APPLICATION DE L'ANALYSE DISCRIMINANTE AU TRAITEMENT DES NON-RÉPONSES AUX ENQUÊTES

Henri MARIOTTE

Présentation du problème

Importance du traitement des non-réponses

La contribution actuelle présente une partie des travaux effectués pendant un stage de deux mois dans la division Méthodes Statistiques et Sondages. Le contenu de cette étude tournait autour de l'analyse discriminante, méthode d'analyse des données classique mais pas toujours très utilisée, notamment à l'INSEE. Nous allons examiner ici les résultats issus de la mise en oeuvre de cette technique statistique. Pour cela, nous allons décrire une application de l'Analyse Discriminante au traitement des non-réponses. Ce phénomène est courant dans toutes les enquêtes, et son étude est la source de nombreux travaux, surtout aux États-Unis, au Canada ou en Suède. Ici nous allons explorer une voie de recherche assez originale.

Parmi les nombreux problèmes que posent la réalisation et le dépouillement des enquêtes, périodiques ou exceptionnelles, que fait l'INSEE, l'un des plus cruciaux est le traitement des non-réponses ou erreur de réponses, qu'elles soient dues aux enquêtés eux-mêmes, aux enquêteurs, ou encore à toute la chaîne de traitement informatique. Il est nécessaire de pouvoir mesurer les effets de ces réponses manquantes, tant sur les résultats que sur la précision de ces résultats. On cherche à connaître le biais introduit par les non-réponses, pour le corriger, ainsi que l'influence que celles-ci peuvent avoir sur la variance des variables étudiées. Plusieurs méthodes sont employées ou explorées pour résoudre cette question, avec deux axes essentiels :

- pondérer les individus pour tenir compte des individus manquants, en se calant sur un certain nombre de variables sociodémographiques importantes, telles que le sexe, l'âge, la catégorie socioprofessionnelle, de façon à ce que les répartitions de ces variables sur l'échantillon et dans la population générale (connue principalement par la dernière enquête-emploi) soient aussi proches que possible. Chacun des individus

présents dans l'échantillon compte pour une partie des individus absents pour cause de non-réponse. Cette méthode, stratification ou redressement *a posteriori*, comparable à la méthode des quotas, suppose le paramètre d'intérêt bien corrélé avec les variables de stratification, mais très peu avec le fait d'être ou non répondant. Il sera très utile, en particulier, pour la correction rendue nécessaire par l'absence complète de certains questionnaires, non-réponses globales.

- affecter aux individus non répondants, par certaines méthodes, une valeur sur une ou des variables, en recherchant par exemple la valeur la plus probable. La valeur affectée peut être issue d'un tirage aléatoire parmi tous les répondants, ou seulement un sous-groupe : c'est le principe du Hot-Deck. Le sous-groupe peut être choisi sur la base des réponses à d'autres questions, que l'on pense liées (corrélées) à la question étudiée, telles que les variables sociodémographiques ou géographiques, notamment. Une autre façon de procéder consiste à postuler un modèle linéaire, qui relie la variable étudiée et d'autres variables (les mêmes que pour le Hot-Deck, par exemple). La valeur attribuée aux non répondants sera issue de ce modèle.

C'est par une méthode de type imputation que nous chercherons ici à traiter les non-réponses. Quand la variable étudiée est de type qualitatif, on se trouve typiquement devant un cas d'application de l'analyse discriminante. Aussi, nous présenterons le Hot-Deck et l'utiliserons avant de développer la méthode de l'analyse discriminante, pour en comparer les résultats.

Le hot-deck

Le Hot-Deck est la procédure de correction la plus habituelle, car simple d'emploi et relativement efficace. Elle consiste à affecter une valeur donnée sur une (ou plusieurs) variables à un individu pour lequel on ne connaît pas la vraie valeur. Il en existe des variantes plus ou moins frustes ou élaborées. Mais le principe général est de déterminer un autre individu, "complet", c'est-à-dire ayant répondu à la question traitée, et d'affecter sa réponse à l'observation en cours de traitement. La manière la plus simple de procéder est de choisir le précédent dans la liste, en considérant que l'ordre d'apparition (de saisie) n'est pas aléatoire, et qu'il y a un lien entre un individu et ces successeurs ou ses prédécesseurs. Mais cette méthode peut être améliorée et précisée en ne conservant, comme choix d'individus "complets", que ceux ayant certaines caractéristiques en commun avec l'individu dont on effectue le traitement. On optera, le plus souvent, pour des variables du type sexe, âge, niveau de diplôme, catégories socioprofessionnelles. Mais pour certaines études ou variables plus spécifiques, on pourra choisir, pour affecter une valeur particulière, la région d'habitation, le type (taille, centre ou banlieue) d'unité urbaine de résidence, le revenu (encore que ce soit souvent la variable la plus mal connue, car comportant beaucoup de refus de répondre). On choisira donc le précédent individu ayant en commun 2, 3 ou même 4 caractéristi-

ques avec notre non-répondant. Le nombre choisi pourra être fonction de la richesse de l'échantillon ainsi que de la qualité, et de la fiabilité recherchée. Il n'y a en principe comme limite au nombre de variables choisies que la possibilité de trouver d'autres individus identiques sur celles-ci, donc la taille de l'échantillon, ou la lourdeur de la manipulation mise en œuvre. On peut, de plus, estimer qu'il y a un nombre au-delà duquel l'amélioration devient marginale. On constate que les hypothèses de base de cette méthode sont les mêmes que celles de la stratification *a posteriori*, à savoir bonne corrélation avec les variables choisies pour la correction et indépendance vis-à-vis du fait d'accepter ou de refuser de répondre, ou de ne pas être exploitable statistiquement pour toute autre sorte de raison. On pourra résumer le principe général de cette méthode en deux étapes:

- la détermination d'un voisinage adéquat de l'individu que l'on veut compléter, constitué d'observations renseignées pour la variable d'intérêt, pour un certain nombre de variables bien choisies;

- le tirage au hasard de l'une des valeurs observées dans ce voisinage, par exemple le choix de l'individu précédent dans le fichier, et appartenant à ce voisinage.

