Troisième partie

La comparabilité des données : enjeux et méthodes de correction

Introduction de la partie III

Si le fait de proposer des médias différents permet d'augmenter le taux de réponse global, la comparabilité des données reste un exercice difficile. La comparaison du nombre moyen de déplacements quotidiens déclaré 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'étude 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.

Un des objectifs de cette partie est de montrer qu'il est possible de distinguer l'effet dû au mode d'enquête de celui lié aux différences socio-économiques 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 le 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. Le second objectif est de mieux comprendre les déterminants de la mobilité, en distinguant ceux qui sont à l'origine de la décision de se déplacer de ceux qui impactent le nombre de déplacements réalisé par les personnes mobiles. Afin d'étudier séparément ces deux effets, nous utilisons un modèle à obstacle (ou modèle Hurdle).

Dans le chapitre 7, nous mettons en évidence l'impact du mode d'enquête (web ou face-à-face) sur la mesure des comportements de mobilité. Nous proposons une formalisation du problème du biais de sélection de l'échantillon, et développons un modèle explicatif des comportements de mobilité permettant de s'affranchir de ce biais. Le chapitre 8 est consacré à une analyse comparative des déterminants de l'immobilité, d'une part, des facteurs qui influencent la mobilité des personnes mobiles, d'autre part. Nous mettons ainsi en évidence l'intérêt des modèles à obstacle (Hurdle) pour modéliser les différentes facettes d'un phénomène complexe : la mobilité quotidienne.

Chapitre 7 : Impact du mode d'enquête sur la mesure des comportements de mobilité

"*Il faut utiliser les modèles, non y croire.*"
Henri Theil (1924-2000).

Dans le chapitre précédent, nous avons montré que la qualité des réponses, et plus précisément le niveau de mobilité quotidienne, varie entre les deux enquêtes web et face-à-face. Le recueil d'informations via différentes sources peut générer des résultats parfois peu comparables et 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. Ce biais de sélection est l'objet d'une importante littérature, théorique et empirique (Winship et Mare, 1992), mais les applications aux enquêtes transport sont rares à ce jour. Les données de l'enquête ménages déplacements de Lyon nous permettent de mettre en évidence le problème de sélection, les non-répondants à l'enquête standard en face-à-face choisissant de remplir ou pas le questionnaire sur le web.

Ces différences constatées entre les deux enquêtes peuvent provenir d'un effet média ou du fait que tous les répondants n'ont pas la possibilité de répondre en ligne, les effets de mode et de sélection étant confondus (De Leeuw, 2005). Il est alors nécessaire de rechercher, toutes choses égales par ailleurs, quel est l'effet marginal du web sur les résultats. Autrement dit, identifier les propriétés propres au web en tant que mode d'enquête. Selon Lozar Manfreda et Vehovar (2002b), Willke *et al.* (1999) et Wydra (1999), les effets de mode surviennent surtout lorsque le web est comparé à des enquêtes administrées en face-à-face. Il est donc essentiel d'établir une procédure pour pouvoir comparer les résultats des deux enquêtes, le choix du média ayant une influence non négligeable sur la mobilité individuelle (Van Evert et Moritz, 2000; Moritz et Brög, 1999).

L'objet de ce chapitre est d'évaluer plus précisément l'effet du mode d'enquête sur le nombre de déplacements quotidiens déclaré par les répondants. L'évaluation nécessite de tester l'existence d'un biais de sélection de l'échantillon, puis le mesurer et le neutraliser, le cas échéant. La section I présente le

problème de l'évaluation, en le transposant à l'analyse comparative du mode d'enquête sur la mobilité quotidienne. Il s'agit de mettre en évidence les enjeux du biais de sélection susceptible de se produire dans les protocoles d'enquête mixte. Suivent les développements théoriques relatifs au biais de sélection, plus spécifiquement ceux concernant la méthode en deux étapes d'Heckman (1979) et celle des variables instrumentales (Heckman et Navarro-Lozano, 2004). L'intérêt et les limites respectives de ces deux méthodes sont expliqués (section II). Puis, nous recensons les variables disponibles pour l'élaboration d'un modèle explicatif de la mobilité et estimons un premier modèle qui ne tient pas compte du biais de sélection (section III). Nous développons ensuite un modèle économétrique permettant de s'affranchir de l'effet de sélection des individus, et l'appliquons aux données de l'enquête ménages déplacements de Lyon. Nous justifions la spécification du modèle retenu et les traitements réalisés pour améliorer la robustesse des résultats. Nous présentons l'ensemble des régressions et interprétons les paramètres d'intérêt du modèle, en particulier la significativité et le signe de l'inverse du ratio de Mills (section IV). Enfin, nous isolons et quantifions la part du différentiel de mobilité liée au mode d'enquête et celle venant du biais de sélection (section IV). Seuls les individus mobiles (qui se sont déplacés durant la période de référence), de 18 ans et plus sont concernés dans ce chapitre [111].

I La formalisation du problème

Dans l'enquête ménages déplacements de Lyon réalisée en 2006, les réponses des enquêtes web et face-à-face ne sont pas comparables. L'analyse des données montre que les répondants web déclarent en moyenne moins de déplacements quotidiens que les répondants à l'enquête en face-à-face. Ce constat peut notamment s'expliquer par les différences socio-économiques, qui caractérisent les deux échantillons. En redressant l'enquête en face-à-face, on a cherché à isoler une population dont les caractéristiques socio-économiques sont proches de celles de l'enquête web et dont les effectifs sont suffisant pour avoir des résultats statistiquement significatifs. Les différences en termes de nombre de déplacements subsistent et ont même tendance à s'amplifier un peu par rapport à l'ensemble de la population. Cet écart s'explique à la fois par une plus forte immobilité des internautes actifs et par une moindre déclaration de déplacements et de sorties. Deux interprétations sont possibles de ces résultats. La première conduit à une sous-estimation de la mobilité imputable au média web. Plusieurs facteurs différencient les modes et peuvent être à l'origine de la diversité des réponses [112], comme ceux qui influencent la transmission de l'information, et les effets de l'interviewer. Les différences observées dans les

[111] Dans l'enquête web, une seule personne remplit le questionnaire. Il s'agit le plus souvent du destinataire de la lettre avis, c'est-à-dire le chef de ménage ou son conjoint, ce qui explique que les individus de moins de 18 ans soient fortement sous-représentés.

[112] Ces différences concernent principalement les questions sensibles, d'attitude, demandant un effort de mémoire important au répondant, ayant beaucoup de réponses possibles et les répondants peu familiers avec internet (Lozar Manfreda et Vehovar, 2002a).

réponses des enquêtes web et face-à-face peuvent également venir du mode d'administration du questionnaire (auto-administré, vs. administré), de l'informatisation du questionnaire, ou être plus spécifiques (Willke *et al.*, 1999). Par exemple, le manque d'ergonomie du questionnaire web peut rendre pénible la saisie des déplacements [113], sans compter les biais relatifs à l'aisance des répondants face à un ordinateur. Cette première interprétation peut être étayée par le fait que la sous-estimation concerne surtout des déplacements courts, tant en temps qu'en distance, principalement effectués à pied, pour des motifs comme les loisirs ou l'accompagnement, qui peuvent être perçus comme moins importants par les internautes (Bonnel et Le Nir, 1998). Avec une meilleure ergonomie, il serait possible de réduire cette omission, mais des contraintes techniques n'ont pas permis d'exploiter toutes les possibilités interactives du média web.

La seconde interprétation concerne les caractéristiques socio-économiques des internautes. Il peut en effet 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. Selon la méthodologie utilisée, les individus choisissent de répondre ou non en face-à-face, et dans la négative de remplir ou pas le questionnaire en ligne. La présence des répondants dans un groupe est probablement déterminée par des facteurs extérieurs, qui peuvent impacter la variable d'intérêt du modèle étudié. Dit autrement, il est fort probable que des caractéristiques socio-économiques, 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; Ressource System Group, 2002). 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. Il est fort probable que seuls les individus intéressés par le thème de l'enquête et ayant un accès privé à l'ordinateur (et à une connexion internet) répondent à l'enquête web (Stanton, 1998; Morrel-Samuels, 2003). Le choix de ce média d'enquête dépend donc en grande partie des facteurs socio-économiques et de l'aisance avec l'ordinateur [114].

Si le mode retenu pour remplir le questionnaire est lié à la mobilité, alors les internautes pourraient 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, biaise l'effet réel du média. Ignorer l'existence du biais de sélection, en particulier pour les répondants web, peut donc avoir des conséquences sur la validité d'un modèle explicatif des comportements de mobilité.

Le choix du mode peut modifier les réponses aux questions formulées iden-

[113] Les enquêtes assistées par ordinateur génèrent globalement moins d'erreurs, mais également moins d'activités enregistrées.

[114] Plus le répondant est jeune, a fait des hautes études et utilise un e-mail, plus il a de chance de répondre sur le web (Romano et Himmelmann, 2002).

tiquement entre deux enquêtes, et le web comme nouveau mode de collecte introduit des effets qui sont encore peu connus (Dillman et Christian, 2005). Il est alors nécessaire de scinder la différence de mobilité observée entre les répondants web et face-à-face en deux parties : l'effet de sélection et l'effet du mode d'enquête [115].

II Le biais de sélection de l'échantillon

On sait 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). Cependant, les premiers développements économétriques des conséquences de cette sélection des individus datent de 1974, avec les travaux d'Heckman (1979). L'exemple typique 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), bien que l'utilisation la plus fréquente des modèles d'auto-sélection concerne l'évaluation d'un traitement ou d'une formation.

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 [116]. 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 sociodémographiques 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. Le modèle traditionnellement utilisé pour mettre en évidence le problème d'auto-sélection des répondants est le modèle Tobit II [117], appelé plus couramment modèle de sélection [118] de l'échantillon.

II.1 L'endogénéité du mode d'enquête au niveau de mobilité

Nous présentons des méthodes permettant de formaliser l'endogénéité du mode d'enquête au niveau de mobilité.

[115] Si la période d'enquête n'est pas la même, alors les effets de temps se mélangent aux effets de mode (De Leeuw, 2005).

[116] Le biais de sélection peut être soit positif soit négatif, c'est-à-dire que si on n'en tient pas compte, l'effet estimé du mode d'enquête peut être soit supérieur soit inférieur à son véritable effet.

[117] Cette classification est due à Amemiya (1984).

[118] 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.

II.1.1 Un effet direct du mode d'enquête non observable

La décision de répondre ou non à l'enquêteur au domicile, puis, le cas échéant, de se soumettre ou pas à l'enquête en ligne, est prise par chaque individu contacté pour répondre à l'étude. Le fait de répondre en ligne est une variable aléatoire notée I_i, qui prend la valeur 1 si l'individu i a répondu au questionnaire web et la valeur 0 s'il a répondu à l'enquête en face-à-face. Le nombre de déplacements réalisé par l'individu i (Y_i) représente la variable d'intérêt du modèle (c'est à partir du niveau de mobilité individuelle déclaré que nous souhaitons évaluer l'effet du mode d'enquête). Il s'agit d'une variable générale, qui tient compte des caractéristiques positives (comme l'anonymat des réponses) et négatives (par exemple, la lourdeur dans la saisie des réponses) du média web.

Nous sommes amenés à distinguer deux variables latentes, Y_{0i} et Y_{1i}, qui correspondent au nombre de déplacements déclaré par un individu i, conditionnellement au média de réponse utilisé (respectivement $I_i = 0$ ou $I_i = 1$). Le problème de l'évaluation provient du fait qu'il n'est jamais possible d'observer simultanément, pour un individu i, le nombre de déplacements déclaré en face-à-face et sur le web. Si l'individu i a répondu sur le web, alors le nombre de déplacements saisi sur le web (Y_{1i}) est observé, alors que le nombre de déplacements déclaré en face-à-face (Y_{0i}) est inconnu. Dans ce cas, la valeur non observée (Y_{0i}) est qualifiée de résultat contrefactuel. Symétriquement, pour un individu i ayant répondu en face-à-face, le nombre de déplacements recueilli par l'enquêteur (Y_{0i}) est observé, alors que le nombre de déplacements qu'aurait déclaré cet individu sur le web (Y_{1i})est inconnu. C'est la valeur non observée Y_{1i}, qui correspond dans ce cas au résultat contrefactuel.

Nous pouvons dire, à partir du raisonnement précédent, que la variable de résultat observée (Y_i) est liée aux variables latentes de résultat Y_{0i} et Y_{1i}, par l'équation :

$$Y_i = I_i * Y_{1i} + (1 - I_i) * Y_{0i} \tag{23}$$

L'effet causal du mode d'enquête sur le niveau de mobilité déclaré correspond à l'écart entre les variables Y_{1i} et Y_{0i}. Il représente la différence entre le nombre de déplacements saisi par l'internaute et celui que ce même individu aurait déclaré à un enquêteur présent à son domicile. Le problème fondamental de l'évaluation est que cette différence n'est pas directement observable. Une des solutions pour déterminer cet effet consiste à poser des hypothèses sur la loi jointe du quadruplet $(Y_{1i}, Y_{0i}, Z_i, X_i)$, avec X_i les variables explicatives du niveau de mobilité individuelle pour l'individu i et Z_i les variables explicatives du fait de répondre en ligne pour l'individu i. L'estimation de l'effet du média web sur la mobilité correspond en effet à l'espérance de la différence des deux variables latentes : $E(Y_{1i} - Y_{0i})$. L'effet moyen du média web sur le nombre de déplacements déclaré s'écrit :

$$\Delta_{I1} = E(Y_{1i} - Y_{0i} \mid I_i = 1) \tag{24}$$

Et l'effet marginal du média web sur le nombre de déplacements déclaré s'écrit :

$$ME_I = E(Y_{1i} - Y_{0i} \mid X_i, Z_i) \tag{25}$$

II.1.2 Application du modèle de Rubin

Par analogie avec le modèle de Rubin (1974), on peut considérer que le mode d'enquête (web) s'apparente à un traitement à évaluer. Il existe un problème d'auto-sélection des individus, lorsqu'un traitement exerce un effet différent sur la population des traités de l'effet qu'il n'aurait exercé sur la population des non traités si celle-ci avait été traitée. Cet effet provient des différences existant entre les populations de traités et de non traités, c'est-à-dire entre les répondants web et face-à-face. Si le nombre de déplacements déclaré par un individu i est indépendant de la probabilité pour ce même individu de répondre en ligne (si $Y_{1i}, Y_{0i} \coprod I_i$), alors :

$$\Delta_{I1} = E(Y_{1i} \mid I_i = 1) - E(Y_{0i} \mid I_i = 0) \tag{26}$$
$$\Delta_{I1} = E(Y_i \mid I_i = 1) - E(Y_i \mid I_i = 0) \tag{27}$$

A contrario, nous pouvons supposer que l'effet du média pour les individus répondant en ligne ($E(Y_{1i} \mid I_i = 1)$) est différent de l'effet qu'engendrerait ce média pour un individu répondant en face-à-face ($E(Y_{1i} \mid I_i = 0)$). Une première source d'endogénéité du mode d'enquête au niveau de mobilité individuelle peut provenir de variables omises, qui sont corrélées à la fois au nombre de déplacements quotidiens et à la propension individuelle à refuser une interview en face-à-face, puis à répondre sur le web. Par exemple, si le chef de ménage a un emploi qualifié, alors ses horaires de travail ne permettent pas à un enquêteur d'interroger l'ensemble des personnes au domicile, mais sa familiarité avec l'utilisation des nouvelles technologies (notamment pour son activité professionnelle) l'encourage à répondre en ligne. Il s'agit de variables cachées, dont on a de bonnes raisons de penser qu'elles existent, mais qu'on ne peut pas mettre en évidence. L'incapacité à isoler ces effets produit des estimations non convergentes du coefficient de la variable mode de réponse, dans le modèle de régression explicatif du nombre de déplacements déclaré.

II.2 Les méthodes d'estimation disponibles

Afin de corriger ces biais, il faut développer un modèle économétrique plus spécifique. Le principe est d'utiliser les informations dont on dispose sur les individus non traités pour construire pour chaque individu traité un contrefactuel, c'est-à-dire une estimation de ce qu'aurait été sa situation s'il n'avait pas été traité (Brodaty *et al.*, 2007). Les méthodes d'évaluation procèdent toujours en plusieurs étapes. Par une fonction de sélection, on estime pour chaque individu la probabilité de recevoir un traitement. Puis, on régresse la variable d'intérêt, en intégrant dans le modèle des estimateurs corrigés. Deux grands

types de variables sont considérés dans l'équation de sélection : les variables observables et les variables inobservables.

II.2.1 La sélection sur variables observables

Les méthodes qui s'appuient sur les hypothèses d'indépendance conditionnelle à des caractéristiques observables sont appelées les méthodes de sélection par appariement. Il s'agit de contrôler les performances d'un traitement par la probabilité de le recevoir, conditionnellement à des variables observables. Le principe de cette méthode est d'utiliser les informations dont on dispose sur les individus non traités (répondants web) pour construire pour chaque individu traité (répondants face-à-face) un contrefactuel, c'est-à-dire une estimation de ce qu'aurait été sa situation s'il n'avait pas été traité. L'hypothèse de sélection sur variables observables, selon laquelle, conditionnellement aux variables observables Z_i (variables explicatives du fait de répondre en ligne pour un individu i), le fait pour un individu de répondre sur le web est rendue indépendante, est très forte. Nous avons :

$$(Y_{0i}, Y_{1i}) \coprod I_1 \mid P(Z_i) \tag{28}$$

II.2.2 La prise en compte de variables inobservables

Une autre solution pour résoudre le problème de sélection est de recourir à une modélisation jointe des résultats potentiels (Y_{0i}, Y_{1i}) et de l'affectation au traitement (I_i), en faisant l'hypothèse que ces trois variables dépendent de termes d'erreur inobservables, potentiellement corrélés entre eux. Cette méthode suppose, contrairement à la méthode par appariement, que les résultats (le nombre de déplacements déclarés) et les choix (répondre ou pas sur le web), soient additivement dissociables. Cette méthode est par définition plus globale et produit ainsi des estimations plus robustes que la méthode par appariement, car lorsque des variables inobservables existent, elles biaisent les méthodes d'appariement sur variables observables (Heckman et Navarro-Lozano, 2004). Cette hypothèse de modélisation du biais de sélection apparaît beaucoup plus pertinente pour évaluer l'effet du média web, car les critères déterminants le choix d'un mode de réponse ne sont pas tous connus de l'analyste. La méthode de Heckman (1979) à deux étapes et la méthode des variables instrumentales reposent toutes deux sur une modélisation du lien entre les variables inobservables et les variables observables.

II.2.3 La méthode des variables instrumentales

La méthode des variables instrumentales, abordée notamment par (Heckman et Robb, 1985), permet de mettre en évidence le biais de sélection. Elle consiste à rechercher une ou plusieurs variables qui influent sur le choix du mode d'enquête, mais pas sur le nombre de déplacements déclaré. Ces variables, non corrélées au terme d'erreur de l'équation du modèle de mobilité

quotidienne, peuvent servir à l'estimation, sans qu'il y ait un biais. Plus formellement, la méthode des variables instrumentales cherche à estimer le nombre de déplacements quotidiens déclarés, sachant les variables explicatives de la mobilité X_{1i} et X_{2i} et les variables de sélection Z_i. Soit :

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

Cette méthode diffère de l'estimation en deux étapes d'Heckman, qui suppose une condition supplémentaire de dépendance du nombre de déplacements déclarés à l'utilisation du média web pour la saisie des réponses :

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

Soit le modèle de régression linéaire explicatif de la mobilité quotidienne :

$$Y_i = \sum_k \beta_k X_{ki} + \delta I_i + u_i \tag{29}$$

avec Y_i le nombre moyen de déplacements réalisé par l'individu i, X_i un vecteur de variables indépendantes ou explicatives de la mobilité et I_i le fait de répondre à l'enquête en ligne. Le fait de répondre en ligne peut être expliqué par le modèle suivant :

$$I_i = 1, \; ssi \; I_i^* > 0$$

$$I_i = 0, \; ssi \; I_i^* \leq 0$$

avec :

$$I_i^* = \sum_m \delta_m Z_{mi} + \epsilon_i \tag{30}$$

La variable I_i^* est une variable latente et Z_{mi} un vecteur de variables explicatives, dites de conditionnement. Nous supposons que ϵ_i suit une loi normale et est liée à u_i par une loi normale bivariée. Pour corriger le biais de sélection par la méthode des variables instrumentales, nous remplaçons dans la fonction d'intérêt la variable I_i par l'estimation de la propension à répondre sur le web (I_i^*), effectuée à l'aide d'un modèle probit. Soit :

$$Y_i = \sum_k \beta'_k X_{ki} + \delta' I_i^*) + u'_i \tag{31}$$

avec u'_i les résidus corrigés du biais de sélection de la régression linéaire. Les estimations des paramètres par la méthode des moindres carrés ordinaires sont convergentes.

Il faut noter que si l'effet des variables inobservables sur la propension à répondre en ligne est strictement équivalent pour les répondants web et les répondants face-à-face, alors la propension à répondre en ligne est un instrument valide, et on peut utiliser la méthode des variables instrumentales. Si en revanche l'effet des variables inobservables sur la propension à répondre en ligne est différent pour les répondants web et les répondants face-à-face, mais que

le fait de répondre en ligne est rendu indépendant des variables inobservables par des variables observables, alors la méthodes des variables instrumentales s'apparente à une méthode de traitement du biais de sélection par appariement.

La méthode des variables instrumentales s'effectue en une seule étape, et ne nécessite pas de modélisation détaillée du processus de participation à l'enquête web. En revanche, si les estimations sont sans biais (lorsque le modèle est bien défini), cette méthode ne donne pas une estimation de l'ampleur du biais de sélection. Par ailleurs, la recherche des variables instrumentales exige un examen approfondi du processus de sélection des individus et les caractéristiques sociodémographiques traditionnellement recueillies dans les enquêtes sont rarement suffisantes, car souvent liées aux résultats du modèle estimé. L'identification de variables instrumentales pertinentes est donc une tâche difficile. C'est pourquoi nous préférons utiliser la méthode en deux étapes (section II.2.2).

II.3 La méthode en deux étapes

Cette procédure consiste à estimer, dans une première équation, la probabilité pour un individu i de répondre en ligne puis, dans une seconde étape, les paramètres de l'équation d'intérêt, en s'affranchissant de l'endogénéité éventuelle du mode d'enquête à la mobilité.

