets) à optimiser.

Parce qu'elle permet un meilleur contrôle de l'incertitude, à la fois au niveau de l'estimation (variance de effets estimés) et de l'information a priori (intervalle de sensibilité), tout en étant plus flexible que les autres méthodes proposées (choix de l'information a priori), nous pensons qu'elle apporte une contribution substantielle au problème de la modélisation APC.

G.2. Comparaison des résultats obtenus

Malgré les différences méthodologiques résumées dans la sous-section précédente, il est intéressant de noter que quelle que soit la méthode utilisée, nous avons des résultats concordants concernant la mesure des effets APC sur notre base de données.

Les graphiques G.2.1 à G.2.3. ci-après résument les résultats suivant les trois méthodes.

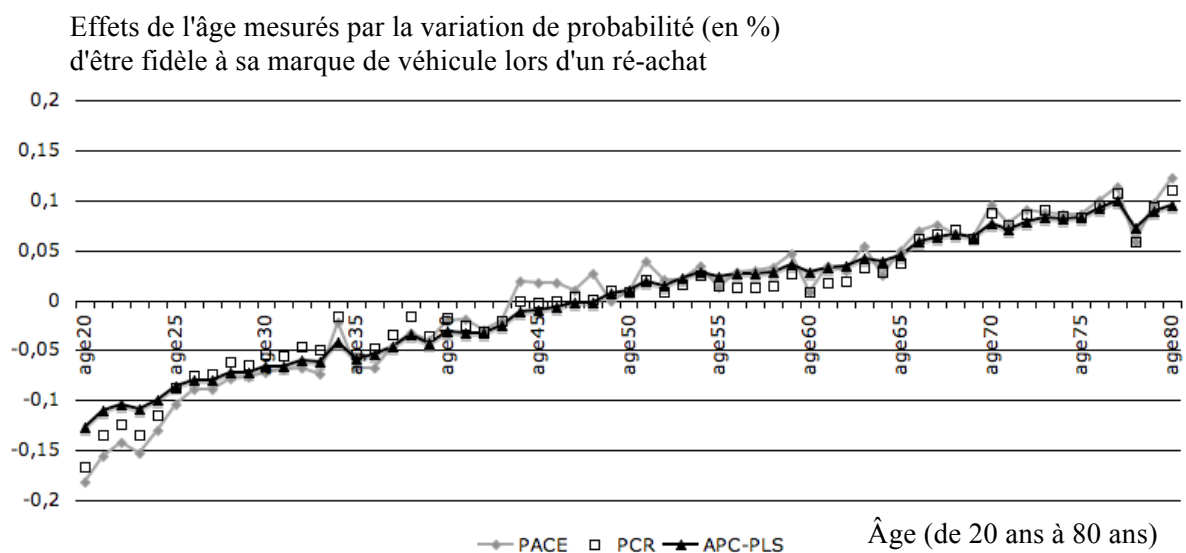


Figure G.2.1. Effets d'âge obtenus par les méthodes PACE, PCR et APC-PLS

Effets de période mesurés par la variation de probabilité (en %) d'être fidèle à sa marque de véhicule lors d'un ré-achat

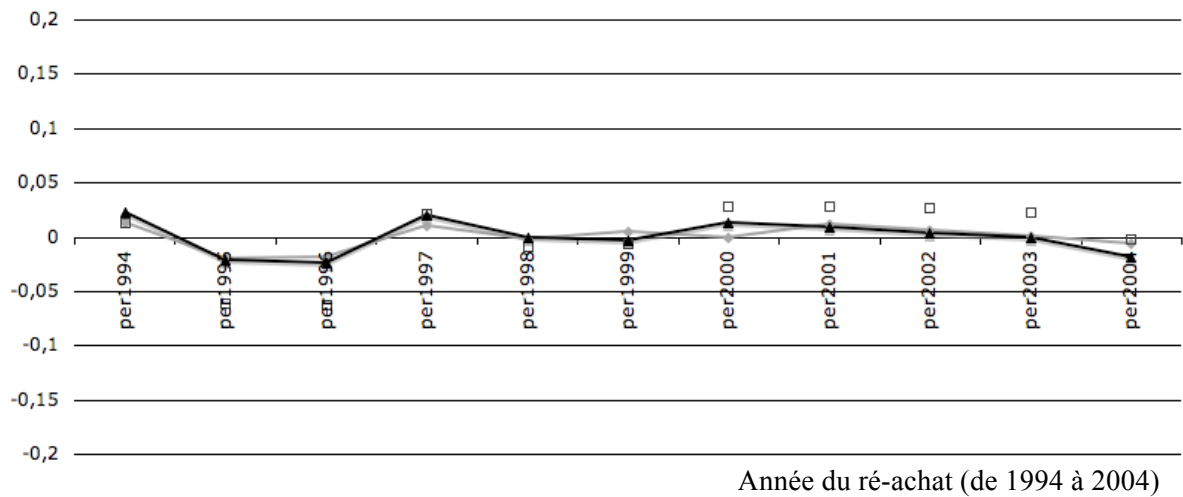


Figure G.2.2. Effets de période obtenus par les méthodes PACE, PCR et APC-PLS

Effets de cohorte mesurés par la variation de probabilité (en %) d'être fidèle à sa marque de véhicule lors d'un ré-achat

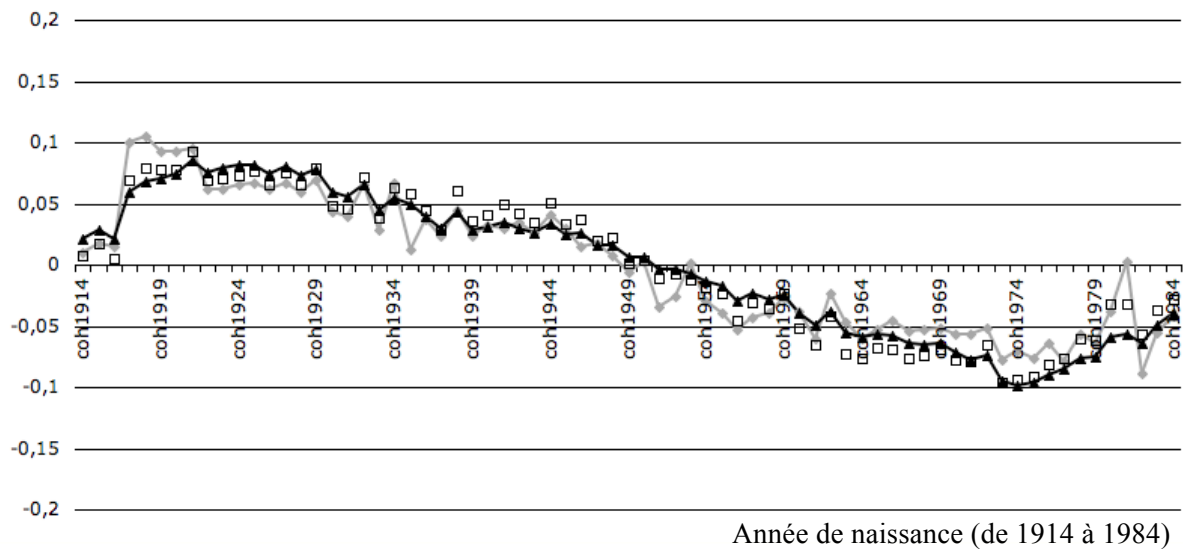


Figure G.2.3. Effets de cohorte obtenus par les méthodes PACE, PCR et APC-PLS

Cette concordance conforte notre confiance en la fiabilité des résultats proposés et donc sur la validation des hypothèses (H1) et (H2) vis-à-vis des influences respectives de l'âge et de la cohorte sur la fidélité à la marque. D'autant que des résultats comparables ont été trouvés à la fois sur les marchés français et allemand.

G.3. De l'usage de l'information a priori

Les méthodes antérieures présentées se répartissent en deux catégories.

- Les méthodes issues de la régression linéaire multiple à savoir l'omission d'une variable et les contraintes d'égalité entre effets successifs (et plus largement entre effets par le biais de redécoupage de variables en tranches de valeurs). Elles visent à contourner le problème de la multicollinéarité parfaite en introduisant l'information a priori *avant* l'estimation des effets, durant la phase de spécification. Pour l'omission d'une variable l'information a priori utilisée est « tous les effets de cette variable sont nuls » et pour les contraintes d'égalité, l'information a priori est « ces deux effets sont égaux ». Le problème de ces méthodes, comme on a pu l'observer, est que la colinéarité reste forte entre les variables restantes, ce qui ne permet pas d'appliquer l'estimation par les moindres carrés de manière adaptée ;

- Les méthodes utilisant un mode d'estimation différent des moindres carrés à savoir les méthodes PCR et PACE. Pour ces méthodes, aucune information a priori n'est introduite avant l'estimation (dans la phase de spécification) et aucune information ne peut être ajoutée après estimation (dans la phase de validation). En fait, l'information a priori est implicitement (PCR) ou explicitement (PACE) incorporée dans le mode d'estimation lui-même et ne peut en aucun cas être contrôlée par l'utilisateur. L'introduction de l'information a priori se fait donc *pendant* l'estimation elle-même.

En ce qui concerne notre méthode, l'APC-PLS, l'information a priori est introduite a posteriori, *après* estimation, durant la phase de validation. En effet, nous pensons que la multicollinéarité parfaite n'est :

- ni un problème de spécification à traiter avant l'estimation car toute hypothèse faite avant l'estimation va potentiellement biaiser ou dégrader l'estimation qui suivra comme on a pu l'observer sur les cas présentés ;

- ni un problème d'estimation car il existe des modes d'estimation autres que les moindres carrés permettant d'avoir une estimation plus précise et plus fiable que celle obtenue par les moindres carrés dans le cas de la multicollinéarité parfaite.

Nous considérons plutôt qu'il s'agit d'un problème de validation et d'interprétation des résultats, devant intervenir après estimation.

En effet, l'interprétation que l'on fera du premier jeu d'effets obtenu dépendra de l'information a priori que l'on veut utiliser, et la robustesse des nouveaux effets ainsi obtenue se fera en utilisant des informations a priori avoisinant l'information initialement utilisée.

Cette utilisation *a posteriori* de l'information a priori est une caractéristique très forte de notre méthodologie et n'a jamais, à notre connaissance, été proposée auparavant dans la littérature traitant de la modélisation APC.

La figure G.3.1. ci-dessous, sur la base d'un schéma de Maddala (2001, p. 15) présentant l'approche économétrique, résume les différences modes d'introduction de l'information a priori présentés ci-avant.

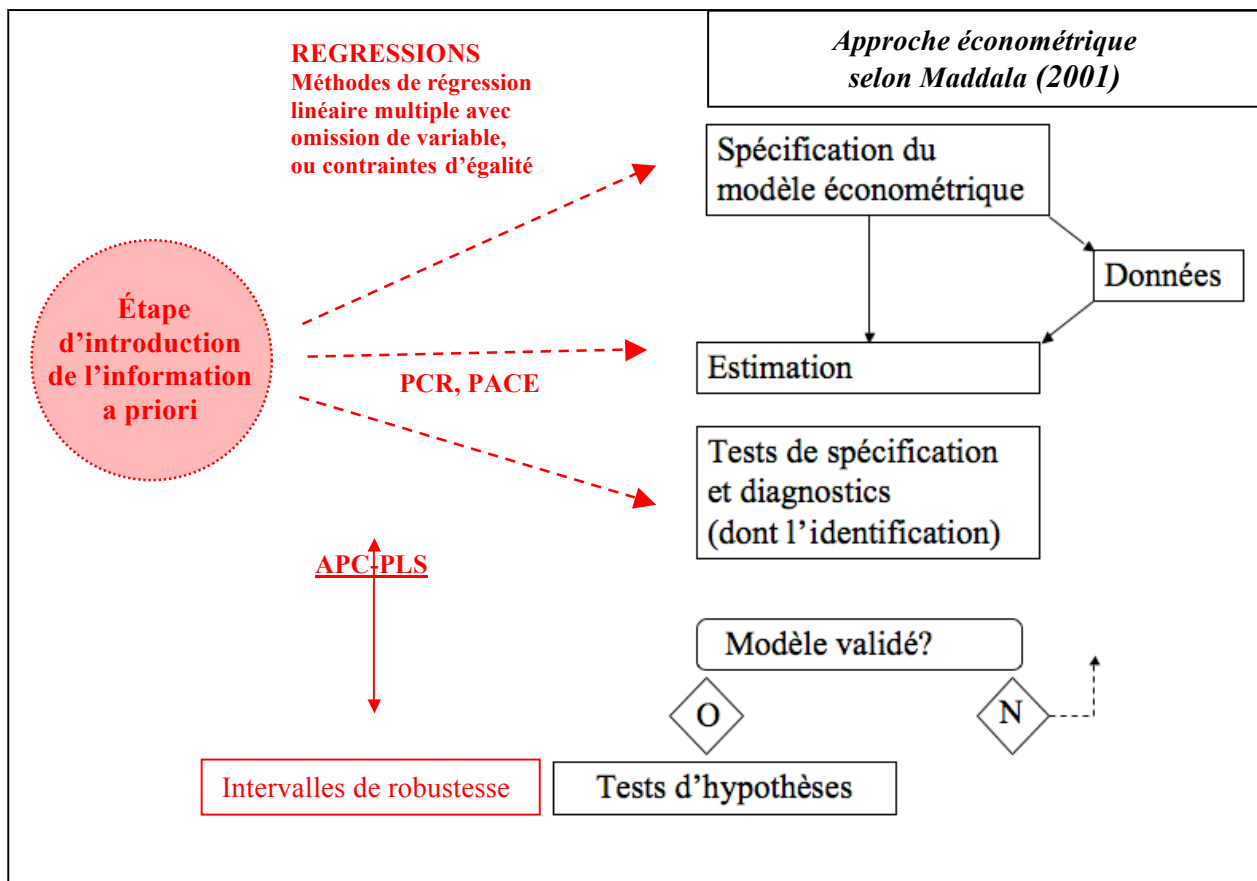


Figure G.3.1. Récapitulatif des modes d'introduction de l'information a priori suivant les différentes méthodologies présentées dans la thèse

G.4. Exemple complémentaire de la méthode de report des écarts

Pour introduire une information a priori, il faut lui donner l'expression d'une contrainte analytique. Cette contrainte analytique sera imposée dans le programme d'optimisation dépendant du paramètre d'optimisation v . On teste toutes les valeurs possibles et on en déduit les valeurs qui correspondent le mieux à la contrainte. Dans une approche similaire à la détermination de ρ dans les modèles de correction d'auto-corrélation, c'est par la procédure qu'on déduit v , qui n'est en aucun cas choisi directement par l'utilisateur.

Pour illustrer une nouvelle fois ce principe, appliquons le à une contrainte analytique intéressante qui est celle de Fukuda (2010).

Fukuda (2010) introduit en analyse APC un critère visant à privilégier les suites de valeurs des effets APC ayant un caractère lisse, c'est-à-dire avec des écarts successifs faibles. Ce critère peut être vu comme la vraisemblance d'un modèle où les écarts consécutifs sont distribués selon une loi normale centrée en 0, dont la variance est contrôlée à priori par des hyperparamètres dont le choix est guidé par l'explication recherchée :

$$\mathcal{F}(A, P, C) = \frac{1}{\sigma_A^2} \sum_{i=1}^{I-1} (A_i - A_{i+1})^2 + \frac{1}{\sigma_P^2} \sum_{j=1}^{J-1} (P_j - P_{j+1})^2 + \frac{1}{\sigma_C^2} \sum_{k=1}^{K-1} (C_k - C_{k+1})^2$$

Les variances sont ici des hyperparamètres non pas estimés à partir des données, mais au contraire choisis pour privilégier les solutions compatibles avec une information a priori (par exemple comme présenté dans la section 2.2 de l'article de Fukuda, on peut fixer la variance des effets d'âge à une valeur faible si l'on souhaite orienter le modèle vers une explication où les effets d'âge sont très lisses).

Comme indiqué à plusieurs reprises dans cette recherche, les problèmes APC ne peuvent être résolus qu'en introduisant une information a priori, et cette information a priori est arbitraire. Reste à l'utilisateur de la choisir au mieux à l'aide des informations exogènes à sa disposition. L'originalité de la méthode de Fukuda repose dans le fait que cette information arbitraire est introduite sous la forme des variances des effets. En appliquant la méthode APC-PLS, nous introduisons l'information a priori lors du choix du critère à optimiser (et v n'est que la

résultante mécanique de ce choix). Si le choix du critère en question se porte sur un critère de type Fukuda, l'utilisateur devra donc choisir les valeurs des variances de la fonction de Fukuda pour introduire l'information a priori désirée.

Lors de la mise en œuvre de la méthode de report des écarts successifs (celle présentée dans notre recherche), la valeur de v peut être choisie pour minimiser ce critère. Non pas directement ou a priori, mais en appliquant la méthode des écarts successifs de manière à ce qu'elle fournisse en sortie la valeur de v minimisant le critère de Fukuda, car encore une fois, v n'est pas une fin en soit, mais simplement un moyen pour optimiser un critère qui a lui une finalité directe. Soient A' , B' et C' les nouvelles valeurs des effets après application de la méthode des écarts successifs. Observons tout d'abord que le critère de Fukuda ne faisant intervenir que les écarts successifs, seul le choix de v (mais pas des nouvelles valeurs initiales ou finales a_1, p_j, c_k) aura une influence sur la valeur du critère après report des écarts. Cette nouvelle valeur sera :

$$\mathcal{F}(A', P', C') = \frac{1}{\sigma_A^2} \sum_{i=1}^{I-1} (A_i - A_{i+1} - v)^2 + \frac{1}{\sigma_P^2} \sum_{j=1}^{J-1} (P_j - P_{j+1} + v)^2 + \frac{1}{\sigma_C^2} \sum_{k=1}^{K-1} (C_k - C_{k+1} - v)^2$$

Cette expression est une fonction quadratique de v , convexe, que l'on peut directement minimiser :

$$\frac{d\mathcal{F}(A', P', C')}{dv} = \frac{2}{\sigma_A^2} ((I-1)v + (A_I - A_1)) + \frac{2}{\sigma_P^2} ((J-1)v - (P_J - P_1)) + \frac{2}{\sigma_C^2} ((K-1)v + (C_K - C_1))$$

On en déduit alors, par optimisation du critère de Fukuda, la valeur de v suivante :

$$v^* = \frac{-\frac{A_I - A_1}{\sigma_A^2} + \frac{P_J - P_1}{\sigma_P^2} - \frac{C_K - C_1}{\sigma_C^2}}{\frac{I-1}{\sigma_A^2} + \frac{J-1}{\sigma_P^2} + \frac{K-1}{\sigma_C^2}}$$

Dans cet exemple de mise en œuvre de la méthode sur le jeu de données France, Fukuda exige de choisir a priori des valeurs pour les variances des effets APC. A but illustratif, nous

choisissons ici pour les hyperparamètres de Fukuda le triplet de valeurs (7 ; 0,5 ; 1) pour les variances des effets respectifs de A, P et C.

La fonction donnant la valeur (à minimiser) du critère de Fukuda en fonction de la valeur de v est représentée sur la figure G.4.1. ci-dessous.

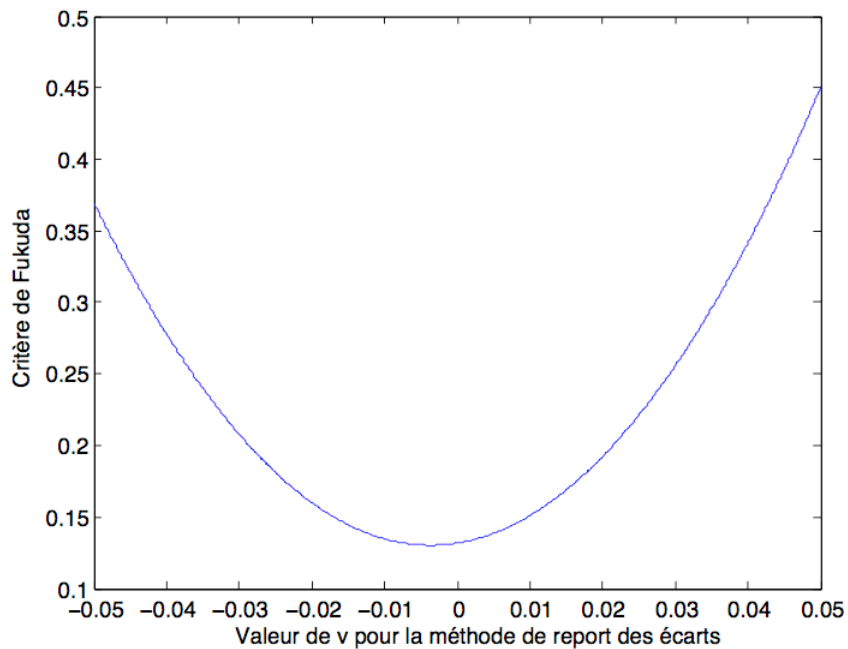


Figure G.4.1. Minimisation du critère de Fukuda

Les effets obtenus avant application du critère de Fukuda (première ligne de graphiques), c'est-à-dire à l'issue de l'application de la régression PLS⁴⁰ qui constitue la première étape de la méthodologie APC-PLS, et après application du critère de Fukuda, seconde étape de la méthodologie APC-PLS (seconde ligne de graphiques), sont représentés sur la figure G.4.2. ci-dessous.