Méthodologie employée

Choix de variables

L'enquête sur laquelle sera effectué tout le traitement des non-réponses appartient à la série des enquêtes mensuelles de conjoncture de l'INSEE, auprès des ménages. L'étude porte sur un échantillon de 1 452 personnes, issues de l'échantillon-maître, réparties de façon "représentatives" géographiquement et socialement. Elle utilise des variables descriptives sociodémographiques, géographiques..., et des questions d'opinion sur la situation économique vécue et ressentie par les ménages. Ces questions ont pour objectif de mettre en parallèle les situations telles qu'elles sont estimées par les organismes de prévision, et telles qu'elles sont ressenties par le public, afin de pouvoir affiner les prévisions à court terme de l'INSEE en matière de conjoncture économique et sociale.
Les variables utilisées lors de l'analyse sont les suivantes :

- d'une part des variables décrivant le sexe, l'âge, la catégorie socioprofessionnelle, le niveau de diplôme, le type d'emploi (temps plein ou partiel) ou de non emploi

(1) L'échantillon-maître est une base de sondage constituée de tous les logements, auxquels sont ajoutés annuellement les logements neufs.

(malade, chômeur, retraité, inactif, femme au foyer) de l'enquêté, l'âge du conjoint éventuel, le revenu, le nombre de personnes, le nombre d'actifs, le nombre d'enfants du ménage, le nombre de pièces du logement, la date d'achèvement de l'immeuble, la région d'habitation ;

- d'autre part 18 questions d'opinion du répondant concernant le niveau de vie, les prix, sa situation financière, ses prévisions d'achats, de logement ou d'équipement, son sentiment quant à l'opportunité d'achats importants ou d'épargne. Ces variables comportent de 2 à 5 modalités de réponses et un nombre de non-réponses (ou réponses "ne sait pas") variant entre 4 à 151, et atteint même 484 pour une question particulière, dont la réponse dépend de la question précédente.

Les questions dont on étudiera les réponses manquantes seront 3 questions d'opinion, les autres n'ayant souvent qu'une ou deux dizaines de non-réponses. L'amélioration apportée par une procédure plus élaborée d'affectation *a posteriori* ne peut alors qu'être marginale. On étudiera aussi la variable "revenu" du ménage qui comporte 133 refus de répondre. Il peut être intéressant de la corriger du mieux possible pour une meilleure connaissance globale de sa répartition.

Pour chacune de ces variables, on étudiera la meilleure manière d'affecter une valeur fictive en lieu et place de la réponse manquante, en choisissant la plus probable, au vu des résultats d'une analyse statistique. On comparera, d'autre part, les résultats obtenus avec ceux qui sont fournis par un traitement utilisant la méthode du Hot-Deck.

On pourrait penser à une méthode plus fine, mais plus complexe d'affectation, non plus d'une modalité précise et définie, mais d'une probabilité affectée à chaque modalité. On créerait ainsi, de manière fictive, p individus statistiques (pour p modalités possibles de la variable d'intérêt), pondérés par la probabilité correspondante, et remplaçant l'individu de départ pour tout le traitement statistique. On pourrait retrouver des résultats entiers (plus souhaitables pour la publication des résultats) par strates ou groupes pré-déterminés, et non plus par individu. Ceci pourrait donner des résultats plus exacts pour des moyennes, des totaux, des répartitions. Mais on ne sait rien de la variance des estimateurs ainsi construits. Nous examinerons rapidement cette méthode par la suite.

Les procédures utilisées

L'Analyse Discriminante, portant sur des variables quantitatives, postule *a priori* ou estime *a posteriori* les lois de probabilité, modélisant la distribution des individus dans les différents groupes. Pour cela on peut considérer que chaque groupe présente une distribution commune, mais décalée (paramètre de position ou d'échelle), ou une distribution spécifique. On peut ainsi déterminer des probabilités d'appartenance à ces

groupes. Les méthodes d'estimation peuvent être paramétriques (linéaires, quadratiques), ou non paramétriques, mais les variables utilisées pour le calcul sont, en tout état de cause, numériques, ou au moins ordinales. Qu'en est-il si on veut utiliser des variables qualitatives, dont les modalités ne peuvent pas être ordonnées selon un classement évident, telles que la catégorie socioprofessionnelle ou la région d'habitation par exemple? Pour résoudre ce problème, on peut penser, *a priori*, à trois procédures statistiques :

- rendre numériques toutes les variables qualitatives par l'intermédiaire d'une Analyse en Correspondances Multiples. Cette technique statistique permet, en effet, à partir de données proposées sous forme de tableaux de contingence portant sur plusieurs variables qualitatives - tableaux de BURT -, ou de modalités transformées en variables indicatrices pour obtenir un tableau disjonctif classique, d'obtenir des axes factoriels principaux sur lesquels sont représentées des variables qui elles sont numériques. Ces variables sont donc alors directement utilisables, associées à d'autres variables numériques (âge, revenu...), dans une procédure classique d'analyse discriminante ;

- une deuxième façon de procéder, assez proche de la précédente, sera de transformer en axes factoriels, ou facteurs principaux, non pas seulement les variables qualitatives, mais toutes les variables, après avoir découpé les variables numériques en un nombre de modalités comparable à celui des variables qualitatives (on retrouve là une des règles de pratique de l'A.C.M. pour une meilleure qualité et validité des résultats). Cette méthode a été proposée par SAPORTA, sous le non de DISQUAL. On constate que cette méthode donne souvent des résultats meilleurs que la précédente. Cela peut s'expliquer par une plus grande homogénéité des variables soumises à l'analyse (tant pour les variables introduites dans l'A.C.M, que dans les variables factorielles qui en sont issues) ;

- enfin une troisième méthode sera également utilisée, et elle a l'avantage d'une certaine simplicité d'emploi. Pour obtenir des variables numériques en entrée de l'analyse discriminante, on pourra associer à chaque modalité de chaque variable qualitative à étudier la variable indicatrice correspondante, valant 1 pour les observations possédant cette modalité, et 0 pour toutes les autres. On pourrait se limiter, dans cette transformation, aux variables qualitatives, mais il semble préférable, empiriquement et intuitivement, d'effectuer ce travail sur toutes les variables qu'on souhaite introduire dans l'analyse. Cette précaution évitera de travailler sur des variables continues (ou presque) d'une part, et des variables à valeurs 0 ou 1 (binaires) d'autre part. On retrouve là un souci d'homogénéité des données introduites dans l'analyse. On remarquera que cette troisième méthode peut sembler très contestable au point de vue théorique. Comment postuler la normalité d'une variable dichotomique, sans modèle sous-jacent (de type Probit ou Logit)? Ceci paraît une grossière approximation. Ici leur introduction est destinée à en comparer les résultats avec ceux issus des méthodes, plus classiques, du paragraphe précédent. On constate,

alors, empiriquement, qu'il n'y a pas de différence significative. Malgré l'absence de justification théorique, l'utilisation de variables indicatrices se révèle, à l'expérience, tout à fait exploitable. Il est intéressant d'insister sur cette apparente contradiction entre, d'une part, une théorie qui rejette cet usage, et, d'autre part, une pratique qui le rapproche de méthodes plus correctes.

