IMPACT DU MODE D'ENQUÊTE SUR LES COMPORTEMENTS DE MOBILITÉ

Caroline BAYART, Patrick BONNEL

Laboratoire d'Economie des Transports
ENTPE, Université Lumière Lyon 2, CNRS

Résumé

Face à la difficulté de recueillir des données de mobilité précises et représentatives de la population à un coût acceptable, les commanditaires d'étude se tournent vers des protocoles d'enquête plus complexes, associant plusieurs modes. Si le fait de proposer des médias différents permet d'augmenter le taux de réponse global, la comparabilité des données reste un exercice difficile. Les caractéristiques socioéconomiques des répondants varient selon le mode d'enquête, et ces différences peuvent être à l'origine de comportements spécifiques. Nous montrons dans cet article qu'il est possible de distinguer l'effet dû au mode d'enquête de celui lié aux différences socioéconomiques observées entre les échantillons de répondants, et de quantifier l'impact du mode d'enquête sur la mobilité déclarée. Le modèle économétrique envisagé pour cette analyse est emprunté au domaine des variables qualitatives. Plus précisément, il s'agit d'un modèle de sélection de l'échantillon, dont nous estimons les paramètres à l'aide de la procédure en deux étapes, élaborée par James Heckman et d'autres à la fin des années 1970. Les résultats de l'enquête ménages déplacements menée à Lyon en 2006 en face-à-face et par internet illustrent nos propos. L'objectif est de quantifier, pour chaque répondant, l'impact du mode d'enquête sur le comportement de mobilité.

Introduction

La difficulté croissante pour obtenir des données d'enquêtes représentatives de la population visée et la complexité des informations nécessaires à l'alimentation de modèles de plus en plus sophistiqués ne permettent généralement plus de recueillir toutes les données au cours d'une même enquête ou selon une méthodologie unique. De plus en plus d'enquêtes reposent sur des protocoles complexes, qui associent plusieurs modes ou méthodologies. L'objectif est d'améliorer la qualité des données produites en augmentant le taux de couverture de la population cible et le taux de réponse global (Couper, 2000 ; Gunn, 2002 ; Dillman, 2001). Mais proposer plusieurs modes ou méthodes de recueil de données n'est pas sans risque, le recueil d'informations via différentes sources pouvant générer des résultats parfois peu comparables. Le danger lorsque l'on fusionne des bases de données est de générer un biais de sélection des individus qui compromet la pertinence des modèles explicatifs des comportements de mobilité. Ce biais de sélection est l'objet d'une importante littérature, théorique et empirique (Winship & Mare, 1992), mais les applications en sciences sociales et plus spécifiquement aux enquêtes transport sont rares à ce jour.
Des travaux montrent, depuis les années 50, que l'estimation d'une équation sur un sous-échantillon obtenu de façon sélective dans la population peut conduire à des biais (Roy, 1951). Les premiers développements économétriques des conséquences de cette sélection des individus datent de 1974, avec les travaux d'Heckman (1979). L'exemple souvent cité dans la littérature est celui d'une équation de salaire estimée sur les seules femmes actives, alors même que le comportement d'activité relève d'un arbitrage dans lequel le salaire que la personne peut obtenir sur le marché intervient. Depuis, de nombreux articles ont mis en évidence l'importance du biais de sélection dans les enquêtes réalisées en sciences humaines et sociales (Maddala, 1986). On notera par exemple le modèle de migration aux USA analysé par Nakosteen et Zimmer (1980), ou celui du taux d'activité féminin de Mroz (1987). L'utilisation la plus fréquente des modèles d'auto sélection concerne l'évaluation d'un traitement ou d'une formation.

Le laboratoire d'Economie des Transports a proposé de réaliser une enquête par le web auprès des non répondants à l'enquête ménages déplacements réalisée à Lyon en face-à-face en 2006 (Bayart, Bonnel, 2008), c'est-à-dire aux individus qui ont refusé de recevoir un enquêteur à leur domicile ou qu'il n'a pas été possible de joindre durant la première vague d'interviews. Les données de cette enquête nous permettent de mettre en évidence le problème d'autosélection, les non-répondants à l'enquête standard en face-à-face choisissant de remplir ou pas le questionnaire sur le web. L'objet de cet article est de tester l'incidence du mode d'enquête sur le comportement de mobilité des répondants, et de s'efforcer ensuite de le mesurer. Nous revenons d'abord sur la méthodologie retenue pour l'enquête ménages déplacements de Lyon, avant de présenter les premiers résultats comparés de la mobilité des individus selon le mode d'enquête (section 1). Suivent les développements théoriques relatifs au biais de sélection (section 2) et une présentation des variables disponibles pour l'analyse de la mobilité. Le modèle économétrique en deux étapes, permettant de s'affranchir de l'effet de sélection des individus, est appliqué aux données de l'enquête ménages recueillies en 2006 (section 4). L'impact des interactions entre les variables explicatives de la mobilité et le mode d'enquête est quantifié en section 5. Enfin, nous avançons quelques pistes de réflexions pour de futures recherches (section 6).

1. Pour poser le problème

1.1. Méthodologie

La méthodologie des enquêtes ménages déplacements françaises est définie par le CERTU[1] (Certu, 2008). A Lyon, l'enquête est traditionnellement menée en face-à-face. En 2006, une innovation méthodologique a permis de solliciter par courrier les ménages refusant de recevoir un enquêteur à domicile ou ne pouvant être joints, après huit tentatives à des jours et horaires différents, pour remplir le questionnaire en-ligne (figure 1). Quel que soit le mode d'enquête (web ou face-à-face), le questionnaire est structuré en trois parties. Nous distinguons d'abord les questions relatives au ménage, puis celles concernant la personne interrogée, avant de terminer par des questions concernant l'ensemble des déplacements de la veille du jour où l'enquête est réalisée.

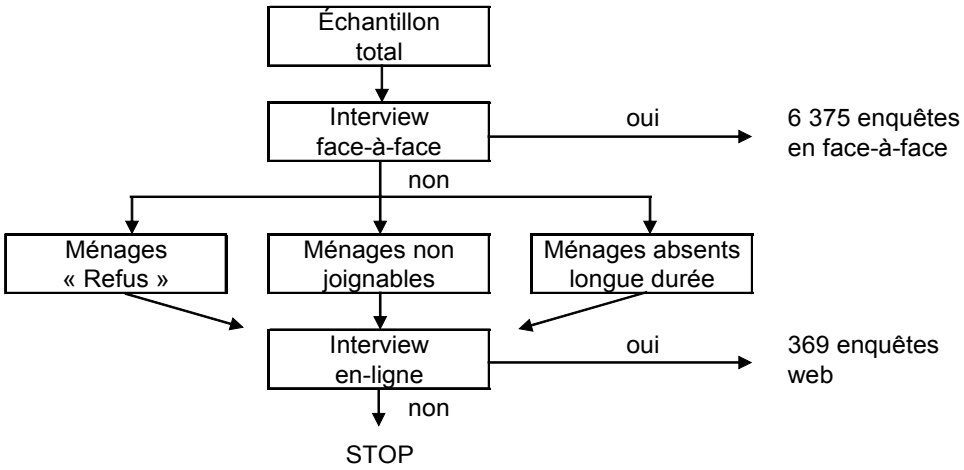

Figure 1 : Schéma de recrutement des ménages

Au cours de la période d'enquête (de novembre 2005 à avril 2006), 11 951 ménages ont été contactés mais seuls 6 375 ont accepté de recevoir un enquêteur à leur domicile sur le territoire du Schéma de COhérence Territorial (S.C.O.T.) de Lyon, déclarant 48 143 déplacements réalisés la veille du jour de l'enquête. Le taux de réponse global de l'enquête en face-à-face est donc égal à 53% sur ce périmètre. Les 4 335 ménages ayant refusé de répondre ou n'ayant pas pu être joint ont été contacté par courrier postal pour les inviter à participer à l'enquête web. Les envois ont été réalisés en deux vagues successives, chacune avec deux relances. Au final, 536 individus ont accédé au site pour tenter de répondre à l'enquête (678 connexions enregistrées), ce qui représente un taux de connexion

[1] Centre d'Etudes sur les Réseaux, les Transports et l'Urbanisme.

de 12,4%. Cependant, tous les ménages qui se sont connectés n'ont pas terminé la saisie (contraintes techniques, difficulté du questionnaire…). Seuls 369 individus, déclarant un total de 1 108 déplacements, ont fourni une réponse suffisamment complète et exploitable, générant un taux de réponse à l'enquête web égal à 8,5%.
En comparant les résultats des deux enquêtes, nous observons que les internautes mobiles de 18 ans et plus se déplacent moins que les individus répondant en face à face (3,78 vs. 4,19 déplacements). Cette moindre mobilité concerne tout à la fois les déplacements et les sorties (enchaînement des déplacements entre une sortie et un retour successif au domicile). Une analyse par mode ou motif montre que le différentiel s'explique principalement par une mobilité marche à pied deux fois plus faible et par une participation aux activités d'accompagnements ou de loisirs nettement plus faible. Ces données sont cohérentes avec l'hypothèse d'une sous-déclaration imputable au média web, dans la mesure où l'on sait que les risques d'omission de déclaration de déplacements concernent surtout les déplacements courts en temps ou en distance et des motifs moins contraints (Bonnel et Le Nir, 1998). On peut toutefois objecter que les différences socio-économiques peuvent expliquer au moins en partie les différences de mobilité, notamment de marche à pied. Dans l'enquête, les internautes ont un niveau d'études et un revenu nettement supérieur, conduisant à une très forte motorisation. Ils sont plus souvent cadres et employés et travaillent plus souvent dans le centre, ce qui conduit à des durées hors domicile plus importantes, réduisant les possibilités de participation à des activités moins contraintes et l'usage de la marche. Toutefois, les différences observées, en nombre de déplacements et de sorties, subsistent même lorsque l'on redresse l'échantillon en face-à-face pour le rendre comparable à l'échantillon web, au regard des principales caractéristiques socioéconomiques des répondants (Bayart et Bonnel, 2009).