II.3.1 Développements théoriques

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 :

$$Y_{1i} = \sum_k \beta_{1k} X_{1ki} + u_{1i} \tag{32}$$

Pour les répondants en face-à-face :

$$Y_{2i} = \sum_k \beta_{2k} X_{2ki} + u_{2i} \tag{33}$$

Avec Y_{1i} et Y_{2i} le nombre moyen de déplacements réalisé par l'individu i, 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 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 [119]. Le modèle ainsi défini n'impose pas que les coefficients des variables explicatives

[119] Notons que seulement un des paramètres Y_{1i} et Y_{2i} est observé, selon que l'individu choisisse de répondre en face-à-face ou sur le web.

X_{1ki} et X_{2ki} soient identiques pour les répondants web et les répondants en face-à-face.

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

$$I_i^* = \sum_m \delta_m Z_{mi} + \epsilon_i \tag{34}$$

$$I_i = 1,\ ssi\ I_i^* > 0$$

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

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

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

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

Nous posons l'hypothèse restrictive qu'il existe des variables de Z_i qui ne sont pas dans X_i. Par ailleurs, nous supposons que les termes d'erreur des équations 32, 33 et 34 (u_1, u_2, et ϵ) suivent une loi normale bivariée de moyennes nulles et de corrélations ρ_1 et ρ_2. La mobilité n'est observable que si les individus répondent en face-à-face ou sur le web. Des perturbations aléatoires vont donc affecter simultanément les variables endogènes des équations de sélection et d'intérêt, et les termes d'erreur des deux équations peuvent être corrélés :

$$(\epsilon, u_{1i}) \sim N(0, 0, \sigma_\epsilon, \sigma_{u1}, \rho_1)$$

$$(\epsilon, u_{2i}) \sim N(0, 0, \sigma_\epsilon, \sigma_{u2}, \rho_2)$$

L'hypothèse centrale est que la distribution normale de ϵ_i est liée à u_{1i} et u_{2i} par une loi normale bivariée et normalement distribuée, où ρ_1 et ρ_2 sont les coefficients de corrélation entre les résidus. Nous faisons l'hypothèse que les éléments inobservés ϵ_i et (u_{1i}, u_{2i}) sont indépendants des variables explicatives X_{1i} et X_{2i} et des variables de conditionnement Z_i.

Prenons tous les individus avec (X_i, Z_i) donné. Formellement, la régression de Y_i sur X_i dans l'échantillon tronqué [120] est :

$$E(Y_{1i} \mid I_i = 1) = E(Y_{1i} \mid X_{1i}, Z_i, I_i = 1) \tag{35}$$

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + E(u_{1i} \mid Z_i, I_i = 1) \tag{36}$$

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

$$E(Y_{2i} \mid I_i = 0) = E(Y_{2i} \mid X_{2i}, Z_i, I_i = 0) \quad (37)$$

$$E(Y_{1i} \mid I_i = 0) = \sum_k \beta_{2k} X_{2ki} + E(u_{2i} \mid Z_i, I_i = 0) \quad (38)$$

Si le choix du mode de réponse des individus est systématiquement corrélé avec leur mobilité, alors les espérances conditionnelles de la mobilité et des termes d'erreur ne sont pas égales à leurs espérances :

$$E(Y_{1i} \mid I_i = 1) \neq E(Y_{1i}) \Leftrightarrow E(u_{1i} \mid I_i = 1) \neq E(u_{1i}) = 0$$
$$E(Y_{2i} \mid I_i = 0) \neq E(Y_{2i}) \Leftrightarrow E(u_{2i} \mid I_i = 0) \neq E(u_{2i}) = 0$$

Nous imposons une normalisation sur la variance de ϵ. Soit $\sigma_\epsilon = 1$ [121]. Cette condition se justifie par le fait que nous n'observons pas la valeur de la variable latente I_i^*, mais seulement son signe (Davidson et McKinnon, 1993). Nous observons donc le résultat du choix individuel de répondre ou non à l'enquête sur le web, mais pas la propension d'un individu à répondre en ligne (I_i^*). Sous l'hypothèse de normalité, nous pouvons écrire [122] :

$$u_{1i} = \rho_1 \sigma_{u1i} \epsilon_i$$
$$u_{2i} = \rho_2 \sigma_{u2i} \epsilon_i$$

σ_{u1i} représente la covariance entre les résidus u_{1i} de la régression linéaire et les résidus ϵ_i ($I_i = 1$), σ_{u2i} représente la covariance entre les résidus u_{2i} de la régression linéaire et les résidus ϵ_i ($I_i = 0$).

En remplaçant dans l'expression 36, nous obtenons :

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + E(u_{1i} \mid Z_i, I_i = 1) \quad (39)$$

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + E(u_{1i} \mid Z_i, I_i^* > 0) \quad (40)$$

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + E(u_{1i} \mid \epsilon_i > -\sum_m \delta_m Z_{mi}) \quad (41)$$

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + E(\rho_1 \sigma_{u1i} \epsilon_i \mid \epsilon_i > -\sum_m \delta_m Z_{mi}) \quad (42)$$

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + \rho_1 \sigma_{u1i} E(\epsilon_i \mid \epsilon_i > -\sum_m \delta_m Z_{mi}) \quad (43)$$

En remplaçant dans l'expression 38, nous obtenons :

$$E(Y_{2i} \mid I_i = 0) = \sum_k \beta_{2k} X_{2ki} + E(u_{2i} \mid Z_i, I_i^* \leq 0) \quad (44)$$

$$E(Y_{2i} \mid I_i = 0) = \sum_k \beta_{2k} X_{2ki} + \rho_2 \sigma_{u2i} E(\epsilon_i \mid \epsilon_i \leq -\sum_m \delta_m Z_{mi}) \quad (45)$$

[121] 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 34 (Cameron et Triverdi, 2005).

[122] En effet, $cov(u, \epsilon) = \rho \sigma_u \sigma_\epsilon = \rho \sigma_u$ et $cov(\rho \sigma_u \epsilon, \epsilon) = \rho \sigma_u V(\epsilon) = \rho \sigma_u$

La troncature sur ϵ entraîne donc une troncature sur Y_1 et Y_2 si, u_1 et ϵ, d'une part, et u_2 et ϵ, d'autre part, sont corrélés ($\rho_1 et \rho_2 \neq 0$). Soit ϕ la fonction de densité et Φ la fonction de répartition de la loi normale. Le détail du calcul de l'espérance d'une loi normale tronquée en s est présenté en annexe XII.

Par définition, nous avons :

$$E(\epsilon_i \mid \epsilon_i > -\sum_m \delta_m Z_{mi}) = \frac{\phi(-\sum_m \delta_m Z_{mi})}{1 - \Phi(-\sum_m \delta_m Z_{mi})} \tag{46}$$

$$E(\epsilon_i \mid \epsilon_i \leq -\sum_m \delta_m Z_{mi}) = -\frac{\phi(-\sum_m \delta_m Z_{mi})}{\Phi(-\sum_m \delta_m Z_{mi})} \tag{47}$$

En remplaçant dans les expressions 43 et 45, nous obtenons [123] :

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + \rho_1 \sigma_{u1i} \frac{\phi(\sum_m \delta_m Z_{mi})}{\Phi(\sum_m \delta_m Z_{mi})} \tag{48}$$

$$E(Y_{1i} \mid I_i = 1) = \sum_k \beta_{1k} X_{1ki} + \rho_1 \sigma_{u1i} \lambda_{1i} \tag{49}$$

$$E(Y_{2i} \mid I_i = 0) = \sum_k \beta_{2k} X_{2ki} + \rho_2 \sigma_{u2i} \frac{-\phi(\sum_m \delta_m Z_{mi})}{1 - \Phi(\sum_m \delta_m Z_{mi})} \tag{50}$$

$$E(Y_{2i} \mid I_i = 0) = \sum_k \beta_{2k} X_{2ki} + \rho_2 \sigma_{u2i} \lambda_{2i} \tag{51}$$

Il s'agit en fait de remplacer ϵ_i par son espérance conditionnelle aux valeurs de δZ_i, c'est-à-dire par $\frac{\phi(\delta Z_i)}{\Phi(\delta Z_i)}$. Les ratios $\lambda_{1i} = \frac{\phi}{\Phi}$ et $\lambda_{2i} = \frac{-\phi}{1-\Phi}$ sont appelés inverse du ratio de Mills [124]. Pour chaque observation, nous calculons l'inverse du ratio de Mills, qui correspond à l'espérance conditionnelle des résidus ϵ_i à $I_i = k$, avec $k = 0, 1$. Cette variable est d'une manière générale la principale source de biais des estimations des coefficients du modèle de régression. La source d'endogénéité du mode d'enquête au niveau de mobilité provient ici de variables omises, qui sont corrélées à la probabilité de choisir le web comme

[123] Rappelons que : $\phi(-\sum_m \delta_m Z_{mi}) = \phi(\sum_m \delta_m Z_{mi})$ et que : $1 - \Phi(-\sum_m \delta_m Z_{mi}) = \Phi(\sum_m \delta_m Z_{mi})$

[124] Ce ratio est appelé également Lambda d'Heckman, car noté $\lambda(x\beta)$ par l'auteur (1979).

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 32 et 33 par la méthode des moindres carrés ordinaires, on omet deux variables :

$$\frac{\phi(\sum_m \delta_m Z_{mi})}{\Phi(\sum_m \delta_m Z_{mi})} = \lambda_{1i} \tag{52}$$

$$\tag{53}$$

$$\frac{-\phi(\sum_m \delta_m Z_{mi})}{1-\Phi(\sum_m \delta_m Z_{mi})} = \lambda_{2i} \tag{54}$$

Dans ce cas, on peut s'attendre à ce que le modèle soit biaisé, puisque les estimations de β_{1k} et β_{2k} sont 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 et Link, 1984).

La méthode en deux étapes permet donc d'estimer les fonctions de mobilité des échantillons web et face-à-face en s'affranchissant du biais de sélection des individus, grâce à l'intégration de l'inverse du ratio de Mills dans les expressions 32 et 33. Soit le modèle :

Pour les répondants web :

$$Y_{1i} = \sum_k \beta_{1k} X_{1ki} + \rho_1 \sigma_{u1i} \lambda_{1i} + e_{1i} \tag{55}$$

Pour les répondants en face-à-face :

$$Y_{2i} = \sum_k \beta_{2k} X_{2ki} + \rho_2 \sigma_{u2i} \lambda_{2i} + e_{2i} \tag{56}$$

II.4 Test de l'existence significative d'un biais de sélection

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 (web et face-à-face). Les hypothèses sont les suivantes :

Pour l'échantillon web :

$$H0 : \rho_1 \sigma_{u1} = 0$$

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

Pour l'échantillon face-à-face :

$$H0 : \rho_2 \sigma_{u2} = 0$$
$$H1 : \rho_2 \sigma_{u2} \neq 0$$

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_{1k} et β_{2k}. Cela peut s'expliquer par le fait que les facteurs non observés conduisant à une sélection non aléatoire des répondants selon le mode d'enquête ne sont pas quantitativement importants dans l'évaluation du modèle explicatif du nombre de déplacements quotidiens, ou bien que ces facteurs sont quantitativement importants mais ne sont pas en relation directe avec les résultats du modèle étudié.

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_{1k} et β_{2k}. Il est alors possible d'obtenir une estimation de l'ampleur du biais de sélection et d'estimer l'effet véritable du mode d'enquête sur le nombre de déplacements déclaré quotidiennement par les répondants.

II.4.1 Sens et intensité du biais de sélection

Il est difficile d'estimer le signe et l'importance de la corrélation entre les deux termes d'erreur 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_1 ou λ_2) sont par définition positifs. En revanche, $\rho_1 \sigma_{u1}$ et $\rho_2 \sigma_{u2}$ peuvent prendre tous les signes. Si $\rho_1 \sigma_{u1}$ 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 déclaré 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.

Notons qu'il est possible d'obtenir des estimateurs efficients du coefficient de corrélation ρ (entre les résidus u_{1i} et ϵ_i d'une part, u_{2i} et ϵ_i d'autre part), ainsi que de la covariance σ, dont le produit est le paramètre de l'inverse du ratio de Mills. Même si l'hypothèse $\rho = 0$ ne peut être acceptée, les estimateurs par la méthode des moindres carrés (dans la deuxième étape) restent non biaisés. L'efficience des estimation des paramètres dépend de l'hypothèse de la distribution normale et bivariée des résidus des deux étapes.

II.4.2 Conclusion : Les limites de la méthode en deux étapes

La première étape de la méthode consiste à estimer l'équation de sélection à l'aide d'un modèle probit, pour obtenir des estimations des δ_m. 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_{1k} et $\rho_1\sigma_{u1}$, par une régression des moindres carrés ordinaires de Y_1 sur X_{1k} et λ_1, et les paramètres β_{2k} et $\rho_2\sigma_{u2}$, par, une régression des moindres carrés ordinaires de Y_2 sur X_{2k} et λ_2. Les équations du modèle de mobilité Y_1 et Y_2 contiennent donc non seulement le vecteur de variables explicatives (respectivement X_{1k} et X_{2k}), mais aussi une nouvelle variable construite, ou inverse du ratio de Mills λ_1 et λ_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.

Si la méthode en deux étapes d'Heckman est largement utilisée pour traiter du biais de sélection, certaines critiques ont été formulées à son égard. Le modèle d'Heckman souffre de plusieurs difficultés. La première concerne l'hypothèse de normalité (Lee, 1982). En effet, les estimations des paramètres semblent très sensibles aux distributions des termes d'erreurs des équations de sélection et d'intérêt du 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 et Mare, 1992).

Par ailleurs, on peut expliquer une différence entre les estimations des paramètres par la méthode en deux étapes et celles obtenues par la méthode des moindres carrés ordinaires, par une forte colinéarité entre les régresseurs et les variables manquantes λ_1 et λ_2. La méthode en deux étapes est donc un compromis entre un biais de sélection et une erreur due à l'introduction d'un régresseur fortement corrélé aux variables explicatives du modèle (Stolzenberg et Relles, 1997). Afin de respecter l'indépendance des variables de conditionnement Z_m et des résidus de seconde étape e_1 et e_2, il est préférable de trouver des variables de conditionnement suffisamment indépendantes du niveau de mobilité [125].

Le ratio de Mills est très sensible à la colinéarité pouvant exister entre les deux équations et les variables explicatives de l'équation d'intérêt X_{ki} sont souvent les mêmes que celle de l'équation de sélection Z_{mi}. Ceci peut conduire

[125]Lorsque le vecteur Z est entièrement contenu dans X_1 et X_2, l'identification de σ et ρ repose entièrement sur la non-linéarité de l'inverse du ratio de Mills. La contrainte de non-linéarité peut être relâchée, en trouvant des variables explicatives au sein de Z qui soient distinctes des variables des vecteurs X_1 et X_2.

à des procédures de restriction des variables explicatives, la solution idéale étant de différencier totalement les variables indépendantes de l'équation de sélection de celles de l'équation de niveau.

Un autre inconvénient est que l'inverse du ratio de Mills est traité comme une variable explicative dans la régression de la seconde étape, alors qu'il correspond à une partie des résidus. Il n'est donc pas évident d'interpréter le signe du paramètre de l'inverse du ratio de Mills, en particulier lorsque la régression est réalisée sur l'ensemble des individus (web et face-à-face). La méthode de Heckman reste néanmoins recommandée pour tester l'existence d'un biais de sélection (Davidson et McKinnon, 1993).

Enfin, il est utile de rappeler que la sélection des répondants peut être fondée sur les facteurs observables ou inobservables. C'est la richesse des données d'enquêtes qui détermine quels facteurs sont observés et lesquels ne le sont pas. Ainsi, plus les données disponibles sont riches et plus la part du biais de sélection imputable à des facteurs inobservables est réduite. Bien qu'il existe toujours des facteurs inobservés qui influent sur la sélection pour le mode d'enquête en ligne, cela n'implique pas nécessairement que la simple comparaison des répondants web et des répondants face-à-face mettra en évidence un biais de sélection. Dans notre étude, il y a un biais de sélection si et seulement si les facteurs inobservés qui influent sur la participation ou non à l'enquête web impactent également le nombre de déplacements déclarés par les individus interrogés.

III Modèle explicatif ne tenant pas compte du biais de sélection

Dans cette section, nous cherchons à identifier, par enquête, les principaux facteurs susceptibles d'expliquer les différences de mobilité entre les individus, sans tenir compte de l'existence éventuelle d'un biais de sélection. Nous définissons la mobilité quotidienne comme le nombre de déplacements moyen effectué par les individus durant une journée. Pour mener notre analyse, nous utilisons un modèle de régression multiple. Un modèle est un moyen de représenter une réalité, c'est-à-dire d'en donner une représentation simplifiée et éventuellement de simuler ce qui peut survenir si un facteur est modifié.

Nous présentons d'abord les variables disponibles pour l'analyse (section III.1), avant de procéder à une régression linéaire multivariée, permettant d'analyser l'influence des facteurs retenus sur la mobilité quotidienne (section III.2).

III.1 Les variables disponibles pour l'analyse de la mobilité

Dans les modèles de régression multiple, on cherche à déterminer la valeur d'une variable numérique, dite « à expliquer » ou « dépendante », par les

valeurs de plusieurs autres variables, dites « explicatives ». Le choix des variables explicatives se fait a priori, en fonction des éléments bibliographiques et empiriques publiés sur le sujet et selon le bon sens de l'analyste.

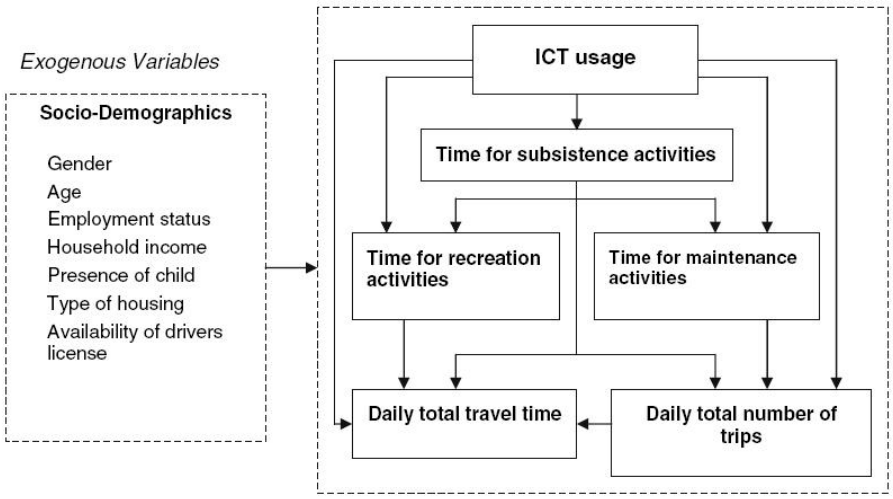

FIG. 59 – Hypothèse de relations entre l'usage des nouvelles technologies, les caractéristiques socio-économiques et le comportement de mobilité des répondants

Source : Wang et Law (2007)

La complexité du phénomène à expliquer, la mobilité quotidienne, conduit à prendre en compte de nombreuses dimensions susceptibles d'influer sur le nombre de déplacements déclaré, comme le montre à titre d'exemple la figure 59. La progression de l'individu dans le cycle de vie dépend en grande partie de son âge, et s'accompagne d'un certain nombre de caractéristiques socio-économiques (passage du permis, entrée dans la vie active, accès à la motorisation, arrivée des enfants, changement de lieu de résidence ...) qui vont impacter plus ou moins son niveau de mobilité. Face au grand nombre de variables disponibles dans l'enquête déplacements, le choix des variables pertinentes à inclure dans le modèle est difficile. Les premières analyses ont mis en évidence l'importance de dix sept d'entre-elles. Quelques statistiques descriptives sont présentées en annexe XIII.

III.1.1 Variables sociodémographiques

Sexe : homme ou femme.

Age : âge de chaque personne.

Nombre d'enfants du ménage : nombre d'enfants de moins de 18 ans présents dans le ménage.

Nombre de personnes du ménage : nombre de personnes qui composent le ménage.

Revenus declarés : variable dichotomique, indiquant si l'individu a déclaré le niveau de revenus annuels nets de son ménage.

Activité : variable combinant le statut du répondant et le lieu de travail des actifs (travailleurs à temps plein ou à temps partiel, apprentis, personnes en formations et en stage). Quatre modalités sont disponibles : les non actifs, les actifs travaillant à Lyon ou à Villeurbanne, les actifs dont le lieu de travail se situe dans la reste de l'agglomération et les actifs n'ayant pas communiqué la zone de leur lieu de travail.

Distance du domicile au centre de l'agglomération : distance entre la zone de résidence du ménage et la zone centrale de l'agglomération (par hypothèse, la zone 10203, Bellecour), exprimée en m.

Densité de la zone de résidence : densité de la zone de résidence du ménage, exprimée en nombre d'habitants par kilomètre carré.

Il semble a priori aisé de mettre en évidence l'effet de certains facteurs tels que le revenu du ménage ou la catégorie socioprofessionnelle du répondant sur la mobilité individuelle. Cependant la variable 'Profession et catégorie socio-professionnelle' est difficile à exploiter dans notre modèle, du fait des effectifs très inégaux et parfois faibles de certaines catégories. Il n'est également pas envisageable d'introduire la variable 'Revenu' dans le modèle, la non-réponse partielle étant très élevée (1 personne sur 3 en face-à-face et 1/4 sur le web). Comme le précisent Bonnafous et Puel (1983), un facteur regroupe les deux variables évoquées ci-dessus et montre un effet non négligeable sur la mobilité. Il s'agit du niveau d'étude ou de diplôme, lié positivement au nombre de déplacements.

Niveau d'étude : niveau d'étude de la personne, en trois modalités : en cours d'études, études non supérieures (allant de pas d'études, jusqu'au BAC) et études supérieures (BAC+2 et plus).

Diplôme : dernier diplôme obtenu par la personne, en deux modalités : non supérieur (allant jusqu'au BAC) et supérieur (BAC+2 et plus).

III.1.2 Variables caractéristiques de l'équipement en télécommunication

Téléphone portable : possession d'un téléphone portable par la personne, à titre personnel ou professionnel.

Liste de téléphone : la variable est scindée en trois modalités, selon que l'individu ne possède pas de téléphone fixe, est inscrit sur la liste rouge ou orange ou est inscrit sur l'annuaire des abonnés.

Connexion internet : possession d'une connexion internet au domicile du ménage.

III.1.3 Variables caractéristiques de la mobilité

Le nombre de voitures du ménage et la possession du permis constituent une bonne indication de la situation des répondants.

Nombre de voitures du ménage : nombre de voitures particulières possédées par le ménage, rapporté au nombre de personnes en âge de conduire (18 ans et plus).

Possession du permis : possession du permis de conduire ou pratique de la conduite accompagnée.

Vendredi : variable qui prend la valeur "1" si le jour de référence pour le recueil des déplacements est le vendredi, et la valeur "0" sinon

III.1.4 Variables caractéristiques du choix du mode de réponse

Mode de réponse : variable qui prend la valeur "1" si le mode de recueil des données est le web, et la valeur "0" pour le face-à-face.

