Une étude exploratoire de la participation aux enquêtes sur access panels en ligne : quels sont les profils des non-répondants ? Quelle est l'influence de la relance et du délai de réponse ?

Philippe JOURDAN
Maître de conférences
IUT Evry
26 rue Richer
75009 Paris
Mél. : philippe.jourdan5@wanadoo.fr

Valérie JOURDAN
Président et directeur général
Panel On The Web SA
36 rue Vivienne
75002 Paris
Mél. : valerie.jourdan@panelontheweb.com

1)- INTRODUCTION

L'évolution des études ad hoc a été marquée au cours des dix dernières années par l'essor des enquêtes menées sur access panels. Initialement cantonnées aux études longitudinales nécessitant des échantillons de tailles importantes interrogés à des fréquences régulières sur leurs habitudes de consommation ou d'achat, les études sur access panels se sont aujourd'hui étendues aux baromètres d'image, aux suivis des retombées de campagnes publicitaires, aux études des valeurs, des opinions, des attitudes et des comportements de populations de référence. En dehors des études longitudinales ou d'usages et attitudes, les access panels sont également utilisés dans le cadre de certaines mesures d'audience ou pour conduire des études classiques d'optimisation du marketing mix : screening de concept, test de concept ou de concept-produit. Les access panels sont également présents dans les secteurs grand public (B to C) et professionnels (B to B).

Par définition, un access panel est un échantillon permanent d'individus recrutés sur l'initiative d'une société d'études et volontaires pour participer à des études administrées le plus souvent sous la forme de questionnaires téléphoniques, en face-à-face, en ligne, etc. Une fraction de cet échantillon peut être rapidement interrogée en fonction d'une problématique de l'annonceur : le recours à un access panel permet donc de disposer d'une population d'étude préqualifiée et prérecrutée, aisément mobilisable dans un délai court ; le caractère volontaire de l'inscription et la rémunération associée à la participation aux enquêtes garantit en outre un taux de questionnaires complets habituellement plus élevé que celui relevé sur des échantillons recrutés pour des besoins ponctuels.

Le développement d'Internet a provoqué l'apparition de nombreux panels d'internautes : en simplifiant le recrutement et la collecte des données d'enquêtes (désormais autoadministrées en ligne), les access panels en ligne présentent d'indéniables atouts, entre autres une optimisation des budgets grâce à la réduction des coûts de la collecte de données, un accroissement de la taille des échantillons d'enquêtes (certains access panels regroupant plusieurs centaines de milliers d'internautes), une qualification élevée des échantillons permettant l'interrogation de cibles rares et une meilleure dispersion géographique des répondants.

Ces avantages s'accompagnent toutefois de nouveaux risques de distorsion des données d'enquêtes. En premier lieu, les access panels en ligne ont été critiqués sur leur faible représentativité d'une population de référence et ce en raison du faible taux de pénétration du

média dans les foyers et de l'inégale répartition de l'accès à Internet selon les régions d'habitation, les tailles d'agglomération, le genre, l'âge ou la catégorie socioprofessionnelle en raison de la nouveauté du média. Ces critiques, pour fondées qu'elles fussent dans un passé récent, devraient logiquement se tempérer en raison de l'accélération de l'équipement Internet des foyers français : au 4ème trimestre 2003, 6,9 millions de foyers français soit 27,7% des foyers et 65% des foyers équipés d'un micro-ordinateur ont accès à Internet depuis leur domicile (*Médiamétrie, 2003*). Ceci étant, un autre facteur de distorsion est souvent négligé : le biais des non-réponses qui, en raison de la démarche volontaire de préinscription caractéristique des access panels, présente des caractéristiques spécifiques.

L'objectif de la recherche est d'explorer quelles sont les variables corrélées à la non-réponse totale et quelle est l'efficacité des méthodes employées en vue d'augmenter le taux de réponse dans le cas d'un access panel grand public en ligne. Nous présentons tout d'abord les déterminants et les conséquences de la non-réponse dans les études par questionnaire avant de détailler la méthodologie de la recherche et d'exposer les résultats exploratoires obtenus.

2)- LA NON-REPONSE DANS LES ETUDES PAR QUESTIONNAIRE

A)- LES DIVERSES FORMES DE LA NON-REPONSE

Comme le souligne fort justement Lebart (*Lebart, 2001*), pour le statisticien la non-réponse doit être considérée comme une méta information dans la mesure où il convient de distinguer la non-réponse totale liée à l'impossibilité de contacter le répondant, de celle, partielle ou totale, qu'explique le refus de répondre, l'incompréhension ou la non-connaissance de la réponse (« ne sais pas ») ou bien encore le fait que le répondant n'est pas concerné par la question (« n'a pas lieu d'être posée »).

Dans le cas général, il convient donc de distinguer plusieurs origines à la non-réponse comme le souligne la liste non exhaustive suivante :

- L'impossibilité technique (mauvais numéro de téléphone, adresse postale incorrecte ou mauvaise adresse électronique) ou physique (absence du domicile, refus de décrocher, boîte aux lettres électronique peu consultée, etc.) de joindre la personne contactée.
- Le refus de répondre qui se traduit par une non-réponse totale ou partielle (abandon en cours de questionnaire, questions non renseignées dans un questionnaire auto-administré, questions jugées trop intrusives ou dérangeantes, etc.).
- L'absence de réponse qu'explique une mauvaise compréhension de la question ou une méconnaissance de la réponse par la personne interrogée. Cette non-réponse peut être reportée dans une modalité de réponse spécifique (« ne sais pas »).
- L'oubli de la réponse involontaire liée à une maladresse de l'enquêteur (question non posée, réponse non reportée, observation incomplète, etc.) ou du répondant lui-même dans le cas d'un questionnaire auto-administré. Le respect des consignes, la qualité de la formation préalable au lancement du terrain ou bien encore la clarté et la lisibilité du support d'interrogation limitent (sans pour autant l'éliminer complètement) les conséquences de ce type de non-réponse.
- L'effacement de la réponse introduite lors du dépouillement des données : il s'agit le plus souvent de suppression involontaire ou accidentelle résultant d'une erreur de transcription, de saisie informatique ou de codification.

- La non-réponse voulue parce que la question n'a pas lieu d'être posée à un sous-ensemble de répondants (question filtrée s'adressant aux hommes seulement par exemple).

Seules les cinq premières sources de non-réponse sont susceptibles d'introduire des biais dans les résultats de l'enquête, en particulier dans le cas d'échantillons formés de manière aléatoire. En effet, la sélection aléatoire d'un échantillon au sein d'une population finie repose sur le principe de l'affectation à chaque individu d'une probabilité connue et non nulle d'être interrogé (*Kalton, 1993*). Or, il est peu probable que l'impossibilité de joindre la personne, le refus volontaire de répondre à l'enquête, l'incompréhension de certaines questions, l'oubli ou l'effacement d'une réponse soient aléatoirement distribués au sein de l'échantillon ; dès lors l'équiprobabilité des observations au sein du fichier des résultats n'est pas garantie. La non-réponse a donc un impact sur la fiabilité des estimateurs.

Dans le cas particulier des enquêtes menées à partir d'access panels d'internautes, deux caractéristiques du mode de collecte de données influent sur la non-réponse attendue.

- En premier lieu, comme dans tout access panel, les répondants sont prérecrutés et par conséquent volontaires pour participer à des études : le taux de participation aux enquêtes subséquentes est donc habituellement élevé (entre 60% et 80%). Ce taux de participation ne doit toutefois pas être confondu avec un taux de réponse. En effet, pour comparer le taux de réponse d'une étude sur access panel avec celui d'une enquête téléphonique, il convient de tenir compte du taux de refus lors du recrutement du répondant. Cette information est peu communiquée par la plupart des instituts. En pratique, cela revient à comptabiliser sur 100 personnes sélectionnées aléatoirement pour participer à un panel, le nombre de ceux qui acceptent (ou refusent) le principe d'adhésion. Le taux de non-réponse (T_{nr}) à une enquête sur access panel se calcule donc à partir du taux d'adhésion (T_{adh}) et du taux de participation (T_{part}) selon la formule suivante :

(1) $$T_{nr} = [1 - (T_{adh} \times T_{part})]$$

Ainsi et sous réserve que le recrutement de l'access panel soit effectué à partir d'une sélection aléatoire des individus depuis une population de référence finie et si l'on admet qu'un tiers des personnes pressenties acceptent d'adhérer au panel et qu'elles seront 7 sur 10 à participer à la première enquête, le taux de réponse n'est en réalité que de : 33% x 70% ≅ 23% !