1.2. Hypothèse de travail

En pratique, le biais de sélection a deux origines. Il s'agit soit d'une auto sélection des personnes interrogées (exclusion de répondants due à la méthode de collecte de données), soit d'une décision de sélection prise par les gestionnaires de l'étude (les questions filtres excluent par définition les répondants pour lesquels la question ne s'applique pas). Dans le cas de l'enquête ménages déplacements de Lyon, les individus choisissent de répondre ou non en face-à-face, et dans la négative de remplir ou pas le questionnaire en-ligne. Les groupes de répondants peuvent différer sur des aspects systématiques et on sait que lorsque des observations sont exclues d'un échantillon de manière non aléatoire, il y a un risque de biais de sélection. Les réponses ne sont pas comparables, puisque la présence des répondants dans un groupe est déterminée par des facteurs extérieurs, qui peuvent également impacter la variable d'intérêt du modèle étudié. Dit autrement, il est fort probable que des caractéristiques socioéconomiques, pas toujours observables, influencent le choix des individus de recevoir un enquêteur à domicile ou de répondre, le cas échéant, sur le web et impactent leurs comportements de mobilité (Berk, 1983 ; Resource system group, 2002). Si le mode retenu pour remplir le questionnaire est lié au phénomène que l'on veut étudier (la mobilité), on dira que la sélection est endogène, par opposition au cas où le mode de réponse est sans rapport avec le nombre de déplacements (mais peut être en rapport avec les variables explicatives) et donc exogène.
Ignorer l'existence du biais de sélection, en particulier pour les répondants web, peut avoir des conséquences sur la validité du modèle. Il peut exister des raisons pour que les individus qui répondent sur le web ne présentent pas un niveau de mobilité équivalent à ceux qui ont répondu en face-à-face et que ces raisons ne soient pas liées au mode de réponse. Ces individus pourraient donc déclarer un nombre de déplacements plus faible, même s'ils ne répondaient pas sur le web, mais en face-à-face. La simple comparaison de la mobilité des répondants web et face-à-face, sans correction du biais de sélection, surestime l'effet réel du média. Evidemment, il est également probable que le mode de recueil des données ait un effet sur le nombre de déplacements saisis. Par exemple, le manque d'ergonomie du questionnaire web, ajouté à son caractère auto-administré peut rendre pénible la saisie des déplacements. Dans ce cas, la différence de mobilité totale observée entre les répondants web et face-à-face se scinde en deux parties : l'effet de sélection et l'effet du mode d'enquête.
Si l'existence du biais de sélection résulte de l'omission d'une variable, il est difficile d'anticiper de quelle manière le modèle va surestimer ou au contraire sous-estimer la relation causale réelle entre les variables explicatives de la mobilité et nombre de déplacements. La direction et la taille du biais dépendent dans ce cas du nombre d'observations exclues (les non internautes par exemple) et de leurs caractéristiques. La situation est particulièrement compliquée dans les modèles multivariés.

2. Le biais de sélection de l'échantillon

Dillman (1978) a mis en évidence le biais de sélection qui résulte de protocoles d'enquêtes où plusieurs modes de recueil de données sont disponibles. Les sciences sociales contiennent par ailleurs plusieurs présentations formelles du biais de sélection (Heckman, 1979 ; Goldberger, 1981). Le recours aux développements économétriques semble intéressant ici pour isoler l'effet des différences sociodémographique de celui du mode d'enquête sur la mobilité quotidienne d'une part, pour quantifier cet effet du mode d'enquête sur le comportement de mobilité des répondants d'autre part.

2.1. Origine du biais de sélection

Considérons l'équation (1) qui permet d'examiner l'effet du mode d'enquête sur le nombre de déplacements quotidiens moyen d'un individu à l'aide d'une régression linéaire :

$$Y_i = \beta X_i + \alpha I_i + u_i, \quad (1)$$

Avec Y_i le nombre moyen de déplacements réalisé par les individus (variable dépendante), X_i un vecteur de variables explicatives et I_i une variable muette indiquant si l'individu a répondu sur internet. On peut se demander si le coefficient α_i mesure l'impact réel du mode d'enquête sur la mobilité quotidienne. La réponse est positive, si l'individu qui choisit de répondre sur le web déclare un nombre de déplacements identiques à celui qu'il aurait déclaré en face-à-face. Cependant, la variable I ne peut pas être considérée comme exogène dans ce modèle. Les individus contactés choisissent de répondre ou pas en face-à-face, et éventuellement acceptent de saisir leur réponses en-ligne, selon certaines caractéristiques qui peuvent également avoir un impact sur leur niveau de mobilité.
Ainsi, les ménages qui répondent en-ligne sont ceux pour lesquels il n'a pas été possible de fixer un rendez-vous avec un enquêteur à domicile. On peut supposer qu'il s'agit de ménages peu disponibles (contraints par leur activité professionnelle par exemple), dont le manque de disponibilité impacte négativement la mobilité, et qui disposent d'un accès à internet pour accéder à l'enquête en-ligne. L'échantillon web n'est donc pas représentatif des habitants de l'agglomération, et cette auto sélection des répondants doit être corrigée à afin d'obtenir des estimations non biaisées des coefficients des variables explicatives du nombre de déplacements quotidiens. En excluant systématiquement des observations de la même manière (individus n'ayant pas accès au web par exemple), on introduit le besoin d'un régresseur additionnel que la méthode des moindres carrées ordinaires ignore (Kmenta, 1971). Le problème de l'évaluation de l'effet du mode d'enquête sur la mobilité provient du fait qu'il n'est pas possible d'observer simultanément la mobilité déclarée sur le web et en face-à-face pour un répondant donné.

2.2. Développements théoriques

Le model traditionnellement utilisé pour mettre en évidence le problème d'auto sélection des répondants est le modèle Tobit II[2], appelé plus couramment modèle de sélection de l'échantillon. En utilisant la méthode d'estimation en deux étapes développée par Heckman en 1979, le modèle explicatif de la mobilité peut se formaliser pour chaque individu i.

Pour les répondants web :

$$Y_{1i} = \beta_1 X_{1i} + u_{1i}, \quad (2)$$

Pour les répondants en face-à-face :

$$Y_{2i} = \beta_2 X_{2i} + u_{2i}, \quad (3)$$

Avec Y_{1i} et Y_{2i} le nombre moyen de déplacements réalisé par les individus, X_{1i} et X_{2i} deux vecteurs de variables indépendantes ou explicatives de la mobilité et u_{1i} et u_{2i} deux termes d'erreur supposés normaux, qui tiennent compte des forces non observées qui pourraient influer sur la mesure des

[2] Cette classification est due à Amemiya (1984)

résultats. En réalité, nous estimerons deux modèles, un sur le sous échantillon des répondants en face-à-face, et un sur le sous échantillon des répondants sur le web.