III.2 Analyse de la mobilité par un modèle de régression linéaire multiple

Considérons l'équation (57), 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 multiple, de la forme :

$$Y_i = \sum_k \beta_k X_{ki} + \alpha I_i + u_i \tag{57}$$

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. Nous formulons dans un premier temps des hypothèses sur la forme des variables explicatives à inclure dans le modèle de régression. Puis, nous traitons le problème de la colinéarité de certaines d'entre-elles, avant d'analyser en détail les résultats des modèles. Nous utiliserons d'abord la base de données dans son intégralité, avant de limiter notre analyse à chacun des deux sous-échantillons (web et face-à-face). Les internautes ont en effet des caractéristiques particulières (Cf. chapitres 5 à 7), qui justifient le développement d'un modèle explicatif plus spécifique.

III.2.1 Hypothèses sur la forme des variables explicatives

Nous savons que le niveau de mobilité est plus faible chez les individus mobiles ayant répondu sur le web (3,78 vs. 4,19 déplacements en face-à-face ; p-value $< 0{,}001\%$). Suite aux analyses exploratoires menées sur les échantillons web et face-à-face, nous pouvons formuler des hypothèses concernant l'impact des variables explicatives retenues sur le nombre de déplacements quotidiens moyen des individus.

Effet du genre : des femmes plus mobiles que les hommes A priori, les femmes se déplacent davantage que les hommes, car elles réalisent quotidiennement un plus grand nombre d'activités. En plus de leur activité professionnelle, elles assurent souvent l'accompagnement des enfants (école, garderie, activités

de loisirs...), ainsi que les achats et démarches administratives du ménage. Cette relation se vérifie dans les deux échantillons web et face-à-face (tableau 60), même si la différence est plus accentuée parmi les internautes.

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Homme	4,08	3,30
Femme	4,28	4,16
P-value	$< 0{,}001\%$	$< 0{,}001\%$

TAB. 60 – Nombre de déplacements moyen par genre et par enquête

Effet de l'âge : une relation non linéaire L'âge est lié au cycle de vie des individus, donc à la composition du ménage.

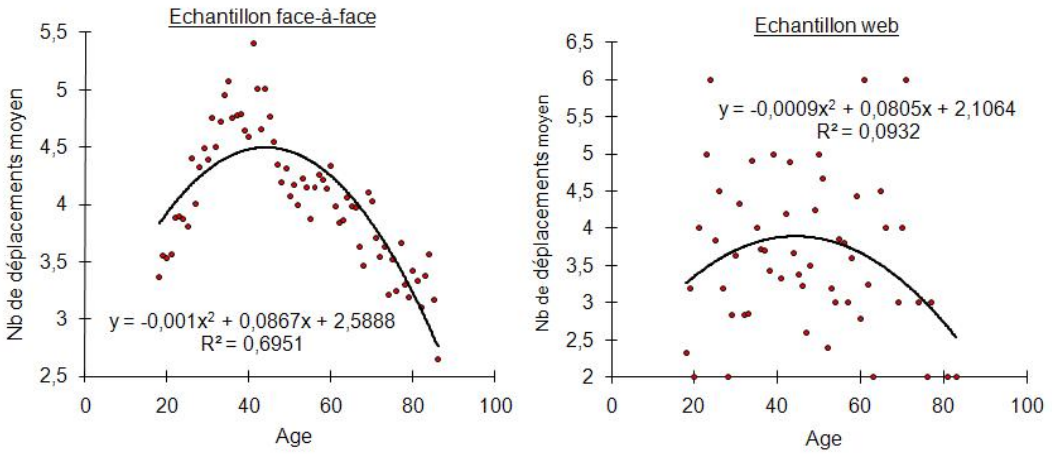

FIG. 60 – Relation entre l'âge des répondants et le nombre de déplacements moyen par enquête

Source : EMD Lyon (2006)

Si nous retenons les personnes âgées de 86 ans maximum dans l'échantillon face-à-face (pour avoir des effectifs suffisants) nous pouvons conclure que le nombre de déplacements augmente avec l'âge jusqu'à un maximum (situé entre 40 et 50 ans), puis décroît par la suite. Nous retrouvons alors une interprétation traditionnelle qui suppose qu'avec l'entrée dans la vie active et l'arrivée des enfants la mobilité s'accroît, mais qu'à partir d'un certain seuil, l'avancée en âge se traduit par une sédentarisation des individus. Cette relation est forte pour l'échantillon en face-à-face ($R^2 = 0,70$) Pour l'échantillon web, les effectifs par âge sont toutefois peu élevés et le nuage de point indique une relation moins significative entre l'âge et le nombre de déplacements ($R^2 = 0,09$) (figure 60).

Effet de la structure du ménage : un impact fort sur la mobilité Nous savons que la mobilité est influencée par la taille du ménage, mais l'impact du nombre de personnes du ménage sur la mobilité est ambigu.

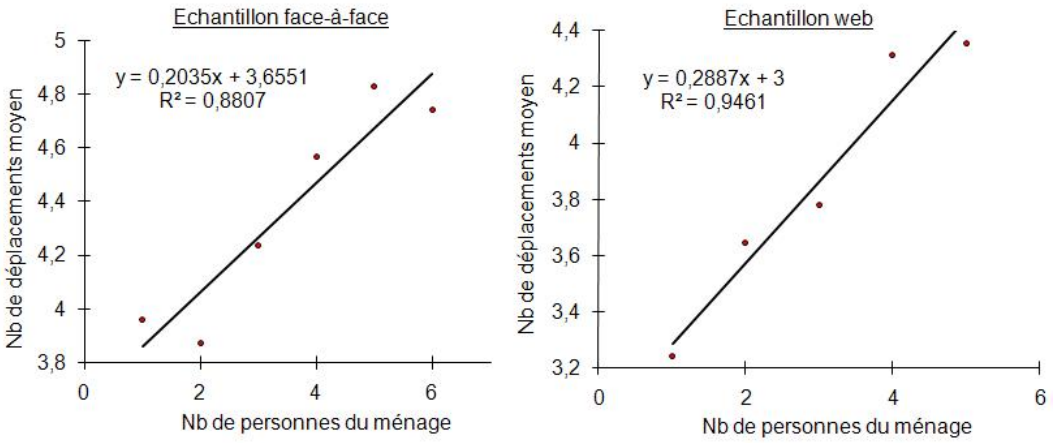

FIG. 61 – Relation entre le nombre de personnes du ménage et le nombre de déplacements moyen par enquête
Source : EMD Lyon (2006)

Si nous retenons les ménages de 1 à 6 personnes en face-à-face et de 1 à 5 personnes sur le web (pour les familles plus nombreuses, les effectifs sont trop faibles pour être considérés), alors le nombre de déplacements augmente avec le nombre de personnes du ménage. La relation est significative, quel que soit l'échantillon ($R^2 = 0,88$ en face-à-face et $R^2 = 0,95$ sur le web) (figure 61). L'augmentation de la taille du ménage s'accompagne donc d'un accroissement des besoins de mobilité. Cette hypothèse se retrouve dans les travaux de (Hanson, 1982), où l'augmentation de la taille des ménages va de pair avec l'augmentation du nombre de déplacements [126]. Nous pouvons toutefois poser l'hypothèse que les déplacements contraints sont mieux répartis lorsque la taille du ménage est importante.

Les programmes d'activités quotidiens déterminent les stratégies de mobilité des individus. Ces activités sont souvent obligatoires, se déroulent en des temps et lieux plus ou moins contraints, et sont déterminées en coordination avec les autres membres du ménage (accompagnement des enfants par exemple) (Kaufmann *et al.*, 2005; Orfeuil, 2000). Les enfants ont beaucoup d'activités qui ne sont pas liées au travail ou à l'école et qui nécessitent la participation d'un adulte de ménage. Les habitudes de déplacements des enfants influencent donc celles des parents et autres adultes du ménage. Par ailleurs, les enfants dont les parents sont actifs et ont un haut niveau de revenus ou d'éducation, ont davantage de chance de participer à des activités extérieures (Mohammadian et Bekhor, 2008).

Les analyses exploratoires laissent penser que le nombre de déplacements

[126] Selon Pouyanne (2004), au lieu de se contenter d'une influence directe de la composition familiale sur la mobilité, on peut adopter l'hypothèse que la composition familiale détermine un type de localisation (zones périphériques et peu denses de l'agglomération), qui vient influer sur les comportements de mobilité. Par ailleurs, les retraités, dont le niveau de mobilité est plus faible que celui observé pour l'ensemble de la population, sont sur-représentés dans les ménages de une à deux personnes. La corrélation entre la taille du ménage et la mobilité n'est donc pas forcément transformable en un lien causal direct.

augmente avec le nombre d'enfants du ménage jusqu'à un certain seuil (entre 2 et 4 enfants), puis décroît ensuite [127]. Cette relation est significative quel que soit l'échantillon considéré ($R^2 = 0,95$ en face-à-face et $R^2 = 0,81$ sur le web) (figure 62).

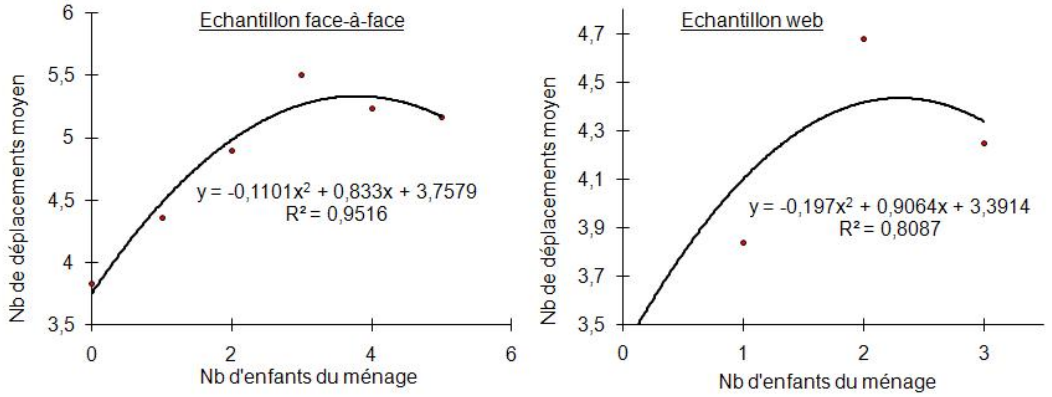

FIG. 62 – Relation entre le nombre d'enfants du ménage et le nombre de déplacements moyen par enquête

Source : EMD Lyon (2006)

Effet du niveau d'éducation : le niveau d'étude ou le diplôme acquis influencent positivement la mobilité Le niveau d'étude est lié au revenu et au statut et semble corrélé positivement avec la mobilité. Il en est de même pour la variable 'Diplômes' [128] (tableau 61). Cette différence est significative quel que soit le mode d'enquête (p-value $<$ 0,001%).

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Niveau d'étude : supérieur	4,53	3,95
Niveau d'étude : non supérieur	4,07	3,23
Niveau d'étude : en cours	3,54	4,43
P-value	$<$ 0,001%	$<$ 0,001%
Diplôme : supérieur	4,39	4,01
Diplôme : non supérieur	4,04	3,20
P-value	$<$ 0,001%	$<$ 0,001%

TAB. 61 – Nombre de déplacements moyen par niveau d'éducation et par enquête

Suivant les modèles utilisés, il conviendra d'utiliser l'une ou l'autre de ces variables. Il faut toutefois noter que la variable 'niveau d'étude' est partielle-

[127] Nous avons sélectionné les ménages de l'enquête en face-à-face de 5 enfants maximum, afin de garder des effectifs suffisants.

[128] La variable 'Diplômes' est plus synthétique que la variable 'niveau d'étude', puisque les individus en cours d'études se voient affecter la modalilité 'supérieur' ou 'non supérieur', en fonction du type de formation suivie.

ment reliée à l'âge, les personnes en cours d'études étant majoritairement des jeunes (moyenne d'âge de 20,7 ans en face-à-face et de 22,2 ans sur le web).

Effet du niveau d'activité : des inactifs davantage mobiles A priori, le niveau d'activité impacte positivement le temps passé sur le lieu de travail et le temps de déplacement quotidien total. En revanche, le temps disponible pour les activités dites de maintenance (achats, démarches ...) et de divertissement (loisirs) est plus restreint [129]. De nombreuses études mettent cependant en évidence une mobilité supérieure pour les actifs, ces derniers cumulant les déplacements liés au travail avec ceux nécessaires au déroulement de la vie quotidienne. L'analyse des données montre que les actifs internautes se déplacent moins que les inactifs, mais la relation est inversée dans le cas de l'enquête en face-à-face. Ces différences sont significatives, quel que soit le mode d'enquête (p-value $<$ 0,001%) (tableau 62). Elles peuvent s'expliquer en partie par la structure des échantillons. Dans l'enquête face-à-face, les retraités, peu mobiles, représentent une large proportion des inactifs. Ceci n'est pas vrai dans l'échantillon web, peu de répondants étant à la retraite. Les inactifs internautes sont donc en âge d'être "actifs" (présence d'enfants...) et ont un niveau de mobilité supérieur.

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Actifs	4,37	3,67
Inactifs	3,96	4,14
P-value	$<$ 0,001%	$<$ 0,001%

TAB. 62 – Nombre de déplacements moyen par niveau d'activité et par enquête

Effet de la localisation résidentielle : des disparités selon le mode d'enquête Les localisations résidentielles possibles pour un type de famille donné obéissent à un système de contraintes (accès aux transports collectifs, budget temps de transport quotidien, motorisation...). Il y a donc une forte dépendance entre la localisation résidentielle des ménages et leur forme de mobilité, relativement aux activités possibles (Orfeuil, 2002a; Ettema *et al.*, 1996) [130]. L'éloignement géographique du domicile par rapport au centre de l'agglomération joue a priori un rôle positif sur la mobilité, puisque la dispersion des activités en périphérie occasionne davantage d'accompagnements. Cependant, certains auteurs avancent que la mobilité est globalement supérieure au centre, en raison de l'importance de la marche à pied, et du rôle des transports en commun (Bonnafous et Puel, 1983).

[129] Les activités liées à l'accompagnement ne devraient pas souffrir du niveau d'activité, les actifs devant souvent gérer le transport de leurs enfants en plus de leur activité professionnelle, bien que les accompagnements peuvent être inclus dans des sorties complexes.

[130] L'impact de l'éloignement du lieu de résidence sur la mobilité peut néanmoins être modéré par l'accès à l'automobile. Ettema *et al.* (2007) précisent qu'en cas de bi-motorisation, les adultes sont moins dépendants et peuvent participer à davantage d'activités hors du domicile sans procéder à des arbitrages à l'intérieur du ménage.

Cette relation contradictoire entre la distance du domicile au centre de l'agglomération et le nombre de déplacements déclaré semble se vérifier dans les échantillons web et face-à-face. Dans l'échantillon en face-à-face, la mobilité des individus diminue, jusqu'à un éloignement de la zone de résidence de 10km du centre de l'agglomération, et augmente ensuite ($R^2 = 0,72$). Nous constatons le contraire dans l'échantillon web ($R^2 = 0,90$). Il est intéressant de noter que, dans les deux cas, cette relation n'est pas linéaire, et s'inverse pour un seuil de 10km (figure 63).

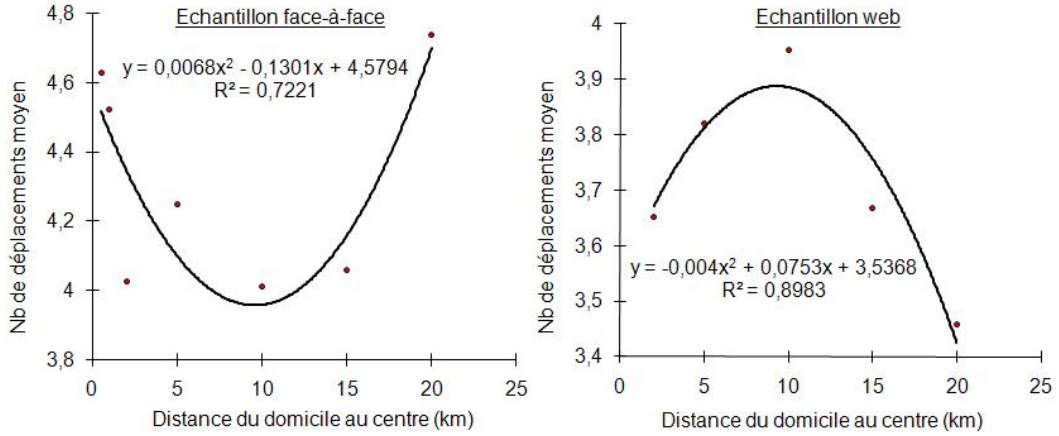

FIG. 63 – Relation entre la distance du domicile au centre de l'agglomération (km) et le nombre de déplacements moyen par enquête
Source : EMD Lyon (2006)

De façon analogue, la densité de la zone d'habitation conditionne la richesse des services et la qualité des transports publics qui peuvent être associés. Si certaines études montrent que le nombre de déplacements a tendance à augmenter quand la densité est faible et à diminuer lorsqu'elle est plus forte, il est cependant probable que les déplacements courts soient sous-comptabilisés dans les zones les plus peuplées, surtout s'ils font partie d'une chaîne de déplacements complexe (Banister, 1997).

Effet de l'équipement du ménage : le rôle non négligeable du téléphone portable Concernant les équipements du ménage, la possession d'un téléphone portable semble liée positivement à la mobilité. Cet équipement permet en effet de rester en relation avec de nombreux contacts personnels, et les personnes avec un haut niveau d'interaction sociale ont tendance à se déplacer davantage. Par ailleurs, dans la sphère professionnelle, le téléphone portable est souvent confié aux salariés ayant à se déplacer dans le cadre de leur mission. Cette relation semble se vérifier de façon significative dans les deux échantillons (tableau 63).

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Téléphone portable : oui	4,31	3,83
Téléphone portable : non	3,84	3,75
P-value	$< 0{,}001\%$	$< 0{,}001\%$

TAB. 63 – Nombre de déplacements moyen par possession d'un téléphone portable et par enquête

Selon Hjorthol (2002), l'accès et l'utilisation de l'ordinateur à la maison n'a pas, a priori, d'effet direct sur la mobilité. Les personnes équipées d'une connexion web à domicile font moins de déplacements pour le travail, mais reprogramment leurs activités et conservent au final le même nombre de déplacements quotidiens. Nous ne prenons pas non plus en compte les variables 'liste de téléphone' et 'connexion internet' dans le modèle explicatif de la mobilité. Nous verrons dans la section suivante que ces variables influencent davantage le choix du web comme média de réponse que le niveau de mobilité individuelle.

Effet de la motorisation : un accès à l'automobile qui encourage la mobilité Hanson (1982) montre que le niveau de motorisation individuel (nombre de voitures du ménage par personne de 18 ans et plus) et le fait de posséder le permis de conduire sont corrélés positivement à la mobilité. Les personnes qui ne possèdent pas le permis ou qui n'ont pas de véhicule à disposition sont en effet plus contraintes dans leurs déplacements, même si d'autres facteurs peuvent intervenir. Ces relations se vérifient dans les deux échantillons. L'impact de la possession du permis sur la mobilité semble toutefois moins important dans l'échantillon web (tableau 64).

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Possession permis : oui	4,30	3,81
Possession permis : non	3,47	3,34
P-value	$< 0{,}001\%$	$< 0{,}001\%$

TAB. 64 – Nombre de déplacements moyen par possession du permis de conduire et par enquête

En ce qui concerne le nombre de voitures à disposition, la relation est très forte dans l'échantillon face-à-face ($R^2 = 0,72$), mais moins prononcée pour l'échantillon web ($R^2 = 0,32$), probablement à cause des plus faibles effectifs (figure 64). La variable 'Nombre de voitures par personne de 18 ans et plus' est traditionnellement liée au niveau de revenu du ménage. Cette variable est donc intéressante, puisque si le revenu est un déterminant majeur de la mobilité (Pouyanne, 2004), le fort taux de non-réponse partielle dans les questionnaire web et face-à-face ne nous permet pas de la prendre en considération.

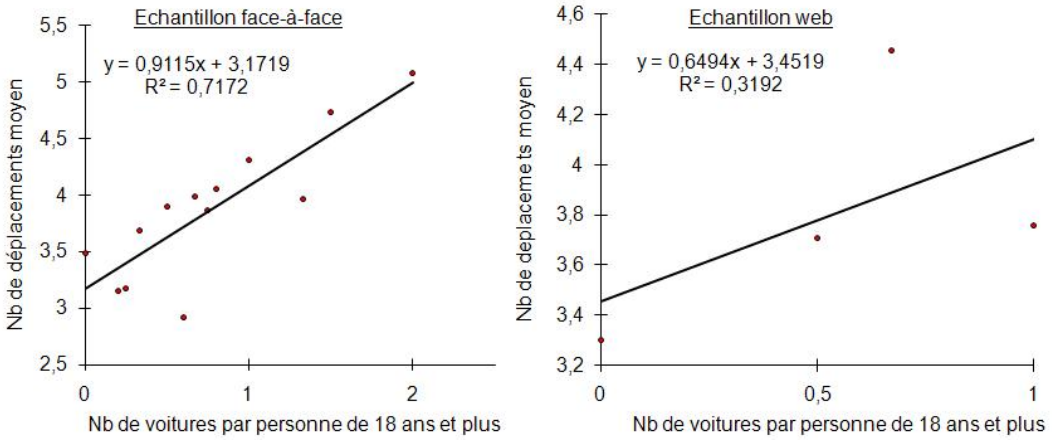

FIG. 64 – Relation entre le nombre de voitures du ménage par personne de 18 ans et plus et le nombre de déplacements moyen par enquête
Source : EMD Lyon (2006)

Le jour des déplacements et le fait de déclarer le niveau de revenu du ménage : deux facteurs explicatifs de la mobilité Nous pouvons supposer que la réduction du temps de travail profite aux autres activités et génère davantage de déplacements pour motifs achats et loisirs. Nos premiers tris confirment cette hypothèse, et montrent un accroissement des déplacements le vendredi. Cette relation est très significative dans l'enquête en face-à-face, mais semble s'inverser dans l'enquête web (tableau 65). Ceci peut s'expliquer par le fait que les internautes qui choisissent de saisir leurs réponses le week-end ou le lundi peuvent avoir oublié leur emploi du temps du vendredi, jour de référence des déplacements.