⁴⁰ A ne pas confondre avec l'application complète de la méthode APC-PLS.

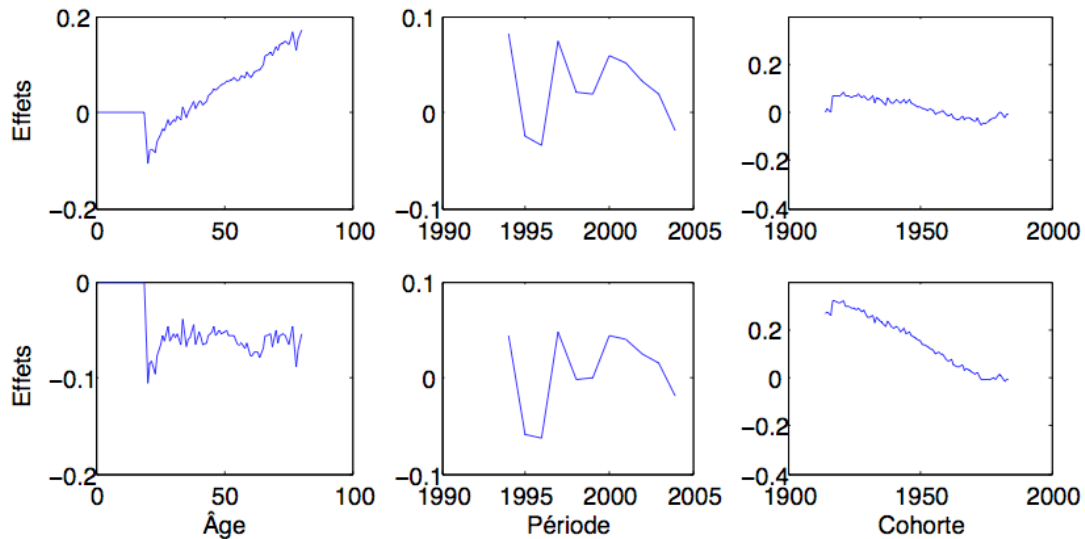


Figure G.4.2. Effets APC avant (première ligne) et après (seconde ligne) minimisation du critère de Fukuda

On observe qu'après optimisation par le critère de Fukuda, les effets d'âge ont été aplatis (ainsi que, très légèrement, ceux des périodes), avec un report sur les effets de cohorte.

On donne dans le tableau G.4.1. ci-dessous, l'écart type de la différence entre les effets d'une année sur l'autre pour les variables APC.

Méthode	Écart-type de ΔA	Écart-type de ΔP	Écart-type de ΔC
PLS	0.014	0.057	0.014
PLS + report des écarts	0.012	0.056	0.015

Tableau G.4.1. Illustration de l'optimisation du critère de Fukuda⁴¹

Cette application donne une illustration supplémentaire de la manière dont v n'est que déduit de la procédure d'optimisation d'un critère choisi par l'utilisateur pour contenir l'information a priori qu'il souhaite imposer au problème APC pour lui donner une solution.

Ce dernier exemple illustre encore une fois la flexibilité de la méthode APC-PLS permettant d'introduire l'information que l'on souhaite, quelle que soit sa forme, du moment qu'elle peut

⁴¹ Encore une fois, ne pas confondre dans ce tableau l'expression PLS (première étape de la méthode APC-PLS) et méthode APC-PLS.

s'exprimer sous la forme d'un critère analytique. Cette flexibilité, point fort de la méthode APC-PLS, réside dans le fait de n'introduire qu'a posteriori (après estimation PLS), l'information a priori.

G.5. Point sur l'interprétation de v dans la méthode de report des écarts

Le paramètre v est interprétable comme la variation absolue des valeurs des écarts entre effets consécutifs. Cette interprétation est assez lourde, mais ne constitue pas un handicap majeur étant donné qu'il s'agit simplement d'un paramètre d'optimisation, qui constitue un instrument de calcul, et non un élément que l'on souhaite interpréter en soi. Le paramètre v n'est jamais fixé a priori, mais il est issu d'une procédure d'optimisation, suivant un critère qui doit lui, par contre être clairement interprétable. Le paramètre v , n'est qu'un moyen, pas une fin.

Par exemple, si on a pour objectif de minimiser les effets de période, il n'existe souvent pas de v correspondant à des effets de période nuls. Si la contrainte analytique consiste à minimiser la somme des valeurs absolues des effets de période, alors on aura des effets de taille minimale, ce qui ne signifie pas nécessairement de taille nulle (comme dans la régression linéaire, la minimisation des moindres carrés n'implique pas l'annulation de la valeur des termes d'erreur). Le seul cas extrêmement improbable pour lequel la minimisation des effets de période aurait un résultat équivalent à l'approche consistant à supprimer tous les effets de période a priori, est celui pour lequel à l'issue de la régression PLS les coefficients de la variable P auraient des valeurs parfaitement proportionnelles.

Cela ne revient donc pas, ni en théorie, ni en pratique, à supprimer la variable période. En théorie, parce que la philosophie de la méthode APC-PLS consiste à ne pas introduire d'information avant l'estimation, mais après. En pratique, parce que l'information sur la variable période ne peut être manipulée qu'à travers une contrainte analytique, et non à travers la spécification du modèle.

H. Limites et futures recherches

H.1. Information a priori utilisée

Comme dans toute approche APC, la validité des résultats dépend en grande partie de la qualité de l'information utilisée. Par rapport aux informations exogènes sur ce marché à notre disposition, l'information choisie, sur les deux marchés considérés, à été que les effets de période entre 1994 et 2004 pouvaient être considérés comme faibles vis-à-vis de la fidélité à la marque. En effet, il n'y a pas eu d'apparition de nouvelles marques fortes, ou de disparition de marques importantes sur ce marché, en comparaison de ce que l'on avait pu observer par le passé dans le tableau C.2.2. avec les arrivées successives des marques européennes, puis asiatiques, dans les décennies 60, 70 et 80. De plus, comme indiqué dans la sous-section E.1.ii, aucun gros choc économique ou nouvelle régulation susceptible de provoquer des changements de gamme et/ou de type de véhicules (et donc indirectement de marque de véhicule) n'a pu être observé durant cette époque. Malgré le soin apporté à la sélection de cette information, celle-ci n'est pas vérifiable et constitue donc une limite intrinsèque à toute les modélisations APC.

Dans les années à venir, avec la multiplication des incitations fiscales en faveur de véhicules moins polluants et l'arrivée de nouveaux types de motorisation (comme l'hybride ou le diesel), les rapports de force entre les constructeurs pourraient changer en Europe. Ainsi les raisons de changement de comportements vis-à-vis de la fidélité à la marque pourraient avoir des liens forts avec la période considérée, indépendants des effets d'âge et de cohorte, qui devraient être pris en compte dans l'information a priori utilisée. Dans ce cas l'information a priori sur les effets de période très faibles ne pourrait probablement pas être utilisée telle quelle.

H.2. Non comparabilité des résultats

Les problématiques APC n'ayant pas de solution unique, un test empirique de comparaison entre méthodes n'est pas possible, ni sur des données réelles ni sur des données simulées par un expérimentateur.

Si l'on simule un jeu d'effets APC donnant des valeurs de Y par addition de ces effets et d'un bruit, les mêmes valeurs de Y peuvent être obtenues par une infinité d'autres jeux d'effets APC. Ainsi :

- Toutes ces explications possibles auront, bien entendu, un pouvoir explicatif équivalent (car fournissant exactement les mêmes évaluations de Y à partir des variables APC). On ne peut donc pas différencier les approches par la performance de prédiction.
- Le fait de retrouver parmi ces jeux d'effets (au pouvoir explicatif équivalent) le jeu initial d'effets choisi en premier lieu par l'expérimentateur ne constitue pas non plus une preuve de performance car ceci est simplement lié : soit au hasard, soit à la correspondance entre la méthode de choix de ce jeu par l'expérimentateur et la méthode d'estimation des effets (comme dans le cas de Rust and Yung dont l'expérimentation, bien qu'intéressante, ne peut de ce fait constituer une preuve).

Si l'on part de données réelles, le modèle sous-jacent n'est pas identifiable. Ce type de données ne peut donc pas non plus être utilisé à des fins de comparaison.

Pour ces deux raisons liées à la nature même de la problématique APC, les recherches en modélisation sur les méthodes APC ne présentent jamais de comparaison chiffrée. Ainsi, les comparaisons entre méthodes APC ne peuvent se faire que sur un plan méthodologique comme nous l'avons fait dans le corps de cette étude et dans la section H.1. précédente.

H.3. Age, période et cohorte : « facteurs » observables

L'âge, la période et la cohorte sont des variables directement observables, mais il n'en reste pas moins qu'elles ne sont qu'à mi-chemin entre le phénomène à expliquer et les causes à l'origine de ce phénomène. La cohorte par exemple est une composition de différentes influences subies simultanément par ses membres et peut ainsi être rapprochée d'un facteur latent « naturel », mais de nature formative (et non réflexive). Il en est bien entendu de même pour l'âge et pour la période.

Du point de vue de la prévision, il est très répandu d'utiliser des facteurs synthétiques ou des proxies. Du point de vue de l'interprétation, deux niveaux de lecture sont possibles. Il est tout à fait possible de raisonner sur les générations, pour comparer deux générations entre elles, ou différentes réactions à un même événement, parce qu'il s'agit d'un concept commun sur lequel tout le monde s'accorde et qui permet de raisonner à un niveau macro. Par contre, dans le cas d'une analyse d'un phénomène précis, à un moment précis, il peut-être intéressant de tenter de dégrouper les différentes composantes de l'identité générationnelle avec un cadre interprétatif plus éloquent.

Cette remarque peut aussi se traduire du point de vue de la modélisation comme le suggère Rogers (1982). Il serait en effet possible d'éliminer ce niveau intermédiaire, et de le remplacer par une série de proxies pour chacune des causes potentielles de l'effet générationnel. Mais cette approche alternative ne permet pas de se dédouaner du contrôle des effets complémentaires (âge et période pour les cohortes) afin d'éviter les biais de variables omises.

Cette approche est celle privilégiée par Boner (1999) dans le cadre de l'analyse du niveau et de la structure de consommation des français. Pour éliminer la variable période, il recherche une variable indiquant le contexte économique de chaque période (car il s'agit d'expliquer la consommation) et influençant également tous les âges et toutes les cohortes. La variable choisie est l'indice INSEE du niveau de vie des ménages. Comme le reconnaît l'auteur, cela revient à supposer que le seul effet de la période sur la consommation serait l'évolution des revenus, et ce pour tous les âges et toutes les périodes. Or cela implique de négliger les différences d'impact suivant l'âge, et entre générations, d'un choc conjoncturel ou encore les évolutions profondes des goûts ou des modes de consommation des ménages au cours du

temps, mais dans une optique d'économie générale, ces hypothèses semblent raisonnables. Cette approche est schématisée dans la figure H.3.1. ci-après.

De la même façon, les différentes composantes de l'âge (biologique, psychologique, sociale) peuvent avoir des effets différents (Rentz, Reynolds et Stout, 1983) et suivant l'objet de l'étude, il peut être intéressant de les distinguer. Ainsi il pourrait parfois paraître plus intéressant de remplacer la mesure de l'âge chronologique par un autre aspect du vieillissement mais cela serait très difficile en pratique au niveau du recueil des données. En effet, pour l'âge perçu, il est difficile d'obtenir des données car cela nécessite d'avoir utilisé lors du recueil un questionnaire adapté, ce qui exclut toute analyse de données secondaires.

Concernant le cycle de vie, cela revient à supposer que le découpage en tranches d'âges des différentes périodes de la vie reste constant suivant les différentes générations, ce qui suppose de passer outre des changements sociaux en la matière (report du départ du foyer pour les jeunes, allongement du temps de travail pour prendre des exemples récents).

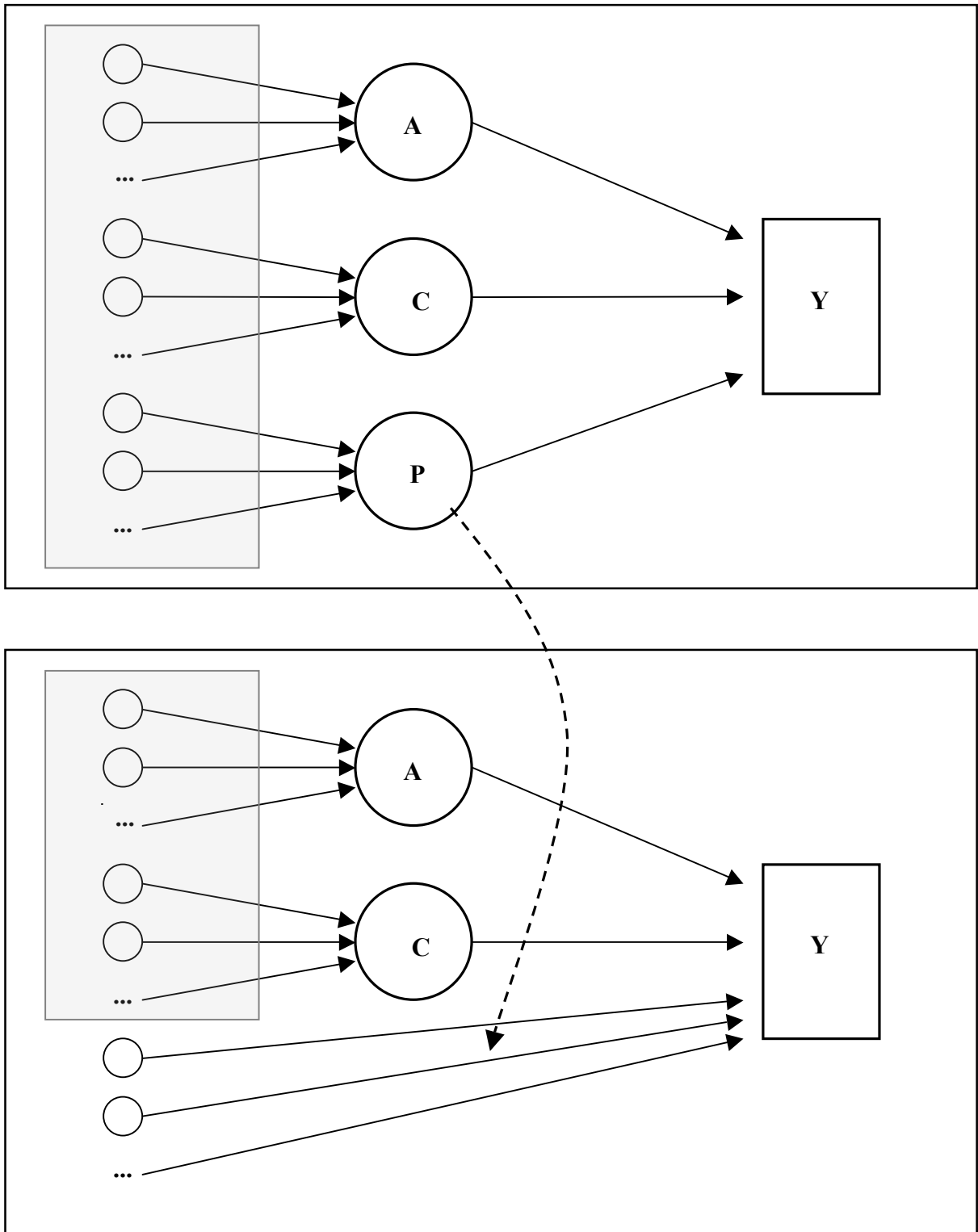


Figure H.3.1. Approche APC par l'utilisation des antécédents d'une des variables

Pour résumer, cette approche des « proxies » semble avoir trois limites :

- la variable période apparaît comme la variable la plus adaptée à l'utilisation de variables de substitution, ou proxies. Or, les deux variables restantes, l'âge et de la cohorte, seraient alors encore mécaniquement fortement colinéaires ;
- le choix de bons proxies reste compliqué pour les effets cohortes, étant donné le peu d'information dont nous disposons à ce jour sur les sources d'influence potentielles liées à la cohorte (c'est-à-dire liées au fait d'avoir connu aux mêmes âges certaines mêmes influences sociales, les mêmes événements historiques, etc) donc un travail de défrichage reste encore à mener sur le choix crucial des bons proxies pour assurer la qualité du modèle ;
- enfin cette approche rend quelque peu difficile la comparaison des effets de cohorte entre différentes études, car nous n'aurons pas nécessairement capté les effets des mêmes proxies, et que ces proxies ne captent qu'une partie des effets de cohorte.

Cependant, cette voie de recherche comporte probablement des ressources à explorer.

Dans cette optique, nous proposons notamment ici une méthodologie en deux temps qui permettrait de combiner une approche utilisant des proxies, avec une approche APC-PLS :

1 - Dans un premier temps, il faudrait utiliser des proxies de la cohorte pour contrôler les effets d'âge et de période, afin de tirer des données *une information a priori* la plus fiable possible (suivant la qualité des proxies) sur les effets d'âge et de période ;

2 - Dans un second temps, il serait alors possible d'utiliser dans la phase de calibrage de l'approche APC-PLS cette information a priori sur les effets d'âge et de période, afin de calibrer au mieux *a posteriori* le modèle obtenu par estimation PLS.

Cette méthodologie en deux temps semble pouvoir concilier le meilleur des deux approches en permettant d'avoir une idée fiable du schéma des effets d'âge et de période grâce aux proxies, puis d'utiliser cette information fiable dans une estimation optimale des effets APC fournis par l'approche APC-PLS. Cette méthodologie sera le sujet de recherches à venir.

H.4. Extension des résultats à des biens de consommation courante

Le marché de l'automobile a été choisi pour les nombreux avantages évoqués dans la sous-section B.2.ii. Ce choix fournit des résultats forts sur le lien entre les variables APC et la fidélité à la marque pour des biens de consommation durables.

À l'avenir, il serait possible de reproduire cette étude pour analyser ce lien pour des produits de consommation courante et de comparer les résultats avec ceux obtenus pour les biens durables. En effet, on peut se demander si dans ce cas les effets APC ont toujours autant d'impact et si les effets de cohorte ne seront pas plus faibles comparés à ceux de l'âge qu'ils ne le sont dans le cas précédent. De plus, cela donnerait un autre contexte dans lequel appliquer la méthode APC-PLS.

Plus généralement, les contextes d'application de ce type d'étude sont nombreux et permettraient d'éclairer sous un nouveau jour les résultats que nous connaissons sur l'effet de l'âge sur les comportements de consommation en les contrôlant et en les comparant aux effets de cohorte. Niveau de satisfaction, taux d'adoption d'un nouveau produit, niveau d'expertise dans une catégorie de produits sont autant de variables qu'il serait intéressant de substituer à notre taux de fidélité pour analyser des dynamiques temporelles des comportements de consommation.

H.5. Substitutions possibles de la technique PLS

Notre approche méthodologique consiste à découpler l'estimation et l'introduction d'une information a priori, et elle a été rendue possible par la technique PLS. Une autre technique, celle de l'Intrinsic Estimator, permet ce type d'estimation, par l'intermédiaire d'une composante principale, et elle a été utilisée par Yang et alii (2004) et (2008) dans le cadre d'analyses APC. Leur approche ne prévoit pas l'introduction d'une information a priori dans une deuxième étape, mais il est tout à fait possible d'utiliser notre méthode des écarts consécutifs pour appliquer la même procédure que pour APC-PLS, à partir d'un jeu de d'effets issu d'une estimation IE. Nous souhaitons à l'avenir pouvoir inventorier les autres méthodes d'estimation pouvant gérer la multicolinéarité parfaite, afin de réaliser des études comparatives sur les résultats finaux obtenus par notre méthodologie en trois temps.