- En second lieu, certaines raisons techniques d'abandon sont évidemment propres au média. Dans la plupart des access panels en ligne, l'internaute est prévenu d'une enquête en cours par un courrier électronique envoyé à sa dernière adresse électronique renseigné. Or les internautes sont amenés à changer fréquemment d'adresse électronique pour des raisons diverses : changement de fournisseur d'accès, volonté de disposer de plusieurs adresses pour différents usages (une adresse professionnelle, personnelle, réservée aux transactions sur le Net, réservée aux e-mails commerciaux, etc.), changement de statut impliquant un changement d'adresse électronique (étudiants, membres d'associations, professionnels, etc.), volonté manifeste d'échapper au harcèlement publicitaire (« spam »), etc. Bien que peu de statistiques soient publiées sur le sujet, le renouvellement des adresses

électroniques serait d'environ 10% par an en France. Il convient donc de distinguer au sein du taux de non-participation deux causes majeures : l'une technique, l'impossibilité quelle qu'en soit la raison de joindre le répondant (adresse désactivée, invalide, mal saisie ou peu voire jamais consultée, etc.), l'autre traduisant le refus de répondre. Dès lors, la formule (1) peut être détaillé comme suit :

(2) $$T_{nr} = [1 - (T_{adh} \times (1- T_{inv}) \times (1-T_{rr}))]$$

avec : T_{nr} = taux de non réponse
T_{adh} = taux d'adhésion au panel
T_{inv} = taux d'adresses invalides
T_{rr}= taux de refus de répondre

B)- LES DETERMINANTS DE LA NON-REPONSE

Quelles sont les principales causes de non-réponse ? Il existe de nombreux ouvrages sur le sujet auxquels nous renvoyons le lecteur intéressé (*Madow et al., 1983 ; Schafer, 1997 ; Little et Rubin, 2002*). De l'ensemble des causes de la non-réponse que nous avons identifiées, celle qui préoccupe le plus le chercheur est naturellement le refus de répondre que Cochran (1977) qualifie de « refus de coopérer » pour quelque motif que ce soit : manque de temps, caractère privé ou intime de l'interrogation, crainte de dévoiler son opinion, etc. La proportion de personnes refusant de répondre à un questionnaire est d'autant plus critique qu'elle est souvent corrélée à l'objet de l'étude (*Wiseman et McDonald, 1980*), à la longueur du questionnaire, à sa complexité, à sa structuration, au mode de collecte de données, à la rémunération accordée (*Brennan, Hoek et Astridge, 1992*) et la force de conviction de l'enquêteur (*O'Muircheartaigh et Campanelli, 1999)*, ces causes étant potentiellement de nature à altérer la validité interne de la collecte de données. Enfin, de nombreuses recherches prouvent depuis longtemps que le refus de répondre n'est pas également distribué au sein de la population : le genre, l'âge, le niveau d'éducation, le revenu sont autant de facteurs corrélés au taux de non-réponse (*Chen, 1996, Marjorie, 1960, Ferber, 1948, etc.*). Ces recherches concernent toutefois les enquêtes par voie postale ou par téléphone, peu d'articles récents traitant des déterminants de la non-réponse dans les enquêtes en ligne (*cf. Bosnjak et Tuten, 2001*).

Dans bien des cas, les instituts d'études n'appliquent aucune procédure particulière pour le traitement des non-réponses : soit les refus de répondre sont considérés comme marginaux, soit l'hypothèse implicite est que la distribution des réponses est identique auprès des répondants et des non-répondants (que la non-réponse soit partielle ou totale). En pratique, il est peu probable que cette assertion soit vérifiée. Litlle et Rubin (1987) ont proposé une classification en 3 familles des non-réponses : dans le premier cas, la non-réponse est distribuée de manière purement aléatoire, à savoir que sa distribution ne dépend d'aucune autre variable étudiée ; dans le second cas, la non-réponse est dépendante des valeurs observées ; enfin dans le troisième cas, la non-réponse est dépendante des valeurs manquantes ou non observées. Ce troisième cas est en pratique le plus difficile à traiter d'un point de vue statistique puisque précisément l'estimation des non-réponses ne peut être approchée qu'à l'aide de variables précisément manquantes (ex. : les classes supérieures sont le plus souvent réticentes à participer à des études sur le revenu, l'investissement ou l'épargne).

C)- LES CONSEQUENCES DE LA NON-REPONSE

Quelles sont les conséquences de la non-réponse sur la fiabilité des résultats d'études ? Supposons que nous cherchions à estimer la valeur du paramètre μ dans la population : la moyenne $\overline{\mu}$ dans la population est la somme des deux moyennes $\overline{\mu}_r$ et $\overline{\mu}_{nr}$ désignant les valeurs obtenues auprès des répondants et estimées auprès des non-répondants que pondèrent le poids relatif de chaque groupe ω_r et ω_{nr} (cf. formule 3). Une transformation simple permet d'évaluer l'erreur d'estimation liée à la prise en compte des seules réponses à l'enquête, soit ($\overline{\mu}_r$ - $\overline{\mu}$) en fonction des autres paramètres (cf. formule 4).

(3) $$\overline{\mu} = \omega_r . \overline{\mu}_r + \omega_{nr} . \overline{\mu}_{nr}$$

(4) $$\overline{\mu}_r - \overline{\mu} = \omega_{nr} . \left(\mu_r - \overline{\mu}_{nr} \right)$$

L'erreur d'estimation est donc fonction de la proportion de non-répondants dans l'enquête et de l'écart des réponses entre les répondants et les non-répondants (*Kalton, 1983*). Il est donc illusoire de penser que cette différence est négligeable ou bien encore qu'elle est également distribuée dans toutes les tranches de la population.

Dans tous les cas, la première précaution consiste à minimiser le taux de non-réponse. Une pratique courante consiste dans le cas d'une étude téléphonique à multiplier les rappels à différentes heures dans la journée ou à différents jours dans la semaine ou bien encore à faire se succéder les modes de collecte de données ; cette dernière façon de procéder pose toutefois le délicat problème de la distorsion des réponses liées à la variation des méthodes de collecte (*Frankel et Frankel, 1977*).

Dans le cas de la non-réponse partielle, certaines approches statistiques permettent d'estimer à partir des valeurs observées les valeurs manquantes. Lorsque les non-réponses touchent un nombre restreint de répondants et ne concernent que quelques questions, les logiciels de traitements statistiques proposent des méthodes d'estimation des valeurs manquantes : ces méthodes s'appuient sur la connaissance des valeurs prises par d'autres variables du questionnaire fortement corrélées avec une valeur de remplacement[1] qui peut prendre la forme du barycentre d'une classe d'observations pour toutes les non-réponses appartenant à cette même classe. Parmi les méthodes les plus couramment utilisées en présence de valeurs manquantes dépendantes de valeurs observées (MAR ou « Missing At Random »), nous pouvons citer l'algorithme EM (« Expectancy-Maximisation ») de Dempster, Laird et Rubin (1977) ou bien encore les méthodes de simulation Monte Carlo par chaînes de Markov (*Robert 1996*). Enfin, s'agissant des modèles opérant variable par variable, les plus usités sont la régression linéaire, l'analyse de la variance et de la covariance, la régression logistique, l'analyse discriminante, les arbres de décision et enfin plus récemment les méthodes neuronales qui généralisent les méthodes précédentes. Pour plus de détails, nous renvoyons le

[1] La non-réponse est en réalité davantage dépendante de la vraie valeur inconnue de la variable traitée que des valeurs des variables observées qui lui sont corrélées. Ainsi le fait que le répondant refuse de déclarer son revenu est fonction du montant de ce dernier ; pour des raisons évidentes, seules les variables fortement corrélées et renseignées sont utilisées pour estimer les valeurs manquantes (par exemples, l'âge et la profession de la personne interrogée dans le cas du revenu). Comme le fait remarquer Lebart (2001), il s'agit « *d'une limitation incontestable mais beaucoup moins lourde de conséquences qu'un modèle d'indépendance totale* ».

lecteur intéressé à l'ouvrage de Schafer (1997) et aux articles de Schafer et Graham (2002), Collins et al. (2001), etc.