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Vendredi : oui	4,40	3,68
Vendredi : non	4,15	3,84
P-value	$< 0{,}001\%$	$< 0{,}001\%$

TAB. 65 – Nombre de déplacements moyen effectué le vendredi par enquête

Par ailleurs, nous observons que la mobilité est plus importante si la personne choisit de communiquer le niveau de revenus annuels nets du ménage (p-value $< 0{,}001\%$). Il est probable que ces répondants doivent faire face à moins de freins pour délivrer des réponses précises concernant leurs déplacements (bien que le fait de dévoiler le niveau de salaire ne soit pas une variable directement liée au niveau de mobilité). Nous vérifions cette relation dans les deux échantillons (tableau 66), les personnes acceptant de déclarer leurs revenus étant plus mobiles que les autres.

Modalités	Nb de déplacements moyen	
	Enquête face-à-face	Enquête web
Revenus déclarés : oui	4,32	3,83
Revenus déclarés : non	3,94	3,61
P-value	$< 0{,}001\%$	$< 0{,}001\%$

TAB. 66 – Nombre de déplacements moyen effectué par enquête, selon que l'individu déclare ou non son niveau de revenus

III.2.2 Le problème de la colinéarité des variables explicatives

L'existence d'une colinéarité [131] entre les variables explicatives d'un modèle de régression linéaire est un problème fréquent. Il est donc nécessaire de s'assurer que les variables explicatives d'une régression multiple ne se répètent pas entre elles, au risque d'aboutir à un modèle instable (des coefficients peuvent avoir un signe opposé à celui attendu), et donc peu utile (Greene, 2002). Pour que celui-ci soit robuste, une des solutions consiste à étudier la matrice des corrélations. Il s'agit plus simplement de la matrice des coefficients de corrélation, calculés sur plusieurs variables prises deux à deux. Si cette dernière indique une très forte liaison entre deux variables explicatives (corrélation positive ou négative), c'est que l'une d'elles est de trop et ne peut être intégrée dans le modèle. Le choix qui consiste à réduire le nombre de variables explicatives d'un modèle pour éviter de rencontrer des problèmes de multicolinéarité n'est cependant pas toujours judicieux, et peut amener à mal spécifier le modèle [132]. L'omission de variables explicatives pertinentes peut en effet biaiser les résultats, puisqu'en surestimant la variance, il est possible de conclure à tort à la significativité de certains coefficients, qui ne le seraient pas si les variables omises avaient été intégrées (Greene, 2002).

Le coefficient de corrélation de Pearson Dans le cadre d'une régression linéaire simple, on 'résume' un nuage de points par une droite, dite de régression. On mesure la qualité de la régression simple par le coefficient de corrélation linéaire de Pearson, rapport de la covariance entre x et y au produit des écarts-types empiriques. Soit :

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{58}$$

Ce coefficient est compris entre -1 et 1. Un signe négatif indique que y varie en sens inverse de x et plus le coefficient est proche de 0, plus les deux variables sont linéairement indépendantes. Le niveau de précision du risque d'erreur sur l'estimation de la corrélation dépend du nombre de degrés de liberté du modèle (Saporta, 2006).

[131] Les vecteurs u et v sont colinéaires s'il existe un réel k tel que u = kv.

[132] Une solution consiste à construire des variables synthétiques, qui agrègent plusieurs variables fortement corrélées.

Quelques exemples de colinéarité des variables explicatives La multicolinéarité est toujours plus ou moins présente dans les jeux de données et la plupart des variables issues de l'enquête ménages déplacements que nous avons retenues pour l'analyse de la mobilité sont de près ou de loin corrélées entre-elles. Par exemple, l'éloignement du lieu de résidence 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...). Pouyanne (2004), constate par ailleurs que la corrélation positive entre le niveau de diplôme et le nombre de déplacements est le fruit de l'effet localisation des diplômés du supérieur dans des zones denses (pour profiter des aménités ou par volonté de se rapprocher des fonctions dites 'supérieures', davantage présentes dans l'hypercentre), où la mobilité est supérieure.

Parmi les variables explicatives sélectionnées dans notre modèle, deux associations semblent particulièrement délicates : la densité de la zone de résidence et son éloignement du centre de l'agglomération (coefficient de Pearson = -0,594 ; p-value $<$ 0,001%), d'une part, le nombre de personnes du ménage et le nombre d'enfants du ménage (coefficient de Pearson = 0,784 ; p-value $<$ 0,001%), d'autre part.

Il existe a priori une colinéarité entre la densité de la zone de résidence et son éloignement par rapport au centre de l'agglomération. La forme de la relation reste ambigüe et ne peut être considérée comme linéaire (le coefficient de détermination n'est pas très important : 0,35 en face-à-face et 0,41 sur le web). Selon les figures 65 et 66, l'élasticité entre la densité de la zone de résidence et son éloignement du centre de l'agglomération n'est pas constante. Quelle que soit l'enquête, la densité de la zone de résidence décroît fortement pour un éloignement du centre de l'agglomération compris entre 0 et 3 kilomètres, puis semble se stabiliser au-delà.

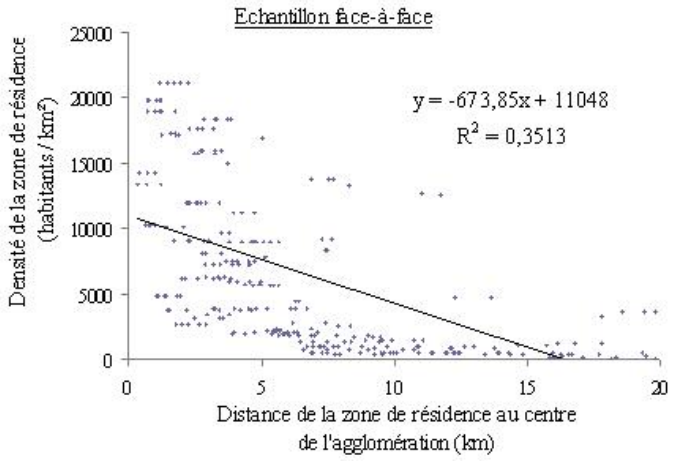

FIG. 65 – Relation entre la distance du domicile au centre de l'agglomération et la densité de la zone de résidence pour l'échantillon face-à-face
Source : EMD Lyon (2006)

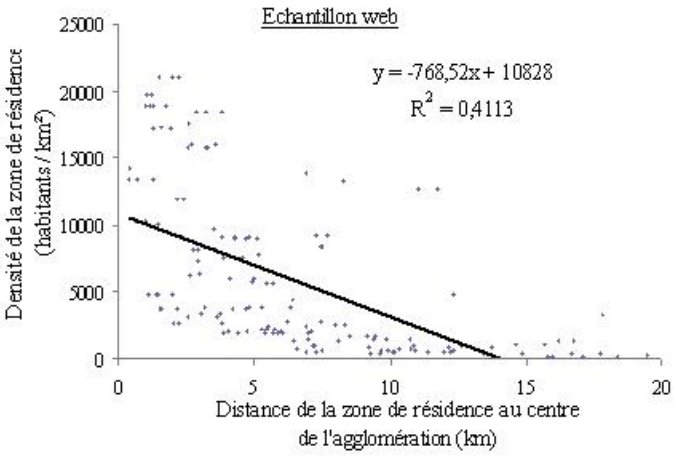

FIG. 66 – Relation entre la distance du domicile au centre de l'agglomération et la densité de la zone de résidence pour l'échantillon web
Source : EMD Lyon (2006)

Par ailleurs, comme le montre la figure 67, les variables 'Nombre de personnes du ménage' et 'Nombre d'enfants de moins de 18 ans du ménage' sont fortement corrélées (R^2 proche de 1 pour les deux échantillons). Cette corrélation peut poser un vrai problème de multicolinéarité dans le cadre des modèles de régression statistique. Suivant les échantillons concernés et les analyses réalisées, nous utiliserons donc l'une ou l'autre de ces variables.

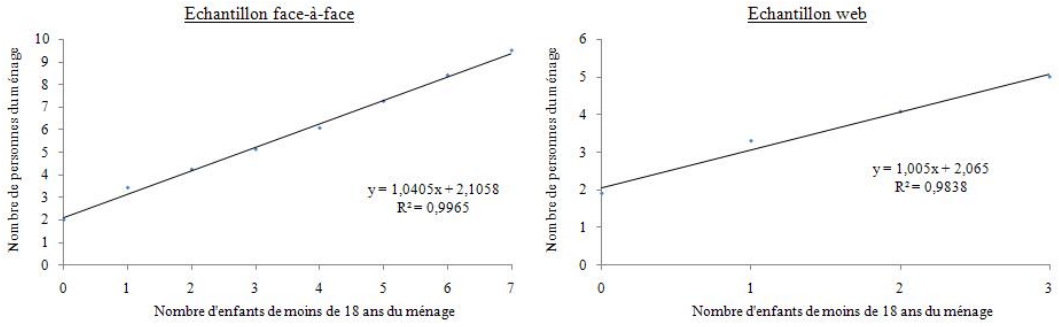

FIG. 67 – Relation entre le nombre d'enfants et le nombre de personnes du ménage
Source : EMD Lyon (2006)

III.2.3 Présentation des résultats du modèle de régression linéaire multiple

Nous présentons les variables explicatives retenues par le modèle, les paramètres estimés, leur écart-type et leur niveau de significativité, ainsi que la valeur du t de Student.

Ensemble des répondants (face-à-face et web) L'introduction des variables explicatives désignées ci-dessus dans le modèle de régression multiple

appliqué à l'ensemble de l'échantillon (web et face-à-face) donne les résultats présentés dans le tableau 67.

Toutes les variables présentées sont significatives à 1% et les coefficients ont un signe conforme à nos prévisions (section 3.2.1). Nous remarquons l'influence particulièrement forte de l'âge, du nombre de voitures à disposition, du niveau d'activité, du nombre d'enfants du ménage et du mode d'enquête. Le coefficient de la variable 'Mode' est négatif (-0,66), ce qui signifie qu'en moyenne un internaute déclare effectuer quotidiennement 0,7 déplacements de moins qu'un répondant en face-à-face possédant les mêmes caractéristiques socio-économiques.

Echantillon total	Coeff.	T-value	Pr(>|t|)	Sign.
Constante	2,45e+00	10,91	< 2e-16	***
Sexe : homme	-2,55e-01	-5,60	2,24e-08	***
Age	4,81e-02	5,49	4,05e-08	***
Age^2	-5,79e-04	-6,55	5,85e-11	***
Possession permis : oui	3,76e-01	5,01	5,45e-07	***
Vendredi : oui	2,01e-01	3,26	<0,01	**
Nb d'enfants / ménage	5,.06e-01	8,98	< 2e-16	***
(Nb d'enfants / ménage)^2	-3,47e-02	-2,04	0,04	*
Nb de voiture / personne	6,05e-01	8,85	< 2e-16	***
Téléphone portable : oui	1,36e-01	2,39	0,02	*
Revenu déclaré : oui	3,29e-01	6,95	3,95e-12	***
niveau d'étude : en cours	-8,26e-01	-7,03	2,23e-12	***
niveau d'étude : non supérieur	-1,65e-01	-3,19	<0,01	**
Mode : web	-6,59e-01	-4,61	4,00e-06	***
Activité : non actif	4,86e-01	7,67	1,85e-14	***
Distance domicile / centre	-1,15e-01	-7,06	1,81e-12	***
(Distance domicile / centre)^2	6,01e-03	7,49	7,60e-14	***

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 67 – Régression linéaire appliquée à l'ensemble de l'échantillon (web et face-à-face)

Critères	Valeur
R2	9,7%
F-stat	67,78
ddl	9946
P-value	< 2,2e-16

TAB. 68 – Significativité de la régression linéaire appliquée à l'ensemble de l'échantillon (web et face-à-face)

Seule 10% de la variance de la variable 'Nombre de déplacements' est expliquée par l'ensemble des variables explicatives qui interviennent dans le modèle (tableau 68). Il est donc fort probable que le niveau de mobilité individuelle soit influencé par d'autres variables, non connues de l'analyste. Cette faible qualité s'explique également par le caractère désagrégé de notre modèle et

surtout par une forte variabilité intra personnelle selon le jour d'enquête. Ce modèle est cependant significatif (p-value < 0,001%).

Echantillon face-à-face Le modèle appliqué au sous-échantillon de l'enquête face-à-face est quasiment le même que le précédent, comme le montre le tableau 69. Les signes des coefficients sont identiques et leurs valeurs absolues sont très proches. Cette similitude entre les deux modèles était prévisible, puisque les observations recueillies en face-à-face représentent 97% du total de notre échantillon. Nous avons pris soin d'enlever la variable 'mode', qui n'a plus de sens ici, tous les individus ayant répondu en face-à-face.

Echantillon face-à-face	Coeff.	T-value	Pr(>|t|)	Sign.
Constante	2,45e+00	10,78	<2e-16	***
Sexe : homme	-2,43e-01	-5,25	1,57e-07	***
Age	4,89e-02	5,52	3,46e-08	***
Age^2	-5,73e-04	-6,42	1,44e-10	***
Possession permis : oui	3,71e-01	4,90	9,82e-07	***
Vendredi : oui	2,13e-01	3,38	<0,01	***
Nb d'enfants / ménage	5,04e-01	8,81	<2e-16	***
(Nb d'enfants / ménage)^2	-3,39e-02	-1,97	0,05	*
Nb de voiture / personne	6,12e-01	8,81	<2e-16	***
Téléphone portable : oui	1,37e-01	2,37	0,02	*
Revenu déclaré : oui	3,32e-01	6,92	4,73e-12	***
niveau d'étude : en cours	-8,28e-01	-2,89	<0,01	**
Activité : non actif	4,74e-01	7,39	1,64e-13	***
Distance domicile / centre	-1,19e-01	-7,15	9,50e-13	***
(Distance domicile / centre)^2	6,22e-03	7,63	2,60e-14	***
Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1				

TAB. 69 – Régression linéaire appliquée à l'échantillon face-à-face

De même que précédemment, environ 10% de la variance de la variable 'Nombre de déplacements' est expliquée par l'ensemble des variables explicatives qui interviennent dans le modèle (tableau 70). Le niveau de significativité du modèle est en revanche très élevé (p-value < 2,2e-16).

Critères	Valeur
R2	9,7%
F-stat	70,56
ddl	9694
P-value	< 2,2e-16

TAB. 70 – Significativité de la régression linéaire appliquée à l'échantillon face-à-face

Echantillon web Beaucoup moins de variables ont des coefficients significatifs dans le modèle de régression appliqué à l'échantillon web (tableau 71).

Ceci s'explique essentiellement par les faibles effectifs, puisque seulement 369 individus ont rempli entièrement le questionnaire sur internet.

Comme nous l'avions supposé dans la section 3.2.1, la taille du ménage, l'inactivité, le niveau de formation et plus encore le nombre de voitures à disposition impactent positivement la mobilité des internautes. En revanche, le fait d'être un homme ou de résider loin du centre-ville pèse sur le nombre de déplacements déclaré. Le test de la variable 'Nombre de personnes' montre l'existence d'un lien très significatif entre la taille du ménage et la mobilité (p-value $<$ 0,001%). Le niveau de formation et d'activité du répondant, ainsi que l'éloignement de son lieu de résidence sont des facteurs moins importants (significativité des coefficients plus faible : p-value $<$ 10%). Etant donné les faibles effectifs de l'échantillon web et le risque d'erreur acceptable, nous les conservons tout de même dans notre modèle.

Echantillon web	Coeff.	T-value	Pr(>|t|)	Sign.
Constante	2,51	5,71	3,03e-08	***
Sexe : homme	-0,79	-3,30	<0,01	**
Nb de voiture / personne	0,74	2,02	0,04	*
Nb de personnes / ménage	0,34	3,59	<0,01	***
Diplômes : supérieur	0,46	1,65	0,10	.
Activité : non actif	0,48	1,75	0,08	.
Distance domicile / centre	-0,05	-1,73	0,08	.

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 71 – Régression linéaire appliquée à l'échantillon web

Toutes les variables présentées sont significatives à 10% et le coefficient de détermination ajusté du modèle est légèrement meilleur (proche de 11%). Le modèle est toujours très significatif (p-value = 3,07e-06).

Critères	Valeur
R2	10,63%
F-stat	6,33
ddl	263
P-value	3,07e-06

TAB. 72 – Significativité de la régression linéaire appliquée à l'échantillon web

III.2.4 Premières conclusions

Nous avons analysé de manière empirique la relation entre la mobilité, définie par le nombre de déplacements déclaré, et un certain nombre de facteurs explicatifs (sociodémographiques, équipement du ménage, motorisation et mode d'enquête), conformément à l'équation (57). Nous pouvons cependant nous demander si le coefficient α de la variable 'Mode' 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 identique 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 leurs réponses en ligne, selon certaines caractéristiques, qui peuvent également avoir un impact sur leur niveau de mobilité. Par exemple, 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'une connexion internet pour accéder à l'enquête en ligne. En excluant systématiquement des individus 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). L'auto-sélection des répondants doit donc être corrigée, afin d'obtenir des estimations non biaisées des coefficients des variables du modèle explicatif de la mobilité.

IV Modèle explicatif incluant le biais de sélection

Nous avons détaillé en section II les fondements économétriques des modèles de régression multivariés qui tiennent compte du biais de sélection. Dans cette section, nous analysons plus précisément les résultats des estimations du modèle d'Heckman en deux étapes, menées sur l'échantillon des répondants à l'enquête ménages déplacements de Lyon (face-à-face et web). Dans la première étape du modèle d'Heckman, nous estimons pour chaque répondant la probabilité de ne pas participer à l'interview en face-à-face et d'utiliser le web pour remplir le questionnaire, à l'aide d'un modèle probit (section III.1). Puis, dans une seconde étape, on estime l'équation de mobilité par la méthode des moindres carrés ordinaires, en incorporant l'espérance conditionnelle aux variables de conditionnement des résidus de la première étape (section III.3). Enfin, nous chercherons à formuler un modèle stable, qui tienne compte des éventuelles interactions entre le mode d'enquête et les variables socio-économiques (section III.4)

IV.1 Première étape : équation de sélection

Nous présentons d'abord les fondements théoriques des modèles probit (section 4.1.1). Puis, nous commentons les résultats des estimations effectuées sur notre échantillon (section 4.1.2), avant de mener une analyse de sensibilité (section 4.1.3).

IV.1.1 Le modèle probit

A l'aide d'un modèle probit, nous cherchons à modéliser le comportement d'individus, qui choisissent ou pas de répondre en face-à-face et, dans la négative, de remplir ou non le questionnaire web. La variable à expliquer y_i est

une variable binaire, qui peut prendre deux valeurs : 0 si l'individu répond en face-à-face ou 1 s'il répond sur le web. Soit U_{1i} l'utilité de l'individu i s'il refuse de recevoir un enquêteur à domicile et choisit de répondre en ligne, et U_{0i} son utilité s'il choisit de répondre en face-à-face. Ces niveaux d'utilité ne sont pas directement observés, mais dépendent de caractéristqiues socio-économiques et d'équipements en moyens de communication des ménages. Supposons que les niveaux d'utilité U_{1i} et U_{0i} sont expliqués par un modèle linéaire, dont les variables explicatives sont x_{1i} et x_{2i}, avec ϵ_{1i} et ϵ_{0i} des aléas, tel que :

$$U_{1i} = \alpha_1 + \alpha_{11}x_{1i} + \alpha_{21}x_{2i} + \epsilon_{1i} \quad (59)$$
$$U_{0i} = \alpha_0 + \alpha_{10}x_{1i} + \alpha_{20}x_{2i} + \epsilon_{0i} \quad (60)$$

L'écart entre U_{0i} et U_{1i} est une variable latente, en fonction de laquelle la décision de répondre ou pas sur le web est prise. La règle de décision est la suivante : l'individu choisit de répondre sur le web si U_{1i} est supérieure à U_{0i}. Soit P_i la probabilité pour un répondant de saisir ses réponses en ligne :

$$P_i = P(y_i = 1) = P(U_{0i} < U_{1i}) \quad (61)$$
$$P_i = P(\epsilon_{0i} - \epsilon_{1i}) \leq (\alpha_1 - \alpha_0) + x_{1i}(\alpha_{11} - \alpha_{10}) + x_{2i}(\alpha_{21} - \alpha_{20}) \quad (62)$$
$$P_i = P(\epsilon_i \leq x_i^{'}\beta) = F(x_i^{'}\beta) \quad (63)$$

avec F la fonction de répartition, ϵ_i un aléa égal à $\epsilon_{0i} - \epsilon_{1i}$, $x_i^{'}$ le vecteur des variables explicatives et β le vecteur des paramètres à estimer, tel que :

$$\beta = \begin{pmatrix} \alpha_1 - \alpha_0 \\ \alpha_{11} - \alpha_{10} \\ \alpha_{21} - \alpha_{20} \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

Dans les modèles désagrégés, il est nécessaire de faire des hypothèses sur les termes résiduels. Les modèles probabilistes les plus usuels sont les modèles Logit et Probit. Le modèle Logit impose des hypothèses très restrictives sur la répartition des ϵ, mais sa formulation mathématique est simple. A l'inverse, le modèle Probit ne nécessite aucune hypothèse sur la répartition des termes d'erreur, mais reste un modèle complexe lorsque le nombre d'alternatives est supérieur à deux [133].

Le modèle de choix est binaire, puisque nous cherchons à modéliser la probabilité qu'un individu réponde sur le web. La décision est représentée par la variable y_i, qui ne peut prendre que deux valeurs. Soit :

[133]Dans le modèle probit, les résidus suivent une distribution normale, alors que dans le modèle logit, ils suivent une distribution de Gumbel. Ces distributions sont proches dans leur forme, et les modèles probit et logit donnent généralement des résultats similaires. Cependant, les valeurs estimées des paramètres ne sont pas directement comparables, puisque les variances des lois logistiques et normales ne sont pas identiques.

$$P_i = P(y_i = 1) = F(x_i^{'}\beta) \tag{64}$$

La fonction F est une fonction de répartition, et on note f sa fonction de densité. Dans un modèle probit, on suppose que la fonction de répartition est une loi normale[134]. On note Φ la fonction de répartition et ϕ la fonction de densité de la loi normale centrée réduite. La probabilité P_i devient :

$$P_i = \int_{-\infty}^{x_i^{'}\beta} \frac{1}{2\pi} e^{-\frac{t^2}{2}} dt = \Phi(x_i^{'}\beta) \tag{65}$$

Les valeurs des coefficients β sont celles qui maximisent la probabilité qu'un événement survienne (refus du face-à-face puis choix du web pour répondre au questionnaire). L'estimation est faite par la méthode du maximum de vraisemblance, dont le principe est assez simple. Pour chaque individu, il est possible de calculer la probabilité de répondre par un média donné (web ou face-à-face), et de connaître les paramètres de la fonction qui renvoie ces probabilités. Pour une observation donnée, la vraisemblance est la probabilité que le modèle renvoie le bon résultat. La méthode du maximum de vraisemblance consiste à trouver pour un modèle donné, la valeur des paramètres qui a le plus de chances de conduire à la situation observée. On va chercher la probabilité d'obtenir le bon résultat pour l'ensemble de notre échantillon, en formant le produit des probabilités individuelles. Il faut ensuite estimer les coefficients de la fonction qui maximise cette vraisemblance. La fonction de vraisemblance est donnée par :

$$L = P(Y_1 = y_1, \ldots, Y_n = y_n) \tag{66}$$

Les variables aléatoires Y_i étant distribuées de manière indépendantes, cette fonction peut s'écrire :

$$L = \Pi_{i=1}^{n} (P_i)^{y_i} (1 - P_i)^{1-y_i} \tag{67}$$

Car y_i prend la valeur 1 avec une probabilité P_i et la valeur 0 avec une probabilité $1 - P_i$. L'expression logarithmique de la fonction de vraisemblance est donnée par :

$$Log(L) = \sum_{i=1}^{n} y_i * Log(P_i) + (1 - y_i) * Log(1 - P_i) \tag{68}$$

$$Log(L) = \sum_{i=1}^{n} y_i * Log[F(x_i^{'}\beta)] + (1 - y_i) * Log[1 - F(x_i^{'}\beta)] \tag{69}$$

Le vecteur des paramètres β est obtenu en maximisant la fonction Log(L) par rapport à β. Il est possible d'appliquer les tests standard de contraintes sur les paramètres, et de tester la significativité du modèle avec le test du ratio

[134]Dans le modèle logit, on suppose que la fonction de répartition est une loi logistique

de vraisemblance [135]. La mesure de la qualité de l'ajustement se fait au niveau global, par une analyse du R^2 [136] et au niveau de chaque coefficient par une comparaison entre les choix prédits et les choix réels.

