Techniques d'enquête, décembre 2001
Vol. 27, N° 2, pp. 143-155
Statistique Canada, N° 12-001 au catalogue

Effet de l'intensité des efforts en vue de joindre les répondants : enquête torontoise sur le tabagisme

Louis T. Mariano et Joseph B. Kadane [1]

Résumé

Dans une enquête téléphonique, le nombre d'appels est utilisé comme indicateur de la difficulté à joindre le répondant. Ceci permet, dans un modèle de non-réponse, une division probabiliste des non-répondants en deux catégories : réfractaires (ceux qui refuseront toujours de répondre) et non-réfractaires (ceux qui ne sont pas disponibles pour répondre). Cela permet en outre d'estimer stochastiquement les opinions de ce dernier groupe de non-répondants et d'évaluer si la non-réponse est ignorable aux fins des inférences au sujet de la variable dépendante. Nous avons appliqué ces idées aux données d'une enquête dans la région métropolitaine de Toronto ayant porté sur les attitudes à l'égard de l'usage du tabac en milieu de travail. À l'aide d'un modèle bayesien, nous échantillonnons la distribution postérieure des paramètres du modèle par les méthodes de Monte Carlo à chaîne de Markov. Les résultats révèlent que la non-réponse n'est pas ignorable et que ceux qui n'ont pas répondu étaient deux fois plus susceptibles d'accepter le libre usage du tabac en milieu de travail que ceux qui ont répondu.

Mots clés : Rappels, nombre de; analyse bayesienne; méthode de Monte Carlo à chaîne de Markov; non-réponse informative; non-réponse ignorable.

1. Introduction

Compte tenu des réalités de la non-réponse dans toute enquête, il est bon de juger comment on tiendra compte de cette dernière dans l'interprétation des données recueillies. Rubin (1976) énonce des conditions nécessaires et suffisantes pour qu'une telle analyse se confonde, des points de vue fréquentistes, de vraisemblance et bayesiens respectivement, avec une analyse reposant sur un modèle qui comporte un mécanisme de données manquantes. C'est en s'appuyant sur ce cadre que Little et Rubin (1987) ont enrichi une vaste documentation spécialisée d'une modélisation « informative et non-ignorable » de la non-réponse.

En cernant l'interaction enquête-enquêté, on peut affiner l'analyse de l'importance des données manquantes dans une enquête. Pour notre propos, nous citerons l'exemple d'une enquête sur les attitudes des Torontois à l'égard de l'usage du tabac en milieu de travail. On avait choisi des numéros de téléphone au hasard et, pour joindre les gens ainsi visés, on avait fait au moins 12 tentatives d'entrée en communication. Les données relatives aux répondants nous renseignent uniquement sur le nombre d'appels jusqu'à achèvement d'interview et ne précisent pas les moments où les tentatives infructueuses ont eu lieu. Même avec ces données moins riches sur la difficulté de joindre les répondants, nous constatons que le nombre d'appels infructueux est une indication importante au moment de considérer les résultats de l'enquête.

L'utilisation des données sur le nombre de tentatives d'entrée en communication avec l'enquêté sélectionné n'a rien d'unique. Potthoff, Manton et Woodbury (1993) exposent une méthode de correction de biais d'enquête par indisponibilité qui prévoit une pondération en fonction du nombre de rappels. Notre analyse vise aussi le biais par indisponibilité, mais avec de grandes différences. Au lieu de supposer qu'il n'y a pas de refus, nous tenons compte de leur éventuelle existence dans une modélisation du mécanisme qui cause la non-réponse. Dans l'analyse qui suit, nous évaluons le rapport entre la non-réponse et la variable dépendante d'intérêt dans cette enquête avec les autres variables explicatives, et ce, après une pondération en fonction tant de la taille des ménages que de caractéristiques démographiques appropriées de la population. Nous nous trouvons donc à nous demander non seulement s'il y a erreur par indisponibilité, mais aussi si une stratification des répondants selon la taille des ménages et la structure âge-sexe actuelle peut écarter la nécessité de prendre l'erreur en compte par un mécanisme de description de la non-réponse. À noter que, dans ce cas, nous nous alignons sur les groupes de Pederson, Bull et Ashley (1996) dans les analyses originales publiées de l'ensemble de données. Des méthodes de correction de cellules plus complexes sont possibles (Little 1996; Eltinge et Yansaneh 1997, mentions bibliographiques de ces documents, *etc.*).

Dans l'ordre de présentation de notre article, la section 2 renseigne plus en détail sur l'enquête, la section 3 expose la méthodologie employée, les sections 4 et 5 examinent respectivement les modèles « données manquantes aléatoires » et « données manquantes non-ignorables », la section 6 décrit les distributions *a priori* choisies pour l'analyse principale, la section 7 explique les résultats de cette analyse et la section 8 tire des conclusions.

1. Louis T. Mariano est un candidat au doctorat, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213; Joseph B. Kadane est Leonard J. Savage University Professor of Statistics and Social Sciences, Department of Statistics, Carnegie Mellon University, Pittsburg, PA 15213.

2. Enquête

Dans la municipalité de Toronto, un règlement sur l'usage du tabac en milieu de travail est entré en vigueur le 1er mars 1988. Depuis janvier de cette même année, on a effectué six enquêtes pour évaluer les attitudes de la population à l'égard du tabagisme, la sensibilisation au danger pour la santé de l'usage du tabac et l'incidence du règlement sur les résidents de la région métropolitaine de Toronto. Les données de notre analyse viennent de la troisième de ces enquêtes. Northrup (1993) livre des indications techniques sur cette dernière. Pour plus de clarté, précisons que nos données d'analyse sont celles de la troisième reprise de l'enquête et que les données des deux premières sont alors celles des phases I et II.

Northrup (1993) indique que les données d'intérêt, qui ont été mises à la disposition des intéressés par l'Institute for Social Research (ISR) de l'Université York, ont été recueillies auprès de 1 429 résidents de la région métropolitaine de Toronto en décembre 1992 et mars 1993. Aux fins de l'enquête, il y a eu sélection probabiliste en deux degrés des répondants. Au premier degré, on a fait de la « composition aléatoire » et, au deuxième degré, on s'est reporté au jour de naissance le plus récent pour sélectionner un adulte après entrée en communication avec le domicile admissible. On a alors pondéré les réponses selon le nombre d'adultes dans les ménages. Dans l'analyse qui suit, nous avons également appliqué une poststratification par groupe âge-sexe selon les données du recensement en vue d'une correction de sous-représentation de sous-populations. Au stade de la collecte de données, le nombre de lignes téléphoniques des ménages n'a pas été pris en considération.

Le nombre de tentatives d'entrée en communication figure comme variable dans l'ensemble de données. Il n'y a pas de valeurs manquantes pour cette variable. Northrup (1993) explique que les 1 429 réponses ont été tirées d'un échantillon de 5 702 numéros de téléphone générés par la technique de composition aléatoire. Sur ces 5 702 ménages, 2 286 ont été jugés admissibles et 3 150, inadmissibles après vérification. On n'a pu déterminer l'admissibilité des 266 ménages restants. L'ISR a supposé que le taux d'admissibilité des ménages était le même pour ces 266 numéros de téléphone que pour le reste de l'échantillon de composition aléatoire. Ce taux implique un total estimatif de 2 398 ménages échantillonnés et un taux de réponse de 60 %. Ainsi, on estime à 969 le nombre de ménages sélectionnés qui n'ont pas répondu. Chacun a reçu 12 appels au minimum le jour, le soir et le week-end avant d'être classé comme « ménage non répondant ».

Aux fins de la présente analyse, la variable dépendante est l'opinion d'une personne sur la réglementation de l'usage du tabac en milieu de travail selon une des trois catégories suivantes : la catégorie 0 correspond à la permission de fumer seulement dans des zones réservées, la catégorie 1, à l'interdiction totale du tabac, et la catégorie 2, à une permission totale. Pour chaque sujet soumis à l'enquête, soit $Y_i \in \{0, 1, 2\}$ l'opinion du sujet i.

Les données portent sur les réponses à 50 questions et sur 18 autres variables de caractérisation des sujets. Voici quelques indicateurs employés :

- « Connaissance des risques » est un résultat entier variant de 0 à 12 pour la connaissance des risques et des effets du tabagisme passif;
- « État de fumeur » indique si le sujet fume actuellement (S), a déjà fumé (SQ) ou n'a jamais fumé (NS);
- « État de réaction » indique si la fumée secondaire dérange le sujet : « dérange toujours » (b.A), « dérange d'habitude » (b.USUL) et « ne dérange pas » (b.NO);
- « Âge » : (âge en années – 50) / 10.

Pederson, Bull, Ashley et Lefcoe (1989) ont élaboré un indicateur de connaissance des effets sur la santé du tabagisme passif à l'aide des réponses à six questions de l'enquête où on mesurait la connaissance qu'avait le sujet des effets de la fumée secondaire. Avec les questions de Pederson et coll., on a créé pour les données de phase III l'indicateur ici rebaptisé « Connaissance des risques ». Une note plus élevée pour la connaissance des risques indique que le sujet connaît mieux les dangers du tabagisme passif. Nous avons enfin modifié et remis à l'échelle la variable « Âge » pour ainsi nous aligner sur le traitement de l'âge par Bull (1994) dans l'analyse des données des phases I et II.