3)- METHODOLOGIE DE LA RECHERCHE

A)- LA NON-REPONSE DANS UN ACCESS PANEL

S'agissant d'un access panel en ligne, l'invitation à répondre au questionnaire d'étude est adressée aux répondants pressentis sous la forme d'un courrier électronique. Le lien vers le questionnaire est le plus souvent directement inclus dans le corps du courrier. Le répondant peut également être relancé à mi-période de la collecte de données : la relance est donc un moyen économique utilisé pour diminuer le taux de non-réponse. Enfin, le taux de non-réponse est naturellement fonction du délai de réponse accordé aux panélistes. Ce délai est le plus souvent fixé de manière empirique (en fonction de la distribution observée des réponses dans le temps) en tenant compte également d'un rétroplanning dicté par la date impérative de remise des résultats au client.

L'access panel dispose donc de deux leviers aisés à mettre en œuvre pour diminuer le taux de non-réponse : accorder un délai raisonnable aux répondants pour répondre (la distribution des réponses dans le temps est une fonction croissante et asymptotique du délai accordé) et relancer de manière incitative et sélective les non-répondants. S'agissant d'une étude en ligne, l'impact budgétaire de ces deux actions correctrices est moindre que dans le cas d'une étude offline (téléphone ou face-à-face) dont le coût du terrain est proportionnel aux cadences journalières effectuées.

C'est pourquoi, la recherche menée se donne trois objectifs :
- [1]- Identifier les profils des non-répondants ;
- [2]- Explorer l'influence du délai sur la réponse ;
- [3]- Explorer l'effet de la relance sur la réponse.

B)- PRESENTATION DE LA COLLECTE DE DONNEES

Les panélistes étudiés sont les personnes régulièrement inscrites dans le panel de la société Panel On The Web et invitées à répondre à un questionnaire en ligne au cours de l'un des trois mois suivants, juin, juillet ou août 2003. Pour neutraliser l'impact du sujet ou de la durée du questionnaire, nous avons retenu une même étude barométrique menée chaque mois auprès d'un échantillon d'individus représentatifs des Internautes français âgés de 15 ans et plus. Chaque répondant a donc été retenu pour une seule des trois vagues. Le fait de disposer de trois vagues successives nous permet de disposer d'un échantillon agrégé de grande taille (4.467 observations) et de neutraliser le cas échéant l'effet d'un mois atypique sur le taux de participation.

La neutralité du sujet (étude de notoriété pour une catégorie de sites grand public) et les caractéristiques du questionnaire autoadministré (nombre de questions limitées à 10 dont une majorité de questions fermées, durée d'administration de 5 minutes en moyenne, exclusion des questions à caractère privé ou intime) nous apportent la certitude raisonnable que le sujet et le questionnaire n'interfèrent pas sur le phénomène étudié, la participation du panéliste à l'étude. Pour les mêmes raisons, une rémunération forfaitaire d'un montant de 1 € a été attribuée à chaque répondant pour chaque vague.

C)- LE MODELE THEORIQUE DE LA RECHERCHE

Pour répondre aux trois objectifs assignés à l'étude, la première des variables expliquées retenues est la participation (PART) du panéliste à la vague d'enquête pour laquelle il est sollicité. Pour les raisons invoquées au paragraphe 2)-A), nous employons les termes de participation (et de taux de participation) plutôt que ceux de réponse (et de taux de réponse) à l'enquête. Deux autres variables sont également étudiées : le délai de réponse (DREP) mesuré auprès des répondants uniquement et l'effet de la relance (RELA). Les variables expliquées retenues se divisent en trois grandes familles selon qu'elles concernent les trois facettes du répondant en tant qu'individu, Internaute ou panéliste :

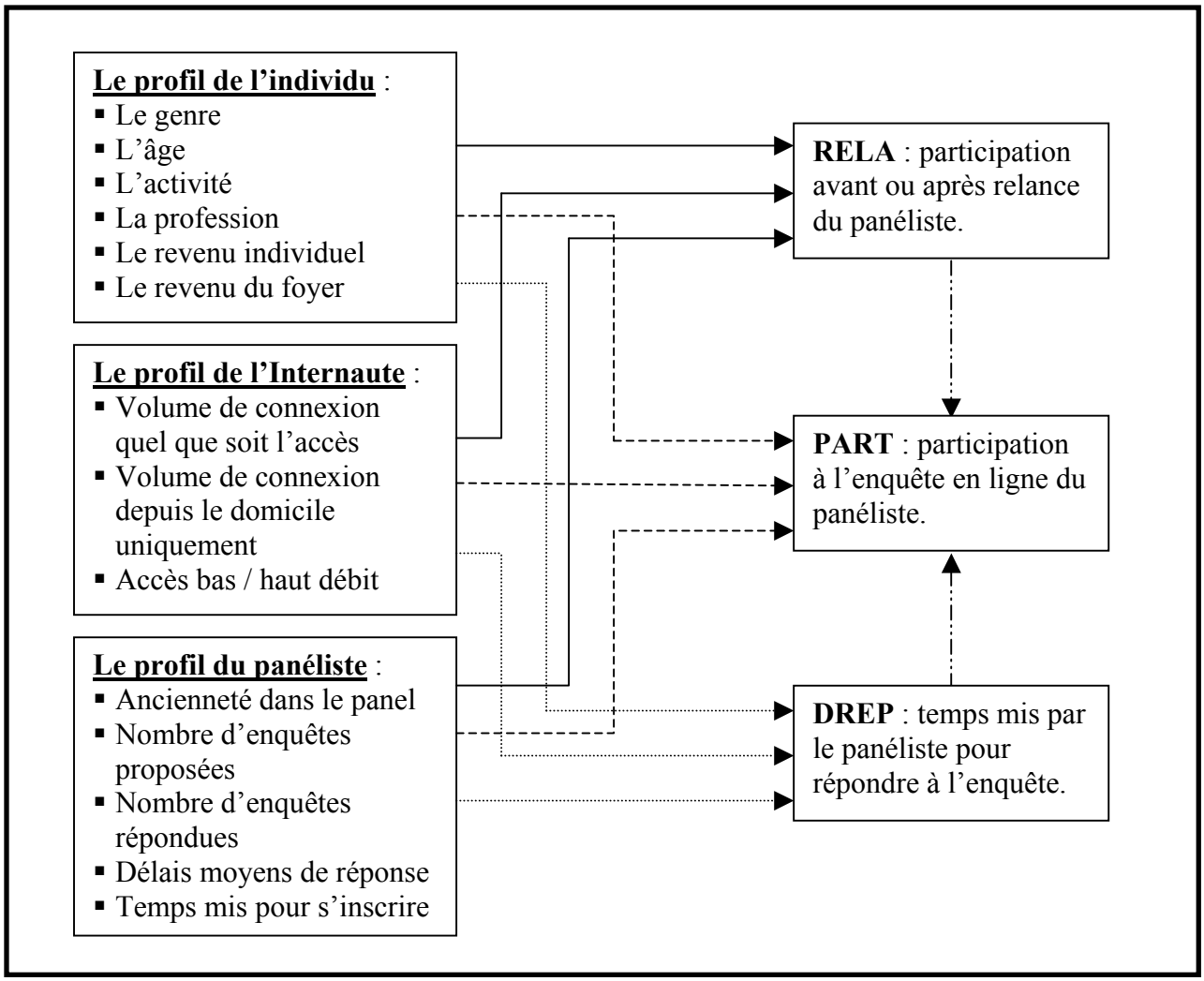
- [1]- Le profil de l'individu : le genre, l'âge, l'activité, la profession, le niveau de formation, le revenu individuel, le revenu du foyer. Ces variables renseignées lors de la phase de recrutement (et mis à jour annuellement) ont été retenues parce qu'elles sont les plus utilisées pour cibler et échantillonner une population à des fins d'études. De plus, elles sont également le plus souvent fortement corrélées aux phénomènes marketing (image de marque, opinion et attitude, achat, consommation, etc.).