Soit la fonction de sélection[3] traduisant la probabilité pour un individu i de répondre sur le web :

$$I^*_i = \delta Z_i + \varepsilon_i, \qquad (4)$$

$I_i = 1$, ssi $I^*_i > 0$,

$I_i = 0$, ssi $I^*_i \leq 0$

Avec I^*_i la variable de sélection, Z_i un ensemble de variables déterminantes du choix du web et ε_i les termes d'erreur supposés normaux. La mobilité observée pour tout individu i se définit comme :

$Y_i = Y_{1i}$, ssi $I_i = 1$

$Y_i = Y_{2i}$, ssi $I_i = 0$

Notons que seulement un des paramètres Y_{1i} et Y_{2i} est observé, selon que l'individu réponde en face-à-face ou sur le web. Contrairement à l'équation (1), le modèle ainsi défini n'impose pas que les coefficients des variables explicatives X_{1i} et Z_i soient identiques pour les répondants web et les répondants en face-à-face. Par ailleurs, nous supposons que les termes d'erreur des équations (2), (3) et (4) (u_1, u_2, et ε) suivent une loi normale bivariée de moyennes nulles et de corrélations ρ_1 et ρ_2. En effet, la mobilité n'est observable que si les individus répondent en face-à-face ou sur le web. Des perturbations aléatoires vont affecter simultanément les variables endogènes des équations de sélection et d'intérêt et les termes d'erreur des deux équations peuvent être corrélés :

$(\varepsilon i, u_1) \sim N(0, 0, \sigma_\varepsilon, \sigma_{u1}, \rho_1)$

$(\varepsilon_i, u_2) \sim N(0, 0, \sigma_\varepsilon, \sigma_{u2}, \rho_2)$

Prenons tous les individus avec (X_i, Z_i) donné. Formellement, la régression de Y_i sur X_i dans l'échantillon tronqué[4] est :

$E(Y_{1i} \mid I_i = 1) = E(Y_{1i} \mid X_{1i}, Z_i, I_i = 1)$

$$E(Y_{1i} \mid I_i = 1) = \beta_1 X_{1i} + E(u_{1i} \mid Z_i, I_i = 1) \qquad (5)$$

$E(Y_{2i} \mid I_i = 0) = E(Y_{2i} \mid X_{2i}, Z_i, I_i = 0)$

$$E(Y_{2i} \mid I_i = 0) = \beta_2 X_2 i + E(u_{2i} \mid Z_i, I_i = 0) \qquad (6)$$

Nous imposons une normalisation sur la variance de ε. Soit $\sigma_\varepsilon = 1$[5]. Sous l'hypothèse de normalité, nous pouvons écrire[6] :

$u_{1i} = \rho_1 \sigma_{u1i} \varepsilon_i$

$u_{2i} = \rho_2 \sigma_{u2i} \varepsilon_i$

En remplaçant dans l'expression (5), nous obtenons :

$E(Y_{1i} \mid I_i = 1) = \beta_1 X_{1i} + E(u_{1i} \mid Z_i, I^*_i > 0)$

$$E(Y_{1i} \mid I_i = 1) = \beta_1 X_{1i} + \rho_1 \sigma_{u1i} E(\varepsilon_i \mid \varepsilon_i > -\delta_i Z_i) \qquad (7)$$

En remplaçant dans l'expression (6), nous obtenons :

[3] Un modèle de sélection est un modèle dans lequel la variable dépendante y n'est pas toujours observée. Le critère de sélection ne porte pas directement sur la valeur de y, mais est défini par une équation auxiliaire.

[4] Un échantillon est tronqué lorsque les observations sont faites seulement pour certains individus, constituant un sous-ensemble de la population observée (Tobin, 1958).

[5] Compte-tenu de la nature des données, seul le signe de I^*_i est observable, et non sa valeur, ce qui empêche l'estimation de la variance de l'équation (4) (Cameron & Triveti, 2005).

[6] En effet, $\text{cov}(u, \varepsilon) = \rho \sigma_u \sigma_\varepsilon = \rho \sigma_u$ et $\text{cov}(\rho \sigma_u \varepsilon, \varepsilon) = \rho \sigma_u V(\varepsilon) = \rho \sigma_u$

$$E(Y_{2i} \mid I_i = 0) = \beta_2 X_{2i} + \rho_2 \sigma_{u2i} E(\varepsilon_i \mid \varepsilon_i < -\delta_i Z_i) \quad (8)$$

La troncature sur ε_i entraîne donc une troncature sur Y_1 et Y_2 si respectivement u_1 et ε et u_2 et ε sont corrélés (ρ_1 et $\rho_2 \neq 0$). Soit l'espérance d'une loi normale tronquée en s :

$$E(\varepsilon_i \mid \varepsilon_i > s) = \int_{z_i=s}^{\infty} \varepsilon_i f(\varepsilon_i | \varepsilon_i > s) . d\varepsilon$$

$$E(\varepsilon_i \mid \varepsilon_i > s) = \int_{z_i=s}^{\infty} \varepsilon_i \frac{\phi(s)}{[1-\Phi(s)]} . d\varepsilon$$

$$E(\varepsilon_i \mid \varepsilon_i > s) = \frac{\phi(s)}{1-\Phi(s)} = \frac{\phi(s)}{\Phi(-s)} \quad (9)$$

Par analogie :

$$E(\varepsilon_i \mid \varepsilon_i \leq s) = -\frac{\phi(s)}{\Phi(s)} \quad (10)$$

En remplaçant dans les expressions (7) et (8), nous obtenons[7] :

$$E(Y_{1i} \mid I_i = 1) = \beta_1 X_{1i} + \rho_1 \sigma_{u1i} \frac{\phi(\delta_i Z_i)}{\Phi(\delta_i Z_i)} = \beta_1 X_{1i} + \rho_1 \sigma_{u1i} \lambda_{1i} \quad (11)$$

$$E(Y_{2i} \mid I_i = 0) = \beta_2 X_{2i} + \rho_2 \sigma_{u2i} \frac{-\phi(\delta_i Z_i)}{1-\Phi(\delta_i Z_i)} = \beta_2 X_{2i} + \rho_2 \sigma_{u2i} \lambda_{2i} \quad (12)$$

Les fonctions ø et Φ sont respectivement les fonctions de densité et de répartition de la loi normale. Les ratios λ_{1i} et λ_{2i} sont appelés inverse du ratio de Mills[8].

La source d'endogénéité du mode d'enquête au niveau de mobilité peut provenir de variables omises, qui sont corrélées à la probabilité de choisir le web comme média d'enquête et au nombre de déplacements saisis. Le biais de sélection correspond donc à un biais de valeur manquante. En effet, si on estime les expressions (2) et (3) par la méthode des moindres carrés ordinaires, on omet deux variables (respectivement $\frac{\phi(\delta_i Z_i)}{\Phi(\delta_i Z_i)} = \lambda_{1i}$ et $\frac{-\phi(\delta_i Z_i)}{1-\Phi(\delta_i Z_i)} = \lambda_{2i}$) et on peut s'attendre à ce que le modèle soit biaisé (les estimations de β_1 et β_2 seront non convergentes). Il est d'ailleurs probable que l'ampleur, le signe et la significativité des coefficients diffèrent lorsqu'ils sont estimés par la méthode en deux étapes. Ces différences dépendent des coefficients $\rho_1 \sigma_{u1}$ et $\rho_2 \sigma_{u2}$ et des coefficients des variables concernées estimés dans le modèle de sélection (Hoffman and Link, 1984). Les variables λ_{1i} et λ_{2i} représentent respectivement pour chaque observation l'espérance conditionnelle des résidus ε_i à $I_i = 1$ et $I_i = 0$. Elles capturent également les valeurs espérées des équations d'intérêt, une fois passé l'effet de sélection. Ces variables sont d'une manière générale la principale source de biais des estimations des coefficients du modèle de régression.

Il est maintenant possible d'estimer les fonctions de mobilité en s'affranchissant du biais de sélection des individus[9], en incluant l'inverse du ratio de Mills dans les expressions (2) et (3). Soit le modèle :

Pour les répondants web :

$$Y_{1i} = \beta_1 X_{1i} + \rho_1 \sigma_{u1i} \lambda_{1i} + e_{1i}, \quad (13)$$

[7] Rappelons que : $\phi(-\delta_i Z_i) = \phi(\delta_i Z_i)$ et que $1 - \Phi(\delta_i Z_i) = \Phi(\delta_i Z_i)$.

[8] Ce ratio est appelé également Lambda d'Heckman, car noté λ(x β) par l'auteur (1979).

[9] La correction du biais de sélection des individus implique que les espérances des termes d'erreur soient maintenant nulles : $E(e_{1i} \mid I_i = 1) = E(e_{1i}) = E(e_{2i} \mid I_i = 0) = E(e_{2i}) = 0$.

Pour les répondants en face-à-face :

$$Y_{2i} = \beta_2 X_{2i} + \rho_2 \sigma_{u2i} \lambda_{2i} + e_{2i}, \quad (14)$$

2.3. Interprétations

Les paramètres du modèle de sélection de l'échantillon peuvent être estimés par la méthode du maximum de vraisemblance. Cependant, la procédure d'estimation en deux étapes d'Heckman (1979) est davantage utilisée. La première étape consiste à estimer l'équation de sélection à l'aide d'un modèle probit, pour obtenir des estimations des δ_i. Pour chaque observation sélectionnée, le modèle calcule la valeur λ_{1i} ou λ_{2i} (inverse du ratio de Mills). Dans une seconde étape, on estime les paramètres β_1 et $\rho_1\sigma_{u1}$ par une régression des moindres carrés ordinaires de Y_{1i} sur X_{1i} et λ_{1i} et les paramètres β_2 et $\rho_2\sigma_{u2}$ par une régression des moindres carrés ordinaires de Y_{2i} sur X_{2i}et λ_{2i}. Les équations du modèle de mobilité Y_{1i} et Y_{2i} contiennent donc non seulement le vecteur de variables explicatives (respectivement X_{1i} et X_{2i}), mais aussi une nouvelle variable construite ou inverse du ratio de Mills. L'existence d'un biais de sélection est testée par l'hypothèse que le coefficient estimé de l'inverse du ratio de Mills est nul dans chaque groupe. Les hypothèses sont les suivantes :

Pour l'échantillon web : $H_0 : \rho_1\sigma_{u1} = 0$

$H1 : \rho_1\sigma_{u1} \neq 0$

Pour l'échantillon face-à-face : $H_0 : \rho_2\sigma_{u2} = 0$

$H_1 : \rho_2\sigma_{u2} \neq 0$

Si la t-value est inférieure à la valeur critique, il n'est pas possible de rejeter l'hypothèse nulle d'absence de corrélation entre les termes d'erreur des équations de sélection et d'intérêt. Dans ce cas, il n'y a pas de biais de sélection significatif dans le modèle, et on peut appliquer la méthode des moindres carrés ordinaires pour estimer directement les coefficients β_1 et β_2. Dans le cas contraire, l'effet de sélection est significatif et le choix du mode apparaît, sous ces hypothèses, endogène au niveau de mobilité. La méthode en deux étapes permet d'obtenir des estimations non biaisées des coefficients β_1 et β_2.