Parmi ces trois méthodes, seront le plus explorées et le plus utilisées, dans ce travail, la seconde et la troisième. Mais dans chacune, il y a lieu de déterminer et sélectionner un certain nombre d'axes principaux, ou de variables indicatrices, sur lesquelles portera l'analyse discriminante ultérieure. Outre la procédure de sélection discriminante pas à pas, dont le principe est celui d'une régression linéaire sur les variables choisies, avec l'introduction de covariables, deux autres pistes de recherche peuvent donner des résultats utiles : on pourra traiter en observations supplémentaires les individus ayant une réponse manquante sur la variable d'intérêt ou en variable supplémentaire la variable indicatrice valant 1 pour tous ces individus (il faudra alors faire porter l'A.C.M. sur toutes les observations), et rechercher les axes de l'Analyse des Correspondances Multiples sur lesquels ces individus, ou cette variable, sont le mieux représentés. L'hypothèse sous-jacente sera de considérer que les axes principaux qui discriminent le mieux ces observations sont ceux qui les représentent le mieux à la sortie de l'A.C.M. Mais il faudra vérifier cette hypothèse, qui peut paraître nécessaire mais non suffisante, par la suite. De manière symétrique, on pourra chercher à déterminer les axes factoriels qui représentent le mieux les variables indicatrices issues des individus renseignés sur la variable étudiée. Ceci a l'avantage d'être cohérent avec la méthode de régression pas à pas, puisqu'on cherche, par deux moyens différents, à expliquer au mieux les modalités de notre variable par des variables numériques pertinentes. On constate empiriquement que les résultats de cette méthode sont assez proches de ceux issus de la régression pas à pas.

Résultats et critères d'appréciation

Le revenu : choix des variables

Comme dans beaucoup d'enquêtes, réalisées par l'INSEE ou d'autres organismes, la variable "revenu" fait apparaître beaucoup de valeurs manquantes 133 sur 1452, soit 9,2%. C'est souvent, pour des raisons sociologiques et "fiscales" sans doute, une des caractéristiques sur laquelle il y a le plus de rétention d'information de la part des ménages. Elle est donc souvent assez difficile à connaître avec précision. Aussi a-t-il paru utile de la traiter dans le cadre de ce travail, pour la corriger du mieux possible. Et ceci peut être d'autant plus utile que cette variable est souvent bien explicative, même si d'autres le sont plus, dans l'étude de beaucoup de phénomènes sociaux et économiques.

Dans un premier temps, à l'issue de l'A.C.M., quels axes factoriels faut-il choisir ? La première méthode, la plus habituelle, est la procédure d'analyse discriminante pas à pas, et elle nous donne, dans l'ordre les axes 1, 4, 3, 2, 5, 26, 28, 6 et 7 qu'on a introduits dans l'analyse. Les statistiques de Fisher associées au modèle, avec covariables (variables factorielles déjà sélectionnées), ou sans covariables, sont les suivantes :

Axe	Statistique de Fisher de la régression	
	avec covariables	sans covariables
1	84,7	84,7
4	17,6	14,3
3	15,4	11,3
2	13,2	9,9
5	8,2	6,5
26	5,9	4,0
28	5,1	4,4
6	4,3	2,9
7	4,2	3,5
13	4,1	3,0
(8		3,0)

On peut remarquer que tous les premiers axes viennent au début de la sélection, même si c'est dans le désordre. Ceci peut paraître logique car, si on a bien choisi ses variables explicatives au départ, les modalités qui jouent un rôle complémentaire et important seront bien représentées sur les premiers axes, à quelques exceptions près. C'est là un résultat assez général, qu'on retrouve dans les exemples traités, sauf un. Seul change le nombre de ces axes qui peut être de 3 ou plus. Mais le premier axe factoriel est toujours parmi les premiers axes sélectionnés. Le fait de ne choisir que des axes très loin des premiers, dans l'ordre des valeurs propres décroissantes, pourrait même être un indice du fait que les variables qualitatives choisies pour l'étude ne sont pas pertinentes. Si on avait examiné une régression de chacune des variables factorielles sur les 9 variables indicatrices définies par les modalités de la variable "revenu", sans covariables, on aurait sensiblement le même classement à une exception près. Ce résultat est donné par la première étape de la procédure pas à pas, et s'explique par l'orthogonalité des axes factoriels. On a ainsi un moyen plus rapide de choisir les variables utiles.

Si on cherche à déterminer les axes représentant le mieux les individus supplémentaires, à valeurs manquantes, on peut procéder de la manière suivante : sommer sur tous ces individus leur qualité de représentation, proportionnelle au cosinus carré, ce qui nous donne des résultats très différents. En effet si les axes 1 et 2 sortent d'abord, c'est ensuite l'axe 8 qui donne les meilleures valeurs. Ceci peut s'expliquer, car on n'utilise pas le même critère dans les deux cas : d'un côté on recherche les variables qui discriminent le mieux les modalités de la variable "revenu" pour ses valeur non-manquantes, de l'autre on cherche les axes qui représentent le mieux les individus à valeurs manquantes. Il n'y a aucune raison pour qu'on trouve des résultats convergents. Cette dernière méthode semble donc moins bien convenir pour chercher une règle d'affectation, puisqu'alors on a besoin d'utiliser les individus renseignés, donc de bien les discriminer.