- [2]- Le profil de l'Internaute : volume de connexion sur le Net quel que soit l'accès, volume de connexion sur le Net depuis le domicile uniquement. Ces deux variables ont été retenues en raison de la distorsion habituellement prêtée aux access panels en ligne : une étude de Hoppe et Lamp (2001) citée par Jolibert et Jolibert (2003) montrent que l'intensité de l'utilisation d'Internet est un facteur influant sur les résultats d'une étude en ligne. Nous y ajoutons une variable sur le type de connexion possédé à domicile (bas ou haut débit) car la rapidité d'accès au réseau Internet et la généralisation de la connexion illimitée pour un prix forfaitaire dans le haut débit exercent un impact notable sur les usages du Web : ainsi les utilisateurs du haut débit sont de grands consommateurs d'Internet qui dépassent nettement les autres internautes sur l'ensemble des usages ; l'audiovisuel, les jeux, les téléchargements et la publication de pages personnelles attirent plus fortement ces utilisateurs et plus généralement toutes les pratiques qui mobilisent d'importantes ressources (source : *Médiamétrie*, étude sur les usages du haut débit, 2001).

- [3]- Le profil du panéliste : ancienneté dans le panel, nombre d'enquêtes proposées au panéliste, nombre d'enquêtes répondues, le délai moyen de réponse aux enquêtes, temps mis pour compléter son inscription. Dans le dispositif de Panel On The Web, le répondant dispose d'un mois pour remplir le questionnaire de recrutement. Il est donc intéressant d'insérer le critère du temps mis pour s'inscrire afin d'isoler quelles sont les variables comportementales prédictives du comportement du futur panéliste. De même, il convient de s'interroger dans quelle mesure le comportement passé du panéliste (en particulier son assiduité à répondre aux études qui lui sont proposées) est un bon prédicteur de son comportement futur.

Si l'on prend en compte, l'ensemble de ces variables, le modèle théorique de la recherche peut être résumé par le schéma suivant (cf. figure 1). A ce stade de nos travaux, nous ne prétendons pas proposer (et tester) un modèle causal complet. Notre ambition est davantage d'explorer les antécédents des trois variables expliquées retenues : la participation à l'enquête (PART), le fait de participer avant ou après relance (RELA) et enfin le délai moyen de réponse à l'enquête (DREP).

Figure 1 : modèle de la recherche



4)- LES RESULTATS DE LA RECHERCHE

Les résultats de la recherche portent sur les corrélations existantes entre les trois grandes familles de variables explicatives retenues – le profil de l'individu, de l'internaute et du panéliste – et les trois variables expliquées – la participation à l'enquête en ligne (PART), la participation avant ou après relance (RELA) et enfin le temps mis par le panéliste pour répondre à l'enquête (DREP). Nous avons délibérément choisi de nous intéresser à deux autres variables (RELA et DREP) en plus de la participation (PART) afin de mettre en évidence quels sont les effets de la relance sur le taux de participation et sa distribution au sein de l'échantillon ou bien quel délai minimum faut-il laisser aux panélistes pour assurer une distribution hétérogène des répondants au sein de l'échantillon final. Ces deux questions sont d'un grand intérêt pour le praticien en études marketing qui doit fixer le délai optimum pour sa collecte de données et choisir (ou renoncer) à relancer les participants à l'étude.

Il s'agit à ce stade de premiers résultats exploratoires destinés par la suite à faire l'objet d'une modélisation au moyen des équations structurelles afin de valider le schéma théorique de la recherche présenté ci-dessus (cf. figure 1).

A)- LES VARIABLES CORRELEES A LA PARTICIPATION

Les périodes d'interrogation – juin, juillet et août 2003 – expliquent un taux moyen de participation plus faible que celui habituellement constaté sur les études menées par access panels en ligne : 57% des panélistes invités ont participé à l'enquête contre 65% en moyenne (source Panel On The Web). Les taux de participation sont également significativement différents d'un mois sur l'autre : la participation la plus forte est relevée pour le mois de juin (64%) contre 56% pour le mois de juillet et 51% seulement pour le mois d'août (χ^2 : 50,21 ; p = 0,00), une distribution qui s'explique en raison d'une prise de congés (entraînant une absence du domicile) plus forte en août et en juillet qu'en juin. Par la suite, les trois vagues ont été agrégées, bien que l'on ne puisse pas totalement écarter l'hypothèse que le caractère « exceptionnel » de la période puisse exercer une certaine influence sur les résultats rapportés (seule une réplication de la recherche un autre trimestre permettrait d'infirmer cette hypothèse).

Le tableau 1 reprend les variables corrélées (ou non) à la participation à l'enquête (PART). Les variables au sein des trois familles sont classées par ordre décroissant de significativité statistique ; nous rappelons aussi quelles sont les variables testées qui ne s'avèrent pas statistiquement corrélées à la participation. Rappelons que la participation est prise en compte dès lors que le répondant a répondu à l'enquête dans le délai imparti (21 jours) et éventuellement à l'issue de la seule relance. Les outils développés par la société Panel On The Web permettent à chaque panéliste de déclarer une période d'indisponibilité pendant laquelle il ne peut pas (ou ne souhaite pas) être interrogé.

Tableau 1 – Variables corrélées (ou non) à la participation à l'enquête (PART)

Variables	χ^2	p
Le profil de l'individu		
▪ Age	64.69	0.000 (*)
▪ Genre	11.63	0.000 (*)
▪ L'activité	75.19	0.000 (*)
▪ Revenu net mensuel du répondant	57.88	0.000 (*)
▪ Niveau d'étude	11.60	0.114 (n.s)
▪ La profession	13.83	0.129 (n.s)
▪ Revenu net mensuel du foyer	14.30	0.216 (n.s)
Le profil de l'internaute [2]		
▪ Volume de connexion Internet domicile	16.17	0.027 (*)
▪ Volume de connexion Internet tous accès	7.47	0.382 (n.s)
Le profil du panéliste		
▪ Ancienneté dans le panel	124,29	0.000 (*)
▪ Nombre d'enquêtes proposées	97.60	0.000 (*)
▪ Nombre d'enquêtes répondues	563.26	0.000 (*)
▪ Délai moyen de réponses aux enquêtes	30.20	0.000 (*)
▪ Temps mis pour s'inscrire	20.45	0.004 (*)

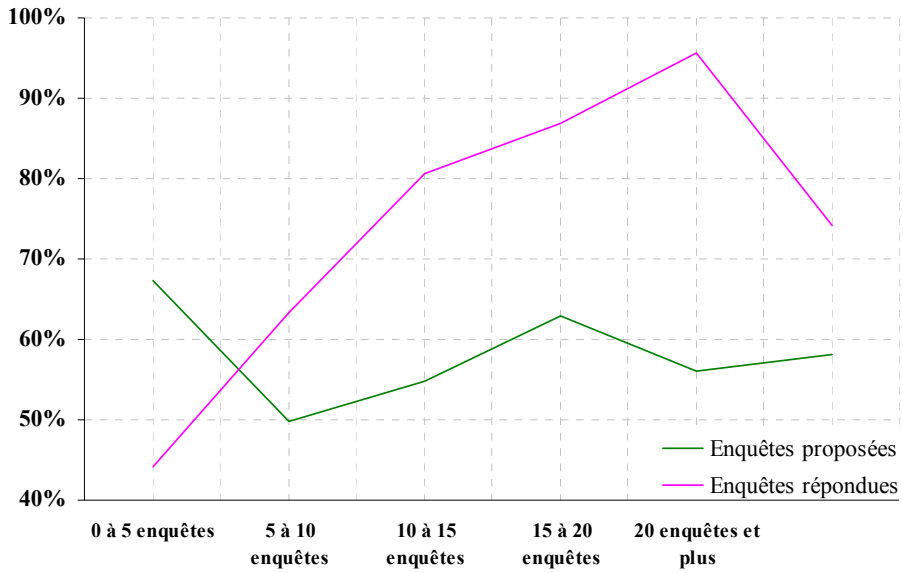
(*) Significatif à un seuil de significativité de 5%
(n.s) Non significatif à un seuil de significativité de 5%

[2] Le type d'accès Internet (haut débit ou bas débit) est une information qualifiée lors de l'enquête et elle n'est donc disponible que pour les seuls répondants à l'étude (cf. paragraphe b et c). Toutefois compte-tenu de sa forte corrélation avec le volume de connexion Internet depuis le domicile, on peut faire l'hypothèse que le type d'accès est également corrélé à la participation, une assertion qu'il appartiendra de valider lors d'une prochaine réplication de l'étude.