IV.1.2 Estimation du modèle de sélection des répondants

Deux types de variables sont introduites dans notre modèle en qualité de variables explicatives : des caractéristiques sociodémographiques (âge, occupation, nombre de personnes du ménage, diplôme, lieu de travail ...), et 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 73.

	Coefficient	Ecart-type	Pr(>|z|)	Signif.
Constante	-3,71	3,47e-01	<2e-16	***
Age	0,05	1,44e-02	<0,001	***
Age^2	-4,78e-04	1,62e-04	<0,01	**
Connexion internet : oui	0,49	8,65e-02	1,03e-08	***
Téléphone portable : oui	0,26	9,19e-02	<0,01	**
Liste téléphone : oui	0,37	6,60e-02	1,68e-08	***
Liste téléphone : pas de téléphone	0,25	1,18e-01	0,04	*
Nb de personnes / ménage	-0,10	2,49e-02	5,10e-05	***
Lieu de travail : non précisé	-0,99	3,64e-01	<0,01	**
Lieu de travail : périphérie	-0,24	7,39e-02	<0,01	**
Lieu de travail : inactif	-0,24	8,80e-02	<0,01	**
Diplôme : supérieur	0,41	6,67e-02	8,16e-10	***
Densité de la zone de résidence	-2,17e-05	6,68e-06	<0,01	**
Revenu déclaré : oui	0,40	7,54e-02	1,04e-07	***
Vendredi : oui	0,47	6,66e-02	1,89e-12	***
Distance domicile / centre	-1, 90e-05	7,90e-03	0,01	*

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 73 – Equation de sélection - Modèle probit

Les variables retenues dans le modèle sont significatives à 5% (probabilité $< 5\%$ de rejeter à tort 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 [137]. L'ensemble des signes des coefficients sont ceux que nous espérions.

[135] $LR = -2(LogL(\beta^*_{MV}) - LogL(\beta_{MV}))$, avec $LogL(\beta^*_{MV})$ la valeur de la fonction LogL pour un modèle non contraint (contenant l'ensemble des variables explicatives) et $LogL(\beta_{MV})$ la valeur de la fonction LogL pour un modèle contraint (suppression d'une ou plusieurs variables explicatives)

[136] Suggéré par Mc Fadden

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

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). 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 abonné à France Télécom, donc de ne plus figurer sur l'annuaire. Il est probable que les internautes affectionnent les nouvelles technologies, et certains ne disposent plus de ligne fixe, mais d'un téléphone mobile uniquement. 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 de diplôme 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 familiarisées 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 à internet sur leur lieu de travail, 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.

Du fait de la colinéarité entre les deux variables 'Densité de la zone de résidence' et 'Distance du domicile par rapport au centre', leur introduction simultanée pourrait rendre les coefficients instables. Cependant, la corrélation existant entre ces variables n'affecte pas le résultat, puisque si on exclut la variable 'Distance du domicile par rapport au centre', le coefficient attaché à la variable 'Densité de la zone de résidence' reste significatif et l'impact marginal est identique. Le résultat n'est donc pas dû à des problèmes de multicolinéarité, puisque l'omission d'une variable caractérisant le répondant ne le modifie pas.

Critères	Valeur
R2	15,2%
AIC	1994,2
ddl	9941

TAB. 74 – Indicateurs de la qualité du modèle de sélection

Le calcul du coefficient de détermination ajusté du modèle (R^2) [138] nous renvoie la valeur 0,152 (tableau 74). 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.

IV.1.3 Analyse de sensibilité

L'estimation d'un modèle probit permet d'évaluer l'influence des différentes variables explicatives sur la survenue de l'événement à expliquer [139]. Afin d'étudier le choix du mode de réponse à l'enquête, nous allons nous intéresser à quelques profils types d'individus. L'objectif est de savoir, pour chaque profil, quelle est la variation de probabilité de l'événement Y_i=1 (réponse en ligne), en cas de variation d'une des variables exogènes. Nous considérons pour ces simulations que le jour de référence des déplacements n'est pas le vendredi.

- Profil 1 : individu d'âge moyen (46 ans) n'ayant pas fait d'études supérieures, appartenant à un ménage de deux personnes, travaillant au centre de l'agglomération, ne possédant pas de connexion internet ni de ligne fixe, mais un téléphone portable, n'ayant pas déclaré son niveau de revenu annuel et résidant à 2 kilomètres du centre de l'agglomération, dans une ville de densité moyenne.

- Profil 2 : individu d'âge moyen (46 ans) ayant fait des études supérieures, appartenant à un ménage de quatre personnes, travaillant au centre de l'agglomération, possédant une connexion internet et un téléphone portable, inscrit sur la liste rouge ou orange ayant déclaré son niveau de revenu annuel et résidant à 5 kilomètres du centre de l'agglomération dans une ville de densité moyenne.

Les probabilités de répondre sur le web ou en face-à-face de ces différents profils sont présentées dans le tableau 75.

[138] Le coefficient de détermination est un indicateur qui permet de juger la qualité d'une régression linéaire, simple ou multiple. D'une valeur comprise entre 0 et 1, il mesure l'adéquation entre le modèle et les données observées. Son défaut étant de croître avec le nombre de variables explicatives, on s'intéresse davantage au coefficient de détermination ajusté.

[139] Les effets marginaux ne sont plus égaux aux valeurs des paramètres, comme c'est le cas dans les modèles linéaires.

TAB. 75 – Probabilités de choix du mode de réponse

	Web	Face-à-face
Profil 1	9%	91%
Profil 2	43%	57%

Les personnes correspondant au profil 2 utilisent davantage le web pour répondre au questionnaire que celles correspondant au profil 1. Ces dernières n'ont pas de connexion internet au domicile et n'ont pas effectué d'études supérieures, ce qui réduit leur degré probable d'aisance avec la navigation sur le web, ainsi que leur faculté à se connecter une vingtaine de minutes pour remplir un questionnaire en ligne.
Si l'on change le profil 1, pour considérer les individus célibataires, alors la probabilité de répondre en ligne atteint 11% : la taille du ménage impacte négativement la probabilitré de répondre sur internet. Il en est de même pour le niveau d'étude, puisque le fait de prendre les personnes n'ayant pas effectué d'études supérieures dans le profil 2 donne une probabilité de répondre en ligne de 28%. Mais ce sont les variables d'équipement en moyens de communication qui ont l'effet le plus important. Ainsi, si les individus du profil 2 ne disposent plus d'une connexion internet à domicile, la part du web chute à 25%.

L'âge est une variable quantitative continue, dont la valeur moyenne dans l'échantillon est égale à 46 ans. Il est intéressant de connaitre l'influence marginale de cette variable sur chaque profil, c'est-à-dire : de combien augmente la probabilité de répondre sur le web si l'âge augmente de dix ans ? Le coefficient de l'âge est sensiblement positif. En revanche, son terme quadratique, qui permet de rendre compte de la relation non linéaire de l'âge avec la probabilité de répondre sur le web est négatif. Une croissance de l'âge jusqu'à 51 ans augmente la probabilité de répondre au questionnaire en ligne, mais l'effet s'inverse au-delà de cette limite. Ainsi, une diminution de l'âge de 1 an diminue la probabilité de répondre sur le web de 1 point en ce qui concerne le profil 1 (P=8%) et de 2 points en ce qui concerne le profil 2 (P=41%). Nous allons à présent nous intéresser à la seconde étape du modèle d'Heckman.

IV.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).

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 donc réintroduites dans la seconde étape que si elles semblent jouer un rôle significatif sur le nombre de déplacements déclaré. 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_1 et λ_2. 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.

En réalité, nous estimerons deux modèles, un sur le sous-échantillon des répondants en face-à-face (section 4.2.1) et un sur le sous-échantillon des répondants sur le web (section 4.2.2), afin de tester l'existence et le sens d'un biais de sélection dans chaque sous-échantillon de répondants.

IV.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 (tableau 76), puisque l'ensemble des coefficients des variables explicatives prennent les signes attendus.

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.

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 déclaré.

Echantillon face-à-face	Rég. sans correction			Rég. avec correction		
	Coeff.	Pr(>|z|)	Sign.	Coeff.	Pr(>|z|)	Sign.
Constante	2,01	<2e-16	***	2,64	<2e-16	***
Sexe : homme	-0,24	2,14e-07	***	-0,24	2,41e-07	***
Age	0,07	1,38e-15	***	0,06	1,73e-11	***
Age^2	-7,30e-04	<2e-16	***	-6,40e-04	2,44e-12	***
Possession permis : oui	0,47	4,61e-10	***	0,45	4,23e-09	***
Nb d'enfants / ménage	0,71	<2e-16	***	0,71	<2e-16	***
(Nb d'enfants / ménage)^2	-0,01	0,0267	*	-0,01	0,05	*
Nb de voitures / personne	0,57	<2e-16	***	0,53	2,00e-14	***
Nb de personnes / ménage	-0,19	5,13e-12	***	-0,18	4,78e-11	***
Revenu déclaré : oui	0,32	1,33e-11	***	0,27	2,90e-07	***
Activité : non actif	0,29	6,85e-07	***	0,30	2,30e-07	***
Distance domicile / centre	6,70e-06	0,1478		9,50e-06	0,04	*
Mills	NA	NA		-0,18	<0,01	**

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 76 – Analyse de la mobilité (Echantillon face-à-face)

Environ 9% de la variance de la variable 'Nombre de déplacements' est expliquée par l'ensemble des variables explicatives qui interviennent dans le modèle (tableau 77). Le niveau de significativité du modèle est en revanche très élevé (p-value < 2,2e-16).

Critères	Valeur
R2	9,2%
F-stat	81,63
ddl	9697
P-value	< 2,2e-16

TAB. 77 – Significativité de la régression utilisée dans la seconde étape (échantillon face-à-face)

IV.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 78. Peu de coefficients sont significatifs, ce qui s'explique notamment par les différences d'effectifs entre les deux échantillons de répondants. La comparaison des deux modèles estimés (web et face-à-face) montre que l'ordre de grandeur des coefficients est le même, mais que celui de leurs écarts-types varie fortement [140], 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 écarts-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

[140] La variance des coefficients estimés est : $V(\beta) = \frac{1}{n} * \frac{s^2}{V(x)}$, 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 socio-économiques sur la mobilité des répondants web. 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.

Le coefficient de l'inverse du ratio de Mills est significatif. Il y a donc des variables qui influent sur le choix de remplir le 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.

Echantillon web	Rég. sans correction			Rég. avec correction		
	Coeff.	Pr(>|z|)	Sign.	Coeff.	Pr(>|z|)	Sign.
Constante	0,89	0,50		3,04	0,08	.
Sexe : homme	-0,85	7.22e-04	***	-0,87	<0,001	**
Age	0,08	0,21		0.04	0,54	
Age^2	-8,57e-04	0,22		-4,38e-04	0,55	
Possession permis : oui	0,26	0,64		0,45	0,44	
Nb d'enfants / ménage	0,40	0,42		0,47	0,36	
(Nb d'enfants / ménage)^2	-0,12	0,53		-0,15	0,43	
Nb de voitures / personne	0,68	0,08	.	0,69	0,08	.
Nb de personnes / ménage	0,27	0,08	.	0,30	0,06	.
Revenu déclaré : oui	0,36	0,23		0,17	0,65	
Activité : non actif	0,74	0,03	*	0,82	0,02	*
Distance domicile / centre	-5,70-05	0,03	*	-5,20e-05	0,06	.
Mills	NA	NA		-0,73	1,46e-03	**

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 78 – Analyse de la mobilité (Echantillon web)

Globalement le modèle de régression est meilleur, puisque nous avons 4% de variance expliquée en plus (R^2 = 15%, vs. 11% sans l'introduction du biais de sélection) (tableau 79).

Critères	Valeur
R^2	15,2%
F-stat	3,50
ddl	234
P-value	8,78e-05

TAB. 79 – Significativité de la régression utilisée dans la seconde étape (échantillon web)

IV.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 de l'équation de sélection et ceux de l'équation d'intérêt (respectivement celle appliquée à l'échantillon web et celle appliquée à l'échantillon 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 à temps partiel n'est pas une variable explicative du modèle de mobilité. Pourtant, cette caractéristique implique une moindre présence au travail, ce qui peut avoir un effet sur la disponibilité des individus (capacité à recevoir un enquêteur à domicile) et leur niveau de mobilité.

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. Soit σ_{u1} = 1,88 et σ_{u2} = 2,23. Ensuite, l'estimation du paramètre ρ permet d'évaluer la corrélation entre l'inverse du ratio de Mills et le niveau de mobilité, c'est-à-dire la force de l'endogénéité du mode d'enquête au nombre de déplacements déclarés :

$$\rho_1 = -0,73/1,88 = -0,39$$

$$\rho_2 = -0,18/2,23 = -0,08$$

A partir de ces résultats, nous pouvons conclure que le nombre de déplacements déclaré par les répondants web apparaît fortement corrélé à l'inverse du ratio de Mills (ρ_1 = -0,39). La corrélation est plus faible dans le groupe des répondants en face-à-face (ρ_2 = -0,08). Il y a donc des facteurs qui incitent les individus à ne pas recevoir un enquêteur à domicile, puis à accepter de répondre en ligne, et qui influencent leur niveau de mobilité.

IV.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 sur la mobilité.

IV.3.1 Comparaison d'un modèle contraint et non contraint

Nous comparons deux modèles de régression multiple explicatifs du nombre moyen de déplacements déclaré : un modèle contraint et un modèle non contraint. Le modèle contraint ne considère comme facteurs explicatifs 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. Nous cherchons à savoir si les interactions entre les variables explicatives et le mode d'enquête, retirées dans le modèle contraint, ont un 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.

IV.3.2 Test du rapport des vraisemblances

Afin de pouvoir conclure sur la significativité des coeffcients et comparer les moèles, il faut utiliser des tests statistiques. Les principaux sont le test de Wald et le test de rapport des vraisemblances, qui donnent souvent des résultats proches. Ces tests suivent le même principe : il s'agit de comparer l'information apportée par un modèle non contraint, contenant un certain nombre de variables explicatives, et celle apportée par un modèle contraint (contenant un sous-ensemble des variables explicatives issues du modèle non contraint). Si la différence entre les deux modèles n'est pas significative, alors les variables explicatives retirées dans le modèle contraint n'ont aucun impact sur la variable à expliquer. Les hypothèses sont les suivantes :

H_0 = le supplément d'information apporté par les variables explicatives du modèle non contraint n'est pas significatif au seuil fixé (généralement $\alpha = 5\%$), ce qui se traduit par la nullité de tous les coefficients concernés.
H_1 = Au moins une variable explicative ajoutée dans le modèle non contraint a une influence significative sur le phénomène étudié.

Dans le test de rapport de vraisemblance, on rejette l'hypothèse nulle si la vraisemblance sous l'hypothèse alternative est significativement supérieure à la vraisemblance sous l'hypothèse nulle. La statistique du test est la suivante :

$$\wedge = [-2 * Log(vraisemblance_{mc})] - [-2 * Log(vraisemblance_{mnc})]$$

avec mc le modèle contraint et mnc le modèle non contraint. Elle suit une loi du Chi-deux à (n-1) degrés de liberté, avec n le nombre total de variables explicatives.

IV.3.3 Application aux données de l'enquête

Nous raisonnons ici sur la totalité de l'échantillon (web et face-à-face), l'introduction de la variable 'Mode' dans le modèle permettant de distinguer les répondants. Soit les hypothèses :

H_0 : les deux modèles contraint et non contraint 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 contraint et non contraint 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 le tableau 80, ceux du modèle contraint (par les variables retenues dans le modèle à deux étapes) dans le tableau 82.

Modèle non contraint	Coefficients	Pr(>|z|)	Signif.
Constante	2,64	<2e-16	***
Sexe : homme	-0,240	2,19e-07	***
Age	0,06	1,47e-11	***
Age^2	-6.36e-04	2,05e-12	***
Possession permis : oui	0,44	3,74e-09	***
Nb d'enfants / ménage	0,71	<2e-16	***
(Nb d'enfants / ménage)^2	-0,03	0,046	*
Nb de voitures / personne	0,53	1,63e-14	***
Nb de personnes / ménage	0,18	4,09e-11	***
Revenu déclaré : oui	0,27	2,63e-07	***
Activité : non actif	0,30	2,09e-07	***
Distance domicile / centre	9.50e-03	0,04	*
Mills	-0,18	<0.01	**
Mode	0,40	0,84	
(Sexe : homme) * mode	-0,63	0,04	*
(Age) * mode	-0,02	0,82	
(Age^2) * mode	1.98e-04	0,82	
(Possession permis : oui) * mode	0,01	0,99	
(Nb d'enfants / ménage) * mode	-0,24	0,69	
((Nb d'enfants / ménage)^2) * mode	-0,12	0,61	
(Nb de voitures / personne) * mode	0,16	0,74	
(Nb de personnes / ménage) * mode	0,48	0,01	**
(Revenu déclaré : oui) * mode	-0,10	0,81	
(Activité : non actif) * mode	0,52	0,23	
(Distance domicile / centre) * mode	0,06	0,06	.
(Mills) * mode	-0,55	0,18	

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 80 – Modèle non contraint

Critères	Valeur
R^2	9,33%
F-stat	40,89
ddl	9931
P-value	<2,2e-16

TAB. 81 – Significativité du modèle non contraint

Modèle contraint	Coefficients	Pr(>|z|)	Signif.
Constante	2,54	<2e-16	***
Sexe : homme	-0,25	4,19e-08	***
Age	0,06	7,72e-12	***
Age^2	-6,41e-04	1,02e-12	***
Possession permis : oui	0,45	1,84e-09	***
Nb d'enfants / ménage	0,71	<2e-16	***
(Nb d'enfants / ménage)^2	-0,04	0,04	*
Nb de voitures / personne	0,53	1,61e-14	***
Nb de personnes / ménage	-0,17	3,38e-10	***
Revenu déclaré : oui	0,26	3,62e-07	***
Activité : non actif	0,32	3,20e-08	***
Distance domicile / centre	8,44e-06	0,07	.
Mills	-0,15	4,85e-03	**

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 82 – Modèle contraint

Critères	Valeur
R^2	8,99%
F-stat	81,84
ddl	9944
P-value	<2,2e-16

TAB. 83 – Significativité du modèle contraint

Le test de rapport de vraisemblance appliqué aux deux modèles, contraint et non contraint, renvoie une probabilité de rejeter à tort l'hypothèse d'absence d'interaction égale à 3,12e-04. La p-value est inférieure à la valeur critique de 5%. Nous pouvons donc rejeter l'hypothèse nulle d'équivalence des modèles contraint et non contraint 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 effectué par les répondants. Trois variables semblent interagir significativement 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.

IV.3.4 Formulation d'un modèle stable

Les résultats du modèle de régression appliqué aux variables explicatives et interactions significatives sont présentés dans le tableau 84. On applique ensuite le test de vraisemblance aux modèles non contraint (contenant l'ensemble des interactions entre les variables explicatives de la mobilité et le mode d'enquête) et non contraint simplifié (qui ne laisse que les trois interactions significatives comme variables explicatives).

La probabilité de rejeter à tort l'hypothèse selon laquelle ces modèles sont équivalents est de 42,83%. Les interactions entre les variables et le mode d'enquête supprimées n'ont aucun pouvoir explicatif significatif de la mobilité quotidienne et nous conservons la formulation du modèle non contraint simplifié ci-dessous.

Modèle stable	Coefficients	Pr(>|z|)	Signif.
Constante	2,65	<2e-16	***
Sexe : homme	-0,24	2,08e-07	***
Age	0,06	1,03e-11	***
Age^2	-6,36e-04	1,53e-12	***
Possession permis : oui	0,45	2,63e-09	***
Nb d'enfants / ménage	0,69	<2e-16	***
(Nb d'enfants / ménage)^2	-0,03	0,051	.
Nb de voitures / personne	0,53	8,07e-15	***
Nb de personnes / ménage	-0,17	1,93e-10	***
Revenu déclaré : oui	0,27	2,84e-07	***
Activité : non actif	0,32	5,13e-08	***
Distance domicile / centre	9,54e-06	0,042	*
Mills	-0,19	6,15e-04	***
Mode	-0,50	0,18	
(Sexe : homme) * mode	-0,55	0,06	.
(Nb de personnes / ménage) * mode	0,19	0,10	
(Distance domicile / centre) * mode	-5,9e-05	0,05	.

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 84 – Modèle stable

Critères	Valeur
R^2	9,20%
F-stat	63,33
ddl	9940
P-value	<2,2e-16

TAB. 85 – Significativité du modèle stable

V Quantification de l'impact du mode d'enquête sur la mobilité

Ce paragraphe illustre de quelle manière les apports des techniques économétriques permettent de comprendre les différences de mobilité observées entre deux échantillons. Nous y détaillons l'impact des variables explicatives sur le nombre de déplacements déclaré, après correction du biais de sélection. Notre regard porte essentiellement sur les variables qui interagissent significativement avec le mode de réponse choisi. L'échantillon web étant 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 = 18%).