Le paramètre qui fait que le modèle de sélection proposé par Heckman diffère d'un modèle probit suivi d'un modèle de régression linéaire est l'existence d'un coefficient de corrélation (ou covariance) entre les termes d'erreur des équations de sélection et d'intérêt (Verbeek, 2004). Il est alors nécessaire de supposer que des facteurs inobservables jouent à la fois sur le niveau de mobilité des individus et sur le média utilisé pour répondre à l'enquête. L'existence d'un biais de sélection peut être motivée par différents facteurs, mais beaucoup d'auteurs raisonnent en termes de maximisation de l'utilité individuelle (Gronau, 1974).

Il est difficile d'estimer le signe et l'importance de la corrélation entre les deux termes d'erreur[10]. L'interprétation des coefficients $\rho_1\sigma_{u1}$ et $\rho_2\sigma_{u2}$ est donc complexe, mais intéressante. Les inverses du ratio de Mills (λ_{1i} et λ_{2i}) sont par définition positifs. En revanche, $\rho_1\sigma_{u1}$ et $\rho_2\sigma_{u2}$ peuvent prendre tous les signes. Si $\rho_1\sigma_{u1i}$ est négatif, alors les termes d'erreurs des équations de sélection et d'intérêt sont négativement corrélés. Il existe des facteurs inobservés qui font que les individus ne peuvent répondre en face-à-face, mais remplissent le questionnaire en ligne, et qui impactent négativement la mobilité. Cela signifie que le nombre de déplacements déclarés pourrait être en moyenne significativement plus élevé si ces individus avaient répondu en face-à-face. Par analogie, si $\rho_2\sigma_{u2}$ est négatif, le nombre de déplacements déclarés par les individus en face-à-face aurait été en moyenne significativement inférieur si ces derniers avaient choisi le web comme mode d'enquête. L'interprétation est inversée si $\rho_1\sigma_{u1}$ et $\rho_2\sigma_{u2}$ sont positifs.

Certaines critiques ont été formulées à l'égard du modèle de sélection, concernant particulièrement l'hypothèse de normalité (Lee, 1982). En effet, les estimations des paramètres semblent très sensibles à la distribution retenue dans le modèle. La littérature propose des approches alternatives, fondées sur des estimateurs non paramétriques permettant de s'affranchir de l'hypothèse de normalité. Les résultats obtenus diffèrent cependant peu du modèle paramétrique de Heckman (Greene, 2002), et les hypothèses moins fortes du modèle génèrent des résultats moins robustes (Winship & Mare, 1992). Par ailleurs, on peut expliquer une différence entre les estimations des paramètres par la

[10] Un moyen est de trouver une variable omise dans les deux équations, qui ne soit pas corrélée avec les régresseurs.

méthode en deux étapes et celles obtenues par la méthode des moindres carrés ordinaires par une forte colinéarité entre les régresseurs et la variable manquante λ (les variables explicatives de l'équation d'intérêt X_i sont souvent les mêmes que celle de l'équation de sélection Z_i). La méthode en deux étape est donc un compromis entre un biais de sélection d'une part et une erreur due à l'introduction d'un régresseur fortement corrélé aux variables explicatives du modèle, d'autre part[11] (Stolzenberg & Relles, 1997).

3. Les variables disponibles pour l'analyse de la mobilité

Face au grand nombre de variables disponibles, le choix des variables pertinentes à inclure dans le modèle est complexe. Trois types de variables sont renseignés dans l'enquête : des variables sociodémographiques, des variables concernant le niveau d'équipement en télécommunication des ménages et des personnes et des variables caractéristiques de la mobilité individuelle. Les analyses ont mis en évidence l'importance de douze d'entre-elles. Elles sont présentées dans les tableaux suivants, accompagnées de quelques statistiques descriptives : nombre d'observations, valeur minimum et maximum, moyenne et écart-type pour les variables continues (tableau 1), modalités et effectifs, pour les variables nominales (tableau 2).

Variables	Observations	Moyenne	Ecart-type	Minimum	Maximum
Age	10 081	46,21	17,51	18	98
Nb d'enfants du ménage	10 081	0,68	1,02	0	7
Nb de voitures du ménage	10 081	0,68	0,39	0	4
Nb de personnes du ménage	10 081	2,79	1,41	1	10
Distance du domicile au centre de l'agglomération (m)	10 081	7 607	5 163	377	21 937
Densité de la zone de résidence (habitants / km²)	10 081	5 899	5 876	134	21 059

Tableau 1 : Statistiques descriptives des variables continues (échantillon face-à-face et web)

Suites aux analyses exploratoires menées sur l'échantillon, nous pouvons formuler des hypothèses concernant l'impact de ces variables explicatives sur le nombre de déplacements quotidiens moyen des individus. A priori, les femmes se déplacent davantage que les hommes, car elles bénéficient souvent d'une activité professionnelle plus flexible et assurent l'accompagnement des enfants (école, garderie, activités de loisirs…), ainsi que les achats et démarches administratives du ménage. L'âge est lié au statut des individus et à leur cycle de vie, donc à la composition du ménage. On observe que le nombre de déplacements augmente avec l'âge jusqu'à un maximum (situé entre 40 et 50 ans), puis décroît par la suite. L'impact du nombre de personnes du ménage sur la mobilité est ambigu. Nous pouvons toutefois poser l'hypothèse que les déplacements contraints sont mieux répartis lorsque la taille du ménage augmente. Par ailleurs, les analyses exploratoires nous laissent penser que le nombre de déplacements augmente avec le nombre d'enfants du ménage. Le niveau d'étude est quant à lui lié au revenu et au statut. Il est corrélé positivement avec la mobilité. Il en est de même pour l'éloignement géographique du domicile (par rapport au centre ville), puisque les activités sont dispersées en périphérie, ce qui occasionne davantage d'accompagnements. Cet éloignement en périphérie dépend d'un ensemble de variables liées au cycle de vie du ménage (effet de l'âge, du revenu, de la taille du ménage, de sa motorisation…).
Concernant les équipements du ménage, la possession d'un téléphone portable semble liée positivement à la mobilité (nombreux contacts personnels, ou nécessité de se déplacer pour motif professionnel). La variable 'liste de téléphone' est scindée en trois modalités, selon que l'individu ne possède pas de téléphone fixe, soit inscrit sur la liste rouge ou orange ou soit inscrit sur l'annuaire des abonnés. Les packages de téléphonie haut débit, couplés à l'abonnement internet, permettent de ne plus être abonnés à France Télécom, et donc de ne plus figurer sur l'annuaire. Il est probable que les

[11] Afin d'éviter une trop forte colinéarité entre X_{1i} et λ_{1i} d'une part et entre X_{2i} et λ_{2i} d'autre part, il est recommandé qu'au moins une variable explicative de l'équation de sélection ne figure pas dans les équations d'intérêt.

internautes affectionnent les nouvelles technologies, et certains ne disposent plus de ligne fixe, mais d'un téléphone mobile uniquement. Le niveau de motorisation (nombre de voitures du ménage par personne de 18 ans et plus) et le fait de posséder le permis de conduire sont corrélés à la mobilité (les personnes qui n'en possèdent pas sont plus contraintes dans leurs déplacements), même si d'autres facteurs interviennent. Nos premiers tris ont également montré un accroissement des déplacements le vendredi. L'hypothèse est que la réduction de temps de travail (RTT) profite aux déplacements d'achats et de loisirs. Par ailleurs, on observe que la mobilité est plus importante si la personne choisit de communiquer le niveau de revenus annuels nets du ménage. Il est possible que ces répondants doivent faire face à moins de freins pour délivrer des réponses précises, bien que le fait de dévoiler le niveau de salaire ne soit pas une variable directement liée au niveau de mobilité. Enfin, nous avons vu que la mobilité est plus faible chez les individus ayant répondu sur le web.