Une méthode qui peut sembler plus efficace que la précédente sera de traiter les variables indicatrices issues des modalités du "revenu", soit 9 variables, une par modalité, et une pour les réponses absentes. On va alors chercher les axes qui représentent au mieux les 9 premières d'une part, et la dernière (non-réponses) d'autre part. Examinons donc ses résultats : ils sont très cohérents, pour les premiers, avec la procédure de sélection discriminante pas à pas, avec un ordre légèrement différent, et un axe qui s'ajoute, encore l'axe factoriel 8. Si on ajoute les qualités de représentation (en fait les cosinus carrés) des 9 variables indicatrices on trouve l'ordre suivant : axes 1, 4, 2, 3, 5, 24, 27, 8, 7, 28 et 10.

Axe	Somme des cosinus carrés des variables indicatrices rev1-rev9
1	0,342
4	0,0865
2	0,0684
3	0,0553
5	0,0452
24	0,0249
27	0,0210
8	0,0209
7	0,0193
28	0,0190
10	0,0190

On retrouve, de façon différente, à peu près les mêmes axes, qui représentent le mieux les modalités du "revenu", et on peut donc penser, à juste titre sur cet exemple, qu'ils les discrimineront mieux. Si on examine les axes qui représentent le mieux la modalité réponse manquante, on ne trouve pas du tout les mêmes axes mis à part l'axe 1, mais les axes 16, 14, 11, 8... On confirme par là ce qui était vu précédemment sur la différence qu'il y a entre chercher à bien représenter les modalités renseignées et la modalité réponse absente.

Si on veut pratiquer une analyse discriminante directement sur les variables indicatrices, quelles sont celles qui sont sélectionnées en priorité? On trouve les variables suivantes :

Variable	Nom développé	Statistique de Fisher
ACTI	Actif	40,4
CSUP	Cadre supérieur	28,5
PLIB	Profession libérale	22,1
HOM_CP	Homme en couple	20,5
CMOY	Profession intermédiaire	21,8
AGRI	Agriculteur	9,6
ZEA1	Région parisienne	7,7
PRIM	Études primaires	6,8
AGE1	Moins de 30 ans	6,1
RETR	Retraité	4,6
NP1	Ménage d'une personne	4,8
ARTI	Artisan ou Commerçant	4,4
FEM_S	Femme seule	
SUPE	Études supérieures	moins de 3
PART	Emploi à temps partiel	probabilités
FOYE	Personne au foyer	
INAC	Inactif	entre 10^{-3} et $2x10^{-2}$

Le revenu : comparatif des résultats

Tout d'abord, nous allons évoquer, pour l'éliminer, une piste qui semblait intéressante. Une analyse non paramétrique donnait des résultats très intéressants, avec un taux d'erreur d'affectation de moins de 12%! Mais dès qu'on prenait le soin d'estimer le taux d'erreur par une validation croisée (où l'individu traité est éliminé de l'échantillon pour le calcul de la probabilité), ce taux d'erreur dépassait 80%. La densité construite était complètement *ad hoc*, et ne convenait pas au traitement du problème. On peut ainsi renouveler une règle impérative : en cas d'analyse non paramétrique (et aussi d'analyse paramétrique), même sur échantillon nombreux (1319 observations renseignées), il est nécessaire d'effectuer une validation croisée, ou sur échantillon test (mais ici les données n'étaient pas assez nombreuses). On constate alors la limite d'utilisation des méthodes non paramétriques. Outre la difficulté de choisir les bons paramètres initiaux, on se trouve confronté à la lourdeur informatique de cette procédure.

Nous en resterons donc aux méthodes paramétriques, mais parmi celles-ci, l'une s'est avérée inexploitable, à savoir la méthode quadratique sur variables indicatrices. En effet les modalités 1, 2, 3 et 9 de la variable "revenu" étant de faible effectif, les matrices de variances de ces groupes connaissaient des problèmes de colinéarité, car certaines variables disjonctives y étaient presque constantes. Ainsi les résultats d'affectation s'en ressentaient : plus de 80% de taux d'erreur. Nous en restons donc à trois méthodes dont nous allons comparer les affectations et les taux d'erreur issus de méthodes de validation croisée. Dans les tableaux récapitulatifs suivants, les modalités voisines sont les 2 plus proches (ou la plus proche pour les valeurs extrêmes) de la modalité initiale de l'individu étudié.

Méthode linéaire sur variables indicatrices en % :

Bien classés	(34,6%)
Classés dans les modalités voisines	(35,4%)

Méthode linéaire sur variables factorielles en % :

Bien classés	(34,6%)
Classés dans les modalités voisines	(35,9%)

Méthode quadratique sur variables factorielles en % :

Bien classés	(33,4%)
Classés dans les modalités voisines	(34,7%)

Ces résultats sont très proches, mais on peut essayer de les comparer de plus près (observation par observation, par tris croisés). On pourra ainsi les améliorer un peu. Sur

1319 observations, 158 ont trois affectations différentes par les trois méthodes, 617 ont deux résultats communs et 544 ont trois résultats communs.

Si on compare la vraie valeur avec l'affectation déterminée, pour ces 544 observations, on a :

Bien classés	242 soit (44,5%)
Classés dans les modalités voisines	181 soit (33,3%)

On constate donc dans le cas de trois affectations identiques un meilleur taux de bonne réponse, et c'est rassurant quant à la qualité des procédures statistiques et du présent travail. Le même travail sur les 617 individus, en affectant la valeur commune à deux affectations, donne comme résultats :

Bien classés	198 soit 32,1%
Classés dans les modalités voisines	211 soit 34,2%

Enfin pour les individus restants, on peut comparer les résultats obtenus (moins bons que les précédents, et ceci paraît logique):

Méthode	Bien classés	Modalités voisines
1	33 soit 20,9%	49 soit 31,0%
2	33 soit 20,9%	64 soit 40.5%
3	27 soit 17,1%	56 soit 35.4%

Donc si on utilise la procédure d'affectation suivante : choix de la valeur commune, dès qu'il en existe une (qu'elle soit citée 2 ou 3 fois), méthode 2, c'est-à-dire linéaire sur variables factorielles dans les autres cas, on obtiendra globalement les résultats suivants:

Bien classés	471 soit (35,7%)
Classés dans les modalités voisines	456 soit (34,6%)

Le résultat est certes amélioré par cette procédure de traitement, mais ce petit progrès de 1.1% sur les bien classés (15 individus sur 1319) justifie-t-il une opération un peu complexe? Le choix immédiat d'une des deux méthodes linéaires peut s'avérer préférable dans sa simplicité. D'autre part le passage par une A.C.M. peut lui aussi être remis en question, dans la mesure ou l'analyse discriminante linéaire sur les variables indicatrices donne presque le même résultat que si on utilise les variables factorielles. La seule différence est de 0.5%, soit 6 observations. La méthode la plus simple s'avère presque aussi efficace que la plus complexe: différence de bien classés de 1,1% (15 individus).