V.1 Les différentes composantes de l'influence du mode d'enquête sur la mobilité

Nous allons détailler deux composantes de l'influence du mode d'enquête sur la mobilité. D'abord l'influence directe du média, ensuite l'impact des intéractions entre le mode et certaines variables explicatives.

V.1.1 Impact direct du mode d'enquête

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, ce qui confirme nos analyses exploratoires.

V.1.2 Impact des interactions entre le mode et certaines variables explicatives

Par ailleurs, trois variables semblent interagir avec le mode : le sexe, le 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 être formalisée par la figure 68.

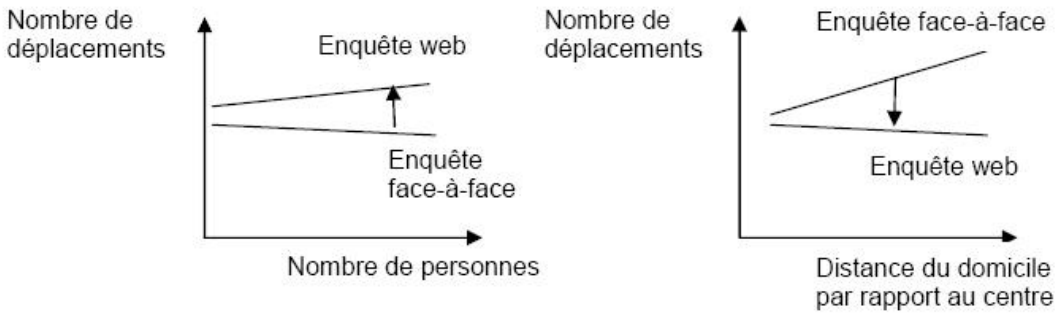

FIG. 68 – Relation bivariée entre le nombre de personnes, l'éloignement du domicile par rapport au centre et le nombre de déplacements des individus

La distance de la zone de résidence au centre de l'agglomération Le coefficient de la variable 'distance du domicile au centre * 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. -4,95e-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 genre Le fait d'être un homme impacte négativement la mobilité des répondants web : l'impact est de -0,24 déplacements par jour pour les répondants en face-à-face, vs. - 0,79 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 cadres), leur laisse encore moins de latitude pour leurs déplacements. Rappelons ici que 3/4 des répondants web sont des actifs.

La taille du ménage 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 -0,17 par personne pour les répondants en face-à-face, vs. 0,02 pour les répondants web. Ces derniers disposent d'un revenu annuel supérieur. Il est donc possible que le nombre de personnes accroisse les besoins et les déplacements (pour motif achat, loisir...) des membres du ménage.

V.2 Les profils types de répondants

Les coefficients des variables :
- Mode ;
- (Sexe : homme) * mode ;
- (Nb de personnes / ménage) * mode ;
- (Distance domicile / centre) * mode.

permettent 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 mètres du centre de l'agglomération, qui a son permis, 1 seule voiture à disposition 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.

D'une manière générale, lorsque les valeurs des variables qui interagissent positivement avec le mode d'enquête (la variable 'mode' prend la valeur 1 si l'individu répond sur le web et la valeur 0 s'il répond en face-à-face) sont élevées, la différence de mobilité s'atténue entre les répondants web et face-à-face. A contrario, lorsque les valeurs des variables qui interagissent négativement avec le mode d'enquête web sont élevées, cette différence de mobilité s'accroît.

VI Conclusion

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. A l'aide de techniques économétriques, nous avons montré qu'il est possible de tester l'existence d'un biais de sélection des répondants lors d'un protocole d'enquête mixte, de le quantifier, puis d'isoler l'impact des différences socio-économiques propres aux enquêtés de l'effet média. La méthode que nous avons utilisée se base sur la procédure d'estimation en deux étapes, empruntée à Heckman. Dans la première étape, on estime par un modèle probit la probabilité pour un individu de répondre en ligne en fonction de ses caractéristiques socio-économiques et des équipements du ménage. Puis, dans une seconde étape, on estime par la méthode des moindres carrés ordinaires (MCO) l'équation d'intérêt, en incorporant l'espérance conditionnelle aux variables de conditionnement des résidus de la première étape. Les coefficients estimés des variables explicatives du modèle ne sont alors plus biaisés et reflètent l'impact des caractéristiques de l'individu et de son ménage sur la mobilité, indépendamment du choix du mode de recueil de données.

Les analyses montrent que la différence de mobilité observée entre les échantillons web et face-à-face ne peut être totalement imputée à l'effet du mode d'enquête, mais s'explique par deux effets : l'effet lié aux différences socio-économiques entre les populations et l'effet lié au mode de recueil de données. Le biais de sélection est statistiquement significatif, quel que soit l'échantillon considéré, c'est-à-dire qu'il existe des variables qui influent à la fois sur le choix de remplir le questionnaire en ligne et sur la mobilité des répondants. Le signe du coefficient du ratio de Mills indique que la mobilité des internautes pourrait être en moyenne significativement plus élevée si ces derniers avaient répondu

en face à face. Inversement, le nombre de déplacements déclaré par les individus soumis à l'enquête en face-à-face pourrait être plus faible si ces derniers avaient saisi leurs réponses sur le web. L'effet mode est également significatif, mettant en évidence l'effet du média web sur le nombre de déplacements déclaré. L'analyse précise que l'effet du web n'est pas uniforme sur la population des internautes. Les variables observées qui interagissent directement avec le mode d'enquête sont : le sexe, le nombre de personnes du ménage, ainsi que la distance entre le domicile et le centre de l'agglomération.

L'état actuel de diffusion du web au sein de la population française ne permet pas la réalisation d'une enquête web exclusive auprès de l'ensemble de la population. L'échantillon ainsi constitué ne serait pas représentatif et les résultats biaisés, ainsi que l'illustre notre exemple. L'utilisation du web dans les protocoles d'enquêtes mixtes est en revanche intéressante, mais impose d'analyser le biais de sélection des individus. La question est évidemment 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.

Des développements complémentaires vont permettre d'approfondir l'analyse comparative de la mobilité selon le mode d'enquête. Nous avons montré dans les chapitres précédents que la sous-mobilité des répondants web s'explique en partie par une immobilité plus importante. Un moyen de mieux appréhender les facteurs à l'origine de cette sous-mobilité est d'utiliser un modèle explicatif de l'immobilité selon le mode d'enquête. Dans le chapitre suivant, nous mobilisons des techniques qui ont l'avantage de séparer les facteurs explicatifs de la décision de se déplacer de ceux qui influent sur le niveau de mobilité.

Chapitre 8 : Intérêt du modèle "Hurdle" pour la comprehénsion des comportements de mobilité

"*Le Temps est l'image mobile de l'éternité immobile.*"
Platon (-428,-347).

Dans le chapitre précédent, nous avons cherché à expliquer le nombre de déplacements moyen quotidiennement effectués par les habitants de l'agglomération lyonnaise. Cependant, il s'avère qu'une part non négligeable de l'échantillon a déclaré être immobile le jour de référence de l'enquête. Dans le cadre de deux estimations économétriques, nous allons nous intéresser aux facteurs impactant la décision de se déplacer, d'une part, le niveau de mobilité, d'autre part. Il s'agit de donner une estimation de la mobilité quotidienne moyenne des individus, ainsi que la fonction reliant cette dernière aux caractéristitiques socio-économiques, à la motorisation et à l'équipement des ménages en moyens de communication en considérant que, pour une part non négligeable de la population enquêtée, la mobilité quotidienne est nulle (aucun déplacement déclaré). Notre démarche va notamment consister à mieux caractériser les déterminants de l'immobilité.

La variable dépendante de nos modèles, le nombre de déplacements quotidiens déclaré, est quantitative mais ne peut prendre que des valeurs entières. Le modéle développé fera donc référence aux techniques employées pour la régression sur les variables de comptage. Par ailleurs, l'explication du niveau de mobilité ne se fera que pour les individus mobiles, c'est-à-dire ayant effectué au moins un déplacement dans le périmètre de l'enquête durant la période de référence. La probabilité pour un individu de se situer dans l'intervalle des personnes mobiles sera prise en compte avec des notions similaires à celles utilisées dans les modèles qualitatifs. Selon Thomas (2000), il existe trois façons de considérer les variables qualitatives en statistique : les incorporer comme variables explicatives dans un modèle de régression, étudier leur corrélation ou les traiter comme des variables dépendantes. Pour ce dernier cas, il faut distinguer les modèles à variables qualitatives binaires, comme les modèles Logit et Probit, des modèles à variable dépendante limitée, où la variable expliquée n'est observée que sur un intervalle, comme les modèles Tobit. La particula-

rité des modèles qualitatifs réside dans le fait que les techniques d'inférences par la méthode des moindre carrés ne sont pas adaptées. Nous appliquons à la place la méthode du maximum de vraisemblance. De nombreux problèmes rencontrés lors de la régression linéaire (multicolinéarité, hétéroscédasticité...) sont toujours présents, mais peu de travaux traitent des effets du non respect de certaines hypothèses sur la qualité de l'estimation des modèles à variable dépendante limitée (Maddala, 1986).

Afin d'examiner les déterminants de la mobilité chez les répondants web et face-à-face, nous utilisons un modèle de Poisson à obstacle, appelé plus couramment dans la littérature "Poisson Hurdle model". Il est alors possible de distinguer les facteurs incitant les personnes à se déplacer au moins une fois durant la période de référence de l'enquête de ceux influençant le niveau de mobilité des personnes mobiles. L'objectif est de répondre séparément à ces deux questions : comment expliquer que certains individus se déplacent alors que d'autre restent immobiles durant la période de référence ? Une fois la décision de mobilité prise, qu'est-ce qui pousse certaines personnes à se déplacer davantage que d'autre durant la même période ? De nombreux exemples relatifs aux sciences sociales sont détaillés dans la littérature, comme les travaux de Bohara et Krieg (1996) concernant le taux de migration aux USA, ceux de King (1989) sur l'implication des nations dans les conflits internationaux ou ceux de Bounie *et al.* (2006) sur les dépenses effectuées en France par carte de crédit.

I Pour poser le problème

Le tableau 86 donne la distribution des individus par enquête (web et face-à-face) selon le nombre de déplacements quoitidiens déclaré. Nous constatons que, globalement, les répondants en face-à-face se déplacent davantage que les répondants web. Par ailleurs, nous observons une forte proportion de personnes immobiles, en particulier dans l'enquête web (environ une personne sur cinq ne se déplace pas durant la période de référence). Dans ce cas, quels sont les déterminants de la mobilité quotidienne ?

Enquête	Obs.	0	1	2	3	4	5	6	7+	Total
Face-à-face	Ind.	1459	110	3854	1081	3096	1106	1151	1414	13271
	%	11%	1%	29%	8%	23%	8%	9%	11%	100%
Web	Ind.	70	12	98	48	65	30	24	22	369
	%	19%	3%	27%	13%	18%	8%	7%	6%	100%

TAB. 86 – Distribution du nombre de déplacements par personne et par enquête

La faible proportion de personnes n'effectuant qu'un seul déplacement, quelle que soit l'enquête, laisse penser qu'une fois la décision de se déplacer prise, les répondants effectuent plusieurs déplacements. Le nombre de déplacements déclaré par les personnes mobiles est toutefois plus important dans l'échantillon en face-à-face. Dès lors, comment peut-on expliquer la fréquence

des déplacements entre les répondants ? Enfin, et plus généralement, les facteurs explicatifs de la décision de mobilité et de la fréquence des déplacements sont-ils identiques, ou bien peut-on isoler, des facteurs spécifiques pour chaque type de comportement ?

Notre analyse tente de mettre en évidence les facteurs motivant les individus à effectuer un premier déplacement, mais également à comprendre pourquoi certains se déplacent davantage que d'autres. Nous pensons que la décision de se déplacer dinstingue fortement les individus. En effet, ceux qui ont besoin de se déplacer en semaine pour effectuer des activités pour leur compte (travail, loisirs ...) ou celui d'autrui (accompagnement ...) ont probablement des caractéristiques différentes de ceux qui déclarent rester à leur lieu de résidence. Plusieurs facteurs peuvent expliquer l'immobilité des répondants, comme la volonté de ne pas communiquer l'information, une maladie ou une absence d'activité le jour de référence des déplacements, ou le mode d'enquête [141]. Par ailleurs, le niveau de mobilité des personnes mobiles est très variable, et nous avons vu dans le chapitre précédent que certains facteurs sociodémographiques inflencent positivement le nombre de déplacements. A contario, d'autres caractéristiques semblent modérer cette mobilité. Ces observations nous amènent à supposer que le modèle statistique qui détermine la probabilité de se déplacer diffère de celui qui détermine la fréquence des déplacements. Dans les modèles standard applicables aux données de comptage, comme le modèle de Poisson, ces deux procédés sont forcés d'être identiques. Ces modèles ne permettent pas une structure conditionnelle et donnent des estimateurs biaisés (Grogger et Carson, 1991). C'est pourquoi, à la place d'appliquer le modèle de Poisson, nous estimons un modèle qui permette de différencier le processus de choix binaire (mobile / immobile) de l'estimation de fréquence de la variable d'intérêt (nombre de déplacements déclaré par personne mobile).

A cause de cette différence supposée, notre modèle est scindé en deux parties. Nous estimons d'abord la probabilité individuelle de se déplacer durant la période de référence. Puis, conditionnellement à cette décision, nous estimons dans un second temps la fréquence des déplacements déclarée. Il est possible de distinguer le modèle économétrique donnant le résultat binaire (se déplace / ne se déplace pas) de celui qui détermine le niveau de mobilité, à l'aide d'une procédure connue sous le nom de 'Hurdle model' ou modèle à obstacle (Mullahy, 1986). Il n'est cependant pas possible de distinguer, avec les données disponibles, l'immobilité réelle (individus ne s'étant pas déplacés) du refus de déclaration (non déclaration des déplacements effectués). Nous estimons la décision de faire un déplacement initial et d'effectuer des déplacements additionnels à l'aide des mêmes variables explicatives dans les deux modèles, ce qui nous permet de comparer directement les coefficients estimés dans les deux équations. Si les modèles génèrent des résultats significativement différents, alors nous pourrons conclure que modéliser simplement la décision de mobilité ommet d'expliquer la motivation de nombreuses personnes de réaliser des déplacements additionnels, ou qu'évaluer simplement le niveau de mobilité

[141] Le caractère auto-administré du web peut encourager certains individus à se déclarer immobiles, pour éviter d'avoir à saisir leurs déplacements.

de permet pas de prendre en compte les facteurs à l'origine de l'immobilité d'un certain nombre de répondants.

II Formalisation économétrique

La méthodologie adaptée est basée sur une spécification de Poisson à obstacle (Hurdle Poisson) [142]. D'abord, la décision de se déplacer est modélisée par un résultat binaire, à savoir si un individu a effectué un déplacement dans le périmètre d'enquête durant la période de référence ou pas. Lorsque la décision de se déplacer est prise, générant la réalisation d'un déplacement, cela s'appelle 'crossing a Hurdle'. Le nombre de déplacements effectué par l'individu est supposé suivre un modèle de Poisson tronqué en zéro (Zeileis *et al.*, 2008).

II.1 La loi de Poisson

La distribution de Poisson est adpatée aux variables entières (Zorn, 1996). En statistique, une variable aléatoire Y suit une loi de Poisson [143] de paramètre réel positif θ, notée Y $\mathrm{P}(\theta)$, si et seulement si elle suit, pour tout entier naturel k, une loi de probabilité définie par :

$$P(Y = k) = \frac{e^{-\theta}\theta^k}{k!} \tag{70}$$

θ est le paramètre unique de la loi de Poisson. Il représente à la fois la moyenne et la variance de la distribution de la variable d'intérêt. Soit :

$$E(Y) = \theta \tag{71}$$
$$V(Y) = \theta \tag{72}$$

L'espérance conditionnelle de la distribution d'une variable aléatoire Y, étant donné un ensemble de facteurs explicatifs x_k est donnée par :

$$E(Y \mid x_k) = e^{(\theta_k x_k)} \tag{73}$$

où $\beta = (\beta_0 \dots \beta_k)$ est un vecteur de paramètres inconnus à estimer et $x = (1 \dots x_k)$ un vecteur de variables explicatives. Nous notons :

$$\theta_i \equiv e^{(\beta x_i)} \tag{74}$$

Pour chaque observation i tirée au hasard dans la population, la probabilité de y_i conditionnelle à l'ensemble de facteurs x_{ki} est donnée par la relation :

[142] *'The idea underlying the Hurdle formulations is that a binomial probability model governs the binary outcome of wether a count variable has zero or a positive realization. If the realization is positive, the 'Hurdle' is crossed and the conditional distribution of the positives is governed by a truncated-at-zero count data model' (Mullahy, 1986).*

[143] Ou loi des événements rares.

$$P(Y = y_i \mid x_{ki}) = \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!} \tag{75}$$

avec y = 0, 1, 2,...

On a alors :

$$lnL[P(y_i)] = -\theta_i + y_i ln\theta_i - lny_i! \tag{76}$$
$$lnL[P(y_i)] = -e^{\beta x_i} + y_i(\beta x_i) - lny_i! \tag{77}$$

Si nous supposons que les observations concernant différents individus i sont mutuellement indépendantes, il est possible d'estimer simplement les paramètres β_k par la méthode du maximum de vraisemblance. La fonction de vraisemblance est en effet égale à la somme du logarithme des probabilités, exprimée en fonction de β :

$$ln(L) = -\sum_{i=1}^{n} e^{\beta x_i} + \sum_{i=1}^{n} y_i(\beta x_i) - \sum_{i=1}^{n} lny_i! \tag{78}$$

On a désormais (k+1) paramètres à estimer et (k+1) conditions du premier orfre, qui peuvent s'écrire sous la forme du vecteur :

$$\frac{\partial ln(L)}{\beta_k} = \sum_{i=1}^{n}(y_i - \theta_i)x_i = 0 \tag{79}$$

Si la distribution de Poisson est appropriée, et en posant l'hyptohèse que la répartition des y_i est aléatoire dans l'échantillon, nous obtenons une etimation non biaisée de β.

II.2 Le modèle de Poisson censuré

Les données utilisées en microéconomie sont souvent caractérisées par une censure de la variable dépendante. Un intervalle de valeurs possibles pour cette variable est alors transformé en une simple valeur c et la variable dépendante est égale à c pour une part significative des observations. Les méthodes de régression classiques ne permettent pas de tenir compte de la différence qualitative qui existe entre les valeurs limites (égales à c) et les valeurs continues de la variable dépendante.

Dans notre modèle, la variable à expliquer y_i est une variable dite de comptage, observable seulement sur un certain intervalle. L'objectif de l'analyse est de connaître les déterminants de la variable d'intérêt y_i, sans éliminer de l'échantillon les individus pour lesquels cette variable est nulle. Dans la littérature, les modèles appropriés pour ce type d'analyse sont les modèles Tobit

[144], utilisés notamment lorsqu'un grand nombre d'observations de la variable à expliquer sont nulles. Si nous conservons dans l'échantillon les observations concernant les individus immobiles, nous appliquons le modèle tobit censuré, et les valeurs observées sont censurées à zéro. Si nous choisissons de supprimer ces observations, alors il faut appliquer le modèle tobit tronqué. On dit qu'un modèle est tronqué si les variables explicatives ne sont pas observables lorsque la variable dépendante passe en-dessous (ou au-dessus) d'un certain seuil c. Ce cas peut se produire, soit si les individus pour lesquels $y_i < c$ ne sont pas intérrogés, soit si les réponses aux variables explicatives x_i n'ont de sens que lorsque $y_i > c$.

On considère une variable aléatoire y_i, distribuée selon une loi de Poisson. Nous savons par hypothèse, que :

$$E(Y \mid x_k) = e^{(\theta_k x_k)}$$

Un échantillon de n observations de couples (y_i, x_{ki}) est constitué, mais la variable y_i n'est pas toujours observable. On dispose à la place de la variable latente y_i^*, qui est reliée à la variable y_i par la relation [145] :

$$y_i^* = y_i, \; ssi \; y_i > c \tag{80}$$
$$y_i^* = c, \; sinon \tag{81}$$

où c est une constante. Si le vecteur x_k est observé pour tout i, que y_i^* soit ou non supérieure à c, alors l'échantillon est censuré et seule la variable dépendante n'est observée que sur un certain intervalle. Si nous éliminons les observations telles que $y_i^* \leq c_i$, alors l'échantillon est tronqué.

Si l'échantillon est censuré en zéro, alors la constante c est nulle et le modèle peut s'écrire :

$$P(y_i = y_i^*) = \frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!} \tag{82}$$

avec $\theta_i = e^{\beta_k x_{ki}}$, x_i un vecteur de variables explicatives et β un vecteur de paramètres.

Si $y_i = 0$, alors nous pouvons poser :

$$P(y_i = 0) = \frac{e^{-\theta_i}\theta_i^0}{0!} \tag{83}$$
$$P(y_i = 0) = e^{-\theta_i} \tag{84}$$

[144] En référence à Tobin qui l'a proposé la première fois pour analyser l'achat de biens durables par les ménages (Tobin, 1958). De nombreux exemples d'applications du modèle Tobit sont présents dans la littérature, comme ceux décrivant le nombre d'heures de travail ou les salaires des personnes selon leur statut (Amemiya, 1984).

[145] Le point de censure peut théoriquement être vers le haut ou vers le bas, selon le problème à modéliser.

En revanche, si $y_i > 0$, alors nous pouvons écrire :

$$P(y_i > 0) = 1 - P(y_i = 0) \tag{85}$$
$$P(y_i > 0) = 1 - e^{-\theta_i} \tag{86}$$

Par ailleurs, nous pouvons calculer la probabilité de y_i, conditionnellement à la réalisation de l'événement $y_i > 0$. Soit :

$$P(y_i \mid y_i > 0) = \frac{\frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}}{1 - e^{-\theta_i}} \tag{87}$$

II.3 Modèle Hurdle

Certaines données empiriques contiennent plus de valeurs nulles qu'il est autorisé pour appliquer le modèle de Poisson. La catégorie de modèle capable de prendre en compte cette propriétés est appelé modèle Hurdle. Il contient deux composantes. Un modèle tronqué en zéro, de type Poisson, est employé pour modéliser les valeurs positives de la variable d'intérêt et une composante de type 'Hurdle' permet de distinguer les valeurs nulles des valeurs positives de la variable d'intérêt. Pour cette dernière, un modèle binomial ou une distribution censurée en zéro de type Poisson peut être employée. Plus formellement, le modèle Hurdle combine un modèle de données de comptage tronqué à gauche en $y = 1$ et un modèle Hurdle en zéro censuré à droite en $y = 1$.