Variables	Observations	Modalités	Effectifs
Sexe	10 081	Homme	4 757
		Femme	5 324
Possession du permis	10 081	Oui	8 671
		Non	1 410
Vendredi	10 081	Oui	1 585
		Non	8 496
Activité	10 081	Actif : travail au centre	2 474
		Actif : travail périphérie	2 923
		Actif : travail non précis	274
		Non actif	4 410
Revenus déclarés	10 081	Oui	6 649
		Non	3 432
Connexion internet	10 067	Oui	5 916
		Non	4 151
Liste de téléphone	10 063	Annuaire des abonnés	6 934
		Liste (rouge, orange)	2 136
		Pas de ligne fixe	993
Téléphone portable	10 064	Oui	7 392
		Non	2 672
Niveau d'études	9 980	Supérieur	3 682
		Non supérieur	5 432
		En cours	866
Diplôme	9 980	Supérieur	4 337
		Non supérieur	5 643
Mode de réponse	10 081	Web	270
		Face-à-face	9 811

Tableau 2 : Statistiques descriptives des variables nominales (échantillon face-à-face et web)

4. Modèle économétrique

Dans notre échantillon, nous posons l'hypothèse que le mode de collecte des données a un impact sur la mobilité des individus. Il s'agit d'abord d'estimer l'équation de 'choix' du média d'enquête à l'aide d'un modèle probit (section 4.1.), puis d'expliquer les différences de mobilité des individus, au moyen d'un modèle spécifique qui permet de s'affranchir des effets du mode d'enquête (section 4.2.).

4.1. Première étape : équation de sélection

Deux types de variables vont être introduites dans notre modèle en qualité de variables explicatives : des caractéristiques sociodémographiques (âge, occupation, nombre de personnes du ménage, niveau d'études, niveau de revenus annuels du ménage, lieu de travail), puis des variables concernant l'équipement en télécommunication des ménages (connexion internet, téléphone portable, inscription

sur l'annuaire des abonnés). Le modèle probit explicatif du choix du web, appliqué aux variables sélectionnées ci-dessus, donne les résultats présentés dans le tableau 3 (9 980 degrés de liberté).

	Coefficient	Ecart-type	Pr(>|z|)	Signif.
Constante	-3.712	3.469e-01	<2e-16	***
Age	0.049	1.437e-02	0.000663	***
(Age)²	-4.78e-04	1.618e-04	0.003136	**
Connexion internet : oui	0.495	8.648e-02	1.03e-08	***
Téléphone portable : oui	0.257	9.186e-02	0.005156	**
Liste téléphone : oui	0.372	6.598e-02	1.68e-08	***
Liste téléphone : pas de téléphone	0.248	1.180e-01	0.035711	*
Nb de personnes / ménage	-0.101	2.494e-02	5.10e-05	***
Lieu de travail : non précisé	-0.993	3.640e-01	0.006401	**
Lieu de travail : périphérie	-0.24	7.392e-02	0.001184	**
Lieu de travail : inactif	-0.244	8.802e-02	0.005611	**
Diplôme : supérieur	0.41	6.674e-02	8.16e-10	***
Densité de la zone de résidence	-2.169e-05	6.684e-06	0.001174	**
Revenu déclaré : oui	0.401	7.543e-02	1.04e-07	***
Vendredi : oui	0.469	6.661e-02	1.89e-12	***
Distance domicile / centre	-1. 9e-05	7.898e-03	0.014361	*

Signif. : 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Tableau 3 : Equation de sélection - Modèle probit

Les coefficients estimés dans le modèle probit sont significatifs (probabilité < 5% de rejeter à tord l'hypothèse ‘le coefficient est nul’). L'étude de leur signe donne une idée de l'influence des divers facteurs sur le choix de répondre par le web : un coefficient positif augmente la probabilité de répondre en-ligne, mais un coefficient négatif diminue la probabilité de répondre sur internet[12]. L’ensemble des signes des coefficients sont ceux que nous espérions.

La possession d'une connexion internet au domicile, d'un téléphone portable, l'inscription sur la liste rouge ou orange ou l'absence de ligne de téléphone fixe au domicile ou le fait de déclarer son niveau de revenus augmentent la probabilité de répondre sur le web. Bien que l'accès au site hébergeant le questionnaire puisse se faire en dehors du domicile, les ménages possédant une connexion internet ont une probabilité plus élevée de répondre sur le web. Les internautes sont également mieux équipés que la moyenne en moyens de communication, ou appartiennent à une catégorie socioprofessionnelle plus élevée (possession d'un téléphone portable personnel ou professionnel). Les packages de téléphonie haut-débit sont accessibles aux personnes possédant une connexion internet. Fortement sollicités par les démarches commerciales, il est probable que les internautes préfèrent s'exclure de l'annuaire des abonnés en s'inscrivant sur la liste rouge ou orange.

La probabilité de répondre en ligne augmente également avec le niveau d'études et l'âge. Inversement, elle diminue avec l'éloignement géographique du lieu de travail des actifs (par rapport au centre de l'agglomération), le fait d'être inactif et le nombre de personnes du ménage. Les personnes diplômées sont plus familières avec internet. Elles ont probablement utilisé cet outil durant leurs études et l’utilisent encore au domicile ou sur leur lieu de travail. L'âge est lié au cycle de vie des individus. Il est probable qu'avec la progression dans la vie professionnelle, les revenus et l'équipement en moyens de communication s'améliorent. En revanche, internet est une technologie relativement récente et non maîtrisée par tous ; certaines études montrent que l'usage d'internet diminue avec l'âge. L'introduction d’un terme quadratique permet de prendre en compte un impact non linéaire de l'âge. Par ailleurs, la localisation des entreprises est liée aux types d'emplois. Les bureaux, où travaillent des personnes qualifiées qui disposent plus facilement d’un accès sur leur lieu de travail,

[12] Nous travaillons en écart par rapport à une personne de référence, ou à une modalité en ce qui concerne les variables nominales.

sont davantage présents dans le centre de l'agglomération. Les inactifs ont en revanche du temps à consacrer à un enquêteur à domicile, et sont donc moins tentés par un questionnaire web.
Le calcul du coefficient de détermination du modèle (R^2) nous renvoie la valeur 0,12. Environ 12% de la variation dans le choix du mode de réponse (web vs. face-à-face) est expliqué par les variables explicatives retenues dans le modèle. Ce résultat peut s'expliquer par l'absence d'une variable explicative importante dans le modèle de sélection. Si cette variable influe également sur le comportement de mobilité, alors il est impossible de modéliser le nombre de déplacements quotidiens par une simple régression linéaire. Le modèle probit utilisé ci-dessus apporte une solution, en calculant pour chaque individu un facteur de correction, appelé inverse du ratio de Mills.

4.2. Deuxième étape : équation d'intérêt

La deuxième étape consiste à expliquer les différences de comportements en termes de mobilité, au moyen d'un modèle spécifique qui comprend : une variable dépendante (le nombre moyen de déplacements réalisé par les individus), plusieurs variables indépendantes ou explicatives (les facteurs observés censés avoir un effet sur la mobilité), l'inverse du rapport de Mills (variable obtenue dans la première étape) et un terme d'erreur (pour tenir compte des forces non observées qui pourraient influer sur la mesure des résultats). En réalité, nous estimerons deux modèles, un sur le sous échantillon des répondants en face-à-face, et un sur le sous échantillon des répondants sur le web. L'estimation des coefficients des variables explicatives et de la variable relative au biais de sélection se fait par une régression des moindres carrés.
Nous ne retenons ici que les variables qui impactent directement la mobilité des individus. Les variables sociodémographiques prises en compte dans l'équation de sélection ne sont réintroduites que si elles semblent jouer un rôle significatif sur la mobilité quotidienne. Dans ce cas, l'effet marginal des régresseurs sur la mobilité a deux composantes. Il y a un effet direct sur la moyenne de Y_1 et Y_2, capté par β_{1k} et β_{2k} et un effet indirect dû à leur présence dans λ_{1i} et λ_{2i}. La compensation de ces deux effets permet de mettre en évidence l'impact marginal d'une variation des variables explicatives pour un mode d'enquête donné.

4.2.1. Analyse de la mobilité pour l'échantillon en face-à-face

Le modèle restreint au sous échantillon des individus ayant répondu en face-à-face nous donne des résultats intéressants, puisque l'ensemble des variables sont significatives et que les coefficients prennent les signes attendus. Les données sont présentées dans le tableau 4.