Premier enseignement de la recherche : la participation à une enquête en ligne n'est pas également distribuée selon le profil de l'individu. Le genre, l'âge, l'activité et le revenu du répondant influent significativement sur le taux de participation aux trois enquêtes proposées : les hommes ont une propension plus forte à répondre (59%) que les femmes (54%) ; les internautes entre 35 et 54 ans participent davantage (63%) que les 15-24 ans (48%) ; il en découle que les salariés ont également une participation plus forte (62%) que les étudiants (46%) et que ceux qui déclarent ne disposer d'aucun revenu donne moins souvent leur opinion en ligne (52%) que les autres (le taux de participation le plus élevé étant relevé auprès des panélistes déclarant entre 2400 €uros et 3000 €uros de revenus nets mensuels soit 62%). Inversement, la profession, le niveau d'étude et le revenu du foyer n'ont pas d'influence sur le fait de participer ou non à l'étude. Ces résultats militent en faveur d'un taux de réponse élevé (conditionné en partie par l'usage d'une relance et l'octroi d'un délai raisonnable pour répondre), sous peine de distordre les estimateurs lorsque les réponses au questionnaire sont influencées par les variables de profil individuel ci-dessus.

Deuxième enseignement : le profil du panéliste explique également la propension à participer à l'enquête. Les inscrits les plus récents au panel participent davantage, indice qu'il existe un taux d'usure ou de mortalité naturelle des panélistes (71% de participation pour ceux qui sont inscrits depuis moins d'un an contre 51% seulement pour ceux qui sont inscrits depuis plus d'un an mais moins de deux ans). Le taux de participation s'accroît de nouveau auprès des panélistes inscrits depuis plus de deux ans (55%). Le nombre d'enquêtes proposés et le nombre d'enquêtes auxquels le panéliste a répondu exercent également une influence sur la participation ainsi qu'en atteste la figure 1 ci-dessous. Le taux de participation culmine auprès de ceux qui ont répondu à 20 enquêtes et plus (96%) posant le délicat problème de l'effet de maturation (ou d'apprentissage) ; il est également plus élevé auprès des panélistes plus récents (1 à 5 enquêtes proposées) et plus matures (entre 15 et 20 proposées), décroissant au-delà.

Figure 1- Taux de participation en fonction du nombre d'enquêtes



Enfin et logiquement, la participation à l'enquête dépend du profil de l'internaute et plus particulièrement du volume de connexion au média (depuis le domicile uniquement). Le taux de participation le plus élevé est relevé auprès des panélistes qui se connectent en moyenne entre 5 et 10 heures par semaine (61% contre 57% auprès des autres tranches).

B)- LES VARIABLES CORRELEES A L'EFFET DE LA RELANCE

La pratique courante dans les access panels consiste à relancer de façon sélective les panélistes n'ayant pas répondu à l'étude proposée à mi-période environ de la collecte de données. Le faible coût de la relance par courrier électronique explique la généralisation de cette pratique, bien que peu d'études existent sur l'impact de la relance sur le taux de participation et sur une éventuelle distorsion de l'échantillon qu'elle peut entraîner.

Le tableau 2 reprend les variables corrélées (ou non) à l'effet de la relance (RELA) ou plus précisément au fait de répondre avant ou après avoir été relancé. La plateforme développée par Panel On The Web pour la gestion automatisée des enquêtes permet de ne relancer que les seuls panélistes n'ayant pas répondu à l'enquête ou n'ayant pas terminé leur questionnaire. Pour étudier l'effet sélectif de la relance sur la participation aux enquêtes en fonction du profil de l'individu, de l'internaute ou du panéliste, nous divisons l'échantillon des répondants en deux : les individus qui ont complété leur questionnaire sans être relancés et les autres. A nouveau, les variables au sein des trois familles sont classées par ordre décroissant de significativité statistique.

Tableau 2 – Variables corrélées (ou non) à l'effet de la relance (RELA)

Variables	χ^2	p
Le profil de l'individu		
▪ Genre	3.80	0.051 (n.s)
▪ Age	11.05	0.086 (n.s)
▪ La profession	14.29	0.111 (n.s)
▪ Revenu net mensuel du répondant	11.83	0.756 (n.s)
▪ L'activité	3.88	0.794 (n.s)
▪ Niveau d'étude	3.75	0.810 (n.s)
▪ Revenu net mensuel du foyer	6.32	0.852 (n.s)
Le profil de l'internaute		
▪ Type d'accès Internet (bas / haut débit)	6.90	0.032 (*)
▪ Volume de connexion Internet tous accès	3.59	0.827 (n.s)
▪ Volume de connexion Internet domicile	2.10	0.953 (n.s)
Le profil du panéliste		
▪ Délai moyen de réponses aux enquêtes	539.99	0.000 (*)
▪ Ancienneté dans le panel	9.88	0.019 (*)
▪ Temps mis pour s'inscrire	14.71	0.039 (*)
▪ Nombre d'enquêtes proposées	5.38	0.371 (n.s)
▪ Nombre d'enquêtes répondues	5.04	0.411 (n.s)

(*) Significatif à un seuil de significativité de 5%
(n.s) Non significatif à un seuil de significativité de 5%

Le tableau 2 nous révèle qu'à la différence de la participation, aucun critère qui décrit le profil de l'individu n'est corrélé au fait de répondre avant ou après la relance. Les hommes et les femmes, les plus jeunes ou les plus âgés, les plus ou moins fortunés ou éduqués sont aussi nombreux à répondre soit avant soit après la relance. Ce résultat tend

à démontrer que si la relance accroît le taux de participation à l'étude, il n'a pas d'effet sur la composition à terme de l'échantillon sur les critères démographiques.

Il n'en est cependant pas de même si l'on retient les critères qui définissent le répondant en tant qu'internaute et panéliste. Le type d'accès Internet influe sur l'effet de la relance, les individus connectés en bas débit étant en proportion plus nombreux à répondre après une relance que les internautes haut débit (27% contre 21%). Le volume de connexion à Internet aussi bien depuis son domicile que quel que soit le lieu d'accès au Web n'influe pas sur le fait de répondre avant ou après la relance, un résultat surprenant si on le rapproche du fait que le volume de connexion est corrélé au temps moyen de réponse à l'étude (cf. infra).

Enfin, la relance a un effet sélectif sur la participation selon le profil du panéliste. Selon le temps mis pour s'inscrire lors du recrutement, le temps moyen relevé pour répondre aux études antérieures et l'ancienneté dans le panel, la propension à répondre avant ou après la relance est inégalement distribuée : 80% de ceux qui se sont acquittés de leur inscription en un jour (une seule session le plus souvent), lorsqu'ils répondent à l'étude, le font avant relance contre 76% pour les autres ; 90% de ceux qui répondent aux enquêtes en trois jours maximum n'ont pas attendu la relance pour s'acquitter de l'étude faisant l'objet du test (rappelons qu'en moyenne 21% des répondants ont rempli le questionnaire à l'issue de la relance) ; enfin, les répondants les plus récemment inscrits dans le panel (depuis moins d'un an) sont également les plus réactifs : seulement 18% d'entre eux attendent la relance pour répondre contre 24% de ceux qui ont entre un an et deux ans d'ancienneté.

En conclusion, l'effet de la relance est mitigé : si la participation avant ou après relance est peu corrélée au profil de l'individu en termes de variables sociodémographiques, elle s'avère inégalement distribuée selon le type de connexion Internet (elle accroît plus fortement la participation auprès de ceux qui sont connectés en bas débit) ; elle est en outre dépendante du profil historique du panéliste : la relance est moins systématique auprès des panélistes les plus récents (en raison sans doute d'une motivation à répondre à leurs premières enquêtes plus prononcée, bien qu'aucun effet du nombre d'enquêtes proposées ou répondues ne soit statistiquement vérifié) ; elle est d'autant plus nécessaire que le panéliste affiche par son comportement passé (temps mis pour s'inscrire, temps moyens relevés pour répondre aux études) une inertie plus grande. Ces résultats sont confirmés par l'étude des variables corrélées au temps de réponse.