3. Aperçu de la méthodologie

Notre question fondamentale est la suivante : pouvons-nous ne pas tenir compte de la non-réponse par unité et traiter les données observées comme un sous-échantillon aléatoire de la population? Pour reprendre les termes de Little et Rubin (1987) et de Rubin (1976), s'il est possible de traiter les données d'observation de la variable dépendante d'intérêt comme un sous-échantillon aléatoire, les données manquantes sont alors entièrement aléatoires (« missing completely at random » ou MCAR). S'il est possible de traiter les données d'observation de la variable dépendante d'intérêt comme sous-échantillon aléatoire en conditionnant par les variables explicatives, les données manquantes sont « simplement aléatoires » (« missing at random » ou MAR). Si θ représente les paramètres des données et π, les paramètres du processus générant les données manquantes, Rubin (1976) dit de ces paramètres qu'ils sont distincts si on ne peut lier *a priori* π à θ par des restrictions d'espace paramétrique ni des distributions *a priori* de paramètres. Si le traitement MCAR ou MAR s'applique et que π et θ sont distincts, le mécanisme qui cause la non-réponse est jugé « ignorable » aux fins des inférences au sujet de la distribution de la variable d'intérêt. Si les données manquantes pour la variable dépendante sont

fonction des valeurs de cette variable, le mécanisme est dit « non-ignorable » (NI). Groves et Couper (1998) font observer que, si les probabilités de participation sont fonction de la variable dépendante visée, le biais de non-réponse peut être relativement important même avec un bon taux de réponse.

Soit R_i un indicateur de réponse, $R_i = I_{\{\text{répondant}\}}$ (sujet i) et $R = (R_1, ..., R_n)^T$. Little et Rubin (1987) font voir qu'une méthode possible de prise en compte du mécanisme de non-réponse est l'inclusion dans le modèle de cette variable indicatrice de réponse. On peut qualifier le mécanisme de non-réponse d'ignorable si π et θ sont distincts et que :

$$f(R \mid Y_{\text{obs}}, Y_{\text{mis}}, \pi) = f(R \mid Y_{\text{obs}}, \pi) \qquad (1)$$

où Y_{obs} et Y_{mis} représentent respectivement les données observées et les données manquantes de la variable dépendante d'intérêt.

Dans toute cette analyse, nous emploierons les termes « hypothèse MAR » et « hypothèse NI ». Précisons que, dans le premier cas, il s'agit de l'hypothèse du caractère non informatif du mécanisme de non-réponse aux fins des inférences relatives à la variable dépendante indiquée à la section 2. En d'autres termes, les valeurs observées de cette variable constituent un sous-échantillon aléatoire de la population, peut-être à l'intérieur de postrates, et il est inutile de tenir compte du mécanisme de non-réponse. Dans le second cas, il s'agit de l'hypothèse du caractère non-ignorable de ce même mécanisme et de l'impossibilité de traiter comme sous-échantillon aléatoire les données recueillies pour la variable dépendante. Plus précisément, les inférences au sujet de la population doivent tenir compte du mécanisme de non-réponse.

L'évaluation de l'hypothèse MAR se fait en trois étapes. On examine d'abord ce qu'on peut faire si on retient cette hypothèse. Comme la variable dépendante d'intérêt comporte trois catégories et que certaines des variables explicatives sont quantitatives, on recourt à une régression logistique polytomique. On examine alors et la forme fréquentiste et la forme bayesienne d'un tel modèle.

On élabore ensuite un modèle NI. On modélise le mécanisme de non-réponse en se reportant au nombre de tentatives d'entrée en communication avec les divers sujets. Ici, on examine l'idée d'une fraction de « survivants » dans l'échantillon pour juger s'il est effectivement possible de joindre tous les sujets sélectionnés. Ensuite, on relie le mécanisme de non-réponse à la variable dépendante en intégrant le nombre d'appels au modèle de régression logistique.

Dans l'élaboration du modèle NI, nous employons une technique bayesienne de détermination des valeurs probables des données manquantes compte tenu des données d'observation et des paramètres du modèle. Nous appliquons à cette fin la technique d'augmentation de données où il y a imputation des données manquantes à chaque itération d'une simulation de Monte Carlo à chaîne de Markov (méthode MCCM). Une autre façon de procéder serait de calculer par l'algorithme de maximisation des espérances ME (Dempster, Laird et Rubin 1977) des estimations par la méthode du maximum de vraisemblance (EMV) pour ces mêmes données manquantes.

Nous évaluons enfin l'hypothèse MAR. L'existence de coefficients non nuls pour le nombre d'appels dans le volet « régression logistique » du modèle NI implique que le nombre d'appels fait toute la différence, c'est-à-dire que les opinions de ceux qui n'ont pas répondu aux 12 premières tentatives d'entrée en communication sont susceptibles de différer des opinions de ceux qui ont répondu après quelques tentatives seulement. Dans ce cas, le mécanisme de non-réponse n'est pas indépendant des valeurs des données manquantes et une hypothèse MAR n'a pas sa place. Il s'agit ensuite d'examiner les probabilités de réponse en expression logarithmique pour les trois modèles. Les différences dégagées indiquent l'ordre de grandeur de l'erreur causée par une fausse hypothèse MAR. Ainsi, dans l'évaluation de ces hypothèses, on répond à deux questions : y a-t-il une différence? quelle est l'importance de cette différence?

4. Modèles MAR

4.1 Régression logistique

En nous reportant aux données recueillies auprès des $(m = 1\,429)$ sujets qui ont répondu à l'enquête, nous modélisons par une régression logistique pondérée les opinions de la population au sujet de l'usage du tabac en milieu de travail. Nous avons resserré l'éventail des prédicteurs possibles (selon les questions de l'enquête et les renseignements généraux) par une série de tests de Wald. Nous avons ensuite comparé les modèles possibles par des tests de rapport des vraisemblances, CIA et CIB. Nous avons jugé que le modèle le mieux ajusté était celui qui comprenait des termes additifs pour les variables « Connaissance des risques », « État de fumeur », « État de réaction » et « Âge » (voir la section 2).

Comme chacun des modèles de notre analyse comporte un volet « régression logistique », il serait bon de mieux décrire ici la notation employée. La catégorie 0 « permission de fumer seulement dans des zones réservées » est notre catégorie de référence. On doit se rappeler que $Y_i \in \{0, 1, 2\}$. Pour le modèle MAR, nous nous reportons uniquement aux valeurs observées des opinions des sujets sur l'usage du tabac en milieu de travail, $Y_{\text{obs}} = (Y_1, ..., Y_m)$. Soit $Y_{ij} = I_{\{j\}}(Y_i)$ un indicateur du sujet i répondant dans la catégorie j et soit W_i le poids attribué à chaque sujet. Comme dans les analyses originales publiées de cet ensemble de données (Pederson et coll. (1996)), il y a pondération de ménage (voir Northrup (1993)) et de post-stratification (voir l'annexe A) dans tous les modèles considérés ici.

Nous avons intégré à notre modèle les deux variables explicatives catégoriques « État de fumeur » et « État de

réaction » par les variables indicatrices de deux des trois catégories, l'effet de la troisième catégorie étant alors absorbé dans le terme « ordonnée à l'origine ». Pour l'état de fumeur, les variables indicatrices « S_i » et « SQ_i » désignent respectivement l'usage actuel et antérieur du tabac. Pour l'état de réaction, les variables indicatrices « $b.USUL_i$ » et « $b.NO_i$ » désignent respectivement une réaction habituelle et une absence de réaction du sujet i à la fumée secondaire.

Soit X_i = le vecteur des variables explicatives pour le sujet i,

$$X_i = (\text{K}-\text{risk}_i,\ \text{S}_i,\ \text{SQ}_i,\ \text{b.USUL}_i,\ \text{b.NO}_i,\ \text{Age}_i).$$

Par un modèle logistique multinomial non ordonné, nous considérons $p_j(x_i) = P(Y_{ij} = 1 \mid X_i = x_i)$, c'est-à-dire les probabilités que le sujet i réponde dans la catégorie $j \in \{0, 1, 2\}$, étant donné les variables explicatives observées pour ce sujet. Bien sûr, ce modèle utilise des équations linéaires η_{ij} décrivant en expression logarithmique les probabilités de réponse du sujet i dans la catégorie j par rapport à la catégorie de référence $j = 0$. Ainsi, pour $j = 1, 2$, nous voulons examiner :

$$ln\ \frac{p_j(x_i)}{p_0(x_i)} = \eta_{ij} = \beta_{0j} + X_i\beta_j, \qquad (2)$$

avec $\eta_{i0} = 0$. Les deux équations linéaires résultantes, η_{i1} et η_{i2}, ont chacune sept coefficients, soit une ordonnée à l'origine β_{0j} et les coefficients suivants :

$$\beta_j = (\beta_{\text{K}-\text{risk}_j},\ \beta_{\text{S}_j},\ \beta_{\text{SQ}_j},\ \beta_{\text{b.USUL}_j},\ \beta_{\text{b.NO}_j},\ \beta_{\text{Age}_j}).$$

Le modèle de régression logistique MAR comporte 14 paramètres. Le vecteur de ces 14 paramètres, représenté par $\beta = (\beta_{01}, \beta_1, \beta_{02}, \beta_2)$, a la vraisemblance (ou, plutôt, la pseudo-vraisemblance puisque les poids sont incorporés au moyen de la variable W_i) :

$$L(\beta) \propto \prod_{i=1}^{m} \prod_{j=0}^{2} \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i}. \qquad (3)$$

4.2 Régression logistique bayesienne

Nous reprenons la valeur de vraisemblance de l'équation (3) et les données recueillies auprès des répondants dans une analyse bayesienne. Les quatre variables explicatives de l'analyse fréquentiste précédente sont aussi reprises. Nous avons attribué des distributions *a priori* (voir la section 6) aux paramètres de régression logistique. Nous procédons par simulation MCCM pour la distribution postérieure des paramètres.