Echantillon face-à-face	Régression sans correction			Régression avec correction		
	Coeff.	Pr(>|z|)	Sign.	Coeff.	Pr(>|z|)	Sign.
Constante	2.013	<2e-16	***	2.640	<2e-16	***
Sexe : homme	-0.24	2.14e-07	***	-0.24	2.41e-07	***
Age	0.065	1.38e-15	***	0.057	1.73e-11	***
(Age)²	-7.3e-04	<2e-16	***	-6.4e-04	2.44e-12	***
Possession permis : oui	0.465	4.61e-10	***	0.444	4.23e-09	***
Nb d'enfants / ménage	0.711	<2e-16	***	0.710	<2e-16	***
(Nb d'enfants / ménage)²	-0.004	0.0267	*	-0.003	0.04632	*
Nb de voitures / personne	0.571	<2e-16	***	0.534	2.00e-14	***
Nb de personnes / ménage	-0.188	5.13e-12	***	-0.181	4.78e-11	***
Revenu déclaré : oui	0.325	1.33e-11	***	0.268	2.90e-07	***
Activité : non actif	0.288	6.85e-07	***	0.304	2.30e-07	***
Distance domicile / centre	6.7e-06	0.1478		9.5e-06	0.04358	*
Mills	NA	NA		-0.18	0.00146	**

Signif. : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 4 : Analyse de la mobilité – Echantillon face-à-face

Le fait d'être un homme et d'appartenir à un ménage de taille élevée impacte négativement la propension à se déplacer. A l'inverse, la mobilité semble être une fonction croissante de l'éloignement du domicile par rapport au centre de l'agglomération, du nombre d'enfants et de la volonté de déclarer son niveau de revenus annuels nets. Ces résultats peuvent s'expliquer notamment par les déplacements pour motif " accompagnement ", plus nombreux pour les habitants de la périphérie (dispersion des activités), les ménages avec enfants (vie scolaire et associative) et les femmes. La moindre réticence des personnes qui déclarent leurs revenus à communiquer l'ensemble des activités effectuées la veille de l'interview impacte favorablement la mobilité. Le nombre de déplacements quotidiens moyen augmente également avec la possession du permis de conduire et le nombre moyen de voitures par personne du ménage en âge de conduire. L'effet de la motorisation ou de la possession du permis de conduire est important, puisque le fait de ne pas pouvoir se déplacer en voiture particulière limite les occasions de déplacement.
Enfin, l'introduction de termes quadratiques traduit un impact non linéaire de l'âge et du nombre d'enfants sur la mobilité. Le nombre de déplacements croît jusqu'à l'âge de 40 ans, puis décroît par la suite, à cause probablement de la moindre proportion d'accompagnements.
Le coefficient associé à l'inverse du ratio de Mills est significatif. Il existe donc un biais de sélection des individus. Le signe de ce coefficient est négatif, ce qui signifie qu'en moyenne le choix du web comme mode d'enquête par ces répondants aurait un impact négatif sur le nombre de déplacements déclarés.

4.2.2. Analyse de la mobilité pour l'échantillon web

Les estimations des coefficients du modèle explicatif de la mobilité appliqué à l'échantillon des répondants web sont disponibles dans le tableau 5.

Echantillon web	Régression sans correction			Régression avec correction		
	Coeff.	Pr(>|z|)	Sign.	Coeff.	Pr(>|z|)	Sign.
Constante	0.894	0.501		3.038	0.076	.
Sexe : homme	-0.846	7.22e-04	***	-0.870	0.001	**
Age	0.077	0.212		0.039	0.542	
(Age)²	-8.57e-04	0.220		-4.38e-04	0.547	
Possession permis : oui	0.263	0.638		0.454	0.437	
Nb d'enfants / ménage	0.399	0.418		0.465	0.360	
(Nb d'enfants / ménage)²	-0.121	0.526		-0.151	0.433	
Nb de voitures / personne	0.683	0.077	.	0.690	0.082	.
Nb de personnes / ménage	0.266	0.079	.	0.299	0.056	.
Revenu déclaré : oui	0.360	0.230		0.166	0.647	
Activité : non actif	0.741	0.030	*	0.823	0.023	*
Distance domicile / centre	-5.7e-05	0.030	*	-5.2e-05	0.058	.
Mills	NA	NA		-0.731	1.46e-03	**

Signif. : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 5 : Analyse de la mobilité – Echantillon web

Peu de coefficients sont significatifs, dans le modèle restreint au sous échantillon des individus ayant répondu en ligne. Ceci s'explique notamment par les différences d'effectifs entre les deux échantillons de répondants. La comparaison des deux modèles estimés ci-dessus montre que l'ordre de grandeur des coefficients est le même, mais que celui de leurs écarts-types varie fortement[13], puisque 13 271 individus ont été interrogés en face-à-face contre seulement 369 sur le web (soit un rapport de 1 à 36). L'ordre de grandeur des écart-types des coefficients estimés varie dans un rapport de 1 à 6 entre les deux échantillons web et face-à-face. Les valeurs de la statistique de test sont donc beaucoup

[13] La variance des coefficients estimés est : $V(\beta) = 1/n * s^2/V(x)$, avec n le nombre d'observations, s^2 la variance de l'échantillon et V(x) la variance de la population.

plus faibles dans le cas de l'échantillon web et ne dépassent que rarement le seuil critique de 1,96 permettant de conclure à la significativité statistique des coefficients (au risque $\alpha = 5\%$).
Les estimations non corrigées et corrigées des coefficients diffèrent beaucoup. Les coefficients non corrigés du biais de sélection auraient pu inclure des faux positifs ou des faux négatifs, mais ce n'est pas le cas ici puisqu'après correction les coefficients conservent leur signe. L'ajout de la variable « Mills » permet d'identifier l'impact réel des facteurs socioéconomiques sur la mobilité des répondants web. Globalement le modèle de régression est meilleur, puisque nous avons 2% de variance expliquée en plus ($R^2 = 15\%$, vs. 13%).
Le fait d'être un homme impacte négativement la propension à se déplacer, les femmes effectuant globalement plus de déplacements. En revanche, contrairement à ce que l'on observe dans l'échantillon en face-à-face, certains coefficients ne prennent pas le signe attendu. Ainsi, le fait de résider en périphérie diminue la propension à se déplacer. Les actifs cadres sont fortement représentés dans l'échantillon web. Ils habitent souvent en périphérie, disposent d'un haut niveau de formation et déclarent des revenus élevés, mais leur emploi est chronophage. Ils ont peu de temps libre en semaine pour effectuer des activités non contraintes, ce qui limite les possibilités de déplacements. Par ailleurs, la mobilité semble croître, ici, avec le nombre de personnes du ménage. Les répondants web ont des composantes socioéconomiques plus élevées, qui leur permette de réaliser davantage d'activités.
Le coefficient de l'inverse du ratio de Mills est significatif. Il y a donc des variables qui influent sur le choix de remplir en questionnaire en-ligne et la mobilité des répondants, et un biais d'endogénéité du mode sur la mobilité. Le recours à la méthode d'estimation en deux étapes est justifié, puisque le choix du web apparaît, dans ces conditions, endogène au niveau de mobilité déclaré. Par ailleurs, le signe négatif de l'inverse du ratio de Mills signifie que la mobilité pourrait être en moyenne significativement plus élevée si ces répondants n'avaient pas répondu sur le web.

4.2.3. Interprétation des coefficients des inverses du ratio de Mills

Nous avons vu que les coefficients estimés $\rho_1\sigma_{u1}$ et $\rho_2\sigma_{u2}$ peuvent prendre tous les signes, en fonction des signes de ρ_1 et ρ_2, c'est-à-dire du signe de la corrélation entre les résidus du l'équation de sélection et ceux de l'équation d'intérêt concernant respectivement les répondants web et face-à-face. Les résidus correspondent à des variables non observées et par conséquent non prises en compte dans le modèle, qui peuvent avoir un effet sur la variable à expliquer. Par exemple, le fait de travailler de nuit n'est pas une variable explicative du modèle de mobilité. Pourtant, cette caractéristique implique que les individus sont davantage disponibles en fin de journée pour recevoir un enquêteur à domicile et susceptibles d'effectuer des déplacements la journée et la nuit, d'où une mobilité élevée.

Les coefficients de l'inverse du ratio de Mills correspondent au produit de ρ_1 par σ_{u1} pour l'échantillon web (= -0,731) et de ρ_2 par σ_{u2} pour l'échantillon en face-à-face (= -0,180). La procédure d'estimation en deux étapes d'Heckman ne nous permet pas de distinguer ces valeurs. Il est cependant possible d'estimer les écarts-types des résidus de l'équation d'intérêt. Avec $\sigma_{u1} = 2,2$ et $\sigma_{u2} = 2,4$, nous avons :

$\rho_1 = -0,731/1,91 = -0,383$
$\rho_2 = -0,180/2,23 = -0,081$

L'estimation du paramètre ρ permet d'évaluer la corrélation entre l'inverse du ratio de Mills et le niveau de mobilité. Nous pouvons conclure que le nombre de déplacements, pour les répondants web, apparaît fortement corrélé à l'inverse du ratio de Mills ($\rho_1 = -0,383$). Cette corrélation est plus faible dans le groupe des répondants en face-à-face ($\rho_2 = -0,081$). Ces valeurs permettent d'évaluer la force de l'endogénéité du mode d'enquête au nombre de déplacements déclarés.