Le dernier temps de cette étude sera de comparer les résultats obtenus par analyse discriminante à ceux qu'on obtiendrait par une affectation utilisant le Hot-Deck portant sur quatre variables: sexe, âge, niveau de diplôme et catégorie socio-professionnelle. On pourra estimer ainsi l'apport de l'analyse discriminante. Nous considérerons donc comme valeurs manquantes 10% des individus renseignés (131 individus environ) pour pouvoir ensuite comparer les résultats du Hot-Deck et de l'analyse discriminante aux vraies valeurs, et ainsi déterminer s'il y a apport ou non de l'analyse discriminante. La répartition de la variable "revenu" sur les 133 individus du test pour les trois variables, vraie valeur, valeur corrigée par analyse discriminante (REVDI) et valeur corrigée par le Hot-Deck (REVHD) sont les suivantes :

Groupes	Vraie valeur	REVDI	REVHD
1	2,3	0,0	3,0
2	2,3	0,0	2,3
3	5,3	8,3	5,3
4	13,5	4,5	9,0
5	17,3	25,6	13,5
6	15,8	10,5	11,3
7	15.8	9.0	15.0
8	22.6	36.8	36.1
9	5.3	2.3	4.5

On retrouve sur ces résultats un défaut de l'analyse discriminante. Si on n'introduit pas de probabilités *a priori* (en fait probabilités égales pour chaque groupe), il y a sous-affectation dans les groupes les plus nombreux. En revanche, et c'est le cas ici, quand on introduit des probabilité *à priori* proportionnelles aux effectifs de classe, ce qui est logique puisqu'un individu a plus de chance d'appartenir à un groupe d'effectif

fort, il y a une tendance à la sur-affectation dans les groupes les plus nombreux (et une sous-affectation dans les groupes de faible effectif). On retrouve aussi ce défaut pour la Hot-Deck, mais dans une moindre mesure. Là où l'analyse discriminante semble, empiriquement, plus performante c'est quand on compare la valeur affectée par la procédure de correction et la vraie valeur. Les résultats sont en effet les suivants :

	Discriminante	Hot-Deck
Bien classés	41 (30,8%)	26 (19,5%)
Mal classés	45 (33,8%)	76 (57,1%)
Classés dans les modalités voisines	47 (35,3%)	31 (23,3%)

On constate alors une amélioration spectaculaire par rapport au Hot-Deck, et il pourra donc être utile de poursuivre et préciser l'étude de cette façon de traiter les non-réponses. Dans le but de pallier la sur-affectation dans certains groupes, on pourra travailler sur les probabilités *a priori* qu'il faut introduire sur chaque modalité pour améliorer les résultats et éliminer, au moins en partie, ce défaut. Ceci sera traité rapidement dans une dernière étude. Ainsi pourrait-on introduire cette méthode dans le traitement et le dépouillement de certaines enquêtes de l'INSEE, puisqu'on a pu voir que le traitement se systématisait assez facilement. Mais il y aurait lieu, par des estimations de variance, de déterminer quelle est la mesure de précision obtenue.

Les questions d'opinion

Pour traiter les questions d'opinion, et déterminer au mieux une affectation adéquate, on se limitera à quelques unes d'entre elles: celles qui connaissent les plus forts taux de non-réponses. En effet pour cerner l'amélioration apportée par l'analyse discriminante, il était bon d'avoir un sous-échantillon de valeurs absentes d'effectif suffisant, d'une part, et pour que cette méthode permette vraiment un progrès, il est utile de la faire porter sur au moins une centaine d'individus. Pour des questions ayant 70 non-réponses, au plus, une amélioration de 10% à 20% des affectations, apportée par l'analyse discriminante, par rapport au Hot-Deck ne portera que sur une dizaine d'individus, ce qui peut sembler trop faible pour justifier toute une démarche un peu lourde. D'autre part nous avons éliminé pour cette étude les questions qui avaient une répartition trop inégalitaire entre les différentes modalités. Nous avons vu précédemment que cela posait un problème de sur ou sous-affectation selon qu'on utilise des probabilités a priori proportionnelles ou uniformes pour les diverses modalités. Donc ces questions ont aussi été sorties de la présente étude, qui n'était qu'une esquisse rapide sur ce sujet, et ne permettait donc pas de l'approfondir dans toutes les directions possibles.

Pour toutes ces raisons l'étude a été limitée à trois questions (Q2, Q5, Q6). Ces questions ont les caractéristiques suivantes :
Q2 : Pensez-vous que, d'ici un an, le niveau de vie des Français s'améliorera nettement ou un peu (modalités: 1 et 2), restera stable (3), se dégradera un peu ou nettement (modalités: 4 et 5).

Donc 5 modalités ayant pour répartition :

1	14	1,0%
2	190	13,1%
3	466	32,1%
4	468	32,2%
5	163	11,2%

et 151 réponses "ne sait pas" soit 10,4%.

Q5 : Par rapport à ce qui se passe actuellement, pensez-vous que dans les mois qui viennent la hausse des prix sera plus rapide (modalité 1), aussi rapide (2), moins rapide (3), les prix resteront stables (4) ou diminueront légèrement (5)?

Donc 5 modalités ayant pour répartition:

1	213	14,7%
2	455	31,3%
3	204	14,0%
4	412	28,4%
5	39	2,7%

et 129 réponses "ne sait pas" soit 8,9%.

- Q6 : Pensez-vous que les gens aient intérêt à faire actuellement des achats important (Oui :1, Neutre :2, Non :3)?