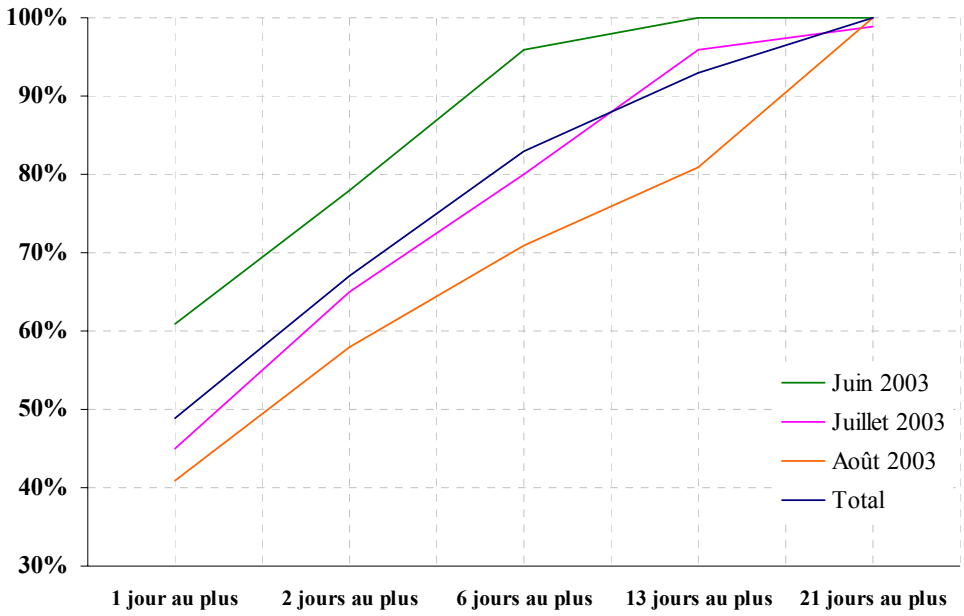
C)- LES VARIABLES CORRELEES AU TEMPS DE REPONSE

Pour répondre aux trois vagues d'enquêtes, les panélistes disposaient de trois semaines à compter de la première notification par courrier électronique. Chaque répondant disposait donc du même délai pour répondre à l'étude. A mi-période, une relance a été faite auprès des non-répondants ou des panélistes qui avaient entamé le questionnaire sans l'achever (dans la mesure où le questionnaire était court, ces derniers étaient marginaux). Trois semaines de collecte de données est un délai plus long que celui habituellement retenu pour les études en ligne ; dans notre cas, ce choix se justifiait par le fait que l'enquête se déroulait pendant la période estivale.

La figure 2 donne la distribution des réponses pendant la durée des enquêtes ; l'examen des quatre courbes confirment l'allure exponentielle et rapidement asymptotique de la

distribution des réponses dans le temps s'agissant d'une étude sur access panel en ligne : en moyenne 50% des répondants remplissent le questionnaire en moins de 24 heures et environ 70% sous deux jours. Bien que les trois courbes présentent une forme proche, la distribution des réponses du mois d'août est plus « étale » que celle du mois de juin et de juillet 2003. Rappelons toutefois qu'ont été retirés de l'échantillon, les personnes ayant déclaré leur indisponibilité pendant la période de l'enquête, raison pour laquelle les trois courbes conservent somme toute une allure comparable.

Figure 2- Courbes de distribution des réponses dans le temps



Le tableau 3 reprend les variables corrélées (ou non) au temps de réponses à l'enquête. Le temps de réponse à l'enquête est une variable ordinale comprenant 5 modalités : moins de 24 heures ; entre 1 et 2 jours ; entre 3 et 6 jours ; entre 6 et 13 jours ; entre 13 et 21 jours. Comme pour les deux tableaux précédents, les variables au sein des trois familles sont classées par ordre décroissant de significativité statistique.

Tableau 3 – Variables corrélées (ou non) au temps de réponse (DREP)

Variables	χ^2	p
Le profil de l'individu		
▪ Genre	8.60	0.071 (n.s)
▪ L'activité	32.92	0.238 (n.s)
▪ Revenu net mensuel du foyer	48.78	0.287 (n.s)
▪ Revenu net mensuel du répondant	69.61	0.294 (n.s)
▪ Age	26.14	0.346 (n.s)
▪ La profession	35.43	0.495 (n.s)
▪ Niveau d'étude	26.46	0.548 (n.s)
Le profil de l'internaute		
▪ Volume de connexion Internet tous accès	44.11	0.003 (*)
▪ Volume de connexion Internet domicile	46.08	0.017 (*)
▪ Type d'accès Internet (bas / haut débit)	16.57	0.035 (*)
Le profil du panéliste		
▪ Délai moyen de réponses aux enquêtes	992.19	0.000 (*)
▪ Temps mis pour s'inscrire	41.56	0.047 (*)
▪ Ancienneté dans le panel	17.24	0.140 (n.s)
▪ Nombre d'enquêtes proposées	20.86	0.405 (n.s)
▪ Nombre d'enquêtes répondues	12.32	0.905 (n.s)

(*) Significatif à un seuil de significativité de 5%
(n.s) Non significatif à un seuil de significativité de 5%

L'examen du tableau 3 nous montre qu'aucune des variables qui décrivent le profil de l'individu n'est corrélée au délai de réponse à l'étude ; il est vrai qu'en moyenne sept panélistes sur dix répondent à l'étude dans les deux premiers jours. Le genre, l'âge, l'activité, la profession, le niveau d'études ou le revenu du répondant et du foyer n'exercent aucune influence sur la propension à répondre immédiatement ou non à l'étude, un résultat qui plaide en faveur d'un des atouts du recours aux access panels en ligne : la réactivité des répondants qui ne paraît pas induire de distorsion sur le profil sociodémographique des répondants.

Il n'en est pas de même toutefois des variables d'usage de l'Internet. Ainsi qu'il est logique, le volume de connexion sur Internet (depuis son domicile et quels que soient les accès) et le type de connexion sur Internet sont positivement corrélés au délai de réponse à l'étude : ceux qui se connectent plus de 10 heures par semaine sur Internet sont 55% à répondre sous 24 heures ; enfin, ceux qui disposent d'une connexion haut débit depuis leur domicile sont 52% à répondre en un jour (contre 46% des connectés bas débit). Ces résultats indiquent qu'il est actuellement nécessaire d'accorder un délai suffisant lorsque le sujet de l'étude est susceptible d'être influencé par le profil de l'internaute (étude sur l'usage de l'Internet, sur les sites fréquentés, sur les centres d'intérêt développés en ligne, sur l'achat en ligne, etc.) sous peine de distordre la représentativité de son échantillon et la qualité de ces estimateurs.

Enfin, s'agissant du comportement passé du panéliste, deux variables seulement se révèlent positivement corrélées au temps mis pour répondre à l'étude : le délai moyen relevé sur l'ensemble des études auxquelles il a déjà répondu et le temps mis pour achever son inscription au panel. Ces deux variables semblent indiquer qu'il existe une certaine constance dans le comportement de l'internaute vis-à-vis des études en ligne : ceux qui répondent le plus tard le font généralement pour l'ensemble des études et ont

été aussi ceux qui ont mis le plus de temps à compléter leur inscription. La mise en place d'un outil de notation de la réactivité du panéliste, appuyé sur ces deux critères est pour cette raison envisagée chez Panel On The Web. Cet outil permettrait de disposer d'un panel très réactif pour des études « express » sous réserve des limites indiquées au paragraphe précédent. Enseignements également intéressants : l'ancienneté dans le panel (entre 1 mois et 4 ans), le nombre d'études proposées et le nombre d'études auxquelles a répondu le panéliste n'influencent pas le délai moyen de réponse (alors que ces mêmes variables influencent la participation à l'étude).

D)- HIERARCHIE DES VARIABLES EXPLICATIVES DE LA PARTICIPATION (PART)

Des trois variables expliquées, l'une d'elle mérite une exploration plus approfondie : le taux de participation en raison d'une part du plus grand nombre de variables explicatives qui lui sont corrélées (10 sur les 14 renseignées) et d'autre part des implications managériales qui en découlent, le taux de participation étant une des variables clés de la gestion d'un access panel. Les tests du χ^2 nous permettent d'établir quelles sont les relations statistiques entre les deux groupes de variables, dépendantes et indépendantes, sans toutefois établir une hiérarchie entre elles et éliminer la part de variance que les variables explicatives partagent entre elles.