Conclusion sur la méthode APC-PLS

A l'issue de cette recherche, nous sommes convaincus que les dynamiques générationnelles sont encore trop peu étudiées en marketing, notamment par rapport à des disciplines connexes comme l'économie ou la sociologie. D'autant plus qu'en période de papy-boom dans les pays occidentaux et de transition démographique dans les nouveaux pays émergents, la compréhension de l'impact des générations sur les choix de consommation constitue un des éléments majeurs de la réussite d'une stratégie marketing.

La difficulté à mesurer simultanément les effets d'âge, de période et de cohorte, liée à la multicollinéarité parfaite entre ces trois variables, pour mieux comprendre l'impact de facteurs temporels sur la fidélité à la marque nous a amené à mettre en place une méthodologie particulière, en rupture avec les méthodologies précédentes. Sa principale particularité réside dans le fait d'utiliser a posteriori, l'information dite a priori, que l'on peut obtenir sur le phénomène étudié. Cela nous a permis d'obtenir une plus grande souplesse d'utilisation et d'interprétation par rapport aux méthodes précédentes. Cependant, nulle méthode ne peut prétendre régler complètement le problème de la multicollinéarité parfaite qui empêche toute détermination d'une solution unique. Pour cette raison, nous avons pris soin d'introduire dans notre méthodologie le concept d'intervalle de sensibilité, permettant de prendre en compte, autant que possible, l'incertitude intrinsèque à ce type de problématique.

Au-delà de ces aspects méthodologiques, cette étude nous a permis de mesurer l'effet relatif de la cohorte par rapport à l'âge au sujet de la fidélité à la marque et d'ainsi de nuancer les effets d'âge, souvent surestimés, par le fait que les effets de cohortes sont le plus souvent ignorés. La composante générationnelle s'est avérée ici essentielle dans l'analyse de la fidélité à une marque automobile (d'une amplitude à peu près égale à la moitié de celle de l'âge dans le cas de la France, et supérieure à celle de l'âge dans le cas de l'Allemagne). Ce résultat est un exemple du pouvoir explicatif de la génération sur les comportements de consommation.

Nous espérons donc, autant d'un point de vue méthodologique que théorique, avoir contribué à faire avancer ce thème qui nous semble essentiel, autant du point de vue de la recherche académique que de la gestion marketing.

Conclusion générale

Au cours de ces travaux de thèse, nous avons pu mesurer l'étendue et la qualité des travaux de recherche dans le champ de la modélisation marketing au cours des dernières décennies. Humblement, nous espérons que les résultats présentés dans cette thèse contribueront à aider le praticien dans certains de ses choix de modélisation.

Notre question de recherche était la suivante :

Comment compléter avantageusement les techniques de modélisation actuelles afin de révéler et de mesurer les effets masqués que sont les effets locaux dans les fonctions de choix binaires et les effets de génération dans l'analyse temporelle des comportements de consommation ?

La problématique consistait donc « démasquer » certains effets masqués par les méthodes traditionnelles et à proposer des moyens de compléter ces méthodes pour extraire un maximum d'informations à partir des données initialement disponibles.

Cette problématique nous a amené à développer deux nouvelles méthodologies, souvent au prix de longs développements mathématiques, et dans des champs connexes à la modélisation marketing. Ces méthodologies AEL et APC-PLS nous ont permis de mettre en évidence le caractère local de certains critères de choix de consommation, notamment dans le cadre de l'analyse des switchers, pour le premier essai ; et de mettre en balance les effets d'âge et de génération dans l'étude de la fidélité à la marque, pour le second essai. Les applications de ces méthodologies sur différentes études de cas ont permis d'illustrer ces méthodologies de manière concrète, et d'apporter de nouveaux éléments dans la connaissance et l'explication des comportements de consommation.

L'analyse des effets locaux a permis de mettre en évidence comment chaque critère de choix peut avoir un effet différent suivant les probabilités de choix des consommateurs. Les méthodes de discrimination traditionnelles permettaient déjà de décrire quels étaient, en moyenne, et sur l'ensemble des individus, les critères de choix les plus importants pour prédire la décision des individus. Notre approche AEL permet elle d'aller plus loin dans l'analyse de ces résultats, car lorsque l'on passe de l'ensemble des individus, à des sous-

populations d'individus définies par les probabilités de choix, les effets des critères de choix peuvent avoir des ampleurs différenciées. La méthodologie proposée permet de mesurer ces différences et de les tester, afin d'apporter une connaissance plus fine de l'importance des critères suivant les types d'individus. Ces effets dits locaux, ont été mis en évidence sur deux bases de données, à savoir le choix entre deux marques nationales, et le choix de voter lors d'un référendum ; et les résultats obtenus ont des implications concrètes sur les critères à privilégier suivant les individus ciblés.

L'analyse Age-Période-Cohorte par la régression PLS (méthodologie APC-PLS) avait elle pour but de révéler le potentiel des effets de génération dans l'analyse temporelle des comportements de consommation et de proposer de nouvelles approches pour mesurer ces effets de génération sans négliger ni les effets d'âge, ni les effets de cohortes. Cette méthodologie propose d'utiliser de manière innovante la régression PLS pour sa capacité à estimer le lien entre le comportement à prédire, et des facteurs explicatifs parfaitement colinéaires (les effets d'âge, de période et de cohorte). La régression PLS est une méthode de modélisation de plus en plus répandue, mais elle n'avait pas encore été utilisée pour modéliser des phénomènes APC. Pour compléter avantageusement cette première étape d'estimation, la méthode de « report des écarts consécutifs » permet d'introduire, après estimation, une information dite « a priori » afin de mettre en cohérence les résultats obtenus avec les informations exogènes à disposition du praticien. La combinaison de ces deux étapes permet donc d'estimer le phénomène étudié sans avoir à modifier ou à dégrader les données initiales, tout en laissant la liberté au praticien d'introduire une information de son choix, du moment que cette information peut prendre la forme d'une expression fonctionnelle à minimiser.

Et au-delà de ces aspects méthodologiques, ce second essai traite en profondeur l'étude de l'évolution temporelle de la fidélité à la marque. Les résultats de cette étude mettent en évidence l'importance des effets de génération que nous cherchions à révéler en complément des résultats déjà connus sur les effets d'âge. Il en ressort que quelque soit la méthodologie employée, quelque soit le marché étudié, le poids de la composante générationnelle dans la fidélité à la marque ne peut être négligé, au risque de surestimer le poids réel des effets d'âge.

De plus, nous espérons avoir fait preuve de curiosité et de créativité, autant méthodologique que marketing, durant ces longues années de recherche.

Tout d'abord en intégrant de nouvelles méthodes, issues d'autres disciplines, dans la panoplie des outils actuellement disponibles en marketing. Nous prôtons en effet une hybridation des méthodologies statistiques. D'après cette expérience, les méthodes de l'économétrie et de l'apprentissage statistique ont, bien qu'aillant connu des développements autonomes, plus de points communs qu'il n'y paraît de prime abord. Il nous semble possible dans bien des cas de les combiner pour tirer de chacune d'elle le meilleur parti en fonction du problème à traiter. Dans un avenir proche, ces méthodes pourraient d'ailleurs se rapprocher et converger vers des méthodologies hybrides du type de celles présentées ici. Ceci demande de la part des modélisateurs une souplesse et une exigence accrues, mais ne peut que contribuer au progrès de la recherche en modélisation marketing.

Nous avons aussi proposé des sujets d'étude que nous jugeons importants pour l'évolution des pratiques marketing. Les effets de cohorte tout d'abord qui nous semblent essentiels à la compréhension des dynamiques sociales qui régissent en partie nos choix de consommation, et qui permettent d'enrichir les interprétations actuelles liées à l'âge des consommateurs. Les effets locaux ensuite, permettant de se détacher des stéréotypes liés à chaque segment et de distinguer une plus grande variété dans les fonctions de décision des consommateurs, en particulier des consommateurs à la marge, pour lesquels les leviers de décision sont moins facilement visibles. Dans les deux cas, il s'agit donc d'aborder les comportements du consommateur dans un contexte particulier : celui de leur génération d'une part, et celui de la prévisibilité de leur comportement d'autre part. Ces contextes avaient déjà été abordés suivant certains aspects par les recherches antérieures mises en valeur dans nos revues bibliographiques, mais nous pensons qu'ils ont encore beaucoup à nous révéler en permettant de compléter les explications plus traditionnelles sur les phénomènes étudiés.

Autre point commun majeur entre ces deux approches méthodologiques, elles fournissent une aide à l'interprétation a posteriori, par rapport à des premières étapes souvent assez usuelles. Cette approche a la particularité de ne pas aborder les problèmes en faisant table rase des habitudes en modélisation, mais de permettre au praticien d'enrichir ses premières analyses dans le cas où le sujet s'y prête, et d'ainsi pouvoir aller plus loin dans l'exploitation des résultats.

Cette caractéristique singulière s'inscrit dans une perspective de diagnostic a posteriori assez répandue en modélisation statistique, consistant à explorer les différentes facettes des modèles

proposés, d'en connaître ainsi d'autres apports potentiels, mais aussi les limites. Ces diagnostics reconsidèrent les résultats obtenus sous un nouveau jour et apportent de nouveaux éléments d'information, eux-mêmes soumis à des examens de contrôle, tels les tests d'égalité pour l'AEL et les analyses de sensibilité pour la méthodologie APC-PLS. Ainsi, de notre point de vue, ces deux effets apportent une contribution intéressante en termes d'interprétabilité, que nous avons essayé d'illustrer au mieux dans cette thèse.

Pour conclure cette thèse, nous tenions à affirmer notre enthousiasme pour aller plus loin dans l'exploration des thématiques abordées, car nous pensons qu'elles ouvrent de nouvelles perspectives de recherche passionnantes pour le futur.

Tout d'abord, d'un point de vue méthodologique, l'hybridation des méthodes pourrait amener à de nouveaux champs d'innovation. L'analyse FANOVA, permettant de considérer sous un angle non-linéaire des effets qui étaient souvent traités de manière linéaire par l'ANOVA, ou encore les SVMs, dont le pouvoir prédictif a déjà été démontré par de récents articles en marketing, mais dont le potentiel n'a encore été qu'effleuré en termes de modélisation, et beaucoup d'autres méthodes issues de l'apprentissage statistique, pourraient être comparées et combinées avec les outils économétriques plus traditionnellement utilisés en marketing.

Du point de vue de l'étude des comportements de consommation, les approches proposées ici peuvent s'appliquer à d'autres sujets de recherche, sur lesquels nous travaillons actuellement. L'AEL serait tout à fait adaptable l'analyse des choix multiples, en complément de l'analyse des choix binaires. Des comparaisons entre populations de « switchers » sur différentes catégories de produits, permettraient de comparer, d'une catégorie à une autre, les pratiques marketing à privilégier. Le poids relatif des effets de génération par rapport aux effets d'âge pourrait être étudié de manière plus systématique, notamment à travers une recherche reprenant sous cet angle une sélection d'articles de recherche portant sur les effets d'âge. Et dans un second temps, l'étude des facteurs permettant d'expliquer plus en détails la nature des effets d'âge ou de génération, amènerait à mieux comprendre pourquoi les poids relatifs de ces effets peuvent varier en fonction des phénomènes et des marchés analysés.

À l'avenir nous poursuivrons donc nos recherches sur ces sujets et espérons être capable de susciter l'intérêt d'autres chercheurs, afin de que de nouveaux effets soient démasqués, et que nouvelles hybridations méthodologiques soient proposées, testées et partagées.

Bibliographie

- Abdi, Hervé (2003), Partial Least Squares Regression, In : Lewis-Beck M., Bryman, A., Futing, T. (Eds), Encyclopedia of Social Sciences Research Methods, Thousand Oaks, Sage.
- Altmann R. (2003), Encouraging Savings Through the Life Cycle, Journal of Financial Services Marketing, 7, 4, 297-309.
- An, J. et Owen, A.B. (2001), Quasi-regression. Journal of Complexity, 17, 588–607.
- Bailey, M. et B. Ulman (2005), Trillion-Dollars Moms: Marketing to a New Generation of Mothers, Editions Kalman Business.
- Baker S.M. et Kennedy P.F. (1994), Death by Nostalgia : A Diagnostic of Context-Specific-Cases, Advances in Consumer Research, 21, ed. C. Allen et D. Roedder-John, Utah, Association of Consumer Research, 169-174.
- Barendregt J.J, Looman C.W.N et Bronnum-Hansen H. (2002), Comparaison de l'Intensité du Tabagisme par Cohorte au Danemark et aux Pays-Bas, Bulletin de l'Organisation Mondiale de la Santé, 80, 1, 26-32.
- Baudelot C. et Gollac M. (1995), Le Salaire du Trentenaire: Question d'Age ou de Génération, Economie et Statistique, 304/305, 17-36.
- Baudelot C. et Glaude M. (1990), Les diplômés paient-ils de moins en moins ? Étude de la relation salaire diplôme 1970-1977-1985, Données sociales, édition 1990, Insee, 103-108.
- Belk, R.W. (1990), The Role of Possessions in Constructing and Maintaining a Sense of Past, Advances in Consumer Research, 17, eds M.E. Goldberg, G. Gorn et R.W. Pollay, Provo, Utah, Association of Consumer Research, 669-676.
- Bennett, Kristin P., Erin J. Bredensteiner (2000), Duality and geometry in support vector machines, Proceedings of the 17th International Conference on Machine Learning, 65–72.

Bennett, Kristin P., Donghui Wu, Leonardo Auslender (1998), On support vector decision trees for database marketing. R.P.I Math Report No. 98-100. Rensselaer Polytechnic Institute, Troy, NY.

Berger (1989), Demographics Cycles, Cohort Size and Earnings, *Demography*, 26, 2, 311-321.

Böckenholt, U. and R.Langeheine (1996), Latent Change in Recurrent Choice Data, *Psychometrika*, 56, 285-302.

Bodier M. (1999), Les effets d'âge et de génération sur le Niveau et la Structure de la Consommation, *Economie et Statistique*, 31, 325/325, 163-180.

Bowman, Adrain W. et Adelchi Azzalini (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Oxford Statistical Science Series, Oxford University Press, USA (November 13, 1997).

Breiman, Leo. 2001. Statistical modeling: The two cultures, *Statistical Science*, 16, 199-231.

Brown, Mary and Carol Orsborn (2006), *Boom: Marketing to the Ultimate Power Consumer – The Baby Boom Woman*, Amacom.

Burges, Christopher J. C. (1998), A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 2(2), 121–167.

Burnett J.J. (1989), Retirement versus Age : Assessing the Efficacy of Retirement as a Segmentation Variable, *Journal of the Academy of Marketing Science*, 14, 4, 333-343.

Burnett J.J. (1991), Assessing the Patronage Behaviour of the Retired Male: a Comparison with Chronological Age, *International Review of Retail, Distribution and Consumer Research*, 6, 3, 243-258.

Calantone, R. J. and A. G. Sawyer (1978), The Stability of Benefit Segments, *Journal of Marketing Research*, 15, 395-404.

- Carlsson, Gosta and Katarina Karlsson (1970), Age, Cohorts and the Generation of Generations, *American Sociological Review*, 710-718.
- Carstensen L.L., Mikels J.A. et Mather M. (2005), Aging and the Intersection of Cognition, Motivation and Emotion, *Handbook of the Psychology of Aging*, eds. J. Birren et K.W. Schaie, San Diego, Academic Press, 6th edition.
- Cazès, P. (1997), Adaptation de la régression PLS au cas de la régression après analyse des correspondances multiples, *Revue de Statistique Appliquée*, 45 no. 2, 89-99.
- Chan, Tat, Chakravarthi Narasimhan, and Qin Zhang (2008), Decomposing Promotional Effects with a Dynamic Structural Model of Flexible Consumption, *Journal of Marketing Research*, 45, 487-498.
- Chandon, P., Wansink, B., & Laurent, G. (2000), A Congruency Framework of Sales Promotion Effectiveness, *Journal of Marketing*, 64,4, 65-81.
- Chauvel L. (1998), *Le Destin des Générations, Structure Sociale et Cohortes en France au XXème Siècle*, PUF.
- Chatiliez E. (2001), *Tangy*, Film, Coproduction Les Productions du Champ Poirier, TF1 Films Productions, TPS Cinéma, Télérama. Diffusion : TFA International.
- Choi, Beomjoon and Rohini Ahluwalia (2006), The Role of Consumer Inferences About Price Discounts in Influencing Switching Behavior. *Advances in Consumer Research*, Vol. 33.
- Cohen, Jacob et Stephen G. West (2003), *Applied Multiple Regression, Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates.
- Cole C. et Balasubramanian (1993), Age Differences in Consumers' Search for Information: Public Policy Implications, *Journal of Consumer Research*, 20, 6, 157-169.
- Cole C., G. Laurent, A. Drolet, J. Ebert, A. Gutchess, R. Lambert-Pandraud, E. Mullet, M. I. Norton et E. Peters (2008), *Decision Making and Brand Choice by Older Consumers*,

Marketing Letters, vol. 19, n° 3/4, pp. 355-365

Coleman, Linda Jane, Marie Hladikova and Maria Savelyeva (2006), The Baby Boomer Market, *Journal of Targeting, Measurement and Analysis for Marketing*, 14-3, 191-209.

Colombo, Richard A. and Donald G. Morrison (1989), A Brand Switching Model with Implications for Marketing Strategies, *Marketing Science*, 8,1.

Craney, Trevor A. and James G. Surlles (2002), Model-Dependent Variance Inflation Factor Cutoff Values, *Quality Engineering*, Vol. 14 Issue 3.

Cristianini, Nello, John Shawe-Taylor (2000), *An Introduction to Support Vector Machines – and Other Kernel-Based Learning Methods*. Cambridge University Press.

Critchley, Frank et Cecilia Vitiello (1991), The Influence of Observation on Misclassification Probability Estimates in Linear Discriminant Analysis, *Biometrika*, 78, 3, 667-690.

Cui, Dapeng and David Curry (2005), Prediction in Marketing Using the Support Vector Machine, *Marketing Science*, 24, 4, 595-615.

Cutler, Stephen J. and Robert L. Kaufman (1975), Cohort Changes in Political Attitudes: Tolerance of Ideological Nonconformity, *Public Opinion Quarterly*, 39-1 (Spring), 69-81.

Danigelis N.L et Cutler S.J (2007), Population Aging, Intracohort Aging, and Sociopolitical Attitudes, *American Sociological Review*, 72, 5, 812-830.

Divard, R. et P. Robert-Demontrond (1997), La nostalgie : un thème récent dans la recherche en marketing, *Recherche et Applications en Marketing*, Vol.12, 24, 41-62.

Dann S. (2007), Branded Generations: Baby Boomers Moving into the Senior Market, *Journal of Product and Brand Management*, 16, 6, 429-431.

Dawson S. et E. Spangenberg (1987), Television and the elderly: chronological ages as a prediction of viewing habits and attitudes toward advertising, *Advances in Consumer Research*, 14, 569.

Doherty Susan and Roy Nelson (2008), Customer loyalty to food retailers in Northern Ireland: 'devoted loyals' or 'promiscuous switchers'? *International Journal of Consumer Studies*, 32, 349-355.

Doksum K., S. Blyth, E. Bradlow, X.-L. Meng and H. Zhao, Correlation Curves as Local Measures of Variance Explained by Regression, *Journal of American Statistical Association*, 89, 426, 571-582.

Drolet A., Williams P. et Lau-Gesk L. (2007), Age-Related Differences in Responses to Affective vs rational Ads for Hedonics vs utilitarian Products, *Marketing Letters*, 18, 4, 211-221.

Egri, Carolyn P. and David A. Ralston (2004), Generation Cohorts and Personal Values: A Comparison of China and the United States, *Organization Science*, 15-2 (March-April), 210-220.

Evgeniou, Theodoros, Massimiliano Pontil, et André Elisseeff (2004), Leave One Out Error, Stability, and Generalization of Voting Combinations of Classifiers, *Machine Learning*, 5,1,71-97.