5. Modèle NI

5.1 Modélisation du mécanisme de non-réponse

Comme les valeurs manquantes ne sont pas nécessairement aléatoires, nous devons tenir compte du mécanisme qui cause la non-réponse. Northrup (1993) indique que les non-répondants à l'enquête ont reçu au moins 12 appels infructueux (dont 3 le jour, 4 le soir et 4 le week-end au moins). Nous devons malheureusement constater que d'autres indications utiles du nombre d'appels n'ont pas été retenues. Nous ignorons lesquels des non-répondants ont reçu plus de 12 appels ou quelles ont été les périodes d'appel (jour, soir ou fin de semaine). Nous ignorons également les détails de la non-réponse, c'est-à-dire si le sujet a refusé de participer après entrée en communication, s'il y a jamais eu enregistrement du message d'appel par un répondeur ou si on n'a pas répondu du tout. La stratification des non-répondants est donc impossible et ceux-ci sont tous considérés comme unités interchangeables dans la présente analyse.

On a tenté un certain nombre de fois d'entrer en communication avec chaque sujet jusqu'à achèvement d'interview ou caractérisation de non-répondant. Dans le cas des répondants, la variable du nombre d'appels (C_i) indique le nombre de tentatives infructueuses jusqu'à la première entrée en communication. Nous pouvons nous attendre, par conséquent, à ce que le nombre d'appels suive une distribution géométrique avec troncation des observations pour les non-répondants. Plus précisément, soit $\pi = P$ (tentative fructueuse). Considérons alors $C_i \sim \text{Geometric}(\pi)$ et $P(C_i = c_i) = \pi(1 - \pi)^{c_i - 1}$. À noter que, si nous avions disposé de données auxiliaires sur le nombre d'appels dans le cas des non-répondants (comme dans Groves et Couper 1998), nous pourrions avoir calculé ici des probabilités conditionnelles de réponse.

Les histogrammes de la figure 1 comparent les données (12 premiers appels) à une distribution géométrique où le paramètre π est de $0{,}225$. La concordance est assez grande. La statistique d'ordre d'échantillon semble indiquer que $\pi \in (0{,}2,\ 0{,}25)$. L'histogramme des données effectives d'enquête révèle qu'il y a moins de sujets joints au premier appel qu'au deuxième. Il est possible qu'on ait fait plus de deuxièmes appels à un moment de la journée où le taux de succès était supérieur.

Supposons que $\pi = 0{,}225$. Selon la propriété « sans mémoire » d'une distribution géométrique, nous pouvons nous attendre à ce que 218 des 969 non-répondants aient effectivement répondu au téléphone à la $13^{\text{ième}}$ tentative. Ainsi, les données sur les 13 premières tentatives d'entrée en communication seraient celles de la figure 2. Nous pouvons nettement voir que cette figure n'a pas le comportement d'une variable aléatoire suivant la distribution géométrique. La question est la suivante : si on avait appelé tous les sujets un nombre illimité de fois, les aurait-on tous joints? Si on répond oui à la question pour l'ensemble de données, on a le problème illustré à la figure 2.

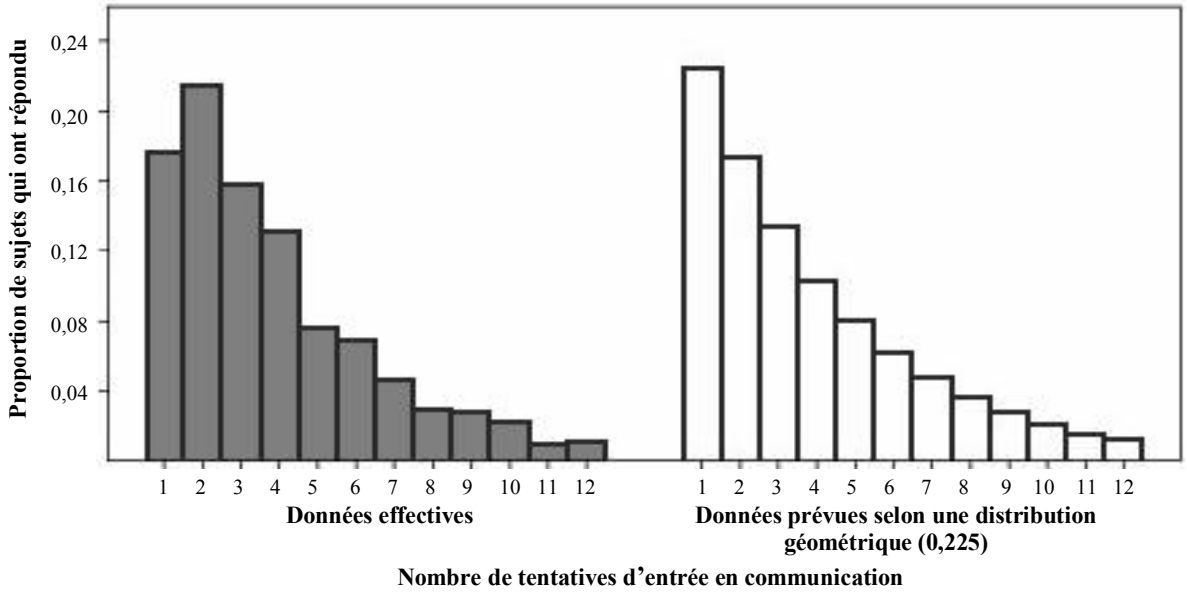

Figure 1 . On compare les données effectives d'enquête sur les entrées en communication aux 12 premières tentatives, d'une part, et les résultats prévus selon une distribution géométrique (paramètre $\pi : 0,225$) du nombre d'appels nécessaires à l'achèvement de l'interview, d'autre part.

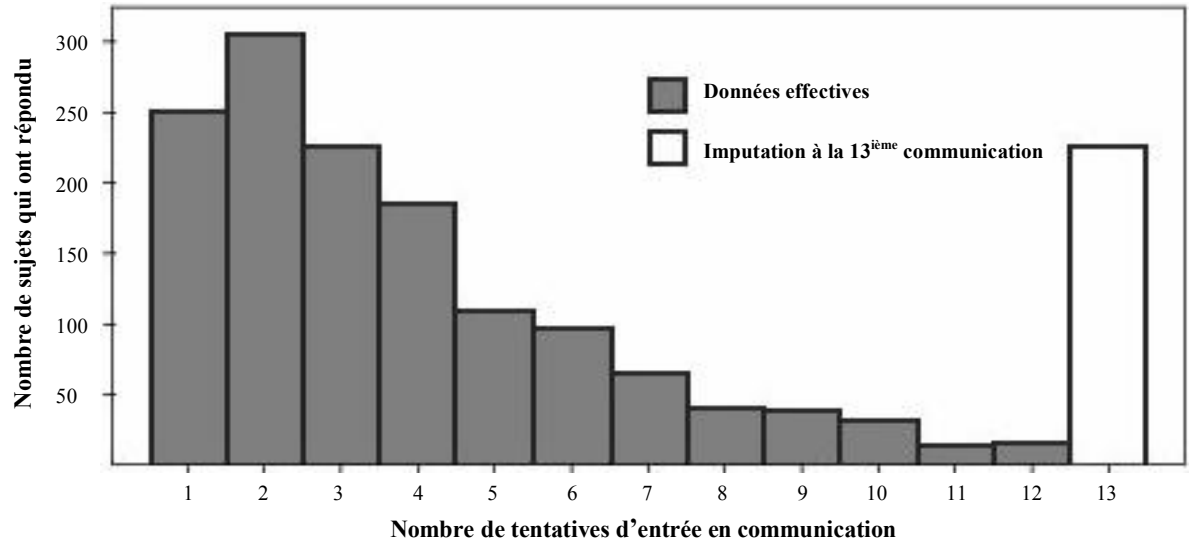

Figure 2 . On décrit le nombre effectif d'entrées en communication pour chacune des 12 premières tentatives, ainsi que le nombre prévu d'entrées en communication à la 13ième tentative. La valeur probable pour la 13ième tentative est tirée d'une distribution géométrique $(\pi : 0,225)$ où on modélise le nombre d'appels jusqu'à achèvement d'interview.