Pour atteindre cet objectif, nous recourons à une segmentation selon la méthode de l'arbre de décision interactif (IDT) proposée par le logiciel SPAD. Cette méthode s'appuie sur une discrimination en terme d'arbre binaire, voie ouverte par Morgan et Sonquist (1963) et Morgan et Messenger (1973) avec la méthode dite AID pour « *Automatic Interaction Detection* ». La segmentation par arbre de décision binaire présente l'avantage d'être peu contrainte par la nature des données : on peut en effet utiliser comme variables explicatives à la fois des variables continues ou ordinales et nominales, le même algorithme étant mis en œuvre pour analyser une variable nominale (analyse discriminante) et une variable continue (régression multiple). Pour plus de détails, nous renvoyons le lecteur intéressé à l'ouvrage de Celeux et Nakache (1994) et à l'article illustré de Gueguen et Nakache (1988). Il existe différents algorithmes de segmentation : les deux plus répandus sont CHAID (*Kass, 1980*), méthode d'induction d'arbre de décision reposant sur un critère de discrimination statistique, la mesure du χ^2 et C&RT, issu d'une monographie (*Breiman et al., 1980*) qui propose une méthode unifiée pour traiter les problèmes de discrimination et de régression à partir de la notion de « pureté », c'est-à-dire d'élagages successifs de l'arbre complètement développé de manière à optimiser le taux de mauvais classement. Nous choisissons l'algorithme de CHAID, plus approprié lorsqu'on veut procéder à une première exploration des données (*Kass, 1980*) et dans lequel la décision de segmenter un sommet dépend d'un test d'indépendance du χ^2 effectué sur le tableau de contingence associé aux feuilles qui seront produites par la segmentation. Si ce test est négatif, le sommet n'est pas segmenté et devient un sommet terminal. L'algorithme est appliqué successivement à deux échantillons : un premier dit « de test » correspond à 33% des individus tirés au hasard permet de déterminer les règles de segmentation qui seront ultérieurement appliquées à la fraction restante de la base (appelé échantillon d'apprentissage).

Nous choisissons de retirer d'une première analyse la variable nombre d'enquêtes proposées dans la mesure où celle-ci s'avère corrélée (et redondante) avec deux autres variables intégrées dans la segmentation : l'ancienneté dans le panel ($\chi^2 = 4098$; p = 0.000) et le nombre d'enquêtes auxquelles le panéliste a répondu ($\chi^2 = 5396$; p = 0.000). La matrice de confusion sur les échantillons de test et d'apprentissage nous permet de valider la qualité de l'arbre de segmentation obtenue : avec respectivement 75% et 76% des individus correctement classés, l'analyse de segmentation se révèle satisfaisante.

La mesure de l'impact de chaque attribut nous permet de connaître le rôle de chaque variable dans l'élaboration de l'arbre. Les valeurs indiquées au tableau 4 sous le libellé « impact » représentent une moyenne pondérée de l'impact de chaque attribut sur toutes les segmentations candidates, sachant que moins d'importance est conférée aux impacts mesurés sur les parties basses (à droite) de l'arbre.

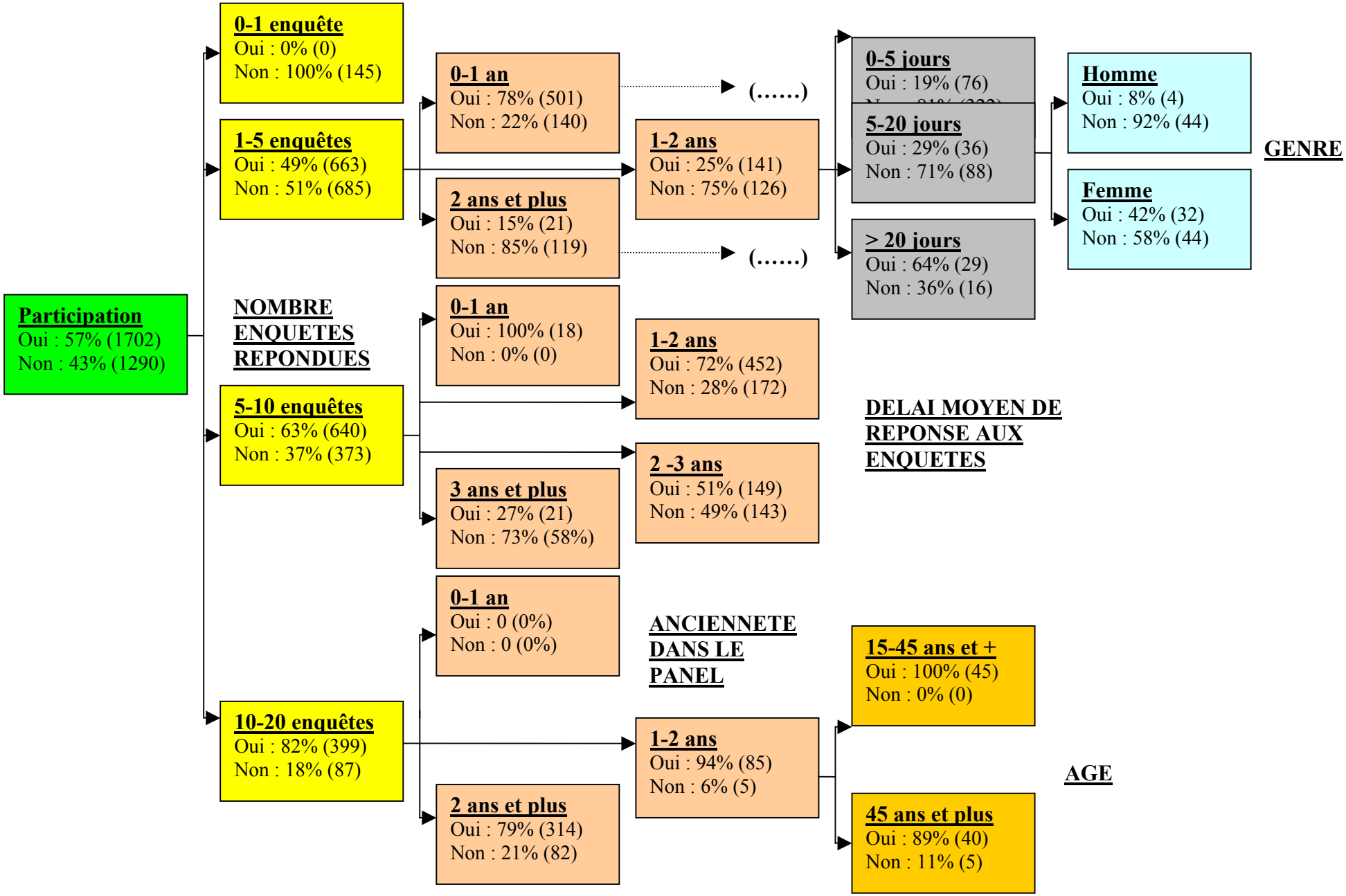
Tableau 4 – Caractérisation de la participation (PART) par les variables explicatives

Variables	χ^2	Valeur test	P	Impact pondéré
▪ Nombre d'enquêtes répondues	563.26	99.99	0.000	0.2020
▪ Délai moyen de réponses aux enquêtes	296.30	99.99	0.000	0.1458
▪ Ancienneté dans le panel	124.29	10.65	0.000	0.1266
▪ L'activité	76.52	7.23	0.000	0.1132
▪ Age	64.69	6.81	0.000	0.0787
▪ Revenu net mensuel du répondant	57.88	4.72	0.000	0.0680
▪ Genre	11.42	3.18	0.001	0.0448
▪ Temps mis pour s'inscrire	21.79	2.55	0.005	0.0390
▪ Volume de connexion Internet domicile	19.35	2.22	0.013	0.0000

(*) Classées par ordre décroissant de la probabilité $p < 0.05$.

L'analyse du tableau que complète l'examen de l'arbre dont une illustration est présentée à la figure 3 nous enseignent que les trois variables les plus fortement corrélées (explicatives) au taux de participation relèvent toutes du profil du panéliste, à savoir le nombre moyen d'enquêtes auxquelles il a répondu, le délai moyen de réponse et son ancienneté dans le panel. Les variables d'activité, d'âge, de revenu individuel et de genre qui caractérisent l'individu ont un impact moins élevé sur le fait de participer à l'étude puis vient ensuite le temps mis pour s'inscrire lors de la phase de recrutement et enfin le volume de connexion à Internet depuis son domicile. Il semble donc exister une propension à participer de manière régulière aux études en ligne en partie indépendante du profil de l'internaute et sans doute davantage liée à une attitude générale sur le fait de donner son opinion ou de se conformer aux règles souscrites lors de son engagement dans la phase d'inscription au panel ; en d'autres termes, la probabilité de participer à une étude est conditionnée par sa participation passée, sa réactivité et son ancienneté dans le panel. Soulignons enfin que le taux de participation le plus élevé est relevé auprès de ceux qui ont répondu à au moins cinq enquêtes (69% contre 44%), qui mettent en moyenne trois jours pour répondre (63%contre 57%) ou bien, dans un autre registre, qui ont moins d'un an d'ancienneté dans le panel (71% contre 53%). Tous critères confondus, le segment qui affiche le taux de participation le plus élevé est celui des panélistes âgés de 15 à 45 ans, ayant de 1 à 2 ans d'ancienneté et ayant déjà répondu à 10 enquêtes au moins et 20 enquêtes au plus.