4.3. Test de stabilité du modèle

Il s'agit maintenant de tester si les estimations des coefficients générés par le modèle en deux étapes sont stables et de quantifier un éventuel impact du mode d'enquête. Nous comparons deux modèles de régression simples, permettant d'expliquer le nombre moyen de déplacements des répondants : un modèle contraint et un modèle non contraint. Le modèle contraint ne considère comme facteurs explicatifs de la mobilité que les variables retenues dans le modèle en deux étapes ci-dessus. Le

modèle non contraint inclut également l'ensemble des interactions entre les variables explicatives et le mode d'enquête (web et face-à-face).
L'hypothèse à tester est la suivante : les interactions entre les variables explicatives et le mode d'enquête, retirées dans le modèle contraint, n'ont aucun pouvoir explicatif significatif sur le nombre de déplacements quotidiens moyen des individus. Dans ce cas, il est impossible de conclure à l'existence d'un effet stable du mode d'enquête sur les réponses des enquêtés. Soit les hypothèses :
H_0 : les deux modèles contraints et non contraints sont équivalents, et les interactions entre les variables explicatives de la mobilité et le mode d'enquête ne sont pas significatives.
H_1 : les deux modèles contraints et non contraints ne sont pas équivalents, et les interactions entre les variables explicatives de la mobilité et le mode d'enquête sont significatives.

Les estimations du modèle de mobilité non contraint appliqué à l'ensemble de l'échantillon sont présentées dans les tableaux 4 et 5, ceux du modèle contraint (par les variables retenues dans le modèle à deux étapes) dans le tableau 6.

Modèle contraint	Coefficients	Pr(>|z|)	Significativité
Constante	2.538	<2e-16	***
Sexe : homme	-0.251	4.19e-08	***
Age	0.058	7.72e-12	***
(Age)²	-6.41e-04	1.02e-12	***
Possession permis : oui	0.450	1.84e-09	***
Nb d'enfants / ménage	0.710	<2e-16	***
(Nb d'enfants / ménage)²	-0.035	0.042	*
Nb de voitures / personne	0.529	1.61e-14	***
Nb de personnes / ménage	-0.170	3.38e-10	***
Revenu déclaré : oui	0.263	3.62e-07	***
Activité : non actif	0.320	3.20e-08	***
Distance domicile / centre	8.44e-06	0.069	.
Mills	-0.155	4.85e-03	**

Tableau 6 : Modèle contraint

Le test de Wald appliqué aux deux modèles, contraint et non contraint, renvoie une probabilité de rejeter à tord l'hypothèse d'absence d'interaction égale à 3.21e-04. La p-value est inférieure à la valeur critique de 5%. Nous pouvons donc rejeter l'hypothèse nulle d'équivalence des modèles contraints et non contraints et conclure qu'il existe une interaction significative entre les variables explicatives de la mobilité et le mode d'enquête considéré, qui permet d'expliquer le nombre de déplacements effectués par les répondants. Trois variables semblent significativement interagir avec le mode de réponse : le sexe, le nombre de personnes composant le ménage et la distance entre le centre de l'agglomération et le lieu de résidence du ménage. En revanche, le coefficient de la variable 'Mills' n'est pas significativement différent entre les deux échantillons. Les résultats du modèle de régression appliqué aux variables explicatives et interactions significatives sont présentés dans le tableau 7.

Modèle final	Coefficients	Pr(>|z|)	Significativité
Constante	2.650	<2e-16	***
Sexe : homme	-0.240	2.08e-07	***
Age	0.057	1.03e-11	***
Age²	-6.36e-04	1.53e-12	***
Possession permis : oui	0.445	2.63e-09	***
Nb d'enfants / ménage	0.694	<2e-16	***
(Nb d'enfants / ménage)²	-0.033	0.051	.
Nb de voitures / personne	0.534	8.07e-15	***
Nb de personnes / ménage	-0.173	1.93e-10	***
Revenu déclaré : oui	0.265	2.84e-07	***
Activité : non actif	0.315	5.13e-08	***
Distance domicile / centre	9.54e-06	0.042	*
Mills	-0.191	6.15e-04	***
Mode	-0.504	0.176	
(Sexe : homme) * mode	-0.547	0.059	.
(Nb de personnes / ménage) * mode	0.187	0.102	
(Distance domicile / centre) * mode	-5.9e-05	0.053	.

Tableau 7 : Modèle final – Interactions significatives

On applique ensuite le test de Wald aux modèles non contraint (contenant l'ensemble des interactions entre les variables explicatives de la mobilité et le mode d'enquête) et contraint simplifié (qui ne laisse que les trois interactions significatives comme variables explicatives). La probabilité de rejeter à tord l'hypothèse selon laquelle ces modèles sont équivalents est de 43%. Les interactions entre les variables et le mode d'enquête supprimées n'ont aucun pouvoir explicatif significatif de la mobilité quotidienne. Nous conservons donc ce modèle simplifié.

5. Impact des interactions entre les variables explicatives de la mobilité et le mode d'enquête après correction du biais de sélection

Ce paragraphe illustre de quelle manière les apports des techniques économétriques permettent de comprendre les différences de mobilité observées entre deux échantillons. Nous allons détailler l'impact des variables explicatives sur le nombre de déplacements déclarés, après correction du biais de sélection. Notre regard se porte essentiellement sur les variables qui interagissent avec le mode de réponse choisi. Mais l'échantillon web est de taille modeste au regard de l'échantillon face-à-face. Nous conservons dans l'analyse les interactions significatives au seuil d'erreur de 10%, exception faite du coefficient de la variable mode (p-value = 17.6%).
Le mode de recueil de données, web ou face-à-face, impacte directement le niveau de mobilité. Si le questionnaire est rempli en-ligne, le nombre de déplacements décroît de 0.5.
Par ailleurs, trois variables semblent interagir avec le mode : le sexe, nombre de personnes du ménage et la distance entre le domicile et le centre de l'agglomération. Pour ces deux dernières, la relation bivariée avec le mode d'enquête peut est formalisée par la figure 3.

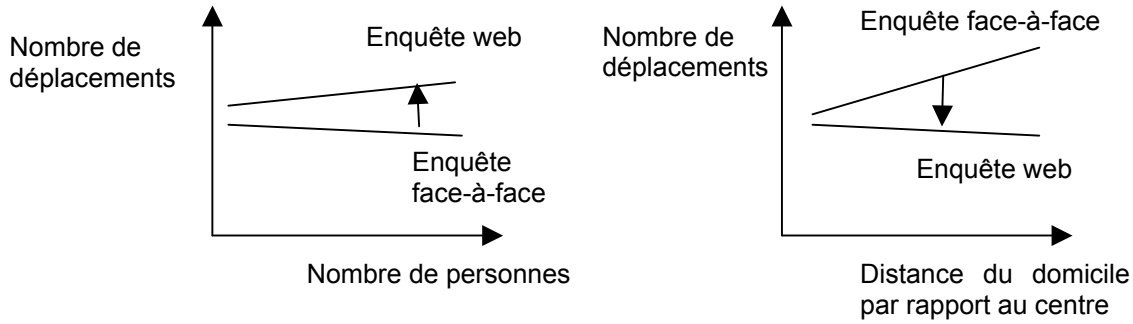

Figure 3 : Relation bivariée entre le nombre de personnes, l'éloignement du domicile et le nombre de déplacements des individus.

Le coefficient de la variable 'distance du domicile au centre x mode' est légèrement négatif (-5.90e-05). Ainsi, la mobilité des répondants web diminue avec l'éloignement de leur lieu de résidence du centre de l'agglomération : l'impact est de 9.54e-06 déplacements pour les répondants en face-à-face, vs. -4.90e-05 déplacements pour les répondants web. Les répondants web occupent davantage d'emplois de cadres et d'employés, situés dans le centre de l'agglomération. Avec l'éloignement du lieu de résidence, la distance domicile-travail augmente. Le temps disponible pour les activités moins contraintes est donc réduit, ce qui impacte négativement leur mobilité. Nous remarquons que ce coefficient est très faible par rapport aux autres. Ceci s'explique par l'unité choisie (m) pour calculer la distance entre le domicile et le centre de l'agglomération.
Le fait d'être un homme impacte négativement la mobilité des répondants web : l'impact est de -0,240 déplacements par jour pour les répondants en face-à-face, vs. - 0,787 pour les répondants web. Comme nous l'avons décrit précédemment, les femmes se déplacent davantage que les hommes, car elles doivent gérer des activités familiales en plus de leur activité professionnelle. Cet effet est amplifié en ce qui concerne les répondants web. Nous pouvons penser que leur niveau d'emploi, plus exigeant (beaucoup de cardes), leur laisse encore moins de latitude pour leurs déplacements. Rappelons ici que 3/4 des répondants web sont des actifs.
A contrario, la mobilité des répondants web augmente avec le nombre de personnes du ménage : l'impact sur le nombre de déplacements quotidiens est de -0,173 par personne pour les répondants en face-à-face, vs. 0,014 pour les répondants web. Ces derniers disposent d'un revenu annuel supérieur. Il est donc probable que nombre de personnes accroisse les besoins et donc les déplacements (pour motif achat, loisir…) des membres du ménage.