Compte tenu de ces données, il est possible d'affirmer que ce ne sont pas tous les sujets sélectionnés qui seront joignables. Maller et Zhou (1996) décrivent les sujets épargnés, c'est-à-dire ceux qui ne subissent pas l'événement d'intérêt. Pour reprendre les termes qu'ils emploient, s'il est impossible d'obtenir une réponse d'un sujet sélectionné après un nombre illimité d'appels, le sujet est caractérisé comme « épargné ». Les sujets « non épargnés » sont alors caractérisés comme « non réfractaires ». L'ensemble des sujets épargnés (c'est-à-dire réfractaires) forme alors la fraction de « survivants » de l'échantillon. Pour employer des termes plus familiers, les sujets épargnés sont ceux qui, joints, ont refusé de répondre, qui auraient refusé s'ils avaient été joints ou qui étaient physiquement ou mentalement incapables de jamais participer. Northrup (1993) indique que ceux qui ont initialement refusé de participer ont ensuite été joints par les intervieweurs les plus expérimentés, aussi posons-nous l'hypothèse ici que tous ceux qui ont continué à refuser n'auraient jamais participé. Le groupe des non-réfractaires comprend les enquêtés qui, joints, auraient répondu et ceux qui étaient physiquement ou mentalement incapables de participer à l'époque de la collecte de données, mais qui auraient été désireux et capables de le faire en tout autre temps.

Soit la variable $Z_i = I_{\{\text{non réfractaire}\}}$ (sujet i) comme indicateur de sujet i non réfractaire et $\rho = P$ (sujet i non réfractaire), c'est-à-dire $Z_i \sim \text{Bernoulli}(\rho)$. Supposons maintenant que le nombre de tentatives d'entrée

en communication avec les sujets non réfractaires suit une distribution géométrique, c'est-à-dire que $C_i \mid Z_i = 1 \sim$ Geometric (π). Se trouve-t-on à éliminer le problème illustré à la figure 2?

Soit R_i un indicateur de réponse du sujet i. On peut prendre le mécanisme de non-réponse en compte par l'intégration de ces indicateurs de réponse au modèle. Il reste que l'introduction de la variable « sujet non réfractaire » implique deux catégories distinctes de non-réponse. Il est donc possible de faire une caractérisation plus fine et d'utiliser tant les indicateurs de sujet non réfractaire $Z = (Z_1, ..., Z_n)^T$ et de réponse R dans un modèle mixte de description de non-réponse. Dans une mise à jour de l'équation (1), le mécanisme de non-réponse est non informatif si et si seulement (π, ρ) est distinct de θ et que

$$f(R, Z \mid Y_{\text{obs}}, Y_{\text{mis}}, \pi, \rho) = f(R, Z \mid Y_{\text{obs}}, \pi, \rho). \quad (4)$$

Soit $C_{\text{obs}} = (C_1, ..., C_m)$ et $Z_{\text{obs}} = (Z_1, ..., Z_m)$ les vecteurs du nombre d'appels et du caractère « non réfractaire » observé de chaque enquêté. Soit $R = (R_1, ..., R_n) =$ le vecteur de réponse pour chaque sujet visé. Chacun des sujets i peut se ranger par sa réponse dans trois catégories qui s'excluent les unes les autres, à savoir A_{obs} – observé, A_{mis} – manquant et A_{imm} – épargné, où :

$A_{\text{obs}} = \{i : i$ était non réfractaire et a répondu

$A_{\text{mis}} = \{i : i$ était non réfractaire, mais n'a pas répondu en 12 tentatives d'entrée en communication$\}$

$A_{\text{imm}} = \{i : i$ était réfractaire$\}$.

Les probabilités d'inclusion d'un sujet dans chacune de ces catégories peuvent ainsi se calculer :

$$P(i \in A_{\text{obs}}) = P(Z_i = 1,\ R_i = 1,\ C_i = c_i) \quad = \rho\,\pi(1-\pi)^{c_i - 1}$$

$$P(i \in A_{\text{mis}}) = P(Z_i = 1,\ R_i = 0,\ C_i > 12) \quad = \rho(1-\pi)^{12}$$

$$P(i \in A_{\text{imm}}) = P(Z_i = 0) \quad = 1 - \rho.$$

Selon ces données, $m = 1\,429$ sujets dans A_{obs} et $n - m = 969$ sujets non répondants dans $A_{\text{mis}} \cup A_{\text{imm}}$; $n = 2\,398$ est le nombre estimatif total de sujets sélectionnés. Ainsi, la densité conjointe de Z_{obs}, R et C_{obs} étant donné ρ et π est la suivante :

$$f(Z_{\text{obs}}, R, C_{\text{obs}} \mid \rho, \pi) \propto \left[\rho^m \pi^m (1-\pi)^{\left(\sum_{i=1}^m c_i\right) - m}\right] \times \left[(1-\rho) + \rho(1-\pi)^{12}\right]^{n-m}. \quad (5)$$

Le modèle mixte décrit par l'équation 5 peut être considéré comme un cas spécial des modèles de non-réponse examinés dans Drew et Fuller (1981).

Il serait bon de vérifier si cette distribution conjointe est une juste représentation des tendances de réponse des sujets « non réfractaires » dans l'ensemble de données. L'estimation EMV de ρ est simplement la proportion de répondants dans l'échantillon, et on se trouve à nettement sous-estimer ρ. Si on pose les distributions *a priori* $U(0,1)$ pour ρ et π à la fois et examine leur distribution conjointe postérieure par simulation MCCM, les médianes postérieures s'établissent à $\rho = 0,636$ et à $\pi = 0,205$, avec des intervalles postérieurs crédibles bilatéraux symétriques de (0,613, 0,659) et (0,191, 0,219) pour ρ et π respectivement. La figure 3 indique à quoi ressemblerait l'ensemble de données après imputation du nombre manquant d'appels pour les non-répondants non réfractaires selon les médianes postérieures. On se trouve à avoir éliminé en majeure partie le problème de la figure 2.

La distribution géométrique paraît suffisante (après prise en compte du caractère non réfractaire), mais un critique s'est interrogé sur l'utilisation de cette distribution sans prise en compte de covariables peut-être utiles. Comme nous l'avons expliqué, on n'a pas recueilli de données pour les covariables qui, selon nous, auraient été des plus intéressantes aux fins de cette analyse. Une autre possibilité de modélisation du mécanisme de réponse des sujets non réfractaires est l'emploi d'une distribution Gamma discrétisée. Si on a besoin de plus de complexité, on peut aussi songer à la distribution v – Poisson (une distribution Poisson biparamétrique qui généralise un certain nombre de distributions discrètes bien connues, dont la distribution géométrique) de Shmueli, Minka, Kadane, Borle et Boatwright (2001).

5.2 Rattachement de la non-réponse à la variable dépendante – modèle NI

Comme la distribution géométrique conditionnelle du nombre d'appels décrit la non-réponse des sujets « non réfractaires », on peut tenir compte de l'effet de cette non-réponse sur la variable dépendante en faisant du nombre d'appels une variable explicative supplémentaire aux fins de l'estimation de vraisemblance en régression logistique. On se trouve ainsi à ajouter deux paramètres au volet « régression logistique » du modèle. Ce sont les coefficients du nombre d'appels β_{call_j} de chacune des équations linéaires η_{ij} décrites à l'équation (2).

L'existence de coefficients non nuls du nombre d'appels indiquerait alors que la variable dépendante n'est pas indépendante du mécanisme de non-réponse et que, par conséquent, celui-ci n'est pas ignorable. Si les coefficients sont nuls, la non-réponse des « non réfractaires » est ignorable. Les conclusions tirées ici s'appuient sur l'hypothèse à la