Donc 3 modalités dont la répartition est:

1	351	24,2
2	536	36,9
3	446	30,7

Sur ces trois questions ont été appliquées les méthodes précédemment décrites, que ce soit dans le choix des variables (indicatrices ou factorielles) à introduire dans l'analyse, ou dans leur utilisation pour l'affectation.
Pour ce qui est du choix, les variables sélectionnées ont été les suivantes :

Question 2 :

Variables indicatrices	Axes factoriels
REV8	1
REV9	2
ZEA1	3
RETR	20
AGE1	21
AGE3	
MALA (En congé de maladie)	

Question 5 :

Variables indicatrices	Axes factoriels
CMOY	5
SECO (Études secondaires)	21
ZEA7	29
REV5	38
HOM_S (Homme vivant seul)	44
NP1	46
NP5	

Question 6 :

Variables indicatrices	Axes factoriels
AGE4	1
AGE5	2
ZEA1	4
ZEA7	5
REV8	10
REV9	26
ONQ (Ouvrier non qualifié)	31
NP5	35
JCHO (Jeune chômeur)	42

Cette énumération rapide nous permet de constater que les axes factoriels sélectionnés contiennent le plus souvent les premiers (1, 2 et 3 pour la question 2; 1, 2, 4 et 5 pour la question 6). Seule la question 5 fait exception, mais on pourra constater par la suite que cette question est la plus difficile des trois à traiter et à corriger. On peut voir là une explication à cette exception: toutes les variables choisies ne permettent que peu de prévoir la modalité choisie par un individu.

En ce qui concerne les variables indicatrices, si certaines variables classiques de description socio-démographique ou économique apparaissent (revenu, catégorie socio-professionnelle, âge, nombre de personnes du ménage ou région d'habitation), ce ne sont pas les mêmes modalités qui sont sélectionnées, à l'exception des hauts revenus ou des ménages nombreux.

Trois affectations "primaires" ont donc été ainsi déterminées en utilisant les méthodes linéaire sur variables indicatrices (1), linéaire sur variables factorielles (2), quadratique sur ces mêmes variables (3). Trois affectations "secondaires" en ont été déduites, selon le processus déjà exposé, à savoir affecter une valeur si elle revenait au moins deux fois, et comparer les trois méthodes pour les observations non traitées ainsi (ayant 3 affectations différentes), d'où les méthodes 4, 5 et 6 définies dans le même ordre. Une fois muni de ces 6 moyens d'affecter une valeur à un individu non renseigné, nous avons comparé leurs résultats respectifs sur les individus renseignés à l'aide d'une resubstitution croisée (l'échantillon test n'était pas ici envisageable du fait d'un trop faible effectif global de 1452 observations). Cette comparaison a permis de sélectionner l'une des 6 façons d'affecter une valeur à un individu donné. Pour les questions 2, 5 et 6, cette procédure a donné les résultats suivants, où le taux de réussite est le taux d'affectation dans la vraie modalité, et le taux d'approximation le taux des individus affectés dans une modalité immédiatement voisine de la vraie :

Question 2 : (réponses renseignées 1301)

Méthode	Taux		
	Réussite	Approximation	Erreur
1	517 (39,7%)	602 (46,3%)	182 (14,0%)
2	517 (39,7%)	613 (47,1%)	171 (13,1%)
3	497 (38,2%)	598 (46,0%	206 (5,8%)
4	518 (39,8%)	609 (46,8%)	174 (13,4%)
5	517 (39,7%)	611 (47,0%)	173 (13,3%)
6	513 (39,4%)	604 (46,4%)	184 (14,1%)

Question 5 : (réponses renseignées 1323)

Méthode	Taux		
	Réussite	Approximation	Erreur
1	463 (35,0%)	374 (28,3%)	486 (36,7%)
2	470 (35,5%)	366 (27,7%)	487 (36,8%)
3	457 (34,5%)	363 (27,4%)	503 (38,0%)
4	472 (35,7%)	365 (27,6%)	486 (36,7%)
5	479 (36,2%)	364 (27,5%)	480 (36,3%)
6	468 (35,4%)	368 (27,8%)	487 (36,8%)

Question 6 : (réponses renseignées 1333)

Méthode	Taux		
	Réussite	Approximation	Erreur
1	612 (45,9%)	608 (45,6%)	113 (8,5%)
2	580 (43,5%)	618 (46,4%)	135 (10,1%)
3	593 (44,5%)	583 (43,7%)	157 (11,8%)
4	589 (44,2%)	611 (45,8%)	133 (10,0%)
5	586 (44,0%)	614 (46,1%)	133 (10,0%)
6	589 (44,2%)	611 (45,8%)	133 (10,0%)

En examinant ces résultats, on constatera que les différentes méthodes donnent souvent des résultats très proches, et que, dans le dernier exemple, la méthode la plus simple, à savoir analyse discriminante sur variables indicatrices, peut être la plus efficace (ici de l'ordre de 2%). On peut penser que point n'est toujours besoin de méthodes sophistiquées, et que des méthodes simples peuvent s'avérer efficaces. Une autre remarque est aussi nécessaire à ce niveau, la méthode d'analyse discriminante calculant une fonction d'affectation quadratique (méthode 3) se révèle le plus souvent moins bonne que les autres de 1 à 2%, cependant elle permet quelquefois de corriger un peu les méthodes linéaires. On retrouve dans cette remarque un fait déjà vu, la méthode quadratique, même si elle fait moins d'hypothèses sur les données, demande l'estimation de plus de paramètres, ce qui explique qu'elle peut se révéler décevante.

Maintenant que nous disposons de ces affectations, nous pouvons dans chaque cas choisir la plus efficace, et comparer ses résultats à ceux que permettrait d'obtenir une correction par la méthode du Hot-Deck. Comme on peut le constater, ces méthodes sont diverses:

– méthode linéaire sur variables factorielles dans le premier cas ;

- méthode "panachée" avec comparaison des trois méthodes "primaires", affectation d'une valeur si elle revient deux fois, et traitement des individus ayant trois valeurs différentes par la méthode linéaire sur variables factorielles (méthode 5) ;

- méthode linéaire sur variables indicatrices dans le troisième cas.