Les coefficients des variables mode, (Sexe:homme)*mode, (Nb de personnes / ménage)*mode et (Distance domicile / centre)*mode permettent ainsi de quantifier, pour chaque répondant, l'impact du mode d'enquête sur le comportement de mobilité.
Soit un homme de 35 ans, actif et possédant le permis de conduire, qui réside à 1 km du centre ville. On fait l'hypothèse qu'il appartient à un ménage composé de 4 personnes, dont 1 enfant de moins de 18 ans, et de 2 voitures et qu'il n'a pas communiqué ses revenus lors de l'enquête. Ce profil de répondant, déclare sur le web 3.99 déplacements, vs. 4.58 en face-à-face.
Si on considère à présent un homme actif de 22 ans, sans enfant, vivant en couple à 500 mètre du centre de l'agglomération, qui a son permis, 1 seule voiture à disposition du ménage et qui a déclaré ses revenus, le nombre de déplacements est égal à 3.97 sur le web et 3.82 en face-à-face. La différence de mobilité est donc fortement atténuée.
A contrario, une femme active de 53 ans ayant son permis, appartenant à un ménage de 5 personnes (4 adultes et un enfant de moins de 18 ans) résidant à 5 km du centre de l'agglomération, qui possède 3 voitures et qui n'a pas déclaré ses revenus déclare, toutes choses égales par ailleurs, 4.44 déplacements sur le web et 4.30 déplacements en face-à-face.
Lorsque les valeurs des variables qui interagissent avec le mode d'enquête sont élevées, la différence de mobilité s'accroît.

6. Pistes de réflexion

La comparaison du nombre de déplacements moyen quotidiens déclarés par les répondants montre que les internautes se déplacent moins que les individus interrogés en face à face. L'hypothèse d'une sous-déclaration imputable au média web est tentante, mais il est également possible que les différences socio-économiques observées entre les deux échantillons expliquent au moins en partie cet écart de mobilité. Les internautes ont un niveau d'études et un revenu nettement supérieurs, conduisant à une très forte motorisation. Ils sont plus souvent cadres et employés et travaillent davantage dans le centre, ce qui conduit à des durées hors domicile plus importantes, réduisant les possibilités de participation à des activités moins contraintes et l'usage de la marche.
Le risque est de confondre le phénomène étudié (la variation de la mobilité individuelle) avec le processus de sélection des répondants. On peut conclure qu'il existe un problème d'auto-sélection, si le choix du web exerce un effet différent sur la mobilité des internautes qu'il n'aurait exercé sur celle des répondants en face-à-face si ces derniers avaient rempli le questionnaire en-ligne. A l'aide de techniques économétriques, il est possible d'isoler l'impact des différences socio-économiques propres aux répondants de l'effet média. La procédure d'estimation en deux étapes empruntée à Heckman permet de s'affranchir du biais d'auto sélection des individus (volonté de répondre en face-à-face ou de remplir le questionnaire en-ligne). Les coefficients estimés des variables du modèle ne sont plus biaisés et reflètent l'impact des variables socioéconomiques sur la mobilité, indépendamment du choix du mode de recueil de données. Dans notre exemple, le biais de sélection est statistiquement significatif. Le mode d'enquête a donc une incidence sur la mobilité qu'une simple régression linéaire ne peut mettre en évidence. L'exercice montre que les variables socioéconomiques qui impactent la mobilité des répondants web sont le sexe, le nombre de personnes du ménage, ainsi que la distance entre le domicile et le centre de l'agglomération. Certaines hypothèses ont été formulées pour tenter d'expliquer ces différences comportementales.
Cet article propose une méthode intéressante pour s'affranchir du biais de sélection des individus dans le cadre d'une enquête mixte. La question est de savoir si ce biais est suffisamment important pour devoir être corrigé. Ceci dépend de la part de la population exclue par un mode de collecte de données et de la précision des données attendue. Dans cette expérience, le relativement faible taux de pénétration du web dans la population et les exigences croissantes des modèles de planification ne permettent pas d'occulter le biais de sélection de l'échantillon. Toutefois, des développements complémentaires permettraient d'approfondir l'analyse comparative de la mobilité selon le mode d'enquête. D'abord, nous avons montré que la sous-mobilité des répondants web concernaient certains types de motifs (non contraints) et de mode de transports (marche à pied) (Bayart et Bonnel, 2008). Une analyse des facteurs explicatifs de ces types de déplacements permettrait de corriger plus finement la mobilité quotidienne des internautes. Par ailleurs, la sous-mobilité des répondants web s'explique par une immobilité plus importante. Certaines techniques ont l'avantage de séparer les facteurs explicatifs de la décision de se déplacer de ceux qui influent sur le niveau de la mobilité. Un modèle explicatif de l'immobilité selon le mode d'enquête serait également un moyen de mieux appréhender les facteurs à l'origine de la sous-mobilité des répondants web.

Bibliographie

Amemiya, T., "Tobit models: a survey", *Journal of Econometrics*, Vol. 24, pp. 3-61 (1984).
Ampt, E.S., « Response Rates - Do they matter? In : « Les enquêtes de déplacements urbains : mesurer le présent, simuler le futur », Bonnel P., Chapleau R., Lee-Gosselin M., Raux C. (Ed), collection Programme Rhône-Alpes Recherches en Sciences Humaines, Lyon, pp. 115-125 (1997).
Bayart, C., Bonnel, P., « Le potentiel du web pour les enquêtes de mobilité », Journées de Méthodologie Statistique, Paris (2009).
Bayart, C., Bonnel, P., « Enquête web auprès des non-répondants de l'enquête ménages déplacements de Lyon 2005-2006 », Rapport pour le PREDIT, Laboratoire d'Economie des Transports, Lyon, 262p (2008).
Berk, R.A., " An introduction to sample selection bias in sociological data", *American Sociologic Review*, Vol. 48, n°3, pp.386-398 (June 1983).
Bonnel, P., Le NIR M., "The quality of survey data: telephone versus face-to-face interviews", *Transportation*, vol. 25, n°2, pp. 147-167 (1998).

Cameron, A.C., Trivedi, P.K., “Microeconometrics Methods and Applications”, New York, Cambridge University Press (2005).
Certu, « L'enquête ménages déplacements standard Certu », Collections du CERTU, éditions du CERTU, Lyon, 204p (2008).
Couper, M.P., “Web surveys: a review of issues and approaches”, *Public Opinion Quarterly*, Vol. 65, n° 2, pp. 230-253 (2000).
Dillman, D.A., Phelps, G., Tortora, R.D., Swift, K., Kohrell, J., Berck, J., “Response Rate and Measurement Differences in Mixed Mode Surveys Using Mail, Telephone, Interactive Voice Response and the Internet”, The American Association for Public Opinion Research (AAPOR), 56th Annual Conference (2001).
Goldberger, A.S., “Linear regression after selection”, *Journal of Econometrics*, Vol. 15, pp. 3579-356 (1981).
Gronau, R., “Wage comparisons: a selectivity bias”, *Journal of Political Economy*, Vol. 82, pp. 1119-143 (1974).
Gunn, H., “Web-based Surveys: Changing the survey process”, *First Monday*, Vol. 7, n°12 (December 2002).
Heckman, J., “Sample selection bias as a specification error”, *Econometrica*, Vol. 47, n°1, pp. 153-161 (January 1979).
Hoffman, D., Link, C.R., “Selectivity bias in Male wage equation: black-white comparison”, *The Review of Economics and Statistics*, Vol. 66, n° 2, pp. 320-324 (May 1984).
Kmenta, J., “Elements of econometrics”, New York: McMillan (1971).
Lee, L.F., “Some approaches to the correction of selectivity bias”, *The Review of Economic Studies*, Vol. 49, n°3, pp. 355-372 (July 1982).
Maddala, G.S., “Limited dependant and qualitative variables in econometrics”, Cambridge University Press, London (1986).
Mroz, T.A., “The sensitivity of an empirical model of married women’s hours of work to economic and statistical assumptions”, *Econometrica*, Vol. 55, n°4, July 1987, pp. 765-799 (1987).
Nakosteen, R.A., Zimmer, M.A., “Migration and income: the question of self-selection”, *Southern Economic Journal*, n°46, pp. 840-851 (1980).
Ressource System Group Inc., Documentation for SBIR Phase II Final Report: “Computer-Based Intelligent Travel Survey System”, DTRS57-00-C-10030, Prepared for the FHWA (October 2002).
Roy, A.D., “Some thoughts on the distribution of earnings”, *Oxford Economic Papers*, Vol. 3, pp. 135–146 (1951).
Stolzenberg, R.M., Relles, D.A., “Tools for intuition about sample selection bias and its correction”, *American Sociological Review*, Vol. 62, n°3, pp. 494-507 (June 1997).
Stopher, P.R.,Greaves, S.P., “Household travel surveys: Where are we going?”, *Transportation Research Part A: Policy and Practice*, Vol. 41, N°5, pp. 367-81 (2006).
Tobin, J., “Estimation of relationship for limited dependent variables”, *Econometrica*, Vol. 26, n°1, pp. 24-36 (January 1958).
Verbeek, M. (2004), “A guide to modern econometrics”, 2nd Edition, John Wiley & Sons Ltd, England, 429p (April 2004).
Winship, C., Mare, R.D., “Models for sample selection bias”, *Annual Review of Sociology*, Vol. 18, pp. 327-350 (1992).