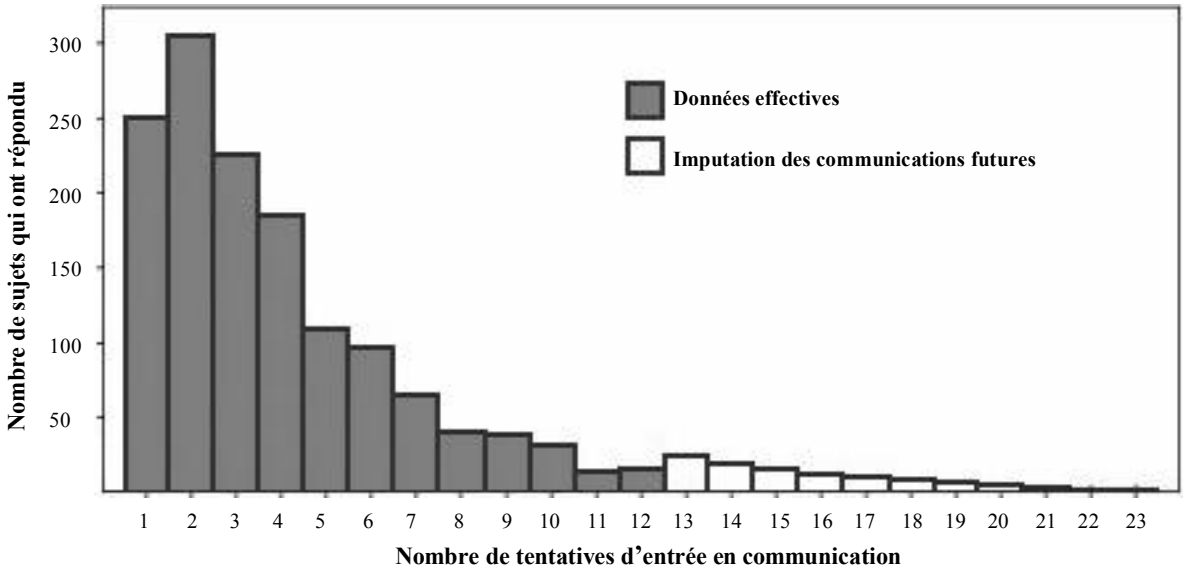

Figure 3 . On décrit le nombre effectif d'entrées en communication pour chacune des 12 premières tentatives, et le nombre prévu d'entrées en communication aux treizième et suivantes. Les valeurs d'imputation sont fondées sur des probabilités de tentative fructueuse de 0,205 et de « sujet non réfractaire » de 0,636.

base du modèle selon laquelle la relation entre le nombre d'appels, la variable dépendante et les autres variables explicatives considérées est la même pour les répondants et les non-répondants non réfractaires. Si on prend le nombre d'appels en compte au volet « régression logistique » du modèle, on se trouve à exclure les sujets épargnés, car il n'y aura jamais de contact avec eux.

La fonction de pseudovraisemblance totale du modèle IN (ou plus précisément du modèle IN des sujets non réfractaires) est le produit des morceaux « non-réponse » et « régression logistique » :

$$L(\rho,\ \pi,\ \beta) \propto \left[\rho^m \pi^m (1-\pi)^{\left(\sum_{i=1}^m c_i\right) - m} \right] \times \left[(1-\rho) + \rho(1-\pi)^{12} \right]^{n-m} \times \left[\prod_{i=1}^{m} \prod_{j=0}^{2} \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} \right]. \qquad (6)$$

À noter que nous incluons ici la variable W_i de pondération de ménage et de poststratification en vue de tenir compte de ce qu'une stratification appropriée des répondants pourrait rendre inutile l'introduction d'un mécanisme de description de la non-réponse.

5.3 Augmentation de données

Tanner et Wong (1987) proposent un calcul par itération des distributions postérieures dans le cas des données manquantes. Cette méthode s'applique chaque fois que, en augmentant l'ensemble de données, l'analyse devient plus simple et que les éléments d'augmentation sont facilement produits. Il faut compléter la notation. Soit S le nombre total de sujets non réfractaires dans l'échantillon. $S = \sum_{i=1}^{n} Z_i$, $S \sim$ Binomial (ρ). Soit X la matrice des variables explicatives (avec le nombre d'appels) pour l'ensemble des sujets sélectionnés aux fins de l'enquête. Soit $Y = (Y_1, \ldots, Y_n)$ le vecteur de leurs réponses. On divise X en $\{X_{\text{obs}}, X_{\text{mis}}, X_{\text{imm}}\}$ et Y en $\{Y_{\text{obs}}, Y_{\text{mis}}, Y_{\text{imm}}\}$. On sait, par la propriété sans mémoire de la distribution géométrique, quelle est la distribution du nombre supplémentaire d'appels dont on a besoin pour joindre les sujets dans A_{mis}. On peut ainsi l'exprimer : $\forall i \in A_{\text{mis}}$, soit $V_i = C_i - 12$, distribuée selon la loi géométrique de paramètre π.

Supposons maintenant que les valeurs vraies de S, X_{mis} et Y_{mis} sont connues. Le calcul de vraisemblance pourrait alors prendre la forme suivante :

$$L(\rho,\ \pi,\ \beta \mid X_{\text{obs}},\ X_{\text{mis}},\ Y_{\text{obs}},\ Y_{\text{mis}},\ S,\ R) \propto \left[(\rho\pi)^s (1-\pi)^{(\sum c_{\text{sus}}) - s} \right] \times \left[(1-\rho)^{n-s} \right] \times \left[\prod_{i=1}^{s} \prod_{j=0}^{2} \left(\frac{e^{\eta_{ij}}}{1 + e^{\eta_{i1}} + e^{\eta_{i2}}} \right)^{y_{ij} w_i} \right], \qquad (7)$$

où $\sum C_{\text{sus}} = \sum C_{\text{obs}} + \sum (V_i + 12)$ est le nombre d'appels dont on aurait eu besoin pour joindre tous les sujets non réfractaires et où les sommes se prennent sur l'ensemble approprié de sujets.

On ignore les valeurs vraies de S, X_{mis}, et Y_{mis}, mais on peut, par ce qu'on sait du comportement de ces variables, imputer des valeurs stochastiquement possibles dans l'algorithme MCCM. Étant donné ρ, on peut tirer une valeur de S d'une distribution binomiale tronquée

(2 398, ρ), où $1\,429 \leq S \leq 2\,398$. Étant donné S, le nombre de sujets dans A_{mis} est connu. Pour chacun de ces sujets, on peut tirer une valeur $V_i \sim \text{Geometric}\ (\pi)$, d'où une imputation pour le nombre d'appels dont on a besoin pour joindre chaque sujet non réfractaire et non joint. On peut alors exploiter les relations entre le nombre d'appels et les autres variables explicatives afin d'imputer des valeurs pour le reste de X_{mis}. Plus précisément, on procède à l'imputation des valeurs manquantes « Âge » et « Connaissance des risques » en effectuant respectivement des régressions du nombre d'appels sur ces variables et en dégageant une prévision des équations linéaires ainsi obtenues. De même, les valeurs manquantes de « État de fumeur » et de « État de réaction » s'imputent par régression logistique dans l'un et l'autre cas avec le nombre d'appels comme variable explicative. Ici, on vérifie les hypothèses du modèle à l'aide des données relatives aux répondants et pose l'hypothèse que les mêmes relations valent pour les non-répondants non réfractaires. Il convient de noter que ces équations de régression ordinaire et de régression logistique s'insèrent dans un contexte bayesien (Gelman, Carlin, Stern et Rubin 1998) et qu'il faut inclure d'autres paramètres, β_I, dans la le processus MCCM qui décrit ces relations (on trouvera plus de détails à l'annexe B). Si nous choisissons ce plan d'imputation, c'est par souci d'efficacité de tout l'algorithme MCCM. Comme solution de rechange, il y aurait l'imputation des valeurs manquantes d'une variable explicative en particulier en conditionnant par toutes les autres variables (Rubin 1996, par exemple). Enfin, Y_{mis} peut faire l'objet d'une prévision par les valeurs d'imputation de X_{mis} et la relation décrite dans le modèle de régression logistique. Par souci d'échangeabilité des non-répondants non réfractaires et faute de données de poststratification, nous appliquons une valeur de pondération de 1,0 à toutes les valeurs imputées Y_{mis}. Comme autre choix, nous pouvons – en dehors de l'âge – imputer le sexe et la taille du ménage pour les non-répondants non réfractaires et appliquer la méthode de pondération que décrit l'annexe A aux valeurs imputées Y_{mis}.

5.4 Échantillonnage de distribution postérieure

Tout l'exercice de simulation MCCM consiste en l'application d'un algorithme Metropolis avec enrichissement à chaque itération par la technique d'augmentation de données que nous venons de décrire. L'algorithme MCCM utilisé est exposé sommairement à l'annexe B. Nous évaluons la convergence par la méthode de Hiedelberger et Welch (1983) décrite dans Cowles et Carlin (1996). MacEachern et Berliner (1994) affirment que, dans des conditions non strictes, le sous-échantillonnage des valeurs de simulation MCCM en fonction de l'autocorrélation donnera des estimateurs moins bons. Voilà pourquoi nous avons utilisé toutes les valeurs de simulation dans l'analyse après la période nécessaire d'itérations.

6. Choix de distributions *a priori*

Dans l'évaluation de distributions *a priori* possibles pour les paramètres des modèles NI et MAR, nous avons tenu compte de l'objectif de comparaison des divers modèles. Le choix de distributions *a priori* pour les paramètres s'est opéré d'un point de vue MAR. Deux possibilités ont été étudiées.

La première s'articule autour de l'exploitation des données des enquêtes des phases I et II. Comme ces enquêtes ont évidemment précédé la phase III (d'où viennent nos données) où l'enquête a été identique, nous pouvons élaborer des distributions *a priori* à partir des données des deux premières phases. On y trouve la même variable dépendante, ainsi que les variables « État de fumeur », « Âge » et « Connaissance des risques ». C'est de ces données que nous avons tiré un modèle de régression logistique afin de décrire le rapport entre les opinions au sujet de l'usage du tabac en milieu de travail et ces trois variables explicatives. Nous avons établi des distributions antérieures normales pour les coefficients des trois à leurs valeurs centrales EMV, mais avec une erreur-type majorée. Nous avons accru les termes d'erreur pour trois raisons :

(i) trois ans s'étaient écoulés entre la phase II et la phase III et les opinions auraient pu changer dans ce laps de temps à cause de l'incidence du règlement municipal;

(ii) les valeurs EMV ont été calculées à l'aide de l'hypothèse MAR même qui était évaluée;

(iii) avant la collecte des données de phase III, il était possible que d'autres variables explicatives figurent dans le modèle et, du fait de leur présence, l'effet des trois variables considérées pourrait être autre.