Evgeniou, Theodoros, Constantinos Boussios, Giorgos Zacharia (2005), Generalized robust conjoint estimation. *Journal of Marketing Science*, 24, 3, 415-429.

Farano, Ken, Tim Romberger, et David Meer (2009), Five Rules for Retailing in Recession, *Harvard Business Review*, 64-72.

Fine J. et Fotso S. (1989), Contribution à l'Etude du Modèle Age-Période-Cohorte, *Revue de Statistique Appliquée*, 37, 3, 39-56.

Fornell C. and J. Cha (1994), Partial Least Square, in *Advanced Methods in Marketing Research*, R. P. Bagozzi (Ed.), Blackwell, Cambridge, USA.

Foster, Dean P., Robert A. Stine, and Richard P. Waterman (1998), *Basic Business Statistics: a Casebook*. Springer.

Fukuda, Kosei, (2010), A Cohort Analysis of Household Vehicle Expenditure in the U.S. and Japan: a Possibility of Generational Marketing, *Marketing Letters*, 21, 1, 53-64.

Ganesh, Jaishankar, Arnold, Mark J. and Reynolds, Kristy E. (2000), Understanding the Customer Base of Service Providers: An Examination of the Differences Between Switchers and Stayers, *Journal of Marketing*, 64, 3, 65-87.

Garthwaite, Paul H. (1994), An Interpretation of Partial Least Squares, *Journal of the American Statistical Association*, March 1994, Vol. 89, No 425, 122-127.

Gauchi, J.P. (1995), Utilisation de la régression PLS pour l'analyse des plans d'expériences en chimie de formulation, *Revue de Statistique Appliquée*, 43, 1, 65- 89.

Guadagni, Peter M., John D.C. Little (1983), A logit Model of Brand Choice Calibrated on Scanner Data, *Marketing Science*, 2, 3, 203–238.

Guiot D. (2001), Tendances d'Age Subjectif: Quelle Validité Prédictive, *Recherche et Applications en Marketing*, 16, 1, 25-43.

Guiot D. (2006), Un Cadre d'Analyse du Processus de Vieillesse et de son Influence sur le Comportement d'Achat du Consommateur Agé, *Recherche et Applications en Marketing*, 21, 1, 57-79.

Guyon, Isabelle, David G. Stork. (1999), Linear discriminant and support vector classifiers. *Advances in Large Margin Classifiers*. MIT Press, 147–169.

Green, William H. (2003), *Econometric Analysis*. Upper Saddle River, New Jersey: Pearson Education Inc.

Hoefding W. (1948), A class of statistics with asymptotically normal distribution, *Annals of Mathematical Statistics*, 19, 293-325.

Holbrook, Morris B. and Robert M. Schindler (1989), Some Exploratory Findings on the Development of the Musical Tastes, *Journal of Consumer Research*, 16 (June), 119-124.

Holbrook, Morris B. and Robert M. Schindler (1994), Age, Sex and Attitude Toward the Past as Predictors of Consumers' Aesthetic Tastes for Cultural Products, *Journal of Marketing Research*, 31 (August), 412-422.

Hooker, Giles (2004), Discovering Additive Structure in Black Box Functions, KDD 04, August 22-25, 2004, Seattle, Washington, USA.

Huang, Jianhua Z. (1998), Projection Estimation in Multiple Regression with Application to Functional Anova Models, *The Annals of Statistics*, 26-1, 242-272.

Humby, Clive (2004), Leveraging returns from a \$100m investment in customer rewards, through communications, price, and promotions. Does Marketing Measure Up? Performance Metrics: Practices and Impacts, MSI Conference.

Hwang J.T. et D. Nettleton (2002), Principal Component Regression with Data-Chosen Components and Related Methods. Non publié.

Hotteling, H. (1929), Stability in Competition, *Economic Journal*, 39, 51-57.

Jacoby, J., Chestnut, R.W. (1978), *Brand Loyalty: Measurement and Management*, John Wiley & Sons, New York, NY.

James, Gareth M. (2002), Generalized Linear Models with Functional Predictors, *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 64-3, 411-432.

Jarque, Carlos M. & Anil K. Bera (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6-3, 255-259.

Jiang, Tao et Art B. Owen (2002), Quasi-regression for visualization and interpretation of black box functions, Stanford Working Paper.

Johnson, Lisa and Cheri Hanson (2006), Mind Your X's and Y's: Satisfying the 10 Carvings of a New Generation of Consumers, Free Press.

Kamakura, W. A., B. Kim and J. Lee (1996), Modeling Preference and Structural Heterogeneity, *Marketing Science*, 15, 152-172.

Kannan, P.K. and Gordon P. Wright (1991), Modeling and Testing Structured Markets : a Nested Logit Approach, *Marketing Science*, 10,1.

Kastenbaum R., Derbin V., Sabatini P. et Arrt S. (1972), The Ages of Me Toward Personal and Interpersonal Definitions of Functional Aging, *Aging and Human Development*, 3, 2, 197-211.

Kessler D. et Masson A. (1984), Petit Guide pour Décomposer l'Evolution d'un Phénomène en termes d'Effets d'Age, de Cohorte et de Moment, in *Cycles de vie et Générations*, Economica, Paris.

Kivetz, R., O. Netzer and V. Srinivasan (2004) Alternative Models for Capturing the Compromise Effect, *Journal of Marketing Research*, 41, 237-257.

Knoke. David and Michael Hout (1974), Social and Demographic Factors in American Political Party Affiliation, *American Sociological Review*, 39, 700-13.

Koças, Cenk and Jonathan D. Bohlmann (2008), Segmented Switchers and Retailer Pricing Strategies, *Journal of Marketing*, 72, 124-142.

Koubi M. (2003), Les Trajectoires Professionnelles: une Analyse par Cohorte, *Economie et Statistique*, 369, 119-147.

Krishnamurthi, Lakshman, Tridib Mazumdar, and S. P. Raj (1992), Asymmetric Response to Price in Consumer Brand Choice and Purchase Quantity Decisions, *Journal of Consumer Research*, Vol. 19.

Kritz G.H. et Arsenault P.M. (2006), Teaching Cohort Analysis: An Important Marketing Management Tool, *Marketing Management Review*, 16, 2, 37-43.

Kung, Wing K. (1996), Diagnosing Influential Observations in Quadratic Discriminant Analysis, *Biometrics*, 52, 4, 1235-1241.

Lambert-Pandraud, R. (2000), Le rachat de la marque précédente par les consommateurs âgés, Une synthèse des recherches en marketing, en psychologie sociale et cognitive, et en gérontologie, *Recherche et Applications en Marketing*, Volume 15, Numéro 4.

Lepisto L. (1997), The Adult Longitudinal Panel: A Research Program to Study the Aging process and its Effect on Consumers across the Life Span, Values, Lifestyles, and Psychographics, éd. L. Kahle, Mahwah, Lawrence Erlbaum, 317-333.

Li, Ker-Chau (1991), Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, 86 316-342.

Lilien, Gary L., Philip Kotler, K. Sridhar Moorthy. (1992), *Marketing Models*. Prentice-Hall.

Little, John D.C. (1970), "Models and Managers: The Concept of a Decision Calculus," *Management Science*, 16, No. 8, B-466-485 (April).

Maddala, G.S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*, *Econometric Society Monographs*, Cambridge University Press. 1983. 401 p.

Maddala, G.S. (2001), *Introduction to Econometrics*, Third Edition, John Wiley and Sons, NJ.

Marconi J. (2000), *Future Marketing : Targeting Seniors, Boomers, and Generations X and Y* Editions McGraw-Hill Companies.

Marquardt Donald W. and Ronald D. Snee, Ridge Regression in Practice, The American Statistician, Vol. 29, No. 1., 3-20.

Mason, K., Massy William F., Winsborough, H.H. and Poole, W. Kenneth (1973), Some Methodological Issues in Cohort Analysis of Archival Data, American Sociological Review, 38, 242-258.

Massy, William F. (1965), Principal Components Regression in Exploratory Statistical Research, Journal of the American Statistical Association, 60, 309, 234-256.

Mathur, Anil, Elaine Sherman, and Leon G. Schiffman (1998), Opportunities for Marketing Travel Services to New Age Elderly, Journal of Services Marketing, 12, 4, 265-277.

Matlab (2007), Matlab, The Optimization Toolbox 3, User's Guide, The MathWorks, Inc.

Meredith, Geoffrey and Charles D. Schewe (2002), Defining Markets, Defining Moments: America's 7 Generational Cohorts, their Shared Experiences and Why Business Should Care, John Wiley and Sons.

Moschis, George P. and Anil Mathur (2006), Older Consumer Responses to Marketing Stimuli: The Power of Subjective Age," Journal of Advertising Research, 46, 3, 339-346.

Naik, Prasad A., Michael R. Hagerty, and Chih-Ling Tsai (2000), A New Dimension Reduction Approach for Data-Rich Marketing Environments: Sliced Inverse Regression, Journal of Marketing Research, 37, 88-101.

Nalbantov, Georgi, Jan C. Bioch, Patrick J. F. Groenen (2006), Solving and Interpreting Binary Classification Problems in Marketing with SVMs, in From Data and Information Analysis to Knowledge Engineering, Springer.

Neugarten B.L. (1979), Time, Age and the Life Cycle, The American journal of Psychiatry, 136, 7, 887-894.

Nguyen C.L. et Roedder John D. (2007), Growing Up in a Material World : Age Differences in Materialism in Children and Adolescents, *Journal of Consumer Research*, 34, 4, 480-493.

Noble S.M. et Schewe C.D. (2003), Cohort Segmentation: An Exploration of its Validity, *Journal of Business Research*, 56, 12, 979-987.

Owen, Art B. (1992), Orthogonal arrays for computer experiments, integration and visualizations. *Statistica Sinica*, 2, 439–452.

Palmore, Erdman (1978), When Can Age, Period and Cohort be Separated ?, *Social Forces*, 57, 282-295.

Payne, A. and P.Frow. (1997), Relationship Marketing: Key Issues for the Utilities Sector, *Journal of Marketing Management*, 13, 463-477.

Peng, Leong Yow and Qing Wang (2006), Impact of Relationship Marketing Tactics (RMTs) on Swtichers and Stayers in a Competitive Service Industry, *Journal of Marketing Management*, 22, 25-59.

Perrien J., Recours F. et Hebel P. (2007), Evolution des Comportements Alimentaires : le Rôle des Générations, *Cahiers d’Economie et Sociologies Rurales*, 82/83, 79-108.

Platt, John C. (1999), Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods, Microsoft Research, Research Paper.

Platt, John C. (2000), Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*. MIT Press, 61–74.

Politz, A. and Deming, W.E. (1953), On the Necessity to Present Consumer Preferences as Predictions, *Journal of Marketing* 18,1,1–5.

Poulsen, C. S. (1990), Mixed Markov and Latent Markov Modelling Applied to Brand Choice Behavior, *International Journal of Research in Marketing*, 7, 5-19.

Ramaswamy, V. (1997), Evolutionary Preference Segmentation with Panel Survey Data : An Application to New Products, *International Journal of Research in Marketing*, 14, 57-80.

Rentz, Joseph O. and Fred D. Reynolds (1980), Separating Age, Cohort and Period Effects in Consumer Behavior, in *Advances in Consumer Research*, Vol. VIII, Kent B. Monroe, ed.

Reynolds, Fred D. and Joseph O. Rentz (1981), Cohort Analysis: An Aid to Strategic Planning, *Journal of Marketing*, 45, 62-70.

Rentz, Joseph O., Fred D. Reynolds and Roy G. Stout (1983), Analyzing Changing Consumption Patterns With Cohort Analysis, in *Journal of Marketing Research*, 20 (February), 12-20.

Rentz, Joseph O. and Fred D. Reynolds (1991), Forecasting the Effects of an Aging Population on Product Consumption: An Age-Period-Cohort Framework, in *Journal of Marketing Research*, 28 (August), 355-360.

Roosen, Charles B. (1995), Visualization and exploration of high-dimensional functions using the functional anova decomposition. Dissertation, Dept. of Statistics, Stanford University.

Rosenberg, M. (1979), *Conceiving the Self*, New York: Basic.

Riley, Mathilda White (1973), Aging and Cohort Succession: Interpretations and Misinterpretations, *Public Opinion Quarterly*, 37, 35-49.

Rindfleisch A. (1994), Cohort Generational Influences on Consumer Socialization, in C.T. Allen and D.R. John, (Eds), *Advances in Consumer Research XXI*, (460-476). Provo, UT: Association for Consumer Research.

RIAA (1998), *Consumer Purchasing Trends : 1998 Consumer Profile*. RIAA Reports.

Rodgers W.L. (1982), Estimable Functions of Age, Period, and Cohort Effects, *American Sociological Review*, 47, 6, 774-787.

Roedder John D. Cole C. (2000), Age Differences in Information Processing : Understanding Deficits in Young and Eledrly Consumers, *Journal of Consumer Research*, 13, 3, 297-315.

Roux S. (2001), Les écarts de revenu salarial entre hommes et femmes en début de carrière, *Insee Première*, 801.

Rust, Roland T. and Kary Wan-Yu Yeung (1995), Tracking the Age Wave: Parsimonious Estimation in Cohort Analysis, *Advances in Consumer Research*, 22, 680-685.

Ryder, Norman B. (1965), The Cohort as a Concept in the Study of Social Change, *American Sociological Review*, 30, 843-861.

Schuman, Howard and Jacqueline Scott (1989), Generations and Collective Memories, *American Sociological Review*, 54, 359-381.

Schweidel, David, Peter Fader and Eric T. Bradlow (2008), Modeling Service Retention Within and Across Cohorts under Limited Information, *Journal of Marketing* (forthcoming).

Shaffer, Greg and Z. John Zhang (1995), Competitive Coupon Targeting, *Marketing Science*, 14, 4.

Silva-Risso, Jorge M., Bucklin, Randolph E., and Morrison, Donald C. (1999) A Decision Support System for Planning Manufacturers' Sales Promotion Calendars, *Marketing Science*, Vol. 18, Issue 3.

Sengupta, Debasis and Pochiraju Bhimasankaram (1997), On the Roles of Observations in Collinearity in the Linear Model, *Journal of the American Statistical Association*, 92, 439, 1024-1032.

Smith, J. Walker and Ann Clurman (1997), *Rocking the Ages, The Yankelovich Report on Generational Marketing*, Harper Business, New York.

Stine, Robert A. (1995) Graphical Interpretation of Variance Inflation Factors, *American Statistician*, 49, 53–56.

Schweidel D.A., Fader P.S et Bradlow E.T. (2008), Understanding Service Retention Within and Across Cohorts Using Limited Information, *Journal of Marketing*, 72, 1, 82-94.

Smith J. W. et A. Clurman (2007), *Generation Ageless: How Baby Boomers Are Changing the Way We Live Today . . . And They're Just Getting Started*, Editions Collins.

Swann, W. B., Jr., Griffin, J.J., Predmore, S., & Gaines, B. (1987), The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology*, 52, 881-889.

Stephens N. (1991), Cognitive Age: A Useful Concept for Advertising?, *Journal of Advertising*, 20, 4, 37-47.

Tenenhaus, M. (1998). *La régression PLS*. Paris: Technip.

Tenenhaus, M. (1999). L'approche PLS. *Revue de Statistique Appliquée*, 47, 5-40.

Tenehaus, M., P. Bastien et V. Esposit-Vinzi (2002). *La régression PLS généralisée*, *Cahier de recherche*, n° 766, Groupe HEC, Jouy-en-Josas, France.

Tenenhaus, M., J. Pagès, L. Ambroisine et C. Guinot (2005), PLS Methodology for Studying Relationships Between Hedonic Judgements and Product Characteristics, *Food Quality and Preference*, 16,4, 315-325.

Uncles M.D. et A.C.S. Erhenberg (1990), Brand Choice among older Consumers, *Journal of Advertising Research*, 30, 4, 19-22.

Vandeschrick, Christophe (2001), The Lexis Diagram, a Misnomer, *Demographic Research*, 4, 97-142.

Valette-Florence P. (1988), Spécialité et apports des méthodes d'analyse multivariée de la deuxième génération, *Recherche et Applications en Marketing*, 3,4.

Valette-Florence P. (1990), Analyse structurelle et analyse topologique: illustration d'une démarche complémentaire, Recherche et Applications en Marketing, 5,1.

Vapnik, Vladimir (1995), The Nature of Statistical Learning Theory. Springer-Verlag.

Vapnik, Vladimir (1998), Statistical Learning Theory. Wiley.

Vapnik, Vladimir (2006), *Empirical Inference Science* (2006) essai philosophique proposé dans la ré-édition de l'ouvrage *Estimation of Dependences Based on Empirical Data* (1982) par Springer-Verlag.

Von freymann J. W. (2006), Age and Generational Cohort Effect in Store Assessments and Choice: A case Study, Marketing Management Journal, 16, 2, 203-222.

Wangenheim, Florian and Tomas Bayon (2004), Satisfaction, loyalty and word-of-mouth within the customer base of a utility provider : Differences between stayers, switchers, and referral switchers, Journal of Consumer Behaviour, 3,1, 211-220.

Wedel M. and W. Kamakura (2000), Market Segmentation, Conceptual and Methodological Foundations, Seconde Edition International Series in Quantitative Marketing, Kluwer Academic Publishing.

Wedel, M., W. A. Kamakura, W. S. DeSarbo and F. ter Hofstede (1995), Implications for Asymmetry, Nonproportionality and Heterogeneity in Brand Switching from Piece-Wise Exponential Hazard Models, Journal of Marketing Research, 32, 457-462.

Weiss, Robert E et R. Deniis Cook, A Graphical Case Statistic for Assessing Posterior Influence, Biometrika, 79, 1, 51-55.

West, Patricia M., Patrick L. Brockett, Linda L. Golden (1997), A Comparative Analysis of Neural Networks and Statistical Methods for Predicting Consumer Choice, Marketing Science, 16, 4, 370-391.

Wilkes R.E. (1992), A Structural Modeling Approach to the Measurement and Meaning of Cognitive Age, *Journal of Consumer Research*, 19, 9, 292-301.

Wold, H. (1980), Model Construction and Evaluation when Theoretical Knowledge is Scarce, in *Evaluation of Econometric Models*, Anonymous New York: Academic Press, 47-74.

Wold, S., Kettaneh, N., et Tjessem K. (1996), Hierarchical multi-block PLS and PC models for easier interpretation and as an alternative to variable selection, *Journal of Chemometrics*, 10, 463-482.

XLStat (2007 et 2008), User Guide. AddinSoft Sarl, Paris, France.

Yang Yang, Wenjiang J. Fu, et Kenneth C. Land (2004), A Methodological Comparison of Age-Period-Cohort Models: The Intrinsic Estimator and Conventional Generalized Linear Models, *Sociological Methodology*, 34:75-110.

Yang Yang, Sam Schulhofer-Wohl, Wenjiang J. Fu, et Kenneth C. Land (2008), The Intrinsic Estimator for Age-Period-Cohort Analysis: What It Is and How To Use It, *American Journal of Sociology*, 113.

Yim, Chi Kin et P. K. Kannan (1999), Consumer Behavioral Loyalty: A Segmentation Model and Analysis, *Journal of Business Research*, 44, 2, 75-92.

Yu Jiangsheng, Fei Cheng, Huilin Xiong, Wanling Qu et Xue-wen Chen (2008); A Bayesian approach to support vector machines for the binary classification, *Neurocomputing*, Volume 72, Issues 1–3, December 2008, 177-185.

Zeuner, G. (1869), *Abhandlungen Aus Der Mathematischen Statistik*. Leipzig: Verlag von Arthur Felix.

Zufryden, F.S. (1986), Multibrand Transition Probabilities as a Function of Explanatory Variables : Estimation by a Least-Squares-Based Approach, *Journal of Marketing Research*, 23, 177-183.

Conférences

Les différentes parties de cette thèse ont été présentées dans les conférences académiques internationales ci-dessous :

Bascoul Ganaël (2007), "Local Effect Analysis to Optimize Direct Marketing Action", 2007 INFORMS Marketing Science Conference, Singapore Management University, Lee Kong Chian School of Business, Singapore.

Bascoul Ganaël (2006), "Higher Repeat Purchasing by Older Consumers: Impact of Aging or Cohort Effect?", 2006 INFORMS Marketing Science Conference, Katz GSB, Pittsburgh, USA.