Soit un modèle de Poisson censuré en zéro, dans lequel la probabilité de survenue de l'événement y est influencé par un vecteur de variables explicatives x_1, à l'aide de la fonction :

$$\theta_i \equiv e^{(\beta_1 x_{1i})} \tag{88}$$

et un modèle de Poisson tronqué en zéro, dans lequel la fréquence de survenue de l'événement y est influencée par un vecteur de variables explicatives x_2, à l'aide de la fonction :

$$\lambda_i \equiv e^{(\beta_2 x_{2i})} \tag{89}$$

La question de la méthode à utiliser pour calculer les valeurs prévues des paramètres β_1 et β_2 reste entière. S'agissant d'un modèle non linéaire, nous allons utiliser la méthode du maximum de vraisemblance pour évaluer les paramètres du modèle. Pour une observation donnée, la vraisemblance est la probabilité que le modèle renvoie le bon résultat. On cherche alors la probabilité d'obtenir le bon résultat pour l'ensemble de l'échantillon en formant le produit des probabilités individuelles. La méthode consiste à estimer les coefficients de la fonction qui maximise cette vraisemblance, au moyen d'un algorithme numérique. Cette spécification présente l'avantage que les composantes tronquées et censurées peuvent être estimées séparément. La fonction de vraisemblance jointe pour le modèle Hurdle de Poisson est :

$$L = \prod_{y_i=O} [P(y_i = 0] \prod_{y_i>O} [1 - P(y_i = 0][P(y_i \mid y_i > 0)] \tag{90}$$

Le premier terme représente le résultat binaire caractérisant la réalisation de y_i et le dernier terme correspond au processus de Poisson, conditionnellement à la réalisation de y_i. Soit :

$$L = \prod_{y_i=O} [e^{-\theta_i}] \prod_{y_i>O} [1 - e^{-\theta_i}][\frac{\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}}{1 - e^{-\lambda_i}}] \tag{91}$$

$$ln(L) = -\sum_{y_i=O} \theta_i + \sum_{y_i>O} ln(1 - e^{-\theta_i}) + \sum_{y_i>O} ln[\frac{\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}}{1 - e^{-\lambda_i}}] \tag{92}$$

Il est possible d'utiliser le même vecteur de variables explicatives pour estimer la réalisation d'une valeur positive de la variable d'intérêt (x_1) et la fréquence de survenue de cette même variable (x_2). Ceci permet de comparer directement les deux modèles. Soit :

$$L = \prod_{y_i=O} [e^{-\theta_i}] \prod_{y_i>O} [1 - e^{-\theta_i}][\frac{\frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}}{1 - e^{-\theta_i}}] \tag{93}$$

$$L = \prod_{y_i=O} [e^{-\theta_i}] \prod_{y_i>O} [\frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}] \tag{94}$$

$$ln(L) = -\sum_{y_i=O} \theta_i + \sum_{y_i>O} ln[\frac{e^{-\theta_i}\theta_i^{y_i}}{y_i!}] \tag{95}$$

II.4 Evaluation de la pertinence du modèle Hurdle

Afin d'évaluer la pertinence de l'utilisation d'un modèle 'Hurdle', nous calculons une statistique qui combine la valeur de Log-vraisemblance du modèle Hurdle et celle du modèle de Poisson simple :

$$stat = 2 * [logLik_{Hurdle} - logLik_{Poisson}] \tag{96}$$

Pour tester si le modèle Hurdle est statistiquement significatif, on compare cette statistique à une valeur lue dans la table du Khi-deux. Les hypothèses sont :

H_0 = l'écart entre la valeur de Log-vraisemblance du modèle Hurdle et celle du modèle de Poisson simple n'est pas significatif. Ceci signifie que le modèle Hurdle n'apporte pas un intérêt majeur dans la compréhension du phénomène étudié.

H_1 = l'écart entre la valeur de Log-vraisemblance du modèle Hurdle et celle du modèle de Poisson simple est significatif. Le modèle Hurdle permet alors de mieux rendre compte de l'influence des facteurs explicatifs sur la variable étudiée.

III Application à l'équation de mobilité quotidienne

Pour illustrer nos propos, nous utilisons les données issues de l'enquête ménages déplacements réalisée à Lyon en 2006. L'analyse est menée sur l'échantillon web, d'une part, et face-à-face, d'autre part. Ces deux populations n'étant pas homogènes en termes de caractéristiques socio-économiques et d'habitude de mobilité, il est probable que l'impact des facteurs ne soit pas le même. Seules les personnes de 18 ans et plus sont considérées, par souci de cohérence avec le chapitre précédent et de comparaison entre les deux échantillons. Une seule personne du ménage remplit le questionnaire en-ligne, et nous avons montré qu'il s'agit surtout du chef de ménage ou de son conjoint. A contrario, dans l'enquête en face-à-face, l'ensemble des menbres du ménage sont interrogés, dont les enfants. Ne pas prendre en compte les moins de 18 ans revient à exclure les scolaires de l'analyse, largement sous représentés dans l'achantillon web, et dont le niveau de mobilité est traditionnellement spécifique.

III.1 Les variables disponibles pour l'analyse

III.1.1 Définition de l'immobilité

Dans une enquête déplacements, une personne immobile est une personne qui a déclaré n'avoir effectué aucun déplacement. Il existe certaines raisons au fait que des individus choisissent de ne pas se déplacer durant la période de référence de l'enquête, et un répondant immobile est un problème pour l'analyste. Cette personne n'a-t-elle réellement pas quitté son domicile, a-t-elle oublié les déplacements effectués ou utilise-t-elle cet échappatoire pour réduire le niveau de pénibilité associé à l'enquête ? Les enquêtes ménages déplacements excluent certaines personnes. On peut se demander combien d'individus résidant à leur domicile sont incapables de se déplacer. Seule une infime proportion de la population enquêtée semble concernée (environ 1%). D'autres personnes sont dans une incapacité temporaire à sortir de chez elles (1%), devaient rester chez elles (1%) ou n'ont pas souhaité sortir à cause de la météo (Madre *et al.*, 2007).

Le nombre d'immobiles peut varier fortement selon les enquêtes, une forte proportion pouvant cependant correspondre à des 'soft refusals'. Il s'agit pour les répondants d'éviter de passer du temps à livrer leurs réponses, en déclarant dès le début de l'interview ne pas avoir effectué de déplacement durant la période de référence. *'Something in the interaction between the survey, the*

survey protocol, and the fieldwork firm invites the respondent . . .to use the soft refusal'(Madre *et al.*, 2007) [146]. Le mode d'enquête peut avoir une influence. D'une façon générale, les enquêtes auto-administrées laissent apparaître davantage de personnes immobiles, l'enquêteur n'étant pas présent pour relancer les répondants [147]. Le web permet de répondre à distance, dans des conditions pas toujours propices à la concentration du répondant. La qualité de la rédaction des questions et le professionnalisme des enquêteurs permettent sans nul doute de réduire les 'refus mous'. Mais les sociétés d'études sont plus souvent tournées vers le taux de réponse global que vers la part d'immobiles dans les enquêtes (Madre *et al.*, 2007). Il est probable que moins un répondant aura passé de temps en déplacements, plus il aura tendance à adopter cette attitude, bien que ce 'refus mou' touche toutes les classes de la population, même les plus mobiles. Il peut donc faire perdre des individus qui se sont beaucoup déplacés, surtout dans une interview auto-administrée (Armoogum *et al.*, 2005). Stopher *et al.* (2004b) rapporte que plus les individus sont sollicités pour participer à l'interview, plus ils risquent de se déclarer immobiles. La part d'immobiles est importante, car elle impacte directement le nombre moyen de déplacements et d'activités déclaré (Madre *et al.*, 2003).

Si les décisions de sous-déclaration sont faites de façon aléatoire par le répondant, et ne sont pas liées au nombre de déplacements ou d'activités pratiqués, alors les résulats modélisés ne sont pas biaisés (Han et Polak, 2003). Madre *et al.* (2003) précisent que le 'refus mous' est un facteur aléatoire qui permet d'utiliser un modèle binomial de détermination de l'immobilité.

III.1.2 Description des variables

Nous avons choisi un ensemble de variables caractérisant l'individu et son ménage, pouvant expliquer potentiellement le niveau de mobilité et plus particulièrement la décision de mobilité. Nous avons mis en évidence dans le chapitre précédent que trois types de variables influent sur la mobilité. Il y a d'une part des variables socio-économiques, telles que l'occupation de l'individu, son âge, le nombre d'enfants du ménage ou la localisation du lieu de résidence. Mais des variables concernant l'équipement du ménage en moyen de communication ou liées à la motorisation ont également un impact non négligeable sur le nombre de déplacements déclaré. Nous avons retenu dix variables dans l'analyse :

Sexe : homme ou femme.

Age : âge de chaque personne.

Nombre d'enfants du ménage : nombre d'enfants de moins de 18 ans présents dans le ménage.

Nombre de personnes du ménage : nombre de personnes qui composent le ménage.

[146] Le taux d'immobiles 'acceptable' dans les enquêtes se situe entre 18% et 12% dans les enquêtes transport portant sur un jour de semaine (Madre *et al.*, 2007).

[147] Madre *et al.* (2007) met en évidence un taux d'immobilité de l'ordre de 20% pour les enquêtes postales, contre 13% à 15% pour les enquêtes téléphoniques et face-à-face.

Revenus declarés : variable dichotomique, indiquant si l'individu a déclaré le niveau de revenu annuel net de son ménage.

Activité : variable renseignant le statut du répondant. Deux modalités sont disponibles : les actifs et les non actifs.

Téléphone portable : possession d'un téléphone portable par la personne, à titre personnel ou professionnel.

Nombre de voitures du ménage : nombre de voitures particulières possédées par le ménage, rapporté au nombre de personnes en âge de conduire (18 ans et plus).

Possession du permis : possession du permis de conduire ou pratique de la conduite accompagnée.

Vendredi : variable qui prend la valeur '1' si le jour de référence pour le recueil des déplacements est le vendredi, et la valeur '0' sinon.

III.1.3 Statistiques descriptives

Les tableaux suivants donnent quelques statistiques descriptives sur les variables retenues : nombre d'observations, valeurs minimum et maximum, moyenne et écart-type pour les variables continues (tableaux 87 et 88), modalités et effectifs, pour les variables nominales (tableaux 89 et 90).

Variables	Obs.	Moyenne	Ecart-type	Minimum	Maximum
Age	11 577	47,34	18,11	18	98
Nb d'enfants du ménage	11 577	0,65	1.02	0	7
Nb de voitures du ménage	11 577	0,66	0.39	0	4
Nb de personnes du ménage	11 577	2,78	1.42	1	10

TAB. 87 – Statistiques descriptives des variables continues (échantillon face-à-face)

Variables	Obs.	Moyenne	Ecart-type	Minimum	Maximum
Age	360	44,06	13,98	18	86
Nb d'enfants du ménage	361	0,62	0,91	0	3
Nb de personnes du ménage	361	2,58	1.28	1	6

TAB. 88 – Statistiques descriptives des variables continues (échantillon web)

Variables	Observations	Modalités	Effectifs
Sexe	11 577	Homme	5 424
		Femme	6 153
Possession du permis	11 577	Oui	9 694
		Non	1 883
Activité	11 577	Actif	6 067
		Non actif	5 510
Revenus déclarés	11 577	Oui	7 521
		Non	4 056
Téléphone portable	11 577	Oui	8 204
		Non	3 373

TAB. 89 – Statistiques descriptives des variables nominales (échantillon face-à-face)

Variables	Observations	Modalités	Effectifs
Activité	361	Actif	260
		Non actif	101
Revenus déclarés	361	Oui	275
		Non	86
Vendredi	361	Oui	150
		Non	211

TAB. 90 – Statistiques descriptives des variables nominales (échantillon web)

III.1.4 Hypothèses sur l'influence des différents facteurs

Nous étudions dans ce paragraphe l'impact des variables explicatives retenues. Les résultats diffèrent parfois de ceux du chapitre précédent, car nous considérons à présent l'ensemble des répondants, et plus seulement les mobiles.

Impact du genre : des hommes plus mobiles que les femmes (échantillon face-à-face) En ce qui concerne le genre, plusieurs enquêtes transport réalisées en France et en Belgique mettent en évidence un taux d'immobiles plus important chez les femmes (Armoogum *et al.*, 2005). Par ailleurs, l'analyse descriptitive de l'échantillon face-à-face montre que les hommes délarent en moyenne davantage de déplacements que les femmes (figure 69 [148]), ce qui vient contredire les résultats du chapitre précédent concernant uniquement les personnes mobiles. Ceci s'explique par un plus fort taux d'immobilité chez les femmes (13,81% vs. 10,01%). Les femmes sont moins souvent titulaires du permis de conduire que les hommes, bien que la proportion de femmes ayant

[148] Ce diagramme appelé boîte à moustaches s'utilise pour comparer un même caractère dans deux populations de tailles différentes. Il s'agit de tracer un rectangle allant du premier quartile au troisième quartile et coupé par la médiane. On ajoute des segments aux extrémités menant jusqu'aux premier et neuvième déciles. Les points en dehors des segments correspondent aux valeurs extrêmes.

le permis soit de plus en plus importante, et ont un accès plus restreint et moins régulier à la voiture particulière (Hine, 2004). Le genre ne semble pas discriminant de la décision de mobilité, pour l'échantillon web.

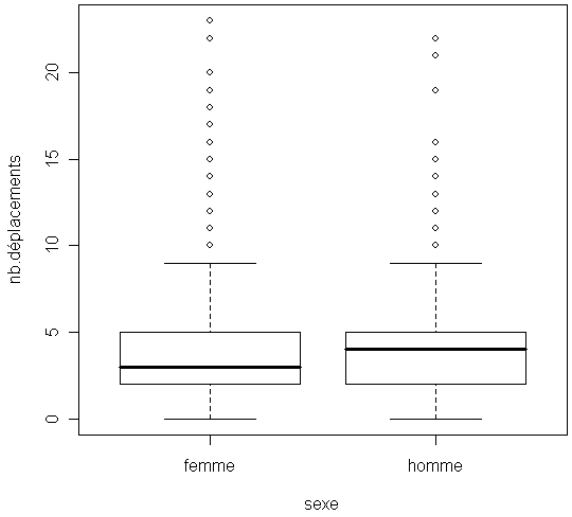

FIG. 69 – Relation entre le genre des répondants et le nombre de déplacements (enquête face-à-face)

Source : EMD Lyon (2006)

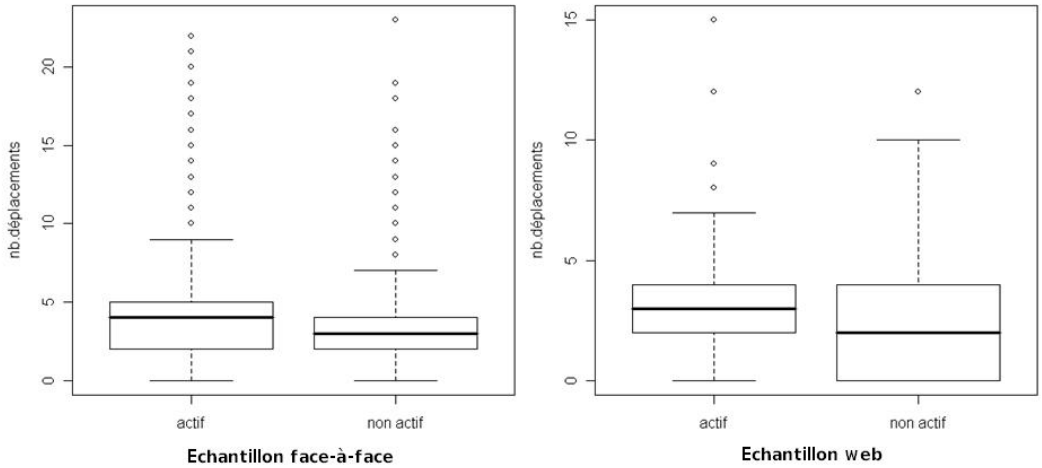

FIG. 70 – Relation entre l'activité des répondants et le nombre de déplacements par enquête

Source : EMD Lyon (2006)

Impact du statut : davantage de personnes immobiles chez les inactifs Le même raisonnement peut-être appliqué à la variable 'activité'. L'analyse descriptive montre, quel que soit l'échantillon, que le nombre moyen de

déplacements est sensiblement plus important chez les actifs (figure 70). Or, nous avons vu précédemment que dans l'échantillon web, parmi les mobiles, le niveau de mobilité était plus important chez les inactifs, ce que confirment d'autres analyses (Madre *et al.* (2003) et Hubert (2003)). De nouveau, cette différence s'explique par un taux d'immobiles soit plus important chez les inactifs (18,95% pour les inactifs et 5,75% pour les actifs en face-à-face, vs. respectivement 30,69% et 15% sur le web). Cette relation est probablement liée à l'âge.

Impact du jour de référence : des immobiles plus nombreux le vendredi Par ailleurs, le taux d'immobiles peut dépendre du jour concerné par le recensement des déplacements, et donc du jour de référence des déplacements. Si les interviews sont conduites en semaine, alors les actifs sont probablement plus difficiles à joindre. Ces derniers répondent davantage le week-end, par le web, au sujet des déplacements de la journée du vendredi. Ce décalage entre la saisie des réponses et la réalisation des déplacements pèse sur la mobilité déclarée.

Le moment de l'interview, ou de saisie des réponses, peut donc biaiser le nombre de déplacements déclaré et impacter le taux d'immobilité (Christensen, 2006). La variable 'vendredi' sera donc prise en compte pour l'analyse de l'échantillon web, en posant l'hypothèse qu'elle pèse sur la probabilité de se déplacer (28,67%, vs. 12,80% d'immobiles) [149].

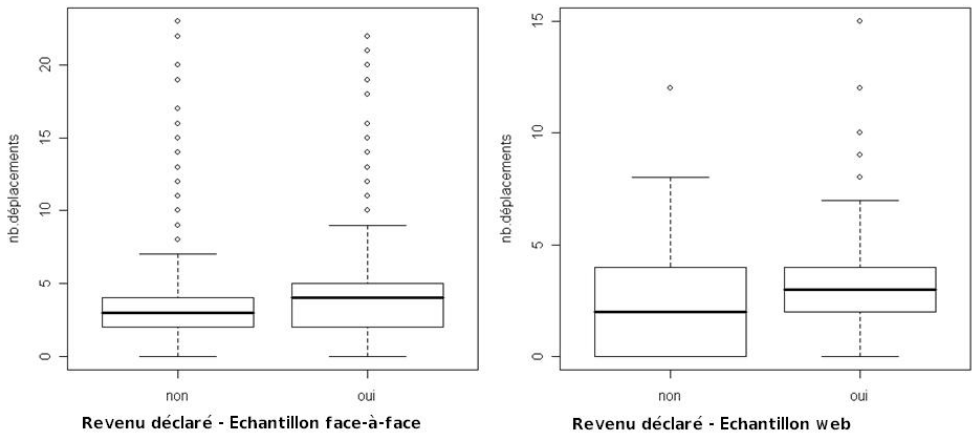

FIG. 71 – Relation entre la déclaration des revenus annuels du ménage et le nombre de déplacements par enquête
Source : EMD Lyon (2006)

Impact de la décision de déclarer les revenus : un facteur générateur de déplacements Le fait de déclarer son revenu à l'enquêteur ou sur son ordinateur peut être interprété comme une volonté de participer à l'enquête

[149] Celle-ci est contraire à la tendance observée dans l'enquête nationale transport, où la plus faible immobilité est observée le vendredi (Armoogum *et al.*, 2005).

et de ne pas vouloir 'retenir' des informations jugées trop personnelles. Nous formulons donc l'hypothèse que cette variable est négativement liée à l'immobilité (11,02%, vs. 13,90% en face-à-face et 16%, vs. 30,24% sur le web) et au contraire propice à un nombre de déplacements déclarés élevé (figure 71).

Impact de la possession d'un téléphone portable et du permis de conduire : une influence non négligeable sur la mobilité Enfin, dans l'enquête face-à-face, deux autres variables semblent impacter le taux d'immobilité : la possession du permis de conduire et celle d'un téléphone portable [150]. Il est probable que ces deux variables influencent positivement la mobilité (figure 72). Si le ménage ne dispose pas d'une voiture, alors le risque d'immobilité est plus important (Madre *et al.*, 2003). Posséder le permis est un atout pour se déplacer, puisque cela rend possible les déplacements en voiture particulière (9,67%, vs. 24,22% d'immobiles), et les possesseurs de téléphone portable sont souvent des personnes actives ou ayant un bon niveau d'intéraction sociale, ce qui génère des activités et donc des déplacements (8,93%, vs. 19,57% d'immobiles).

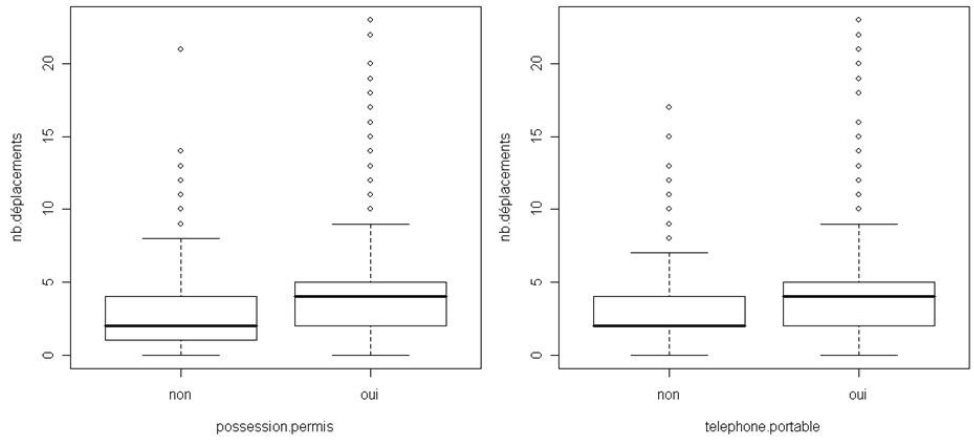

FIG. 72 – Relation entre la possession du permis de conduire ou d'un téléphone portable et le nombre de déplacements pour les répondants en face-à-face
Source : EMD Lyon (2006)

Impact de l'âge : une relation plus ambigue Les récents comparatifs effectués sur les données d'enquêtes transport françaises (Armoogum *et al.* (2005) et Hubert (2003)) montrent un fort taux d'immobiles chez les moins de 20 ans, qui devient faible avec l'entrée dans la vie active, pour remonter ensuite. Le taux d'immobilité est donc minimal pour les classes actives et maximal pour les jeunes et les personnes âgées. L'analyse descriptive sur les échantillons web et face-à-face de l'enquête déplacements de Lyon conduit aux mêmes résultats (figure 73). L'immobilité est de manière générale plus importante chez les personnes âgées, notamment à cause de leur statut de retraité et leurs difficultés

[150] Ces deux variables ne sont pas significatives dans l'échantillon web, puisque la plupart des internautes disposent du permis de conduire et d'un tétéphone portable.