Les variances ont augmenté, mais les moyennes sont restées les mêmes, car on ignorait au départ quel serait le sens de toute variation. Comme les données disponibles des phases I et II ne renseignaient pas sur le nombre d'appels ni sur l'« état de réaction », on a attribué aux coefficients de ces variables une distribution *a priori* normale diffuse (0, 9). Pour plus de clarté, nous appellerons cette première possibilité « distribution *a priori* des phases I et II » dans notre analyse.

La seconde possibilité consiste à attribuer une distribution *a priori* normale (0, 9) aux divers coefficients de régression logistique. Ce choix repose sur les trois raisons mêmes pour lesquelles nous avons accru les termes d'erreur plus haut, c'est-à-dire parce que les variables communes aux enquêtes des phases I et II et de la phase III ne sont pas échangeables. Une élaboration fondée sur les résultats des phases I et II serait peu appropriée. Nous appelons cette seconde possibilité « distribution *a priori* centrée ».

Si nous optons ici pour une distribution normale (0, 9), c'est par commodité. Si on centre la distribution *a priori* à zéro, on prête un même poids à l'un et l'autre sens de la

relation. À notre avis, une variance de neuf convient sans être trop diffuse. L'utilisation d'une distribution *a priori* impropre pourrait donner une simulation de Monte Carlo à chaîne de Markov sans convergence possible. De plus, comme l'indiquent Natarajan et Kass (2000), une distribution propre qui est excessivement diffuse peut se comporter comme une distribution impropre. À la section 7.2, nous évaluons par analyse de sensibilité comment le choix d'une distribution *a priori* influe sur les résultats.

Les paramètres de non-réponse du modèle NI, ρ et π, ont eu le même traitement dans l'une et l'autre de ces possibilités. Nous ne disposions pas d'indications supplémentaires sur les probabilités « sujet joint » ou « sujet non réfractaire ». Ainsi, ρ et π se sont chacun vu attribuer une distribution *a priori* $U(0, 1)$.

Les paramètres d'augmentation de données de chacune des équations de régression logistique β_l ont reçu indépendamment une distribution *a priori* diffuse $(0, 9)$. Pour chaque équation de régression linéaire dans la procédure d'augmentation de données, les coefficients β_r et la variance σ_r^2 ont été fixés à $p(\beta_r, \sigma_r^2) \propto 1 / \sigma_r^2$, c'est-à-dire à la distribution *a priori* non informative type (Gelman et coll. (1998), par exemple). À noter que les formes fermées des distributions postérieures des paramètres de régression linéaire sont connues et peuvent directement s'obtenir.

7. Résultats

Nous examinons d'abord le bien-fondé de l'hypothèse MAR par les coefficients de la variable du nombre d'appels. Nous évaluons ensuite le modèle NI dans sa sensibilité au choix d'une distribution *a priori*. Nous étudions enfin l'ordre de grandeur des effets d'une fausse hypothèse MAR pour cet ensemble de donnée en présentant les changements de probabilités de réponse.

7.1 Coefficients de la variable du nombre d'appels

Pour les distributions *a priori* tant diffuse que centrées, la figure 4 décrit la densité postérieure (ligne unie) et les intervalles crédibles estimatifs à 95 % (pointillés) du coefficient de la variable du nombre d'appels en η_{i1} dans le modèle NI et compare les valeurs au point $\beta_{\text{call}_1} = 0$ (tiretés). Les résultats indiquent clairement que ce coefficient est différent de zéro. Nous relevons aussi un résultat non nul en η_{i2} où, par la distribution *a priori* diffuse, l'intervalle crédible DPH à 95 % pour β_{call_2} est $(-0,03613,\ 0,11595)$.

Le coefficient non nul de C_i démontre une dépendance entre le nombre d'appels et l'opinion du sujet sur l'usage du tabac en milieu de travail. Ainsi, la variable dépendante et le mécanisme de non-réponse ne sont pas indépendants dans les conditions dont parle la section 5.2, d'où l'implication d'une fausseté pour cet ensemble de données de l'hypothèse du caractère aléatoire des observations manquantes avant prise en compte du mécanisme de non-réponse.

La figure 3 semble indiquer que les probabilités d'appel fructueux décroissent à mesure que s'élève le nombre d'appels. Pour vérifier l'hypothèse de l'existence d'une relation linéaire entre le nombre d'appels et les probabilités de réponse en expression logarithmique, nous avons élaboré un autre modèle bayesien NI qui scinde la variable du nombre d'appels en deux, $C_i I_{\{C_i<7\}}$ et $C_i I_{\{C_i\geq 7\}}$, selon que le nombre d'appels est de moins de sept ou non. Nous avons ensuite comparé les distributions postérieures des coefficients de ces deux variables sans découvrir la preuve qu'elles étaient essentiellement différentes. Précisons que, pour η_{i1}, l'intervalle crédible à 95 % pour $C_i I_{\{C_i\geq 7\}}$ contenait le même intervalle pour $C_i I_{\{C_i<7\}}$, et que, pour η_{i2}, les intervalles correspondants étaient largement en chevauchement.

7.2 Sensibilité aux distributions *a priori*

Des distributions *a priori* différentes pour le coefficient du nombre d'appels ou les autres influeraient-elles sur l'effet que nous avons indiqué? Le tableau 1 présente les intervalles crédibles DPH à 95 % pour le coefficient de la variable du nombre d'appels dans la première équation logistique du modèle NI, et ce, pour six distributions *a priori* différentes : distributions diffuses et centrées et quatre autres appelées options 3, 4, 5 et 6. Les options 3 et 4 ressemblent à la distribution centrées sauf que la distribution *a priori* change respectivement à normale $(1, 9)$ et à normale $(-1, 9)$ pour le coefficient du nombre d'appels. L'option 5 utilise des distributions *a priori* normales $(0, 9)$ pour β_{call_1}, β_{age_1}, et $\beta_{\text{b.USUL}_1}$, $(1, 9)$ pour β_{01}, à $(5, 9)$ pour $\beta_{\text{K-risk}_1}$, $(-1, 9)$ pour β_{S_1} et $(-5, 9)$ pour β_{SQ_1} et $\beta_{\text{b.NO}_1}$. Quant à l'option 6, elle prend la distribution *a priori* centrées et réduit toutes les variances de 9 à 2.

Le tableau 1 démontre que, dans les six options, le coefficient de la variable du nombre d'appels diffère nettement de zéro dans la première équation logistique. Le choix d'une distribution *a priori* parmi les six options ne semble pas influer sur la constatation que le mécanisme de non-réponse est non-ignorable pour cet ensemble de données.

Tableau 1
Intervalles crédibles DPH à 95 % pour β_{call_1} dans six distributions *a priori*

Distribution *a priori*	Coefficient du nombre d'appels « C_i » en η_{i1}	
	Intervalles à 95 %	
	Borne inférieure	Borne supérieure
Phase I & II	0,00129	0,07746
Centrée	0,00446	0,07980
Option 3	0,00447	0,07983
Option 4	0,00441	0,07975
Option 5	0,00440	0,07970
Option 6	0,00436	0,07944