Bascoul Ganaël, (2006), "Higher Repeat Purchasing by Older Consumers: Impact of Aging or Cohort Effect?" 2006 EMAC Conference, Athens, Greece.

Bascoul Ganaël (2005), "To Discriminate Between Marginal Consumers, Do Not Analyze Extreme Consumers" 2005 INFORMS Marketing Science Conference, Emory University, Atlanta, Georgia, USA.

Annexes

Annexe 1. Exemple complémentaire dans l'industrie du disque.

Il est intéressant de remarquer que les préférences de consommation culturelles décrites dans le corps du texte pourraient potentiellement s'étendre au support de consommation lui-même. Examinons les chiffres de ventes recueillis par l'Association américaine de l'industrie du disque (RIAA, Recording Industry Association of America) qui regroupe les quatre majors les plus importantes au monde (Universal, Sony, Warner et EMI) que l'on résume dans le tableau B.4.1. ci-après à partir des rapports de 1998 et 2006. La part des personnes de plus de 45 ans dans la consommation de CDs augmente au cours du temps. Ceci peut en bonne partie être dû à l'évolution de la structure démographique (pyramide des âges) suivant ces deux décennies. Mais l'on peut aussi émettre l'hypothèse (sans avoir ici assez d'informations pour le mesurer statistiquement) que les cohortes restent fidèles au support qu'elles ont connu étant jeunes. Ainsi les personnes de plus de 45 restent attachées au support CD, alors que les nouvelles cohortes sont moins sensibles à ce support, comme on peut le supposer ici pour la cohorte des personnes ayant entre 15 et 19 ans en 2006.

Usage du support CD par tranche d'âges	1989	1999	2006
Part des 15-19 ans dans la consommation totale	24.8%	16.8%	12.8%
Part des 45 ans et plus dans la consommation totale	10.2%	16.5%	26.1%

Tendances d'achat de CD par tranche d'âge entre 1989 et 2006

En effet, ces jeunes cohortes ont connu dès leur plus jeune âge la concurrence que constitue le format numérique (avec ses possibilités de téléchargement, légal ou non) face au format CD, avec les habitudes de consommation correspondantes, comme l'achat en ligne et l'écoute nomade sur baladeurs mp3 (e.g. la plateforme iTunes et le baladeur iPod d'Apple, futurs symboles d'une époque).

On peut aussi penser que les plus âgés sont potentiellement plus sensibles à l'illégalité du piratage, ce qui amplifierait ce contraste.

III - VOTRE ACHAT

11 Votre voiture est-elle...

- Une voiture personnelle
- Une voiture à mon nom (acquise et payée par ma société)
- Une voiture au nom de l'entreprise (payée par l'entreprise où je travaille)

12 Auprès de qui avez-vous souscrit ce financement ?

Indiquez dans la première colonne tous les moyens de financement utilisés, indiquez dans la deuxième colonne le moyen principal utilisé

Moyens de financement utilisés (plusieurs réponses possibles)	Moyen principal utilisé (une seule réponse possible)
• Au comptant <input type="checkbox"/>	<input type="checkbox"/>
• A crédit, par l'intermédiaire de l'organisme financier du constructeur <input type="checkbox"/>	<input type="checkbox"/>
• A crédit, par l'intermédiaire du concessionnaire <input type="checkbox"/>	<input type="checkbox"/>
• A crédit, par l'intermédiaire d'un autre organisme financier <input type="checkbox"/>	<input type="checkbox"/>
• Un prêt bancaire <input type="checkbox"/>	<input type="checkbox"/>
• Un prêt (autres sources) <input type="checkbox"/>	<input type="checkbox"/>
• En leasing ou option d'achat <input type="checkbox"/>	<input type="checkbox"/>
• Je ne sais pas, mon employeur a pris en charge le paiement <input type="checkbox"/>	<input type="checkbox"/>

Passez à 14

13 Quel est le prix TOTAL, livraisons et options comprises, de la voiture que vous venez d'acheter ?

(Si vous avez obtenu une reprise, ne pas la déduire du prix que vous avez effectivement payé)

• Environ _____ Francs

13a Lorsque vous avez acheté votre nouvelle voiture, avez-vous bénéficié d'une remise sur le prix de cette voiture ?

OUI → Passez à 13b NON → Passez à 14

13b Quel pourcentage du prix ou quel montant représentait approximativement cette remise ?

_____ % ou _____ Francs

14 Avez-vous bénéficié de conditions spéciales ou avantages en dehors d'une remise sur le prix de la voiture neuve ?

(Plusieurs réponses possibles)

- Aucun
- Bonne reprise
- Financement intéressant
- Réduction collaborateur automobile
- Radio cassette CD offert
- Offres spéciales (livraison gratuite vignette assurance)
- Option Accessoire offert
- Autres (précisez) _____

15 Avez-vous essayé cette nouvelle voiture AVANT de vous décider à l'acheter ?

- OUI
- NON, ce n'était pas nécessaire
- NON, mais j'aurais aimé que cela me soit proposé

16 Pour chacun des usages suivants, indiquez à quelle fréquence vous utilisez VOTRE NOUVELLE VOITURE.

(Une seule réponse par ligne)

	Presque chaque jour	Au moins une fois par semaine	une fois par mois	par an	Jamais Das concerne
1 - Sur autoroute	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2 - Sur route	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 - En ville	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4 - Pour des parcours "tout terrain" (si votre voiture est adaptée)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 - Pour des déplacements domicile/travail	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6 - Pour des déplacements professionnels	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7 - Pour faire des achats	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8 - Pour conduire les enfants à l'école	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9 - Pour tracter une caravane, une remorque	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10 - Avec des adultes à l'arrière	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11 - Avec des enfants à l'arrière	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12 - Avec des gros paquets, des bagages, du matériel de sport dans le coffre ou à l'intérieur de la voiture	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13 - Avec des objets sur la galerie du toit	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14 - Avec le siège arrière rabattu	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15 - Avec des sièges arrière enlevés (si votre voiture est adaptée)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16 - Pour partir en week-end	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17 - Pour partir en vacances	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

17 Quelle est votre SATISFACTION à l'égard du CONCESSIONNAIRE qui vous a vendu votre nouvelle voiture ?

A l'aide de l'échelle ci-dessous, veuillez indiquer votre niveau de satisfaction sur les différentes prestations de votre concessionnaire.

Très satisfait	Plutôt satisfait	Ni satisfait Ni insatisfait	Plutôt insatisfait	Très insatisfait					
10	9	8	7	6	5	4	3	2	1

Ecrire une note de 1 à 10 pour chaque caractéristique

- a Votre satisfaction générale à l'égard du concessionnaire où vous avez acheté votre voiture
- b Votre satisfaction générale sur le déroulement de votre achat
 - Aspect général du hall d'exposition
 - Amabilité du personnel de vente
 - Compétence du personnel de vente
 - Informations fournies sur les différents modèles
 - Informations fournies sur les différents moyens de financement
 - Prix d'achat
- c Votre satisfaction générale concernant la livraison
 - Délai de livraison
 - Etat de la voiture à la livraison
 - Mise en main de la voiture par le vendeur et explications fournies
 - Efficacité du service après-vente pour résoudre les problèmes

18 Depuis la livraison de votre nouvelle voiture, le concessionnaire vous a-t-il contacté pour connaître votre satisfaction ?

- OUI
- NON

19 Qui a décidé de la marque et du modèle de votre nouvelle voiture ?

(Une seule réponse)

- Moi-même
- Moi-même, avec un autre membre de la famille
- Un ou plusieurs autres membres de la famille
- L'entreprise où je travaille a choisi la marque et j'ai choisi moi-même le modèle
- L'entreprise où je travaille a choisi la marque et le modèle

Passez à 20

27

20 Quelles sont les raisons qui vous ont amené à choisir cette voiture plutôt qu'une autre ?

(Inscrivez toutes les raisons même celles qui vous paraissent peu importantes)

21 Avant de vous décider à acheter cette nouvelle voiture, avez-vous hésité avec d'autres modèles de voitures NEUVES ?

OUI → Passez à 22 NON → Passez à 26

22 Quels étaient la marque, le modèle et la version de la voiture avec laquelle vous avez le plus hésité ?

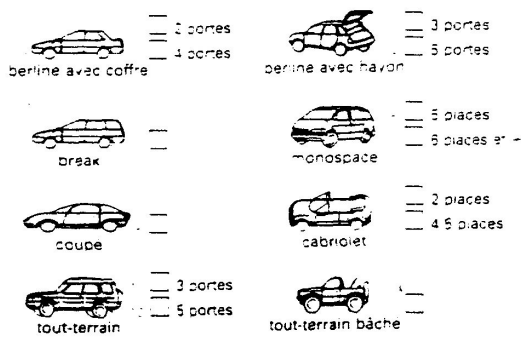
• MARQUE _____ MODÈLE _____

• VERSION _____

• TYPE DE MOTEUR :

Essence Diesel

• CARROSSERIE - NOMBRE DE PORTES :



23 Avez-vous effectué les démarches suivantes pour cet autre modèle ?

- | | | |
|------------------------------------------------------------------|--------------------------|--------------------------|
| | OUI | NON |
| • Êtes-vous allé voir cet autre modèle chez le concessionnaire ? | <input type="checkbox"/> | <input type="checkbox"/> |
| • Le vendeur vous a-t-il donné le prix de ce modèle ? | <input type="checkbox"/> | <input type="checkbox"/> |
| • Le concessionnaire vous a-t-il fait essayer ce modèle ? | <input type="checkbox"/> | <input type="checkbox"/> |

24 Et quelles sont les raisons pour lesquelles vous avez décidé de ne pas acheter cet autre modèle ?

(Inscrivez toutes les raisons)

25 Y a-t-il ensuite d'autres modèles de voitures neuves avec lesquels vous avez également hésité ?

MARQUE _____ MODÈLE _____

MARQUE _____ MODÈLE _____

Je n'ai pas envisagé d'autres modèles

PASSEZ A LA QUESTION 27

26 Si la voiture que vous venez d'acheter n'avait pas existé, quelle marque et quel modèle de voiture neuve auriez-vous acheté le plus probablement ?

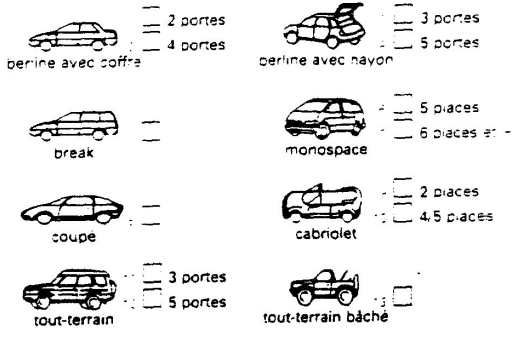
• MARQUE _____ MODÈLE _____

• VERSION _____

• TYPE DE MOTEUR :

Essence Diesel

• CARROSSERIE - NOMBRE DE PORTES :



V - VOTRE OPINION SUR VOTRE NOUVELLE VOITURE

27 Quels sont les points sur lesquels vous êtes satisfait de votre nouvelle voiture ?

28 Et quels sont les points sur lesquels votre nouvelle voiture ne vous donne pas satisfaction ?



NOUS SOUHAITONS CONNAÎTRE VOTRE NIVEAU DE SATISFACTION ...

Pour les deux questions suivantes, nous allons vous demander de nous indiquer votre niveau de satisfaction à l'égard de votre voiture à l'aide de l'échelle ci-dessous :

Très satisfait		Plutôt satisfait		Ni satisfait Ni insatisfait		Plutôt insatisfait		Très insatisfait	
10	9	8	7	6	5	4	3	2	1

29 Pouvez-vous indiquer ci-dessous la note qui caractérise le mieux votre satisfaction à l'égard de votre voiture dans son ensemble ?

Écrivez une note de 1 à 10

30 A l'aide de la même échelle que précédemment, indiquez pour chacune des caractéristiques ci-dessous la note qui, compte tenu de vos propres besoins et attentes, exprime le mieux votre satisfaction ?

ÉCRIRE UNE NOTE DE 1 À 10 POUR CHAQUE CARACTÉRISTIQUE

Ligne	<input type="text"/>	Espaces de rangement pour petits objets	<input type="text"/>	Maniabilité	<input type="text"/>
Qualité en général	<input type="text"/>	Accélération	<input type="text"/>	Confort des sièges avant	<input type="text"/>
Esthétique de l'intérieur	<input type="text"/>	Tenue de route et stabilité	<input type="text"/>	Confort de suspension	<input type="text"/>
Finition de l'intérieur	<input type="text"/>	Freinage	<input type="text"/>	Habitabilité intérieure	<input type="text"/>
Qualité des matériaux de l'habitacle	<input type="text"/>	Changement de vitesses	<input type="text"/>	Capacité du coffre	<input type="text"/>
Niveau de l'équipement	<input type="text"/>	Direction	<input type="text"/>	Facilité à charger et décharger les bagages	<input type="text"/>
Sécurité	<input type="text"/>	Lisibilité des voyants et compteurs	<input type="text"/>	Facilité à entrer et sortir de la voiture	<input type="text"/>
Rapport qualité/prix	<input type="text"/>	Sportivité	<input type="text"/>	Respect de l'environnement	<input type="text"/>
Robustesse mécanique	<input type="text"/>	Prestige/réputation de la marque	<input type="text"/>	Chauffage et ventilation	<input type="text"/>
Robustesse carrosserie	<input type="text"/>	Aspect du tableau de bord	<input type="text"/>	Silence en roulant	<input type="text"/>
Consommation de carburant	<input type="text"/>	Accessibilité aux commandes	<input type="text"/>	Finition de la carrosserie	<input type="text"/>
Visibilité	<input type="text"/>	Protection contre le vol	<input type="text"/>	Absence de défauts jusqu'à présent	<input type="text"/>
				Absence de bruits et de vibrations	<input type="text"/>

VI - VOUS AVEZ VOTRE VOITURE

31) Voici une liste de rubriques pour lesquelles votre voiture a pu avoir des problèmes (en dehors des accidents de circulation). Veuillez cocher les cases correspondant aux problèmes rencontrés et en décrire aussi précisément que possible la nature, l'origine et l'endroit sur la voiture.

Pour chaque case cochée, précisez dans les deux colonnes de droite, si vous avez amené votre voiture à réparer au garage et si la réparation a été réalisée correctement.

La précision de vos réponses aidera le constructeur à éviter des incidents semblables à l'avenir.

RUBRIQUES	Cocher les incidents que vous avez eus	Décrivez de façon détaillée, ce qui vous est arrivé, en précisant si possible, la nature, l'origine et l'endroit sur la voiture.	La voiture a-t-elle été amenée au garage ?	
			OUI	OUI
• Peinture	<input type="checkbox"/> 1		<input type="checkbox"/>	<input type="checkbox"/>
• Aspect de la carrosserie : capot, pare-chocs, portes, vitres et équipements extérieurs	<input type="checkbox"/> 2		<input type="checkbox"/>	<input type="checkbox"/>
• Étanchéité à l'air, à l'eau : fuites, courants d'air	<input type="checkbox"/> 3		<input type="checkbox"/>	<input type="checkbox"/>
• Bruits, grincements, sifflements, vibrations	<input type="checkbox"/> 4		<input type="checkbox"/>	<input type="checkbox"/>
• Ouvertures/fermetures : portes, vitres, coffre, capot, hayon, serrures, toit ouvrant, manipulation des rétroviseurs	<input type="checkbox"/> 5		<input type="checkbox"/>	<input type="checkbox"/>
• Habitacle : tableau de bord, sièges, tapis, commandes, manettes, ventilateur, chauffage	<input type="checkbox"/> 6		<input type="checkbox"/>	<input type="checkbox"/>
• Équipements électriques incluant tous les éclairages, instruments de bord, radio/cassettes/CD, ordinateur	<input type="checkbox"/> 7		<input type="checkbox"/>	<input type="checkbox"/>
• Moteur incluant carburant, démarrage, huile, système de refroidissement	<input type="checkbox"/> 8		<input type="checkbox"/>	<input type="checkbox"/>
• Embrayage, boîte de vitesse, transmission	<input type="checkbox"/> 9		<input type="checkbox"/>	<input type="checkbox"/>
• Freins	<input type="checkbox"/> 10		<input type="checkbox"/>	<input type="checkbox"/>
• Direction, roues	<input type="checkbox"/> 11		<input type="checkbox"/>	<input type="checkbox"/>
• Suspensions, amortisseurs	<input type="checkbox"/> 12		<input type="checkbox"/>	<input type="checkbox"/>
• Échappement	<input type="checkbox"/> 13		<input type="checkbox"/>	<input type="checkbox"/>
• Autres problèmes (précisez)	<input type="checkbox"/> 14		<input type="checkbox"/>	<input type="checkbox"/>
• AUCUN PROBLÈME	<input type="checkbox"/> 15		<input type="checkbox"/>	<input type="checkbox"/>

32) Suite aux incidents cités ci-dessus, votre voiture a-t-elle été immobilisée ?
 NON : OUI, ma voiture a été immobilisée au total jours

VII - LA VOITURE QUE VOUS AVEZ REMPLACÉE

33) Parmi les définitions suivantes, quelle est celle qui correspond à votre situation ?

- Avant d'acquérir la nouvelle voiture vous n'aviez pas de voiture dans le foyer Passer à 44
- La nouvelle voiture a été achetée, en supplément d'une autre voiture sans pour autant la remplacer Passer à 35
- La nouvelle voiture a été achetée pour remplacer une voiture :
 - de même marque Passer à 35
 - de marque différente Passer à 34

34) Pour quelles raisons n'avez-vous pas racheté la même marque que précédemment ?









35) Description de la voiture que vous avez remplacée :

MARQUE _____ MODÈLE _____
 VERSION _____ CYLINDRÉE _____
 PUISSANCE : moteur _____ DIN fiscale _____
 Date de la 1^{re} immatriculation en 19 _____

36) Votre voiture précédente était-elle ?

ACHETÉE : Neuve D'occasion
 VOITURE : Personnelle Société
 AVEC BOÎTE : Manuelle Automatique
 AVEC MOTEUR : Essence Diesel
 AVEC UN TURBOCOMPRESSEUR : Oui Non
 AVEC 4 ROUES MOTRICES : Oui Non

37 Carrosserie / Nombre de portes de votre voiture précédente :

 <input type="checkbox"/> 2 portes <input type="checkbox"/> 4 portes	 <input type="checkbox"/> 3 portes <input type="checkbox"/> 5 portes
 <input type="checkbox"/>	 <input type="checkbox"/> 5 places <input type="checkbox"/> 6 places et +
 <input type="checkbox"/>	 <input type="checkbox"/> 2 places <input type="checkbox"/> 4.5 places
 <input type="checkbox"/> 3 portes <input type="checkbox"/> 5 portes	 <input type="checkbox"/>

38 Combien de temps avez-vous gardé votre précédente voiture ?

Nombre d'années *en* mois de mois

39 Combien de kilomètres avez-vous parcourus au total avec votre précédente voiture ?

(ATTENTION : Pour les voitures achetées d'occasion, soustraire le kilométrage à l'achat du kilométrage à la vente.)

Kilomètres

40 Avez-vous acheté votre voiture précédente (citée en 35) dans le même garage que votre nouvelle voiture ?

OUI NON

41 Qu'avez-vous fait de la voiture que vous avez remplacée ?

- Je l'ai vendue au garagiste à qui j'ai acheté la NOUVELLE
- Je l'ai vendue à un autre garagiste
- Je l'ai vendue à un particulier
- Je l'ai conservée dans le foyer
- Je l'ai mise à la casse
- Autre cas

42 Assuriez-vous l'entretien de votre voiture précédente chez un concessionnaire de la marque ?

OUI NON

43 A l'aide de l'échelle ci-dessous, veuillez indiquer votre satisfaction à l'égard de votre précédente voiture sur les différents points cités.

Très satisfait		Plutôt satisfait		Ni satisfait / Ni insatisfait		Plutôt insatisfait		Très insatisfait	
10	9	8	7	6	5	4	3	2	1

Écrire une note de 1 à 10

- Satisfaction en général de votre voiture
- Qualité en général de votre voiture
- Service après-vente du réseau (Chez le concessionnaire mentionné en 42)

VIII - VOTRE ATTITUDE À L'ÉGARD DE L'AUTOMOBILE

44 Quelle est votre opinion vis-à-vis de chacune des propositions ci-dessous ?

Cochez la case correspondant à votre point de vue pour chaque proposition.