Figure 3 – Arbre de décision interactif simplifié établi à partir de la variable expliquée PART (participation)



5)- CONCLUSIONS ET LIMITES DE LA RECHERCHE

La recherche qui fait l'objet de cet article est une première exploration des données collectées sur la non-réponse (qu'il convient de qualifier plus justement de non-participation) dans le cadre d'une étude menée depuis un access panel en ligne. Il est donc logique à ce stade que les premiers résultats pour intéressants qu'ils soient appellent des réserves et des limites qu'il convient de rappeler.

Un des enseignements majeurs de l'étude est que, pour élevée que soit la participation à une étude menée sur un access panel en ligne, elle ne se distribue pas de façon homogène sur l'ensemble des individus. Les caractéristiques sociodémographiques du répondant - le genre, l'âge, l'activité et le revenu – si elles se révèlent corrélées au fait de répondre à l'étude exercent toutefois une influence moindre que les variables qui témoignent du comportement du panéliste, en particulier le nombre d'enquêtes auxquelles il a répondu, son ancienneté dans le panel et le temps moyen de réponse aux études. Il semble donc exister une propension à répondre de manière constante aux études en ligne, une attitude en partie indépendante des caractéristiques sociodémographiques de l'individu. Quels sont les déterminants de cette attitude ? Une volonté de donner son opinion pour mieux faire connaître ses attentes ? Un réflexe communautaire qu'alimente l'appartenance à la communauté des internautes ? Il est difficile de répondre à cette question qui appelle nécessairement d'autres investigations.

Les premiers résultats sur le taux de participation sont utilement éclairés par l'étude de deux autres variables : le fait de répondre avant ou après une relance et le délai de réponse. Si l'effet de la relance n'est pas dépendant du profil sociodémographique, il est toutefois dépendant du profil du panéliste : l'impact de la relance sur la participation est plus prononcé auprès des panélistes les plus anciens et auprès de ceux qui ont révélé une moindre réactivité dans le passé ; enfin elle est plus efficace auprès des internautes connectés en bas débit qu'en haut débit. Les mêmes conclusions s'appliquent s'agissant des variables corrélées au délai de réponse ; s'y ajoute l'influence du volume de connexion à Internet quel que soit le mode d'accès.

Notre recherche a pris pour variable dépendante la participation et deux autres variables qui lui sont rattachées : le délai de réponse et l'effet de la relance. S'agissant de la problématique plus générale de la non-réponse, nous avons écarté un effet important : celui qu'exerce la non-participation (ou le refus de répondre) sur la qualité des réponses elle-même (que l'on peut assimiler à la fiabilité des estimateurs). La mise en évidence des variables corrélées à la participation (à l'effet de la relance ou au délai de réponse) a un impact managérial d'autant plus prononcé qu'elle produit une distorsion sur la fiabilité des résultats ; en d'autres termes, la non-participation revêt une importance d'autant plus critique qu'il existe, ainsi que nous l'avons rappelé, un écart entre les réponses des répondants et des non-répondants, un sujet que notre recherche n'aborde pas et qui en forme une des limites. Enfin, la période de collecte de données forme aussi une limite de notre recherche : il conviendrait de répliquer le principe de cette étude en dehors de la période d'été pour renforcer la validité externe des conclusions rapportées.

REFERENCES

Bosnjak Michael M., Tuten Tracey L. (2001), « *Classifying Response Behaviors in Web-based Surveys* », Journal of Computer-Mediated Communication, 6, 3, http://www.ascusc.org/jcmc/vol6/issue3/boznjak.html

Breiman L., Friedman J., Olshen R. A. et Stone J.C. (1984), *Classification and regression trees,* California : Wadsworth International, 358 p.

Brennan Mike, Hoek Janet et Astridge Craig (1991), « *The effect of monetary incentives on the response rate and cost-effectiveness of a mail survey* », Journal of the Royal Statistical Society, 33, 1, p. 229-241.

Celeux G., Nakache J.P (1994), « *Analyse Discriminante sur Variables Qualitatives* », Paris, Polytechnica, 270 p.

Chen Henry C. K. (1996), *« Direction, magnitude and implications of non-response bias in mail survey »,* Journal of the Market Research Society, 38, 3, p.267-276.

Collins L. M., Schafer Joseph L. et Kam C. M. (2001), « *A comparison of inclusive and restrictive missing-data strategies in modern missing data-procedures »,* Psychological Methods, 6, p. 330-351.

Dempster A. P., Laird N. M. et Rubin D. B. (1977), « *Maximum likelihood form incomplete data via the EM algorithm »,* Journal of the Royal Statistical Society, 39, 1, p. 1-38.

Ferber Robert (1948), *The problem of bias in mail return : a solution,* Public Opinion Quarterly, 39, 3, p.239-244.

Frankel M. R., Frankel L. R. (1977) « *Some recent developments in sample survey design »,* Journal of Marketing Research, 14, p. 280-293.

Gueguen A., Nakache J. P. (1988), « *Méthode de discrimination basée sur la construction d'un arbre de décision binaire »,* Revue de Statistique Appliquée, 36, 1, p.19-38.

Jolibert Alain, Jolibert Bertrand (2003), *« La validité des panels d'internautes »,* p. 213-220, dans *Savoir gérer : mélanges en l'honneur de Jean-Claude Tarondeau,* coordonnée par Pierre Le Moal, Paris : Vuibert.

Hoppe, Michael et Lamp, Rainer (2001), « *The quality of online panels - a methodological test* », dans Fellows, Deborah S. (Ed.): « *Worldwide Internet Conference and Exhibition Net Effects* », 4, Barcelone, Espagne, 11-13 février 2001, Amsterdam : Esomar, p. 243-262.

Kalton Graham (1983), *Introduction to Survey Sampling,* Quantitative Applications in the Social Sciences Series, Newbury Park, California : Sage Publications, 94 p.

Kass G. V. (1980), « *An exploratory technique for investigating large quantitites of categorical data* », Applied Statistics, 29, 2, p. 119-127.

Lebart Ludovic (2001), *« Introduction au prétraitement des fichiers d'enquêtes : redressement ; données manquantes, fusions / injections »,* dans *Traitement des fichiers d'enquêtes : redressement ; données manquantes, fusions / injection,* Michel Lejeune (éditeur), Presses Universitaires de Grenoble, p. 9-15.

Litlle Roederick J. A., Rubin Donald B., *Satistical analysis with missing data,* 2ème édition, Hoboken, New Jersey : Wiley, 381 p.

Madow William G., Nisselson Harold et Olkin Ingram (1983), *Incomplete data in sample surveys,* 2 vol., New York : Academic Press.

Marjorie Donald N. (1960), « *Implication of non-response for the interpretation of mail questionnaire data »,* Public Opinion Quarterly, 1, p. 99-114.

Morgan J. N., Messenger R. C. (1973), « *THAID : a Sequential Search Program for the Analysis of Nominal Scale Dependent Variables* », Institute for Social Research, université du Michigan.

Morgan J. N., Sonquist J. A. (1963), « *Problems in the Analysis of Survey Data and a Proposal »,* Journal of the American Statistical Association, 58, p. 119-127.

O'Muircheartaigh Colm, Campanelli Pamela (1999), *« A Multilevel exploration of the role of interviewers in survey non-response »,* Journal of the Royal Statistical Society, series A, , 162, 3, p. 437-446.

Robert Christian (1996), *Méthodes de Monte Carlo par chaînes de Markov,* série Statistique Mathématique et Probabilité, Paris : Economica, 340 p.

Schafer Joseph L. (1997), *Analysis of Incomplete Multivariate Data,* Monographs on Statistics and Applied Probability Series, 72, London : Chapman & Hall, 448 p.

Schafer Joseph L., Graham J. W. (2002), « *Missing data : our view of the state of the art »,* Psychological Methods, 7, p. 147-177.

Wiseman Freederick, McDonald Philip (1980), *Towards the development of industry standards for response and non-response rate*, Cambridge, Mass. : Marketing Science Institute.