Une fois ce choix effectué, on va comparer ses résultats avec ceux que donnerait une affectation par le Hot-Deck en utilisant la même procédure que pour la correction du "revenu". On construit un sous-échantillon de non-répondants fictifs parmi les individus renseignés sur la question traitée de l'ordre de 10%, soit 130 observations environ. Cet ensemble d'individus sur lesquels s'effectue tout le traitement, mais dont on connaît la vraie valeur nous permettra d'en tirer quelques conclusions instructives. L'affectation par le Hot-Deck se fait en utilisant la méthode déjà décrite, mais en choisissant comme variables de détermination de voisinage les deux ou trois variables qui semblent être le plus influentes sur la question étudiée. Le critère d'influence choisi est directement issu de la procédure de sélection pas à pas, les variables ayant deux modalités sélectionnées seront choisies pour pratiquer le Hot-Deck.

La comparaison des résultats de ces deux méthodes (Hot-Deck et analyse discriminante) est alors la suivante (on entendra par voisine une affectation d'un individu dans une des modalités immédiatement voisines de la valeur initiale) :

Question 2 : Répartitions par modalités, pour les 1301 individus renseignés pour la question :

Groupe	Vraie valeur	Analyse discriminante	Hot-deck
1	14 (1,1%)	166 (12,8%)	184 (14,1%)
2	190 (14,6%)	166 (12,8%)	184 (14,1%)
3	466 (35,8%)	478 (36,7%)	459 (35,3%)
4	468 (36,0%)	490 (37,7%)	477 (36,7%)
5	163 (12,5%)	154 (11,8%)	168 (12,9%)

Comparaison entre affectations et vraies valeurs pour les 133 individus de l'échantillon test :

Affectation	Analyse discriminante	Hot-deck
Bien classé	53 (39,8%)	47 (35,3%)
Voisin	64 (48,1%)	62 (46,6%)
Mal classé	16 (12,0%)	24 (18,0%)

Question 5 : Répartition par modalités pour les 1323 individus de l'échantillon test :

Groupe	Vraie valeur	Analyse discriminante	Hot-deck
1	213 (16,1%)	202 (15,3%)	227 (17,2%)
2	455 (34,4%)	511 (38,6%)	462 (34,9%)
3	204 (15,4%)	181 (13,7%)	198 (15,0%)
4	412 (31,1%)	398 (30,1%)	401 (30,3%)
5	39 (2,9%)	31 (2,3%)	35 (2,6%)

Comparaison entre affectations et vraies valeurs, pour les 132 individus de l'échantillon-test

	Groupe	
Bien classé	38 (28,8%)	363 (28,8%)
Voisin	32 (24,2%)	525 (32,6%)
Mal classé	62 (47,0%)	51 (38,6%)

Si on compare directement les deux méthodes de correction et affectation, on constate 43 affectations communes et 89 différentes, dont 53 très différentes (plus d'une modalité d'écart).

Question 6 : Répartitions par modalités , pour les 1333 individus renseignés sur la question :

Groupe	Vraie valeur	Proportionnelles	Choisies
1	351 (26,3%)	331 (24,8%)	342 (25,7%)
2	536 (40,2%)	553 (41,5%)	538 (40,4%)
3	446 (33,5%)	449 (33,7.0%)	445 (33,4%)

Comparaison entre affectation et vraie valeur pour les 130 individus du sous-échantillon test :

Affectation	Analyse discriminante	Hot-Deck
Bien classé	62 (47,7%)	43 (33,1%)
Voisin	57 (43,8%)	67 (51,5%)
Mal classé	11 (8,5%)	20 (15,4%)

Une comparaison des deux variables calculées permet de trouver 46 affectations identiques et 84 affectations distinctes, dont 18 très différentes.

Quelques conclusions générales sont utiles au vu de ces résultats. On retrouve, une fois de plus, le défaut de l'analyse discriminante déjà soulevé, celui de la sur-affectation dans les modalités de fort effectif. Le Hot-Deck n'a pas, ou peu, ce défaut. En revanche, et c'est là un point très intéressant, quand on compare la valeur affectée avec l'analyse

discriminante à celle affectée à l'aide du Hot-Deck, on constate souvent une nette amélioration. Si l'analyse discriminante n'est pas une "panacée", puisqu'elle échoue sur la question 5, elle permet un taux d'affectation exacte de près de 50% pour la question 6 (plus 15% par rapport au Hot-Deck), c'est là un très bon résultat qui justifie à lui seul la poursuite et le prolongement de ce travail: c'est peut-être une voie ouverte sur de nouvelles perspectives. Quand à la question 5, on a vu par ailleurs qu'elle était difficile à traiter en utilisant l'analyse discriminante, pour ce cas précis ce n'est probablement pas une méthode adaptée, d'où le résultat bien modeste obtenu.

Les probabilités *a priori*

Une dernière étude comparative a été menée, pour essayer de pallier le problème, maintes fois évoqué, de sur ou sous affectation dans certains groupes. Pour éliminer cette caractéristique, on peut penser à une affectation probabiliste (utilisant les probabilités *a posteriori* calculées). Une seconde méthode est de modifier les probabilités a priori affectées à chacune des modalités de la variable étudiée. Celle-ci a été mise en oeuvre sur la question 6. Des probabilités *a priori* uniformes de 0.333 pour chaque groupe ne permettent pas de retrouver la répartition souhaitée, la répartition trouvée se révélant trop uniforme. Cependant, des probabilités *a priori* proportionnelles aux effectifs de chaque groupe - 0.263, 0.402, 0.335 - amène à une répartition non satisfaisante, par augmentation exagérée des catégories aux effectifs déjà forts. Une solution intermédiaire semble donc utile à tester. Aussi a-t-on choisi, de manière empirique, les probabilités *a priori* suivantes (intermédiaires entre les deux hypothèses précédentes): 0.31, 0.36 et 0.33 pour les trois groupes. Cela amène donc a comparer les nouvelles affectations ainsi déterminées et celles qu'on obtient avec des probabilités proportionnelles.

Comparaison en répartitions par modalités pour les 1333 individus renseignés de l'échantillon :

Groupe	Probabilités		
	Vraie valeur	Analyse discriminante	Hot-deck
1	351 (26,3%)	331 (24,8%)	342 (25,7%)
2	536 (40,2%)	553 (41,5%)	538 (40,4%)
3	446 (33,5%)	449 (33,7%)	449 (33,4%)

On constate donc bien à ce niveau une amélioration sensible de la répartition par groupes. Mais qu'en est-il de la comparaison des affectations par les deux méthodes d'une part, et de la vraie valeur d'autre part?