	Tout à fait d'accord	D'accord	Pas d'accord	Pas d'accord du tout
1 - Avoir une voiture me procure une sensation d'indépendance et de liberté	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2 - Je préfère m'occuper de ma voiture moi-même	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 - Une voiture est, pour moi, un moyen d'exprimer ma personnalité	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4 - Je considère la voiture comme un moyen de transport et non comme un plaisir	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 - En matière d'automobile, je suis toujours à la recherche d'un modèle innovant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6 - Je n'aime pas prêter ma voiture	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7 - Quand j'achète une voiture, j'aime sentir que le constructeur a cherché des solutions techniques originales	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8 - Je ne comprends pas les gens qui sont capables d'acheter une voiture sous l'impulsion du moment, sans prendre le temps de réfléchir	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9 - Mes amis et relations me consultent souvent lorsqu'ils veulent acheter une voiture	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10 - J'aime avoir une voiture qui attire l'attention	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11 - Il faut que la voiture que je conduis corresponde à mon standing	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12 - J'aime contempler les belles voitures	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13 - Lorsque j'achète une voiture, je choisis la moins chère parmi celles qui correspondent à mes besoins	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14 - J'aime conduire vite	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
15 - Selon moi, les voitures sont l'une des plus importantes sources de problèmes pour l'environnement	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
16 - J'aime les voitures qui ont un style inhabituel	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
17 - Une voiture devrait être aussi confortable qu'un salon	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
18 - Je voudrais que ma voiture possède tous les équipements de sécurité possibles même si c'est au détriment d'autres équipements	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

IX - À PROPOS DE VOUS - L'UTILISATEUR PRINCIPAL

45 Êtes-vous...

Un homme Une femme

46 Quel est votre âge ?

Ans

47 Êtes-vous...

(Cochez la case qui correspond le mieux à votre situation)

- Marié, vivant maritalement 1
- Séparé, divorcé, veuf 2
- Célibataire vivant avec les parents / avec un autre membre de la famille 3
- Célibataire vivant seul 4

48 Par quel niveau d'enseignement avez-vous terminé vos études ?

- Primaire 1
- Secondaire 2
- Technique sans bac 3
- Technique avec bac 4
- Supérieur 5

49 Quelles sont vos activités principales pendant vos loisirs ?

50 Avez-vous ?

Au DOMICILE Au TRAVAIL

- Un ordinateur avec CD ROM
- Accès à INTERNET

51 Parmi la liste suivante, quelle est l'activité professionnelle actuelle qui VOUS correspond le mieux, ainsi que celle de votre CONJOINT ?

A VOTRE COMPTE VOUS MÊME VOTRE CONJOINT

- Profession libérale
- Patron, propriétaire d'une entreprise 2
- Artisan, commerçant 3
- Agriculteur exploitant 4

SALARIÉ

- Cadre supérieur, dirigeant d'entreprise, membre d'une profession libérale salariée 5
- Cadre moyen (chef de service / groupe, directeur d'agence...), technicien, agent de maîtrise 6
- Employé (bureaux, commerces, services...) 7
- Ouvrier 8

AUTRE ACTIVITÉ

- Femme au foyer 9
- Étudiant 10
- Retraité 11
- Sans travail 12
- Autre 13

52 Quelle est votre activité professionnelle ? (si vous êtes retraité, mentionnez votre ancienne profession)

(Veuillez donner le maximum de détails)

X - À PROPOS DE VOTRE FOYER

53 Quelle est la composition de votre foyer ?

Nombre de personnes vivant au foyer y compris vous-même _____ sont

_____ adultes (âges de plus de 18 ans)

_____ enfants (âges de moins de 10 ans)

_____ enfants (âges de 10 à 18 ans)

54 Dans votre foyer, combien de personnes y compris vous-même ont leur propre revenu ?

Entourez le chiffre correspondant.

0 1 2 3 4 5 6 et plus

55 Excepté vous-même, y a-t-il d'autres personnes qui conduisent la nouvelle voiture ?

OUI NON

Si OUI, précisez la fréquence d'utilisation de la voiture par les autres membres du foyer (si concerné) :

Aussi souvent que moi Régulièrement mais moins souvent que moi De temps en temps

- Mon conjoint
- Un de mes enfants
- Une autre personne vivant au foyer

XI - LES VOITURES DE VOTRE FOYER

56 Combien utilise-t-on au total de voitures dans votre foyer ?

- Une seule, la nouvelle voiture Passer à 60
- Deux, la nouvelle et une autre voiture Passer à 57
- Trois, la nouvelle et deux autres voitures Passer à 57
- Quatre ou plus, dont la nouvelle voiture Passer à 57

57 Veuillez décrire, ci-dessous les autres voitures utilisées actuellement par vous-même ou un autre membre du foyer. S'il y en a plus de deux, décrivez les deux utilisées le plus souvent (à l'exclusion de la nouvelle voiture citée en 2.)

	VOITURE B	VOITURE C
Marque	_____	_____
Modèle	_____	_____
Cylindrée	_ _ _ cm ³	_ _ _ cm ³
Puissance moteur	_ _ _ DIN	_ _ _ DIN
Puissance fiscale	_ CV	_ CV
Achetée	neuve <input type="checkbox"/> d'occasion <input type="checkbox"/>	neuve <input type="checkbox"/> d'occasion <input type="checkbox"/>
Année de la 1 ^{re} immatriculation	19 _ _	19 _ _

58 Qui conduit le plus fréquemment ces autres voitures ?

(Une seule réponse par voiture)

	VOITURE B	VOITURE C
• Moi-même	<input type="checkbox"/>	<input type="checkbox"/>
• Mon conjoint	<input type="checkbox"/>	<input type="checkbox"/>
• Un de mes enfants	<input type="checkbox"/>	<input type="checkbox"/>
• Mes parents	<input type="checkbox"/>	<input type="checkbox"/>
• Une autre personne vivant au foyer	<input type="checkbox"/>	<input type="checkbox"/>

59 Précisez la voiture que vous considérez comme étant la principale du foyer.

- La voiture que vous venez d'acheter (citée en 2)
- La voiture B (citée en 57)
- La voiture C (citée en 57)
- Aucune en particulier

XII - VOTRE APPRÉCIATION D'ENSEMBLE SUR VOTRE NOUVELLE VOITURE

60 Compte tenu de votre expérience acquise jusqu'à présent, indiquez à l'aide de l'échelle ci-dessous dans quelle mesure recommanderiez-vous à un ami, un parent ou une relation professionnelle, votre nouvelle voiture et/ou le concessionnaire qui vous l'a livrée ?

(Une seule réponse par ligne)

	OUI CERTAINEMENT	OUI PROBABLEMENT	NON PROBABLEMENT PAS	NON CERTAINEMENT PAS
• Je recommanderais ma nouvelle voiture	<input type="checkbox"/>	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4
• Je recommanderais n'importe quel modèle de la marque de ma nouvelle voiture	<input type="checkbox"/>	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4
• Je recommanderais le concessionnaire où j'ai acheté ma nouvelle voiture	<input type="checkbox"/>	<input type="checkbox"/> 2	<input type="checkbox"/> 3	<input type="checkbox"/> 4

XIII - VOTRE APPRÉCIATION D'ENSEMBLE SUR VOTRE NOUVELLE VOITURE

Ces informations sont uniquement demandées pour classer les réponses. Si ces questions vous semblent trop personnelles, vous pouvez ne pas y répondre, néanmoins pour la validité statistique de l'étude, votre réponse est très importante.

61 Votre commune de résidence.

Code postal _____

62 Quel est le total ANNUEL BRUT des revenus de votre foyer ?

Comptez bien les revenus de toutes les personnes de votre foyer y compris vous-même (salaires, primes, pensions, autres...)

- Moins de 75 000 F 1
- De 75 000 à 99 999 F 2
- De 100 000 à 124 999 F 3
- De 125 000 à 149 999 F 4
- De 150 000 à 174 999 F 5
- De 175 000 à 199 999 F 6
- De 200 000 à 249 999 F 7
- De 250 000 à 299 999 F 8
- De 300 000 à 399 999 F 9
- De 400 000 à 499 999 F 10
- De 500 000 à 599 999 F 11
- De 600 000 à 699 999 F 12
- De 700 000 à 799 999 F 13
- Plus de 800 000 F 14

VOS AUTRES REMARQUES OU RÉFLEXIONS SUR VOTRE NOUVELLE VOITURE

Merci de votre précieuse collaboration

N° 78-17 DU 6 JANVIER 1978 - ART. 27 - Les personnes auprès desquelles sont recueillies des informations nominatives doivent être informées : • du caractère obligatoire ou facultatif des réponses ; • des conséquences à leur égard d'un défaut de réponse ; • des personnes physiques ou morales destinataires des informations ; • de l'existence d'un droit d'accès et de rectification. Lorsque de telles informations sont recueillies par voie de questionnaires, ceux-ci doivent porter mention de ces prescriptions.
 DÉCRET DU 11 OCTOBRE 1983 - Ce message a été adressé à l'aide des informations du Fichier Central des Automobiles, autorisé par l'arrêté du 11-10-83 (JO du 25-11-83) que les marques automobiles sont un savoir pour les années d'immatriculation sur Automobile. Si vous ne souhaitez pas être destinataire de ces documents, ou si vous souhaitez annuler votre inscription, contactez le service d'information.

Annexe 3. Diagramme de Lexis des données sur le ré-achat automobile (France)

80	67%	65%	69%	73%	75%	75%	75%	80%	77%	65%	67%	1924
79	68%	59%	68%	74%	80%	73%	77%	75%	68%	66%	66%	1925
78	56%	71%	59%	75%	64%	73%	71%	68%	68%	73%	60%	1926
77	78%	80%	61%	81%	70%	72%	80%	66%	68%	68%	73%	1927
76	76%	68%	66%	73%	64%	74%	75%	70%	69%	74%	72%	1928
75	72%	67%	71%	69%	68%	70%	73%	69%	67%	65%	74%	1929
74	72%	66%	57%	71%	74%	68%	74%	71%	69%	73%	67%	1930
73	70%	69%	69%	75%	66%	60%	74%	68%	76%	72%	63%	1931
72	71%	65%	70%	71%	70%	70%	68%	66%	66%	65%	69%	1932
71	65%	61%	67%	77%	65%	68%	69%	68%	69%	68%	60%	1933
70	68%	73%	66%	70%	66%	65%	68%	67%	68%	69%	69%	1934
69	66%	67%	61%	68%	66%	61%	65%	69%	65%	69%	63%	1935
68	65%	66%	64%	67%	67%	67%	67%	66%	71%	67%	62%	1936
67	66%	70%	66%	64%	62%	68%	61%	66%	68%	64%	65%	1937
66	64%	64%	64%	67%	67%	67%	63%	65%	64%	65%	63%	1938
65	65%	58%	60%	67%	66%	66%	63%	60%	57%	63%	60%	1939
64	59%	58%	64%	59%	61%	64%	67%	59%	62%	60%	59%	1940
63	63%	62%	55%	61%	63%	61%	57%	70%	62%	62%	58%	1941
62	56%	62%	62%	63%	58%	58%	61%	63%	61%	62%	56%	1942
61	54%	58%	56%	61%	60%	71%	57%	59%	58%	63%	60%	1943
60	59%	55%	59%	61%	58%	57%	58%	64%	57%	62%	59%	1944
59	65%	57%	57%	62%	57%	59%	63%	62%	62%	58%	62%	1945
58	61%	52%	54%	65%	59%	60%	60%	59%	60%	61%	58%	1946
57	58%	58%	58%	60%	56%	61%	54%	59%	63%	60%	57%	1947
56	59%	53%	60%	60%	64%	54%	61%	58%	59%	60%	57%	1948
55	58%	54%	61%	56%	58%	61%	59%	58%	60%	59%	55%	1949
54	62%	58%	58%	57%	65%	56%	61%	56%	58%	59%	57%	1950
53	61%	54%	58%	60%	58%	54%	59%	55%	59%	59%	56%	1951
52	60%	55%	57%	52%	59%	60%	55%	57%	55%	57%	54%	1952
51	64%	58%	60%	61%	53%	60%	57%	58%	50%	53%	55%	1953
50	62%	53%	58%	58%	55%	54%	55%	53%	57%	53%	53%	1954
49	58%	59%	55%	57%	53%	55%	53%	56%	51%	53%	54%	1955
48	56%	51%	57%	55%	53%	54%	54%	56%	49%	50%	54%	1956
47	56%	54%	51%	53%	56%	58%	53%	53%	52%	51%	48%	1957
46	57%	44%	52%	53%	52%	50%	51%	52%	56%	56%	52%	1958
45	57%	52%	47%	49%	52%	53%	54%	49%	50%	52%	53%	1959
44	54%	53%	47%	60%	50%	45%	52%	55%	48%	51%	49%	1960
43	50%	51%	45%	54%	50%	43%	50%	44%	52%	52%	45%	1961
42	51%	50%	46%	52%	43%	46%	52%	45%	48%	45%	46%	1962
41	48%	52%	49%	42%	50%	48%	47%	47%	48%	45%	45%	1963
40	50%	49%	45%	49%	51%	50%	45%	48%	49%	45%	43%	1964
Age	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Cohorte
	Période											

(Suite du diagramme de Lexis)

39	47%	46%	48%	46%	49%	46%	40%	50%	41%	45%	42%	1965
38	47%	41%	47%	51%	49%	41%	52%	46%	45%	49%	48%	1966
37	46%	43%	47%	45%	45%	45%	45%	45%	44%	45%	45%	1967
36	46%	46%	40%	40%	48%	45%	42%	43%	41%	44%	42%	1968
35	48%	38%	45%	44%	43%	38%	43%	49%	42%	41%	36%	1969
34	48%	44%	44%	45%	40%	48%	50%	43%	48%	47%	44%	1970
33	46%	42%	36%	45%	39%	42%	38%	42%	44%	42%	46%	1971
32	45%	41%	39%	40%	40%	38%	49%	47%	43%	39%	46%	1972
31	43%	39%	41%	42%	41%	41%	46%	38%	40%	42%	40%	1973
30	46%	38%	38%	39%	37%	46%	43%	39%	45%	39%	41%	1974
29	44%	36%	44%	36%	34%	36%	47%	41%	41%	40%	40%	1975
28	39%	41%	37%	45%	40%	39%	41%	44%	38%	40%	35%	1976
27	39%	36%	38%	45%	36%	42%	30%	35%	44%	44%	34%	1977
26	45%	30%	32%	38%	39%	39%	39%	39%	43%	40%	40%	1978
25	44%	37%	36%	37%	35%	33%	35%	42%	35%	39%	36%	1979
24	41%	31%	36%	31%	32%	33%	34%	33%	37%	34%	39%	1980
23	36%	30%	31%	38%	34%	28%	33%	37%	35%	30%	31%	1981
22	36%	26%	31%	32%	29%	30%	37%	37%	43%	36%	36%	1982
21	36%	29%	23%	40%	44%	29%	32%	33%	35%	26%	33%	1983
20	29%	26%	16%	27%	29%	30%	39%	41%	30%	30%	29%	1984
Age	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	Cohorte
	Période											

Annexe 4. Code MatLab pour la méthode PACE (Rust and Yeung, 1997)

Définition de la fonction de coût à minimiser

function h = costfun(u, X, Y, M1, M2, M3, M4, M5, Aeq, Beq)

Définition des paramètres de la fonction de coût

Le vecteur u contient les paramètres à estimé, à savoir les coefficients des indicatrices d'âge, période et cohorte et le terme constant de la régression, soit un total de 144 éléments.

Le vecteur b est un sous-vecteur de u contenant les coefficients des 143 indicatrices
beta=u(1:143);

Le scalaire μ est le dernier élément du vecteur u , c'est-à-dire le terme constant
mu = u(144);

Calcul du vecteur des résidus

mu_v=mu*ones(size(Y));
mu_diff = X*beta-Y;
epsilon = mu_diff + mu;

Définition des termes de la formule PACE

Groupes de coefficients pour l'âge, la période et la cohorte

A = beta(1:61);
P = beta(62:72);
C = beta(73:end);

Calcul des sommes de carrés par groupe

na = A'*A;
np = P'*P;
nc = C'*C;
ne = epsilon'*epsilon;

Formule PACE

h = M1 * (na+np+nc) + M2*(na*np+np*nc+nc*na) + M3*(na*np*nc) + M4*ne;

Ajout d'une condition pour satisfaire la contrainte d'optimisation :

h = h + M5 * sum(abs(Aeq*u-Beq));

Annexe 5. Tableau récapitulatif des résultats des méthodes PCR, PACE et PLS

	PCR : Effets nets	IC-INF : Borne inférieure de l'intervalle de confiance	IC-SUP : Borne supérieure de l'intervalle de confiance	PACE Effets nets	APC-PLS Effets nets	IC-INF Borne inférieure de l'intervalle de confiance	IC-SUP Borne supérieure de l'intervalle de confiance
<i>Ages</i>	PCR	IC-INF	IC-SUP	PACE	APC-PLS	IC-INF	IC-SUP
age20	-0,167	-0,194	-0,140	-0,181	-0,127	-0,163	-0,091
age21	-0,135	-0,160	-0,109	-0,156	-0,111	-0,139	-0,083
age22	-0,123	-0,148	-0,099	-0,142	-0,104	-0,131	-0,078
age23	-0,135	-0,159	-0,111	-0,153	-0,109	-0,131	-0,087
age24	-0,115	-0,139	-0,092	-0,130	-0,100	-0,122	-0,078
age25	-0,087	-0,110	-0,064	-0,104	-0,087	-0,107	-0,066
age26	-0,075	-0,098	-0,052	-0,088	-0,080	-0,101	-0,059
age27	-0,074	-0,096	-0,051	-0,088	-0,079	-0,102	-0,056
age28	-0,062	-0,084	-0,040	-0,079	-0,073	-0,088	-0,057
age29	-0,064	-0,086	-0,042	-0,076	-0,073	-0,096	-0,049
age30	-0,054	-0,076	-0,033	-0,073	-0,067	-0,087	-0,047
age31	-0,055	-0,077	-0,033	-0,069	-0,066	-0,083	-0,049
age32	-0,047	-0,069	-0,025	-0,068	-0,059	-0,075	-0,044
age33	-0,050	-0,072	-0,028	-0,074	-0,061	-0,077	-0,045
age34	-0,016	-0,038	0,006	-0,022	-0,043	-0,059	-0,026
age35	-0,054	-0,076	-0,031	-0,068	-0,059	-0,075	-0,043
age36	-0,048	-0,070	-0,026	-0,068	-0,055	-0,073	-0,036
age37	-0,035	-0,057	-0,012	-0,047	-0,046	-0,060	-0,032
age38	-0,016	-0,039	0,007	-0,033	-0,035	-0,045	-0,025
age39	-0,036	-0,059	-0,013	-0,040	-0,043	-0,059	-0,027
age40	-0,017	-0,040	0,006	-0,021	-0,031	-0,046	-0,016
age41	-0,025	-0,048	-0,002	-0,019	-0,032	-0,051	-0,014
age42	-0,031	-0,054	-0,008	-0,029	-0,032	-0,048	-0,016
age43	-0,021	-0,044	0,003	-0,021	-0,026	-0,042	-0,009
age44	-0,001	-0,024	0,023	0,019	-0,012	-0,027	0,003
age45	-0,002	-0,026	0,022	0,017	-0,010	-0,023	0,002
age46	0,000	-0,024	0,024	0,018	-0,007	-0,015	0,000
age47	0,004	-0,020	0,028	0,010	-0,003	-0,019	0,013
age48	0,000	-0,024	0,024	0,026	-0,002	-0,011	0,008
age49	0,010	-0,015	0,034	0,000	0,007	-0,007	0,021
age50	0,008	-0,016	0,033	0,009	0,010	-0,004	0,023
age51	0,021	-0,003	0,045	0,039	0,019	0,001	0,037
age52	0,008	-0,016	0,032	0,021	0,014	0,005	0,023
age53	0,017	-0,008	0,041	0,020	0,021	0,010	0,033
age54	0,025	0,001	0,049	0,034	0,028	0,014	0,042
age55	0,014	-0,010	0,038	0,013	0,024	0,017	0,031
age56	0,013	-0,010	0,037	0,027	0,026	0,011	0,042
age57	0,013	-0,011	0,037	0,030	0,027	0,020	0,034
age58	0,015	-0,009	0,038	0,032	0,029	0,011	0,046
age59	0,027	0,004	0,050	0,046	0,036	0,021	0,051
age60	0,008	-0,015	0,031	0,009	0,028	0,015	0,040
age61	0,018	-0,005	0,041	0,034	0,032	0,016	0,048
age62	0,019	-0,004	0,042	0,030	0,034	0,021	0,047
age63	0,032	0,010	0,055	0,054	0,041	0,028	0,055
age64	0,028	0,006	0,051	0,025	0,039	0,027	0,051
age65	0,037	0,015	0,059	0,050	0,045	0,028	0,062
age66	0,061	0,039	0,083	0,069	0,059	0,045	0,073