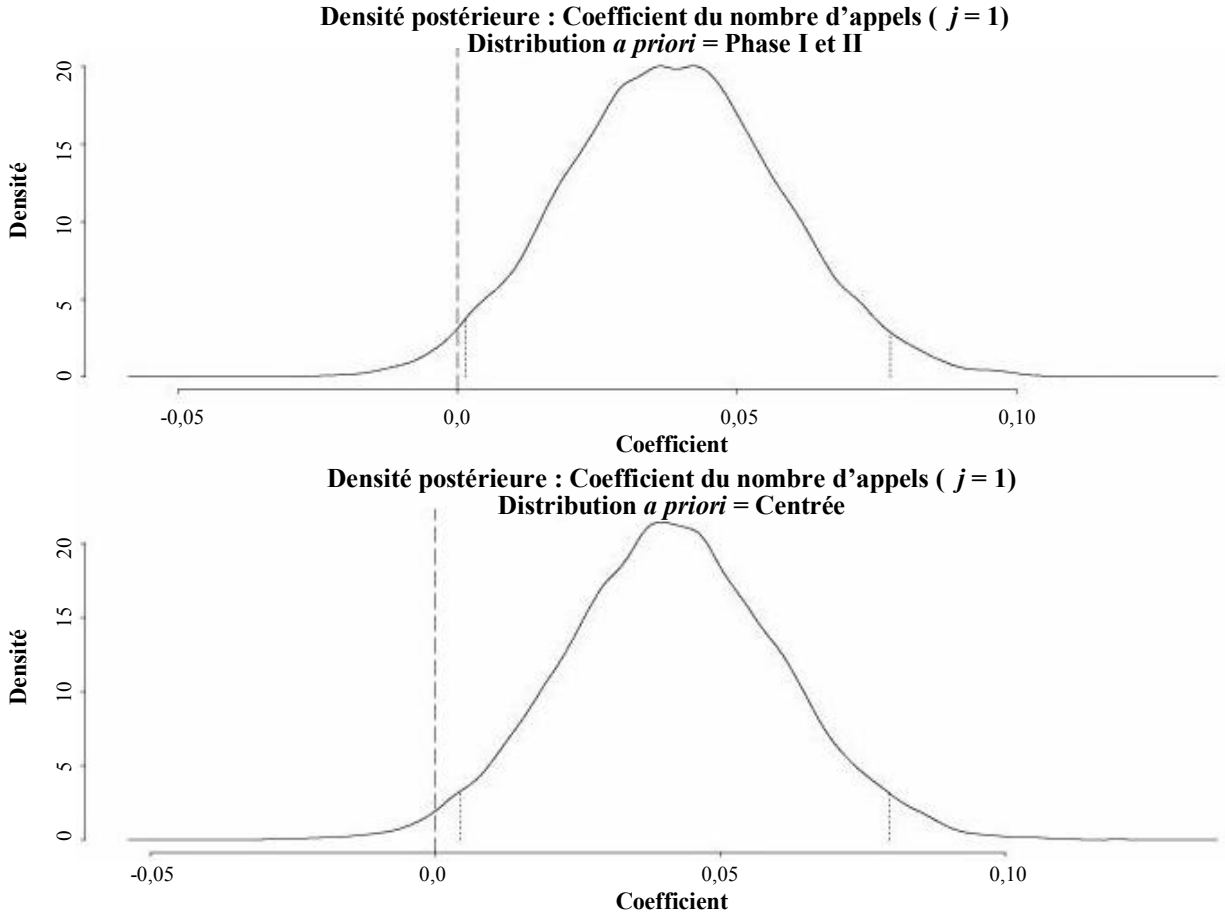

Figure 4. On décrit β_{call}, le coefficient de la variable du nombre d'appels en η_{i1} densité postérieure (ligne unie) et intervalle crédible bilatéral symétriques à 95 % (pointillé) comparativement à $\beta_{call_1} = 0$ (tireté).

7.3 Effet sur les probabilités de réponse

L'hypothèse MAR étant jugée lacunaire, il est bon de s'interroger sur l'utilité de l'erreur que créerait son application. Il est possible d'illustrer l'ordre de grandeur de l'erreur consécutive à l'adoption d'une fausse hypothèse MAR en s'attachant à son effet sur le rapport des probabilités $p_1(x_i)/p_0(x_i)$. Considérons d'abord l'effet sur un profil type de répondant. Le répondant modal est un non-fumeur de 25 à 35 ans que dérange d'habitude la fumée secondaire, qui présente une valeur « Connaissance des risques » de 11 et qui est joignable en deux appels. Nous faisons de ce répondant modal notre sujet 1. Le tableau 2 montre le changement de probabilités postérieures de ce sujet s'il est appelé 13 fois.

La colonne « Sujet 1 » du tableau 2 fait voir une différence considérable de probabilités postérieures si on tient compte du mécanisme de non-réponse. Pour ce profil type de répondant, si le nombre d'appels monte de 2 à 13, les probabilités postérieures de choix de « Interdiction totale de fumer » plutôt que de « Permission de fumer seulement dans des zones réservées » s'accroissent de 52,18 % dans une distribution diffuse et de 57,84 % dans une distribution centrée. C'est la preuve éloquente de l'existence d'un lien entre la variable dépendante et le mécanisme de non-réponse.

S'agit-il de résultats propres à notre sujet modal? Le tableau 2 décrit aussi les effets sur les probabilités de réponse dans le modèle NI pour trois autres profils de sujet en fonction de nos six distributions *a priori*. Le sujet 2 est un non-fumeur de 50 ans que dérange toujours la fumée et qui a une note parfaite « Connaissance des risques ». Le sujet 3 est un ex-fumeur de 27 ans que la fumée ne dérange pas et qui obtient une note de 7 à « Connaissance des risques ». Le sujet 4 est un fumeur de 40 ans que ne dérange pas la fumée et qui reçoit, lui, une note de 3 pour cette même connaissance des risques. Pour des sujets et des distributions multiples, ce tableau dégage invariablement le même résultat. Si on porte le nombre d'appels à plus de 12, on se trouve à augmenter les probabilités postérieures de choix de la catégorie 1 par rapport à la catégorie 0. Pour les divers sujets et distributions du tableau 2, le taux d'accroissement se situe entre 52,18 % et 58,41 %.

On obtient des résultats semblables lorsqu'on examine les probabilités de choix de la catégorie « Permission totale de fumer » par rapport à la catégorie « Permission de fumer seulement dans des zones réservées ». Avec des sujets qui sont un fumeur et un ex-fumeur (sujets 3 et 4), les probabilités postérieures s'élèvent de 46,7 % si on porte le nombre d'appels de 2 à 13 dans une distribution diffuse.

Tableau 2
Comparaison des probabilités de réponse pour quatre sujets types. Nous avons utilisé les médianes postérieures comme estimations ponctuelles des coefficients des modèles bayesiens; nous avons employé la valeur EMV pour le modèle fréquentiste

	Sujet 1	Sujet 2	Sujet 3	Sujet 4
État de fumeur	Non	Non	Ancien Fumeur	Oui
Âge	30	50	27	40
État de réaction	D'habitude	Toujours	Non	Non
Connaissance des risques	11	12	7	3
Probabilités du modèle		Odds Y = 1/Y = 0		
MAR EMV	0,674	2,105	0,457	0,396
MAR distribution antérieure en valeurs diffuses	0,703	4,487	0,209	0,116
IN distribution antérieure en valeurs diffuses : 2 appels	0,640	4,024	0,202	0,108
IN distribution antérieure en valeurs centrales : 2 appels	0,593	4,442	0,162	0,102
Option 3 : 2 appels	0,594	4,449	0,162	0,102
Option 4 : 2 appels	0,592	4,435	0,162	0,101
Option 5 : 2 appels	0,590	4,423	0,161	0,101
Option 6 : 2 appels	0,590	4,426	0,161	0,101
IN distribution antérieure en valeurs diffuses : 13 appels	0,974	6,128	0,308	0,165
IN distribution antérieure en valeurs centrales : 13 appels	0,936	7,013	0,256	0,160
Option 3 : 13 appels	0,937	7,026	0,256	0,161
Option 4 : 13 appels	0,934	7,000	0,255	0,160
Option 5 : 13 appels	0,930	6,975	0,254	0,159
Option 6 : 13 appels	0,931	6,980	0,254	0,160

7.4 Effet sur les probabilités de réponse

La variation des probabilités postérieures que nous venons de décrire s'accompagne d'une variation des probabilités de réponse estimées d'un sujet dans une catégorie. Parmi les répondants, 57,45 % ont choisi la catégorie 0, 40,64 %, la catégorie 1, et 1,91 %, la catégorie 2. Pour le nombre de non-répondants non réfractaires, on relève une médiane postérieure de 469 et un intervalle crédible à 95 % de (25, 944). En moyenne, 55,88 % des non-répondants non réfractaires simulés ont choisi la catégorie 0, 40,03 %, la catégorie 1, et 4,08 %, la catégorie 2. Bien que, pour les catégories 0 et 1, les valeurs moyennes des non-répondants non réfractaires tombent bel et bien dans les intervalles de confiance à 95 % pour les proportions de répondants dans ces catégories, les estimations ponctuelles varient pour chaque catégorie en cas d'inclusion du mécanisme de non-réponse dans le modèle. Dans une comparaison des résultats de la catégorie 2, nous estimons que les non-répondants ont deux fois plus de chances que les répondants de choisir « Permission totale de fumer ». Le petit nombre de sujets dans la catégorie 2 est peu de nature à amener un changement de règlement sur l'usage du tabac en milieu de travail, mais l'augmentation relevée du nombre de nonrépondants dans cette catégorie illustre comment on peut tirer de fausses conclusions au sujet des données si on ne tient pas bien compte des non-répondants.

8. Conclusion

La section 7 démontre que, pour la variable dépendante d'intérêt dans cet ensemble de données, l'affirmation que les observations manquantes sont aléatoires avant prise en compte du mécanisme de non-réponse se révèle erronée. Ceci suppose que la relation entre les variables pertinentes est la même pour tous les sujets non réfractaires. Ajoutons que l'application d'une fausse hypothèse MAR dans l'évaluation de cette variable dépendante risque d'entacher d'une grave erreur le calcul des probabilités postérieures et toute conclusion à en tirer. Pour bien évaluer les opinions sur l'usage du tabac en milieu de travail à Toronto au début de 1993 par la variable dépendante d'intérêt dans le cadre de cette enquête, il est nécessaire d'intégrer le mécanisme de non-réponse à la structure du modèle.