Comparaison à la vraie valeur pour les 130 individus du sous-échantillon test :

Affectation	Probabilités	
	Proportionnelles	Choisies
Bien classé	62 (47,7%)	59 (45,4%)
Voisin	57 (43,8%)	58 (44,6%)
Mal classé	11 (8,5%)	13 (10,0%)

En ce qui concerne la comparaison de la nouvelle fonction d'affectation avec celle donnée par le Hot-Deck, on a des résultats un peu plus convergents que précédemment, à savoir 55 valeurs égales (au lieu de 46) et 75 valeurs différentes, dont 21 très différentes.Il n'y a pas cette fois amélioration du résultat, mais on retrouve des taux de bien classés ou de mal classés assez proches. Ceci permettra de penser que cette méthode, purement empirique, peut être intéressante à creuser, puisqu'elle améliore sensiblement la répartition par modalité sans porter préjudice à la qualité de l'affectation. De plus quel fondement théorique lui apporter? Il y a, là, matière à réflexion et à recherche. Ceci est le dernier prolongement possible de cette étude qui en est restée à un stade seulement exploratoire.

Conclusion et développements

Dans la présentation de ce travail, nous avons effectué une esquisse d'utilisation possible de l'analyse discriminante, à des fins de traitement des non-réponses partielles aux enquêtes. Ces réponses manquantes peuvent être des refus (exemple du "revenu"), des "ignorances" (type de réponses : "ne sait pas"), ou des problèmes de saisie ou de codage et traitement statistique (valeurs aberrantes). Ceci nous permet de décrire une méthode d'analyse discriminante sur variables qualitatives, soit par le biais de variables indicatrices (choix contestable théoriquement mais utilisable en pratique), soit par celui d'une Analyse des Correspondances Multiples et des variables factorielles ainsi déterminées. On constate que cette manière de traiter les non-réponses mérite amplement d'être étudiée plus longuement et plus en détail par la suite. On pourra ainsi détailler les aspects théoriques de ce type d'affectation, mais aussi l'utilisation pratique de ces méthodes (implémentation informatique, estimation de la variance de l'estimateur mis au point, intervalle de confiance qui s'en déduit).
La comparaison avec une des méthodes les plus employées pour le traitement des non-réponses -le Hot-Deck- permet de voir que l'analyse discriminante permet, dans certains cas, une amélioration spectaculaire des résultats d'affectation. Mais il faut rester prudent dans cette affirmation. Il convient de vérifier que les variables étudiées ne sont pas de trop bons exemples, exception parmi un ensemble de cas moins bien traités par cette technique. On a, de plus, constaté que certaines variables semblent se prêter assez mal à un traitement par l'analyse discriminante. Malgré ces réserves, cette technique semble prometteuse et pourrait être testée dans bon nombre d'exemples différents. On pourrait, ainsi, faire des essais à plus grande échelle, sur de nombreux d'échantillons, pour tester les qualités des procédures ainsi mises en oeuvre sur de

nombreux exemples. Si les résultats sont encourageants, on pourra en systématiser l'emploi. A partir de la description de la méthodologie employée, on constate qu'il est assez simple de déterminer un algorithme global de traitement. La méthode DISQUAL est facilement implémentable informatiquement, car le choix des premiers axes factoriels s'impose comme souhaitable.

Toutefois, un autre problème demande à être examiné. Dans ce travail, nous n'avons insisté que sur les résultats d'affectation. D'un point de vue statistique mathématique, nous n'avons vu que l'estimation ponctuelle. Il semble utile d'évoquer l'estimation par intervalle de confiance. Pour cela il nous faut estimer la variance des résultats issus de l'analyse discriminante. Il faut connaître la qualité des résultats offerts, non seulement dans leur exactitude, mais aussi dans leur précision, calcul de variance sur les probabilités d'affectation estimées, et variabilité des affectations ainsi déduites. Nous abordons alors un point crucial de la comparaison entre analyse discriminante et Hot-Deck. Pour déterminer un estimateur de la variance supplémentaire, nous devons estimer la variance supplémentaire due aux non réponses et à leur traitement. Le Hot-Deck est une méthode aléatoire d'affectation, une fois l'échantillon connu, qui permet, donc, par des méthodes de type Bootstrap ou double échantillon, une estimation de cette variance supplémentaire. L'analyse discriminante, telle qu'elle a été appliquée dans cette étude, est une méthode déterministe, conditionnellement à l'échantillon, et ne permet pas de déterminer la variance due à la non réponse. Celle-ci est complètement gommée, ce qui ne permet pas d'estimation fiable par intervalle de confiance. La variance totale de l'estimateur utilisé est, en effet, largement sous-estimée et on ne contrôle plus le risque de première espèce.

Pour pallier ce défaut, on pourra affecter, non plus la valeur la plus probable, mais une valeur aléatoire tenant compte des probabilités a posteriori calculées par l'analyse discriminante. La méthode devenant, alors, aléatoire permettra l'estimation de la variance supplémentaire. Examinons rapidement le principe ce cette façon de procéder. On pourrait, pour chaque individu non renseigné dans l'échantillon, affecter non plus une valeur donnée (la plus probable) calculée par analyse discriminante, mais associer à chacune des p modalités de la variable étudiée la probabilité qui lui est associée. On transformerait ainsi un individu unique en p individus, appartenant chacun à une modalité différente, chaque individu étant pondéré par la probabilité correspondante. Pour retrouver des individus "entiers" on procéderait par sommation (et arrondi) par groupe défini à l'aide de variables bien choisies (variables descriptives classiques pour l'essentiel). Si on ne veut conserver qu'un individu, on peut aussi tirer aléatoirement un des p individus déterminés, avec les probabilités *a posteriori* associées (calculées par la méthode d'analyse discriminante). Ces probabilités permettraient une estimation directe (analytique) de la variance de l'estimateur construit, ou le recours à des méthodes empiriques d'estimation de la variance (de type Bootstrap). Mais il est clair que cette méthode sera, alors, plus pointue à mettre en oeuvre, et que tout un travail, théorique et empirique, est à poursuivre. Il faudra déterminer les meilleures probabilités a priori. Celles-ci joueront, avec cette utilisation de l'analyse discriminante, un rôle très important dans la valeur des probabilités *a posteriori* calculées par la procédure précédente, et donc sur l'estimation de variance qu'on peut en déduire.