age67	0,067	0,045	0,089	0,076	0,062	0,048	0,077
age68	0,070	0,048	0,092	0,066	0,066	0,052	0,080
age69	0,061	0,040	0,083	0,064	0,063	0,048	0,078
age70	0,087	0,065	0,109	0,095	0,077	0,059	0,095
age71	0,075	0,053	0,097	0,077	0,071	0,052	0,090
age72	0,085	0,063	0,108	0,091	0,078	0,062	0,094
age73	0,090	0,068	0,112	0,088	0,082	0,058	0,107
age74	0,084	0,062	0,107	0,086	0,082	0,059	0,105
age75	0,083	0,060	0,106	0,086	0,084	0,066	0,101
age76	0,096	0,072	0,119	0,100	0,092	0,070	0,113
age77	0,107	0,083	0,131	0,113	0,100	0,075	0,125
age78	0,058	0,034	0,083	0,058	0,072	0,052	0,092
age79	0,093	0,067	0,118	0,097	0,088	0,061	0,115
age80	0,110	0,083	0,136	0,123	0,094	0,065	0,124
Cohortes	PCR	IC-INF	IC-SUP	PACE	APC-PLS	IC-INF	IC-SUP
coh1914	0,008	-0,015	0,030	0,010	0,021	-0,021	0,064
coh1915	0,017	-0,006	0,040	0,018	0,028	0,005	0,051
coh1916	0,004	-0,019	0,028	0,015	0,021	-0,012	0,054
coh1917	0,069	0,045	0,092	0,100	0,059	0,025	0,094
coh1918	0,078	0,055	0,102	0,105	0,068	0,040	0,096
coh1919	0,078	0,055	0,102	0,092	0,071	0,029	0,112
coh1920	0,078	0,055	0,101	0,093	0,074	0,044	0,104
coh1921	0,093	0,070	0,116	0,096	0,085	0,061	0,109
coh1922	0,069	0,046	0,092	0,061	0,076	0,044	0,107
coh1923	0,071	0,048	0,094	0,062	0,079	0,057	0,101
coh1924	0,073	0,050	0,096	0,066	0,082	0,062	0,101
coh1925	0,077	0,055	0,100	0,067	0,082	0,063	0,100
coh1926	0,065	0,043	0,087	0,062	0,075	0,053	0,096
coh1927	0,076	0,053	0,098	0,067	0,080	0,057	0,104
coh1928	0,066	0,044	0,088	0,059	0,073	0,057	0,089
coh1929	0,079	0,057	0,101	0,069	0,077	0,054	0,101
coh1930	0,049	0,027	0,071	0,043	0,059	0,039	0,080
coh1931	0,046	0,024	0,068	0,040	0,056	0,038	0,073
coh1932	0,072	0,050	0,094	0,065	0,065	0,045	0,086
coh1933	0,038	0,016	0,060	0,028	0,045	0,026	0,064
coh1934	0,063	0,041	0,086	0,066	0,055	0,035	0,075
coh1935	0,058	0,035	0,080	0,013	0,049	0,029	0,069
coh1936	0,044	0,022	0,067	0,037	0,040	0,026	0,054
coh1937	0,028	0,005	0,051	0,024	0,029	0,013	0,046
coh1938	0,061	0,038	0,084	0,045	0,044	0,024	0,063
coh1939	0,036	0,013	0,059	0,023	0,029	0,015	0,042
coh1940	0,041	0,018	0,065	0,033	0,031	0,020	0,041
coh1941	0,050	0,026	0,073	0,030	0,035	0,021	0,048
coh1942	0,042	0,018	0,065	0,034	0,029	0,019	0,040
coh1943	0,035	0,011	0,059	0,026	0,026	0,013	0,039
coh1944	0,051	0,027	0,075	0,040	0,034	0,019	0,048
coh1945	0,033	0,009	0,057	0,030	0,025	0,013	0,036
coh1946	0,037	0,013	0,061	0,014	0,026	0,013	0,038
coh1947	0,020	-0,004	0,044	0,017	0,016	0,009	0,023
coh1948	0,022	-0,002	0,046	0,008	0,016	0,011	0,022
coh1949	0,002	-0,022	0,026	-0,006	0,006	-0,004	0,015
coh1950	0,004	-0,020	0,028	0,002	0,006	-0,001	0,013
coh1951	-0,012	-0,036	0,013	-0,034	-0,004	-0,013	0,005
coh1952	-0,007	-0,031	0,017	-0,025	-0,004	-0,015	0,008
coh1953	-0,012	-0,036	0,012	0,001	-0,008	-0,019	0,003

coh1954	-0,019	-0,043	0,005	-0,029	-0,013	-0,020	-0,006
coh1955	-0,023	-0,047	0,001	-0,040	-0,017	-0,025	-0,010
coh1956	-0,046	-0,069	-0,022	-0,053	-0,030	-0,050	-0,010
coh1957	-0,031	-0,054	-0,008	-0,043	-0,024	-0,036	-0,012
coh1958	-0,035	-0,059	-0,012	-0,040	-0,029	-0,039	-0,018
coh1959	-0,024	-0,047	-0,001	-0,027	-0,025	-0,040	-0,011
coh1960	-0,051	-0,074	-0,028	-0,039	-0,040	-0,054	-0,026
coh1961	-0,065	-0,088	-0,043	-0,060	-0,049	-0,069	-0,030
coh1962	-0,041	-0,064	-0,019	-0,024	-0,038	-0,053	-0,024
coh1963	-0,073	-0,095	-0,050	-0,047	-0,055	-0,070	-0,041
coh1964	-0,077	-0,099	-0,055	-0,061	-0,059	-0,074	-0,043
coh1965	-0,068	-0,090	-0,046	-0,053	-0,057	-0,081	-0,033
coh1966	-0,069	-0,091	-0,047	-0,046	-0,058	-0,077	-0,039
coh1967	-0,076	-0,098	-0,054	-0,055	-0,065	-0,084	-0,045
coh1968	-0,074	-0,096	-0,052	-0,053	-0,065	-0,088	-0,043
coh1969	-0,069	-0,091	-0,047	-0,052	-0,065	-0,088	-0,041
coh1970	-0,078	-0,100	-0,056	-0,056	-0,072	-0,088	-0,055
coh1971	-0,079	-0,102	-0,057	-0,057	-0,078	-0,101	-0,055
coh1972	-0,066	-0,088	-0,044	-0,052	-0,075	-0,100	-0,050
coh1973	-0,097	-0,119	-0,074	-0,078	-0,094	-0,122	-0,067
coh1974	-0,094	-0,117	-0,071	-0,071	-0,098	-0,124	-0,072
coh1975	-0,091	-0,114	-0,068	-0,077	-0,097	-0,130	-0,063
coh1976	-0,082	-0,105	-0,059	-0,064	-0,090	-0,131	-0,049
coh1977	-0,077	-0,100	-0,054	-0,079	-0,085	-0,118	-0,053
coh1978	-0,061	-0,084	-0,037	-0,056	-0,077	-0,110	-0,044
coh1979	-0,061	-0,085	-0,038	-0,064	-0,076	-0,115	-0,037
coh1980	-0,032	-0,056	-0,009	-0,038	-0,059	-0,091	-0,028
coh1981	-0,033	-0,056	-0,009	0,003	-0,057	-0,096	-0,017
coh1982	-0,056	-0,079	-0,033	-0,089	-0,064	-0,132	0,005
coh1983	-0,037	-0,060	-0,014	-0,055	-0,050	-0,089	-0,011
coh1984	-0,028	-0,051	-0,006	-0,041	-0,040	-0,118	0,039
Périodes	PCR	IC-INF	IC-SUP	PACE	APC-PLS	IC-INF	IC-SUP
per1994	0,013	-0,007	0,032	0,013	0,022	-0,023	0,068
per1995	-0,061	-0,080	-0,042	-0,020	-0,021	-0,052	0,025
per1996	-0,063	-0,082	-0,044	-0,018	-0,025	-0,061	0,011
per1997	0,022	0,003	0,041	0,010	0,020	-0,025	0,067
per1998	-0,009	-0,028	0,010	-0,002	-0,001	-0,036	0,039
per1999	-0,006	-0,025	0,013	0,005	-0,003	-0,042	0,028
per2000	0,028	0,009	0,047	0,000	0,013	-0,026	0,052
per2001	0,028	0,009	0,048	0,012	0,010	-0,037	0,049
per2002	0,026	0,007	0,046	0,006	0,004	-0,041	0,045
per2003	0,023	0,003	0,042	0,001	-0,001	-0,044	0,038
per2004	-0,001	-0,021	0,018	-0,007	-0,018	-0,053	0,023

Annexe 6. Intervalles de sensibilité complémentaires

Dans cette annexe, nous examinons les intervalles de sensibilité pour une valeur de h allant jusqu'à de 0,2%, puis 0,5%.

Le fait que les écarts soient plus grands aux extrêmes qu'au centre des graphiques est lié au mode de calcul des intervalles de sensibilité. Il faut nécessairement un point de départ (ici central) pour appliquer ces perturbations qui se propagent ensuite dans le reste du graphique.

Il est préférable de placer au centre qu'à une extrémité (par exemple à droite), sous peine d'avoir des perturbations exagérées à l'autre extrémité (dans l'exemple à gauche). Une justification complémentaire provient aussi du fait que l'on observe moins souvent les très anciennes et les très jeunes générations, que les générations moyennes, qui seront donc moins sensibles aux perturbations. De même pour les âges, on observe moins de consommateurs très jeunes ou très âgés que de consommateurs d'âge moyen. Enfin, les périodes d'observation moyennes représentent elles aussi plus de générations que les autres.

Cette inégalité de répartition des perturbations représente néanmoins une limite des intervalles de sensibilité.

Une valeur de 0,2% pour le paramètre de variation des écarts consécutifs peut faire assez fortement varier la valeur des effets nets, car cette variation de 0,2% s'ajoute à l'écart entre chacune des modalités consécutives de chaque variable. Dans ce cas l'amplitude des effets nets de période, comme on peut le lire sur la figure 1 est amenée jusqu'à 6% (entre per1994 et per2004), ce qui constitue une valeur très élevée comparée la valeur quasi-nulle proposée par l'IAPI que l'on a choisie.

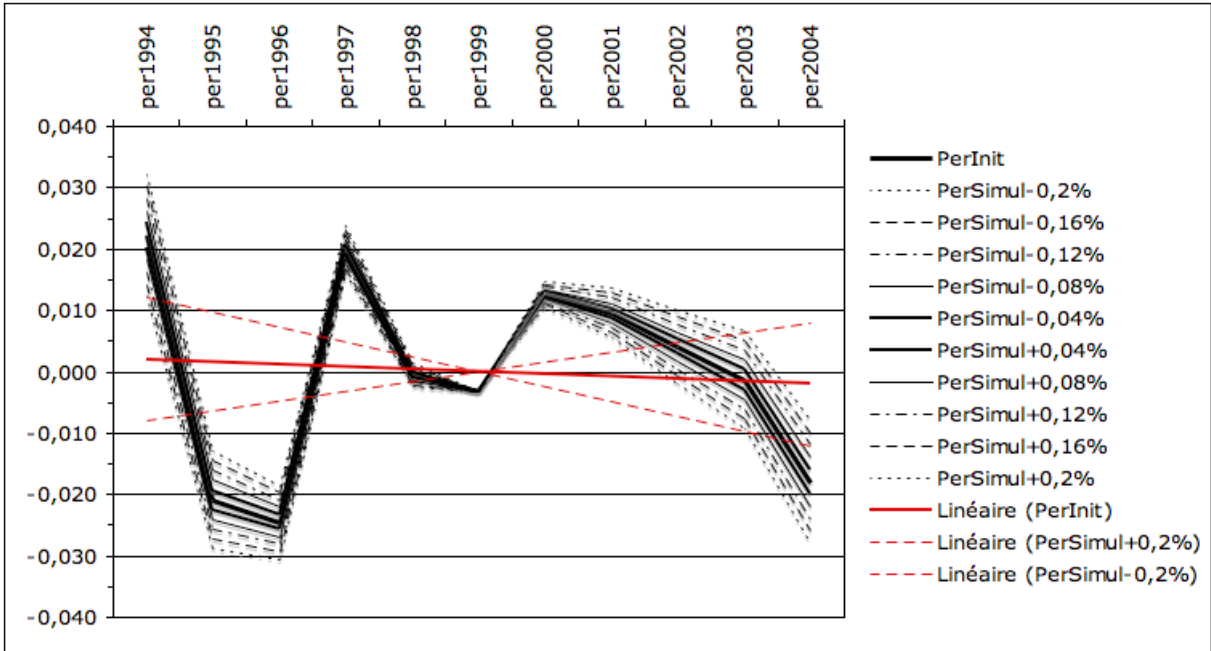


Figure Z.1. Intervalle de sensibilité des effets de période
à 0,04% ; 0,08% ; 0,12% ; 0,16% et 0,20%
(Agrandissement de l'échelle)

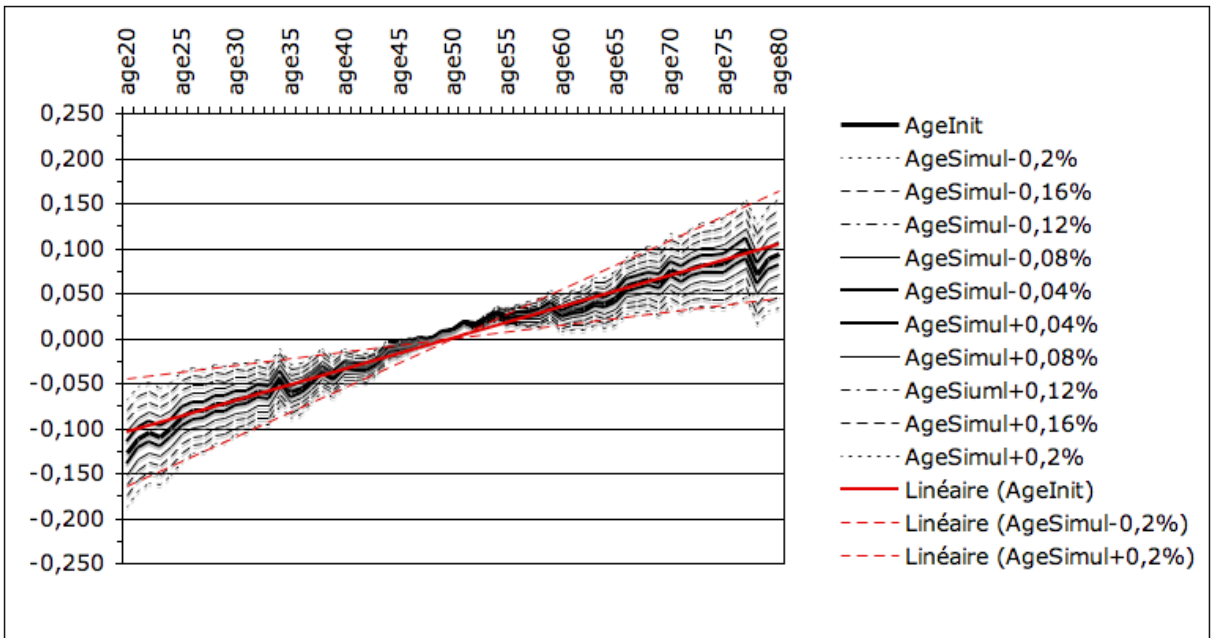


Figure Z.2. Intervalle de sensibilité des effets d'âge
à 0,04% ; 0,08% ; 0,12% ; 0,16% et 0,20%

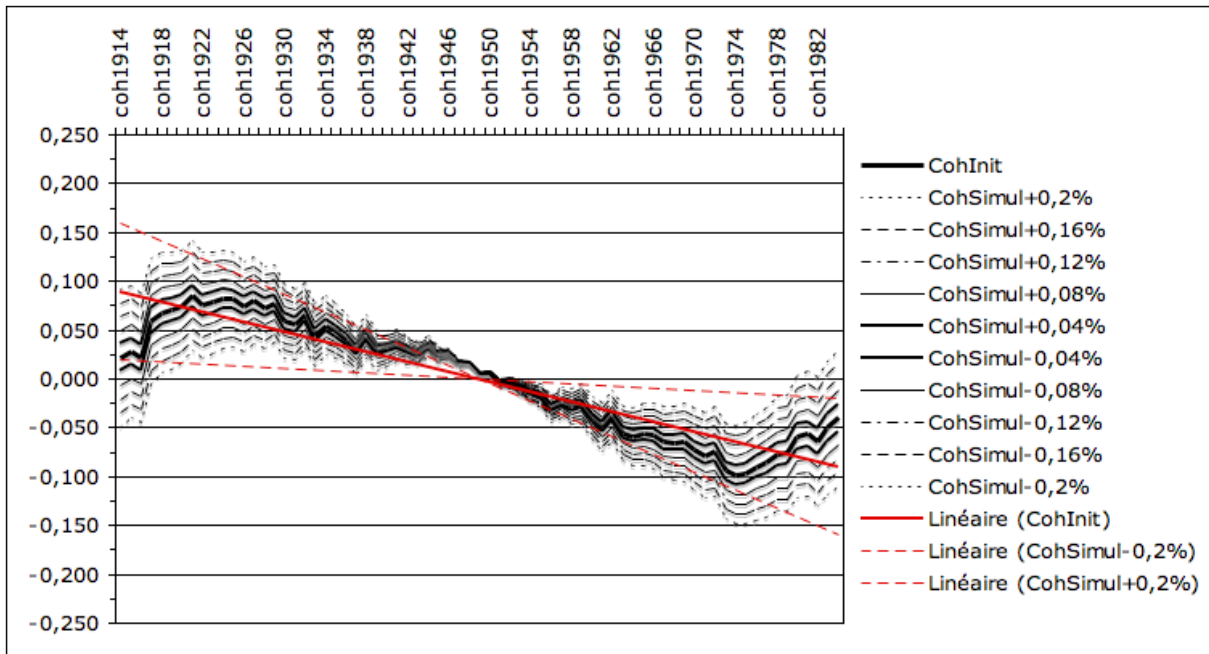


Figure Z.3. Intervalle de sensibilité des effets de cohorte
à 0,04% ; 0,08% ; 0,12% ; 0,16% et 0,20%

Etant donné la distance prise par rapport à l'IAPI, on pourrait craindre de fortes variations dans l'estimation des effets d'âge et de période, comme une inversion de la direction d'un des effets, ou son annulation. Comme le montrent les figures Z.1. et Z.3., on observe de plus grandes variations en amplitude, mais les résultats conservent les mêmes directions.

Pour atteindre les limites des résultats obtenus, nous allons prendre un intervalle de sensibilité à 0,5%. Ce cas correspond aux figures Z.4. à Z.6. suivantes.

On observe qu'en s'éloignant toujours plus loin de l'IAPI, on finit par avoir des résultats contradictoires sur les deux autres effets, et non plus seulement des variations d'amplitude. En effet, au sein de leur intervalle de sensibilité de 0,5%, les effets d'âge et de cohorte peuvent aller jusqu'à prendre des directions opposées (resp. décroissante et croissante).