Dans notre analyse, nous avons utilisé un seul élément d'information, le nombre d'appels. Le traitement aurait pu être plus complet si nous avions disposé de plus de renseignements. Si nous avions connu le nombre exact de tentatives d'entrée en communication avec les non-répondants – au lieu d'un minimum d'appels –, ainsi que l'heure des appels, l'analyse aurait gagné en précision. Qui plus est, si nous avions connu la nature de la non-réponse par refus ou indisponibilité et le nombre effectif de tentatives d'entrée en communication avec les non-répondants, il aurait été possible de mieux caractériser ces derniers. Groves et Couper (1998) signalent que les erreurs statistiques différeront probablement selon qu'il s'agit d'une non-réponse par indisponibilité ou par refus. Comme ils le précisent, un important point en recherche est l'évaluation de l'incidence pour contacter les enquêtés sur l'erreur de mesure.

Les résultats que nous avons présentés ne valent que pour cette variable dépendante évaluant l'usage du tabac en milieu de travail dans ce seul ensemble de données. Comme on peut percevoir que le tabagisme est devenu moins socialement acceptable ces dernières années, on serait fondé à penser que l'erreur de non-réponse due aux questions sur l'usage du tabac pourrait être plus sérieuse que pour d'autres questions. On peut trouver dans Biemer (2001) une comparaison de biais de non-réponse pour diverses questions sur le tabagisme et d'autres sujets. La comparaison n'accrédite pas l'idée que l'erreur de non-réponse est propre aux questions portant sur l'usage du tabac.

De nos résultats, il n'y a rien à tirer comme implications au sujet des mécanismes de non-réponse d'autres enquêtes, mais on peut clairement voir ici que, si on suppose – à tort – que les répondants d'une enquête constituent un sous-échantillon aléatoire d'une population pour les variables

d'intérêt, cette conjecture est peut-être peu judicieuse. Les indications dont on dispose au moment de la collecte des données peuvent permettre d'évaluer si le mécanisme qui cause la non-réponse est ignorable ou non. On peut donc en conclure qu'il serait bon que les gens qui travaillent avec de telles données utilisent les indications disponibles sur la non-réponse dans leur appréciation de cette information et qu'ils communiquent les indications en question aux autres utilisateurs de l'ensemble de données. En règle générale, la collecte et l'analyse de données qui nous disent où et comment on a trouvé les répondants et combien il a été difficile de les joindre sont pour nous une importante voie qui s'ouvre à la méthdologie d'enquête et à la pratique.

Remerciements

Cette étude a été financée par la subvention DMS-9801401 de la National Science Foundation. Les auteurs remercient Shelley Bull de toutes ses observations et ses suggestions utiles et de son aide à l'acquisition des données, tout comme John Eltinge, les critiques anonymes et le rédacteur associé de publication de leurs précieux commentaires.

C'est l'Institute for Social Research de l'Université York qui a fourni les données de l'enquête sur les attitudes à l'égard du règlement sur l'usage du tabac, laquelle a été financée par Santé et Bien-être social Canada. Les données ont été réunies par l'ISR pour le D^{r} Linda Pederson, de l'Université Western Ontario, et les D^{rs} Shelley Bull et Mary Jane Ashley, de l'Université de Toronto. Les responsables de l'enquête, le ministère ontarien de la Santé et l'Institute for Social Research n'assument aucune responsabilité à l'égard des éléments d'analyse et d'interprétation du présent document.

A. Poststratification

HHW_i est le poids de ménage du sujet i comme il est décrit dans Northrup (1993).

- Soit m le nombre de répondants.
- Soit r le nombre cumulatif d'adultes des ménages répondants.
- Soit h_i le nombre d'adultes du ménage du sujet i.
- $HHW_i = h_i \cdot m / r$.

Les proportions de sujets échantillonnés qui appartiennent aux tranches d'âge suivantes ont été calculées pour les répondants des deux sexes : 18 – 24, 25 – 44, 45 – 64 et 65 ans et plus. On a ensuite comparé les pourcentages à la structure par âge-sexe de la population de la région métropolitaine de Toronto.

- Soit p_{1i} la proportion des résidents d'âge adulte de la région métropolitaine de Toronto qui appartiennent à la catégorie d'âge-sexe du sujet i selon les données du recensement de 1991.
- Soit p_{2i} la proportion des répondants appartenant aux catégories d'âge et de sexe du sujet i.
- $W_i = HHW_i \cdot p_{1i} / p_{2i}$, où W_i est le poids final de poststratification utilisé dans l'analyse.

B. Simulation MCCM

La simulation complète MCCM pour le modèle IN consiste en l'application d'un algorithme Metropolis avec pour complément des éléments d'augmentation de données décrits à la section 5.3. Voici un aperçu de cet algorithme. Les variables employées sont définies à la section 5. À chaque itération t,

1. On tire ρ_t pour $Beta(s_{t-1} + 1,2398 - s_{t-1} + 1)$.
2. On impute s_t à partir de $Binomial(\rho_t) \geq 1,429$.
3. On impute C_{mis_t}: on tire les $(s_t - 1,429)\, v_i$ de $Geometric(\pi_{t-1})$ et de $\forall c_i \in c_{\text{mis}},\ c_i = v_i + 12$.
4. On tire π_t de $Beta(s_t + 1,\ \sum c_{\text{sus}_t} - s_t + 1)$.
5. On impute les valeurs du reste de X_{mis} en utilisant les relations avec le nombre d'appels comme le décrit la section 5.3.
6. On met à jour les paramètres supplémentaires utilisés dans l'augmentation de données pour X_{mis}.
   - On met à jour les paramètres de régression linéaire beta sub ret β_r et σ_r en se reportant directement à la forme fermée de leurs distributions postérieures.
   - On met à jour les paramètres de régression logistique β_l par application par pas de l'algorithme Metropolis à chacun.
7. On impute Y_{mis_t}: $\forall y_i \in y_{\text{mis}}$; on tire y_i de $Multinomial(p_0(x_i),\ p_1(x_i),\ p_2(x_i))$.
8. On met à jour chaque β_{kj} par application d'une itération de l'algorithme Metropolis sur la vraisemblance conditionnelle et par application d'une fonction de saut normale.

Bibliographie

Biemer, P.P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics*, 17, 2, 295-320.

Bull, S. (1994). *Case Studies in Biometry*. Analysis of Attitudes toward Workplace Smoking Restrictions, chapître 16, New York : John Wiley & Sons, Inc., 249-270.

Cowles, M.K., et Carlin, B.P. (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91, 883-904.

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

Drew, J.H., et Fuller, W.A. (1981). Nonresponse in complex multiphase surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, VA, 623-628.

Eltinge, J.L., et Yansaneh, I.S. (1997). Méthodes diagnostiques pour la construction de cellules de correction pour la non-réponse, avec applications à la non-réponse aux questions sur le revenu de la U.S. Consumer Expenditure Survey. *Techniques d'enquête*, 23, 37-45.

Gelman, A., Carlin, J.B., Stern, H.S. et Rubin, D.B. (1998). *Bayesian Data Analysis*. Chapître 14, Generalized Linear Models. London : Chapman & Hall.

Groves, R.M., et Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York : John Wiley & Sons, Inc.

Hiedelberger, P., et Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31, 1109-1144.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de la Statistique*, 54, 139-157.

Little, R.J.A., et Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York : John Wiley & Sons, Inc.

MacEachern, S.N., et Berliner, L.M. (1994). Subsampling the Gibbs Sampler. *The American Statistician*, 48, 188-189.

Maller, R., et Zhou, X. (1996). *Survival Analysis with Long Term Survivors*. Chichester, New York : John Wiley & Sons, Inc.

Natarajan, R., et Kass, R.E. (2000). Reference Bayesian methods for generalized linear mixed models. *Journal of the American Statistical Association*, 95, 227-237.

Northrup, D.A. (1993). Attitudes towards workplace smoking legislation: A survey of residents of metropolitan Toronto, Phase III, 1992/93 Documentation techniques. Tech. Rep. Institute for Social Research, York University, non-publiée.

Pederson, L.L., Bull, S.B. et Ashley, M.J. (1996). Smoking in the workplace: Do smoking patterns and attitudes reflect the legislative environment? *Tobacco Control*, 5, 39-45.

Pederson, L.L., Bull, S.B., Ashley, M.J. et Lefcoe, N.M. (1989). A population survey on legislative measures to restrict smoking in Ontario: 3. Variables related to attitudes of smokers and nonsmokers. *American Journal of Preventive Medicine*, 5, 313-322.

Pottoff, R.F., Manton, K.G. et Woodbury, M.A. (1993). Correcting for nonavilability bias in surveys by weighting based on number of callbacks. *Journal of the American Statistical Association*, 88, 1197-1207.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

Shmueli, G., Minka, T.P., Kadane, J.B., Borle, S. et Boatwright, P. (2001). Using computational and mathematical methods to explore a new distribution: The v – Poisson. Rapport Technique 740, Department of Statistics Carnegie Mellon University.

Tanner, M.A. et Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528-549.