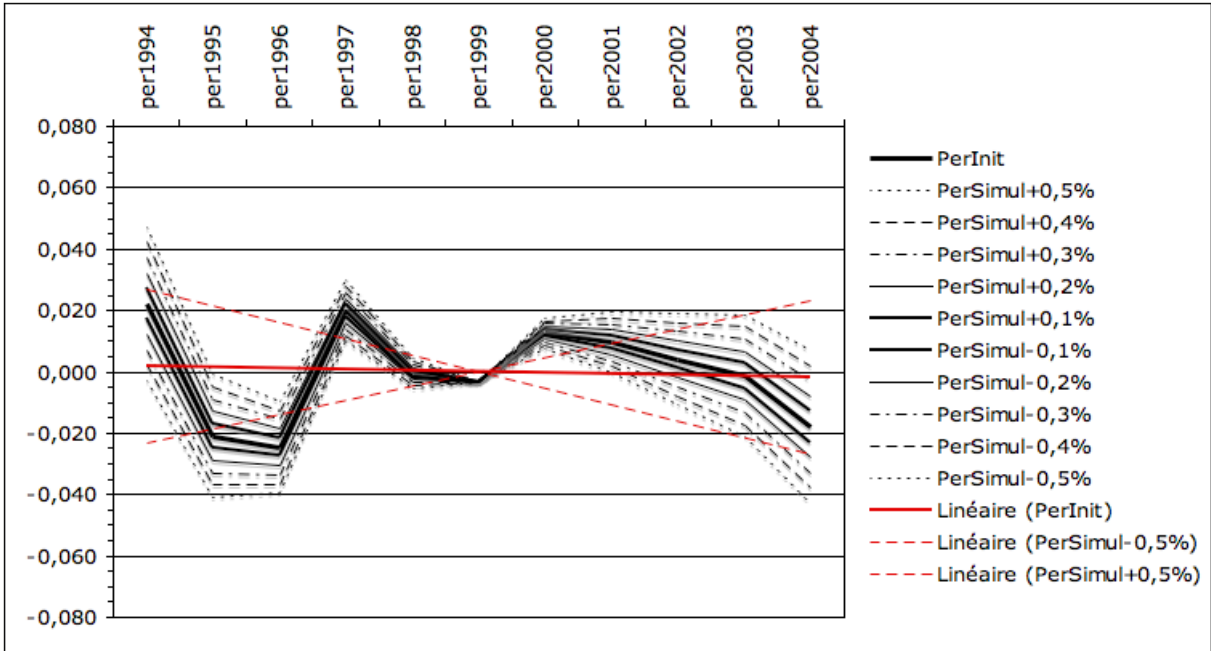


Figure Z.4. Intervalle de sensibilité des effets de période
à 0,1% ; 0,2% ; 0,3% ; 0,4% et 0,5%
(Agrandissement de l'échelle)

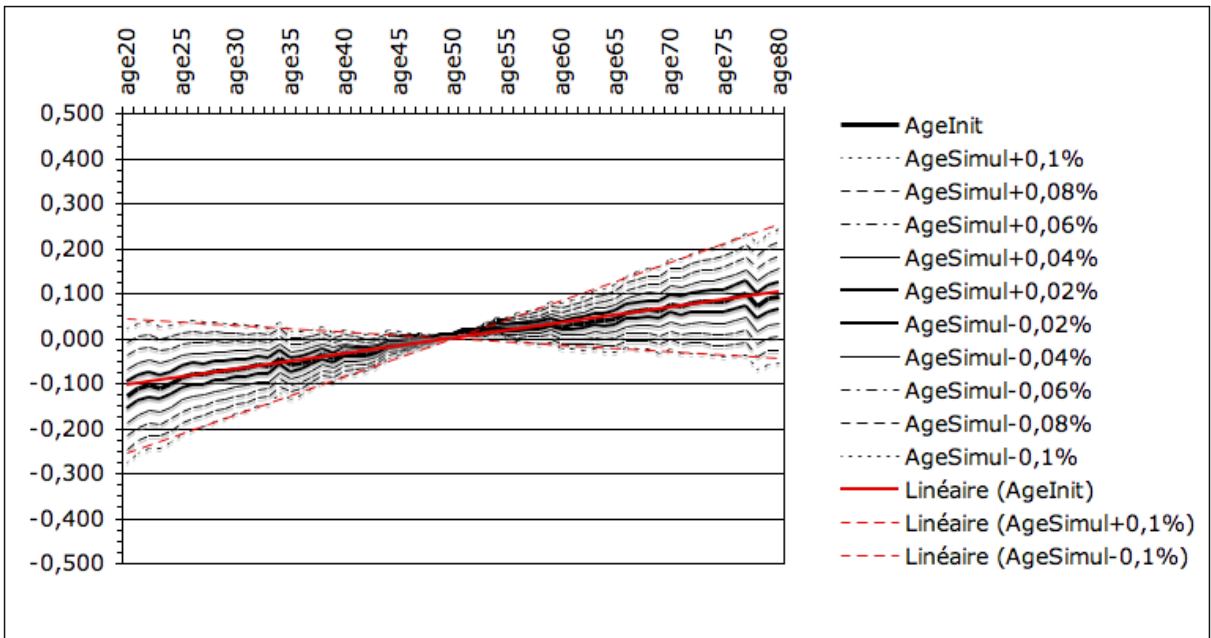
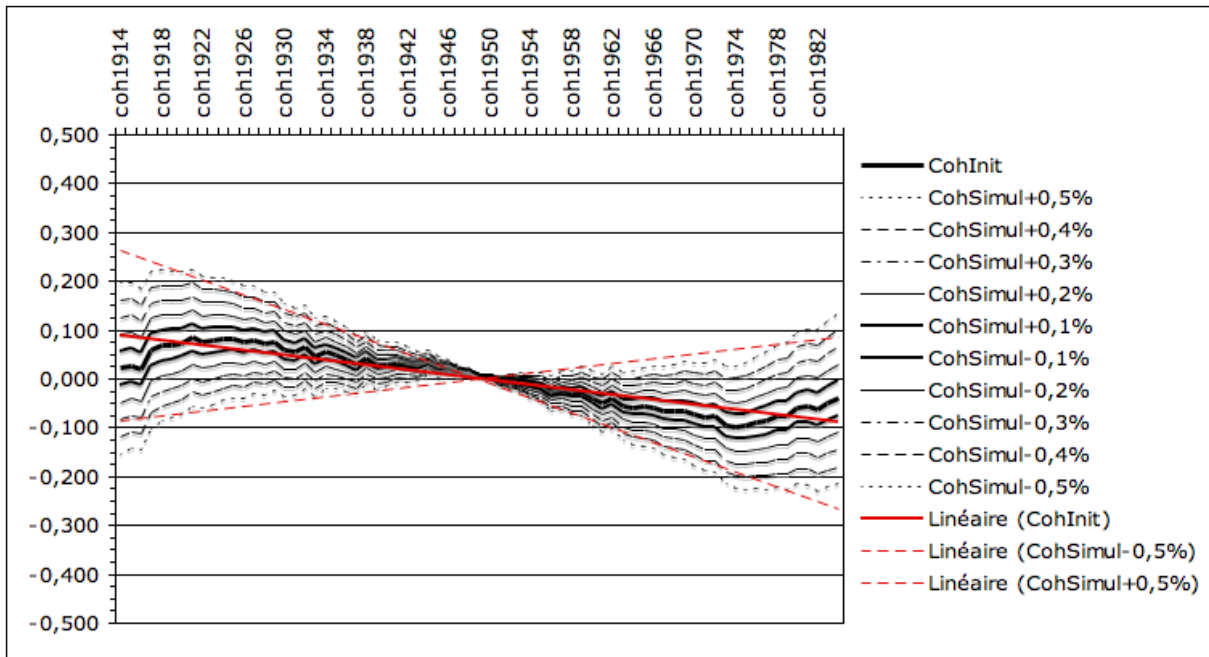


Figure Z.5. Intervalle de sensibilité des effets d'âge
à 0,1% ; 0,2% ; 0,3% ; 0,4% et 0,5%



*Figure Z.6. Intervalle de sensibilité des effets de cohorte
à 0,1% ; 0,2% ; 0,3% ; 0,4% et 0,5%*

À l'issue de cette analyse nous pouvons avoir une idée plus précise de la nature de la sensibilité des résultats obtenus sur les effets APC par rapport à l'information a priori initiale.

Annexe 7. Section du code MatLab pour la méthodologie AEL

Les commentaires de ces codes sont précédés dans le texte du symbole « % ».

N.B. Ces textes ne peuvent pas contenir d'accentuation.

La partie concernant les tests a été placée en italique.

```
%% Configuration
```

```
% 0 Pour resultats avec SVM, 1 pour resultats avec LDA
```

```
SVM = 1;
```

```
% Jeu de donnees a utiliser.
```

```
DATASET = 0;
```

```
%% Chargement des donnees
```

```
% Charge les valeurs des effets obtenues pour chaque variable, dans chaque  
% bin, pour chacune des 30 iterations (Monte Carlo).
```

```
load(sprintf('effects_%d%d.mat', SVM, DATASET));
```

```
if DATASET == 0,
```

```
    labs = {'extr', 'opp', 'camp', 'project', 'constr', 'conf'};
```

```
else,
```

```
    labs = {'MM price', 'CH price', 'Loyal', 'MM disc', 'CH disc'};
```

```
end;
```

```
% N : nombre de variables
```

```
N = length(labs);
```

```
%% Tableau des effets avec signe, moyenne et ecart-type
```

```
f = fopen(sprintf('results/effects_signs_%d%d.tsv', SVM, DATASET), 'w');
```

```
fprintf(f, 'bin');
```

```
for i = 1:N,
```

```

fprintf(f, '\t');
fprintf(f, '%s', labs{i});
end;
fprintf(f, '\n');
for i = 1:8,
    fprintf(f, '%d', i);
    for j = 1:N,
        values = zeros(1, 30);
        for k = 1:30,
            values(k) = effects{k}(j, i);
        end;
        fprintf(f, '\t%.2f ', mean(values));
        fprintf(f, '+/- %.2f', std(values));
    end;
    fprintf(f, '\n');
end;
fclose(f);

```

%% Graphique montrant la distribution des valeurs des effets
%% sur chaque iteration, avec meilleur modele gaussien.

```

clf
for i = 1:8,
    for j = 1:N,
        subplot(8, N, j + (i - 1) * N);
        values = zeros(1, 30);
        for k = 1:30,
            values(k) = effects{k}(j, i);
        end;
        hist([values], 8);
        [counts, centers] = hist([values], 8);
        points = linspace(centers(1), centers(end), 100);
        hold on;
        y = normpdf(points, mean(values), std(values));
    end;
end;

```

```

y = y / sum(y) * length(values) * length(points) / length(counts);
plot(points, y, 'r');
if j == 1,
    ylabel(sprintf('bin %d', i));
end;
if i == 8,
    xlabel(labs{j});
end;
end;
end;

%% Test de normalite de la distribution des effets.

f = fopen(sprintf('results/jarquebera_%d%d.tsv', SVM, DATASET), 'w');
fprintf(f, 'bin');
for i = 1:N,
    fprintf(f, '\t');
    fprintf(f, '%s', labs{i});
end;
fprintf(f, '\n');
for i = 1:8,
    fprintf(f, '%d', i);
    for j = 1:N,
        values = zeros(1, 30);
        for k = 1:30,
            values(k) = effects{k}(j, i);
        end;
        JB = length(values) / 6 * (skewness(values) .^2 + 0.25 *
(kurtosis(values) - 3) .^ 2);
        fprintf(f, '\t%.2f ', JB);
        if JB > 5.99,
            fprintf(f, 'REJECTED ', JB);
        else,

```

```

        fprintf(f, 'NORMAL ', JB);
    end;
end;
fprintf(f, '\n');
end;
fprintf(f, 'Jarque-Bera normality test. chi2cdf(JB, 2) = 5.99\n');
fclose(f);

```

*%% T-test testant si la difference observee dans les valeurs moyennes des
%% effets entre differentes bins est significative.*

```

for j = 1:N,
    f = fopen(sprintf('results/ttest_%d%d_%s.tsv', SVM, DATASET, labs{j}), 'w');
    fprintf(f, 'bin \ \ bin')
    for i = 1:8,
        fprintf(f, '\t%d', i);
    end;
    fprintf(f, '\n');
    for i = 1:8,
        fprintf(f, '%d', i);
        for k = 1:8,
            fprintf(f, '\t');
            values1 = zeros(1, 30);
            for l = 1:30,
                values1(l) = effects{l}(j, i);
            end;
            values2 = zeros(1, 30);
            for l = 1:30,
                values2(l) = effects{l}(j, k);
            end;
            t = mean(values1) - mean(values2);
            aa = var(values1) / length(values1);
            bb = var(values2) / length(values2);

```

```

    sdiff = (aa + bb) ^ 0.5;
    t = t / sdiff;
    df = (aa + bb) .^ 2 / ( aa .^ 2 / (length(values1) - 1) +
bb .^ 2 / (length(values2) - 1));
    criticalpoint = tinv(0.975, df);
    % Uniquement un cote de la diagonale!!!
    if (i > k),
        if abs(t) > criticalpoint,
            fprintf(f, 'T=%0.2f [DIFFERENT %0.2f]', t, criticalpoint);
        else,
            fprintf(f, 'T=%0.2f [REJECTED %0.2f]', t, criticalpoint);
        end;
    else,
        fprintf(f, ' ');
    end;
end;
fprintf(f, '\n');
end;
fclose(f);
end;

```

NB : Lorsque ce programme effectue les tests, de normalité ou d'égalité des moyennes, il se base sur les 30 itérations de MC. Ces itérations de MC, par leur caractère intrinsèquement stochastique, donnent naturellement des estimations très légèrement différentes concernant les valeurs des effets locaux et de leurs écarts-type (et notons qu'il n'y a aucun jeu d'itérations préférable à un autre, toutes les itérations étant issues d'une seule et identique procédure MC). Les valeurs numériques des statistiques de tests peuvent donc très légèrement varier à chaque exécution de ce programme. Cependant, les résultats de tests (validation ou non de l'hypothèse nulle pour un risque donné de 5%) ne varient pas, ce qui constitue le point essentiel d'une telle démarche de test, car les résultats de ces tests sont strictement équivalents.

Annexe 8. Rédaction « pas à pas » des récurrences de la démonstration concernant la méthode de report des écarts successifs.

Par construction de m' , on vérifie que $y(\theta'(A_1, P_j)) = y(\theta(A_1, P_j))$.

Pour une valeur de j fixée et en utilisant la définition de a'_{i+1} et de c'_{k-1} , on montre par récurrence ascendante sur i que cette propriété est vérifiée pour tout A_i .

Tout d'abord, fixons une valeur de j . Nous devons montrer que:

$$y(\theta'(A_i, P_j)) = y(\theta(A_i, P_j)) \text{ pour tout } i.$$

Amorçons la récurrence avec la valeur la plus faible de i , c'est-à-dire 1, car il s'agit d'une récurrence ascendante, et montrons que:

$$y(\theta'(A_1, P_j)) = y(\theta(A_1, P_j))$$

Rappelons que:

$$y(\theta'(A_1, P_j)) = a'_{1j} + p'_{1j} + c'_{1j-1} + m'$$

En remplaçant m' par sa valeur nous obtenons:

$$y(\theta'(A_1, P_j)) = a'_{1j} + p'_{1j} + c'_{1j-1} + m + (a_1 - \alpha) + (p_j - \pi) + (c_k - \gamma)$$

En remplaçant a'_{1j} par sa valeur :

$$y(\theta'(A_1, P_j)) = \alpha + p'_{1j} + c'_{1j-1} + m + (a_1 - \alpha) + (p_j - \pi) + (c_k - \gamma)$$

$$y(\theta'(A_1, P_j)) = a_1 + p'_{1j} + c'_{1j-1} + m + (p_j - \pi) + (c_k - \gamma)$$

Or:

$$y(\theta(A_1, P_j)) = a_1 + p_j + c_{j-1} + m$$

Calculons donc la différence suivante:

$$y(\theta'(A_1, P_j)) - y(\theta(A_1, P_j)) = a_1 + p'_{1j} + c'_{1j-1} + m + (p_j - \pi) + (c_k - \gamma) - a_1 - p_j - c_{j-1} - m$$

Et montrons que cette différence est nulle.

Par annulation deux à deux des termes, nous obtenons la différence:

$$y(\theta'(A_1, P_j)) - y(\theta(A_1, P_j)) = p'_j + c'_{j-1} + (p_j - \pi) + (c_k - \gamma) - p_j - c_{j-1}$$

Que nous pouvons réécrire ainsi en utilisant les définitions de π et de γ :

$$y(\theta'(A_1, P_j)) - y(\theta(A_1, P_j)) = p'_j - p_j + (p_j - p^j) + c'_{j-1} - c_{j-1} + (c_k - c^k)$$

$$y(\theta'(A_1, P_j)) - y(\theta(A_1, P_j)) = p'_j - p_j - (p^j - p_j) + c'_{j-1} - c_{j-1} - (c^k - c_k)$$

Or par construction de p'_j et de c'_{j-1} comme présenté page 202 à la suite de l'équation [21] :

$$p^j - p_j = p'_j - p_j + (J-j) * (-v)$$

$$c^k - c_k = c'_{j-1} - c_{j-1} + (K-j+1) * v$$

Et comme nous l'avons déjà rappelé plus haut, par définition, $K=J-1$, donc:

$$c^k - c_k = c'_{j-1} - c_{j-1} * (J-1-j+1) * v = c'_{j-1} - c_{j-1} + (J-j) * v$$

Ainsi, nous avons:

$$p^j - p_j - (p'_j - p_j) = (J-j) * (-v)$$

$$c^k - c_k - (c'_{j-1} - c_{j-1}) = (J-j) * v$$

En revenant à la différence étudiée dix lignes plus haut, nous pouvons conclure que:

$$y(\theta'(A_1, P_j)) - y(\theta(A_1, P_j)) = (J-j) * (-v) + (J-j) * v = 0$$

Donc:

$$y(\theta'(A_1, P_j)) = y(\theta(A_1, P_j))$$

La récurrence est ainsi amorcée.

Pour compléter cette récurrence, montrons que, pour un i quelconque supérieur ou égale à 1, nous avons la propriété suivante:

$$\text{Si } y(\theta'(A_i, P_j)) = y(\theta(A_i, P_j)) \text{ alors } y(\theta'(A_{i+1}, P_j)) = y(\theta(A_{i+1}, P_j))$$

Ceci consiste à montrer que $y(\theta'(A_{i+1}, P_j)) = y(\theta(A_{i+1}, P_j))$ en admettant l'hypothèse $y(\theta'(A_i, P_j)) = y(\theta(A_i, P_j))$

$$y(\theta'(A_{i+1}, P_j)) = a'_{i+1} + p'_j + c'_{j-i-1} + m'$$

En utilisant, comme indiqué plus haut, les définitions de a'_{i+1} et c'_{j-i-1} de la page 202, nous obtenons :

$$y(\theta'(A_{i+1}, P_j)) = a'_{i+1} + p'_j + c'_{j-i-1} + m' = a'_i + ((a_{i+1} - a_i) + v) + p'_j + c'_{j-i} - ((c_{j-i} - c_{j-i-1}) + v) + m'$$

Or d'après l'hypothèse de récurrence:

$$y(\theta'(A_i, P_j)) = y(\theta(A_i, P_j))$$

Ce qui est équivalent à :

$$a'_i + p'_j + c'_{j-i} + m' = a_i + p_j + c_{j-i} + m$$

En utilisant ainsi cette hypothèse nous en déduisons que:

$$y(\theta'(A_{i+1}, P_j)) = a'_i + ((a_{i+1} - a_i) + v) + p'_j + c'_{j-i} - ((c_{j-i} - c_{j-i-1}) + v) + m' = a_i + ((a_{i+1} - a_i) + v) + p_j + c_{j-i} - ((c_{j-i} - c_{j-i-1}) + v) + m$$

Par simplification des termes égaux deux à deux nous obtenons:

$$y(\theta'(A_{i+1}, P_j)) = a_{i+1} + v + p_j + c_{j-i-1} - v + m = a_{i+1} + p_j + c_{j-i-1} + m = y(\theta(A_{i+1}, P_j))$$

Ce qui conclut la démonstration par récurrence.

Pour une valeur de i fixée et en utilisant les définitions de p'_{j-1} et de c'_{k-1} , on montre par récurrence descendante sur j que cette propriété est vérifiée pour tout P_j .

Cette récurrence a exactement la même architecture que la précédente en fixant un i donné au lieu d'un j donné.

Elle s'effectue avec une amorce en J , et un pas de récurrence passant de j à $j-1$, car elle est descendante.