à se déplacer (Madre *et al.*, 2003).

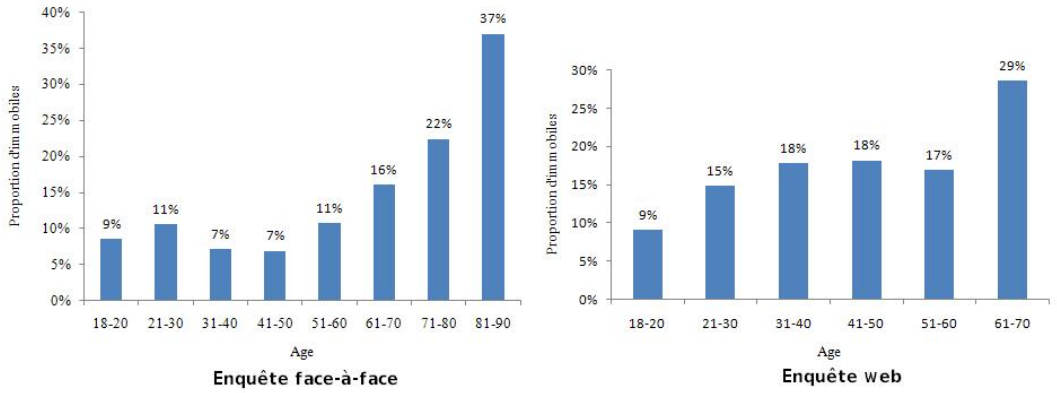

FIG. 73 – Relation entre le taux d'immobiles et l'âge des répondants par enquête

Source : EMD Lyon (2006)

Impact de la taille du ménage : deux facteurs à prendre en considération La relation entre le taux d'immobiles et le nombre de personnes n'est pas linéaire. Nous constatons plus de personnes immobiles parmi les ménages composés de une à deux personnes et ceux de cinq personnes et plus. Inversement, la part des mobiles est plus importante chez les ménages de trois à quatre personnes (figure 74).

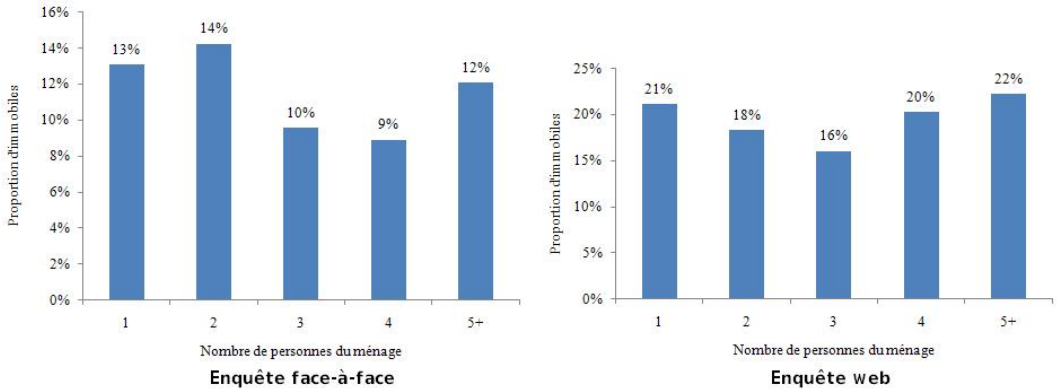

FIG. 74 – Relation entre le taux d'immobiles et le nombre de personnes du ménage par enquête

Source : EMD Lyon (2006)

Cette relation peut être mise en parallèle avec le nombre d'enfants. Les ménages composés de une à deux personnes sont majoritairement des personnes seules ou vivant en couples, sans enfant de moins de 18 ans. Ceux comprenant de 3 à 4 personnes sont pour la plupart des familles avec un à deux enfants. Comme le montre la figure 75, la relation entre le taux d'immobiles et le nombre d'enfants diffère selon le type d'enquête. Sur le web, l'immobilité semble faiblement corrélée aux nombre d'enfants présents dans le ménage (entre 18% et

20% d'immobiles). A contrario, dans l'enquête face-à-face, le taux d'immobiles est élevé chez les ménages sans enfant (14%), puis diminue jusqu'à la présence de deux enfants (7%), avant de remonter à partir de trois enfants.

Il semble donc intéressant de garder ces deux variables dans l'analyse, surtout pour rendre compte du phénomène de mobilité dans l'échantillon web.

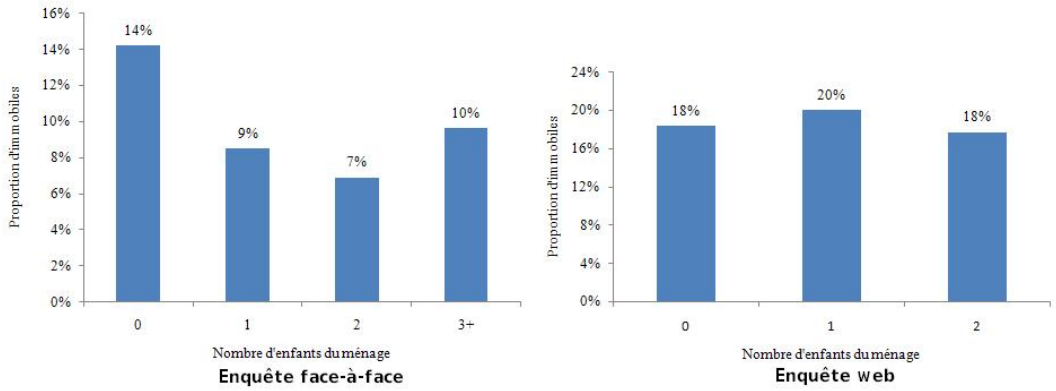

FIG. 75 – Relation entre le taux d'immobiles et le nombre d'enfants du ménage par enquête

Source : EMD Lyon (2006)

Les autres facteurs intéressants pour l'analyse de la mobilité Deux autres facteurs souvent mis en avant dans les modèles explicatifs de l'immobilité sont l'éloignement du lieu de résidence du centre de l'agglomération et le niveau de revenu du ménage. Les habitants de la périphérie sont plus immobiles que ceux du centre-ville (Madre *et al.*, 2007). Mais cette variable n'est pas significative dans nos modèles. Nous avons donc mené l'analyse sans en tenir compte. En ce qui concerne le niveau de revenu, le nombre élevé de non-réponses partielles ne permet pas de prendre en compte un nombre suffisant d'individus (25% dans l'enquête web et 30% dans l'enquête en face-à-face) dans les modèles explicatifs de l'immobilité, bien que les études montrent une forte immobilité chez les ménages déclarant un faible niveau de revenu (Madre *et al.*, 2007).

III.2 La population des répondants face-à-face

Les analyses ci-dessous autorisent la prise en compte des différentes facettes de la mobilité, en distinguant la décision de se déplacer du nombre de déplacements effectivement réalisés (tableaux 91, 92 et 93). L'ensemble des variables explicatives décrites vont permettre d'estimer la probabilité de mobilité, d'une part, et la fréquence des déplacements, d'autre part. Nous mettons plus particulièrement en évidence les effets statistiquement significatifs. Les tableaux suivants indiquent les principaux résultats de nos modèles, appliqués à l'échantillon des répondants en face-à-face. Ils sont suivis de deux indicateurs sur la qualité des modèles : le logarithme du maximum de vraisemblance et l'AIC

(Akaike's information criterion) (Akaike, 1974) [151].

Variables	Estimate	Std. Error	z-value	Pr(>|z|)	Sign.
Constante	1,094	0,031	35,033	< 2e-16	***
Sexe : homme	-0,055	0,010	-5,563	2,65e-08	***
Age	-0,003	0,001	-8,286	< 2e-16	***
Nb de voitures du ménage	0,186	0,014	13,072	< 2e-16	***
Possession du permis : oui	0,245	0,017	14,244	< 2e-16	***
Nb de personnes du ménage	-0,063	0,006	-10,476	< 2e-16	***
Activité : non actif	-0,062	0,011	-5,587	2,31e-08	***
Nb d'enfants du ménage	0,162	0,008	20,868	< 2e-16	***
Revenu déclaré : oui	0,090	0,010	8,558	< 2e-16	***
Téléphone portable : oui	0,069	0,013	5,446	5,17e-08	***

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 91 – Modèle de Poisson simple - Echantillon face-à-face

Variables	Estimate	Std. Error	z-value	Pr(>|z|)	Sign.
Constante	1,157	0,032	35,562	< 2e-16	***
Sexe : homme	-0,073	0,010	-7,094	1,30e-12	***
Age	-0,001	0,001	-1,998	0,046	*
Nb de voitures du ménage	0,151	0,015	10,189	< 2e-16	***
Possession du permis : oui	0,152	0,018	8,334	< 2e-16	***
Nb de personnes du ménage	-0,044	0,006	-6,932	4,14e-12	***
Activité : non actif	0,018	0,011	1,577	0,115	,
Nb d'enfants du ménage	0,149	0,008	18,357	< 2e-16	***
Revenu déclaré : oui	0,082	0,011	7,457	8,82e-14	***
Téléphone portable : oui	0,040	0,013	3,019	0,003	**

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 92 – Modèle de Poisson tronqué - Echantillon face-à-face

[151] AIC = - 2 * log(L) + 2 * k, où L est la vraisemblance maximisée et k le nombre de paramètres dans le modèle. L'AIC représente donc un compromis entre le biais qui diminue avec le nombre de paramètres estimés et la nécessité de décrire les données avec le plus petit nombre de paramètres possible).

Variables	Estimate	Std. Error	z-value	Pr(>|z|)	Sign.
Constante	2,925	0,183	15,982	< 2e-16	***
Sexe : homme	0,144	0,063	2,280	0,023	*
Age	-0,017	0,002	-8,417	< 2e-16	***
Nb de voitures du ménage	0,496	0,092	5,386	7,20e-08	***
Possession du permis : oui	0,537	0,079	6,762	1,36e-11	***
Nb de personnes du ménage	-0,252	0,032	-7,796	6,38e-15	***
Activité : non actif	-0,853	0,073	-11,627	< 2e-16	***
Nb d'enfants du ménage	0,238	0,047	5,032	4,85e-07	***
Revenu déclaré : oui	0,131	0,061	2,127	0,033	*
Téléphone portable : oui	0,207	0,071	2,912	0,004	**

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 93 – Modèle de Hurdle (binomial) - Echantillon face-à-face

Modèle	Poisson simple	Poisson Hurdle
AIC	52817	50458
logLik	-2,640e+04	-2,521e+04
df	10	20

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 94 – Maximum de vraisemblance - Echantillon face-à-face

La simple modélisation du nombre de déplacements par un modèle de Poisson ne permet pas de prendre en considération toutes les facettes du phénomène de mobilité. Le modèle Hurdle apporte une amélioration dans la compréhension des pratiques de mobilité des individus interrogés en face-à-face. En effet, la statistique qui combine la valeur de Log-vraisemblance du modèle Hurdle et celle du modèle de Poisson simple est significative (tableau 95).

Critères	Valeur
Statistique de test [152]	2379,55
Valeur du Chi-deux au risque $\alpha = 5\%$	31,41
Ddl	20
P-value	< 0,001

TAB. 95 – Significativité du modèle Hurdle - Echantillon face-à-face

La différence s'explique notamment par l'impact des variables 'Activité' et 'Sexe'. Le fait d'être inactif joue négativement sur la mobilité dans le modèle de Poisson simple (coefficient = -0,062 ; p-value < 0,001%), alors qu'il joue positivement sur la mobilité dans le modèle de Poisson tronqué (coefficient = 0,018 ; p-value = 11,5%). Le fait d'être un homme impacte négativement le nombre de déplacements déclarés dans le modèle de Poisson simple (coefficient = 0,055 ; p-value < 0,001%), alors qu'il influence positivement la décision de mobilité dans le modèle Hurdle (coefficient = 0,144, p-value = 2,3%).

Une autre conclusion importante est de savoir si les causes de la mobilité initiale (décision de se déplacer ou pas durant la période de référence) sont les mêmes que celles de la fréquence des déplacements, c'est-à-dire si les variables explicatives ont la même influence dans les deux équations. Par exemple, si nous considérons la variable 'Sexe', la décision de mobilité est plus importante pour les hommes (sexe.homme = 0,144, significatif au seuil de 2,3%). En revanche, une fois la décision de se déplacer prise, ce sont les femmes qui se déplacent le plus (sexe : homme = -0,073 dans le modèle de Poisson tronqué, significatif au seuil $<$ 0,001%). La conclusion est la même en ce qui concerne l'activité des répondants. Le fait d'être inactif pénalise la décision de mobilité (activité : inactif = -0.853 dans le modèle binomial, significatif au seuil $<$ 0,001%). Mais lorsqu'une personne inactive est mobile durant la période de référence, alors elle effectue davantage de déplacements qu'une personne active aux caractéristiques socio-économiques similaires (activité : inactif = 0,018 dans le modèle de Poisson tronqué, significatif au seuil de 11,5%). Les variables liées à la motorisation ('nombre de voitures du ménage' et 'permis de conduire'), la possession d'un téléphone portable et le fait de déclarer ses revenus ont un impact positif sur la décision de se déplacer, et le nombre de déplacements déclaré, quel que soit le modèle considéré. A contrario, plus les répondants appartiennent à des ménages de grande taille et plus ils sont âgés, plus leur probabilité de se déplacer et leur niveau de mobilité sont faibles (dans les modèles de Poisson simple et tronqué).

III.3 La population des répondants web

Nous reprenons les traitements de la section précédente, en ciblant à présent les répondants web (tableaux 96, 97 et 98).

Variables	Estimate	Std. Error	z-value	Pr(>|z|)	Sign.
Constante	1,001	0,144	6,951	3,62e-12	***
Age	-0,005	0,002	-2,237	0,025	*
Nb de personnes du ménage	0,098	0,037	2,631	0,009	**
Activité : non actif	-0,037	0,072	-0,523	0,601	
Nb d'enfants du ménage	-0,037	0,053	-0,689	0,491	
Revenu déclaré : oui	0,249	0,078	3,188	0,001	**
Vendredi : oui	-0,231	0,064	-3,622	2,93e-04	***

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 96 – Modèle de Poisson simple - Echantillon web

Variables	Estimate	Std. Error	z-value	Pr(>|z|)	Sign.
Constante	1.069	0.151	7.089	1.35e-12	***
Age	-0.001	0.003	-0.425	0.671	
Nb de personnes du ménage	0.042	0.040	1.055	0.292	
Activité : non actif	0.139	0.075	1.856	0.064	.
Nb d'enfants du ménage	0.097	0.056	1.733	0.083	.
Revenu déclaré : oui	0.097	0.083	1.168	0.243	
Vendredi : oui	-0.057	0.068	-0.848	0.396	

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 97 – Modèle de Poisson tronqué - Echantillon web

Variables	Estimate	Std. Error	z-value	Pr(>|z|)	Sign.
Constante	2,240	0,682	3,284	0,001	**
Age	-0,023	0,011	-2,148	0,032	*
Nb de personnes du ménage	0,350	0,196	1,784	0,074	.
Activité : non actif	-0,752	0,318	-2,363	0,018	*
Nb d'enfants du ménage	-0,722	0,280	-2,578	0,010	**
Revenu déclaré : oui	0,782	0,312	2,510	0,012	*
Vendredi : oui	-0,975	0,289	-3,378	0,001	***

Signif. : 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

TAB. 98 – Modèle de Hurdle (binomial) - Echantillon web

Modèle	Poisson simple	Poisson Hurdle
AIC	1601	1473
logLik	-793,5	-722,7
df	7	14

TAB. 99 – Maximum de vraisemblance - Echantillon web

Moins de variables sont significatives dans les modèles explicatifs de la mobilité appliqués à l'échantillon des répondants web, et tout particulièrement dans le modèle de Poisson tronqué (constante, activité et nombre d'enfants du ménage). Ceci peut s'expliquer notamment par la taille réduite de l'échantillon (361 individus). Dans cet échantillon, une personne sur cinq est déclarée immobile, ce qui rend nécessaire l'identification précise des facteurs déclencheurs de la mobilité. Le modèle binomial explicatif de la décision de mobilité semble performant, puisque l'ensemble des variables sont significatives.

Ici encore, le modèle Hurdle apporte une amélioration dans la comprehension des pratiques de mobilité des individus interrogés. En effet, la statistique qui combine la valeur de Log-vraisemblance du modèle Hurdle et celle du modèle de Poisson simple est significative (tableau 100).

Critères	Valeur
Statistique de test	141,53
Valeur du Chi-deux au risque $\alpha = 5\%$	23,68
Ddl	14
P-value	$< 0{,}001$

TAB. 100 – Significativité du modèle Hurdle - Echantillon face-à-face

La différence s'explique notamment par l'impact de la variable 'activité'. Dans le modèle de Poisson simple, cette variable n'est pas significative (p-value = 60,1%), car il y a un double effet, négatif sur la probabilité de se déplacer (coefficient = -0,752 dans le modèle de Hurdle ; p-value = 1,8%) et positif sur le niveau de mobilité (coefficient = 0,139 dans le modèle de Poisson tronqué ; p-value = 6,4%), les deux étant significatifs. La situation est identique pour la variable 'nombre d'enfants du ménage'. Le fait d'appartenir à une famille possédant de nombreux enfants semble réduire la probabilité de se déplacer (coefficient = -0,722 dans le modèle de Hurdle ; p-value = 1,0%), alors que son influence sur la mobilité est positive (coefficient = 0,097 dans le modèle de Poisson tronqué ; p-value = 8,3%).

Dans le modèle de Poisson simple, la mobilité augmente avec le nombre de personnes du ménage et le fait de déclarer son revenu (coefficients respectifs de 0,098 et 0,249 ; p-value = 9% et 1%). Ces variables jouent en fait sur la probabilité de se déplacer durant la période de référence (coefficients respectifs de 0,350 et 0,782 dans le modèle de Hurdle ; p-value = 7,4% et 1,2%), mais ne sont pas statistiquement significatives dans le modèle de Poisson tronqué explicatif du niveau de mobilité (p-value respectives de 29,2% et 24,3%).

Le jour de référence a également une influence non négligeable, puisque si les déplacements concernent la journée du vendredi, alors la probabilité de se déclarer immobile est relativement forte (coefficient dans le modèle Hurdle = -0,975 ; p-value $< 0{,}001\%$). De même, la probabilité de se déplacer diminue avec l'âge (coefficient dans le modèle Hurdle = -0,975 ; p-value 3,2%). Or, ces variables ne sont pas significatives dans le modèle de Poisson tronqué, explicatif du niveau de mobilité (p-value respectives = 39,6% et 67,1%). Dans le modèle de Poisson simple, cette nuance n'est pas prise en compte, le fait de se déplacer le vendredi et l'âge impactant négativement le nombre de déplacements déclaré (coefficients respectifs de -0,231 et -0,005 ; p-value $<0{,}001\%$ et 2,5%).

III.4 Synthèse

Un enseignement général de ce travail est que les déterminants de la mobilité sont différents de ceux de son intensité. Nous pouvons d'abord dissocier les déterminants exclusifs (variables qui n'affectent que la décision de mobilité ou son intensité) des déterminants qui affectent à la fois la mobilité et la fréquence des déplacements, certaines variables ayant des effets positifs ou négatifs à la fois sur la probabilité de mobilité et sur son intensité.

Variables	Echantillon face-à-face		Echantillon web	
	Décision de mobilité	Niveau de mobilité	Décision de mobilité	Niveau de mobilité
Sexe : homme	+	-	NA	NA
Age	-	-	-	-
Nb de voitures du ménage	+	+	NA	NA
Possession du permis : oui	+	+	NA	NA
Nb de personnes du ménage	-	-	+	+
Activité : non actif	-	+	-	+
Nb d'enfants du ménage	+	+	-	+
Revenu déclaré : oui	+	+	+	+
Téléphone portable : oui	+	+	NA	NA
Vendredi : oui	NA	NA	-	-

Sign. NA : variable non introduite, '+' : variable à effet positif, '-' : variable à effet négatif

TAB. 101 – Mise en perspective, par échantillon, des déterminants de la décision de mobilité et de son intensité

Ensuite, nous pouvons distinguer les facteurs qui influencent uniquement la décision de mobilité et ceux qui ne jouent que sur son intensité. Les résultats sont mis en perspective dans le tableau 101.

Comme nous l'avons vu dans le chapitre précédent, le genre n'est pas une variable qui impacte significativement la mobilité des répondants web. Il en est de même du téléphone portable, des variables de motorisation (possession du permis et nombre de voitures à disposition des personnes de 18 ans et plus). L'échantillon web étant constitué principalement de personnes actives ayant un bon niveau d'éducation et d'emploi, ces variables ne permettent pas de les distinguer. A contrario, le vendredi a un impact très important sur la mobilité des internautes, ce qui s'explique notamment par le choix du jour de connexion. En revanche, cette variable ne semble pas pertinente pour expliquer la mobilité des répondants à l'enquête en face-à-face, puisque les interviews à domicile sont réalisées de manière équivalente selon les différents jours de la semaine.

IV Conclusion

La littérature s'est souvent interrogée sur les déterminants du niveau de mobilité, sans déterminer si les facteurs qui conditionnent la mobilité sont les mêmes que ceux qui influencent son intensité. L'objectif de ce chapitre est de connaître les déterminants de la mobilité déclarée, sans éliminer de l'échantillon les individus qui sont restés immobiles durant la période de référence et de comparer les résultats obtenus sur les deux échantillons, web et face-à-face. En distinguant la décision de mobilité et la fréquence des déplacements, nous sommes capables de cerner efficacement le rôle que les différents facteurs jouent sur la mobilité quotidienne.

Tout au long de ce chapitre, nous avons montré qu'un modèle de Poisson simple n'était pas approprié pour rendre compte des facteurs qui influencent la mobilité quotidienne des individus. Nous avons donc utilisé un modèle de Poisson de type Hurdle. Les analyses économétriques montrent que les facteurs qui incitent à la mobilité ne sont pas forcément ceux qui impactent positivement le niveau de mobilité. Nos résultats confirment l'hypothèse que modéliser uniquement la décision d'effectuer un seul déplacement laisse de côté d'importantes informations concernant les multiples déplacements effectués par les répondants. Dans l'échantillon face-à-face par exemple (le plus représentatif), il ressort d'un modèle de Poisson simple que le fait d'être un homme influence négativement le niveau de mobilité. Le modèle de Poisson Hurdle met en évidence un effet plus complexe : les hommes ont une probabilité plus forte de réaliser un premier déplacement, mais une fois la décision de mobilité prise, alors leur niveau de mobilité est plus faible que celui des femmes. D'un autre côté, ne considérer qu'un modèle explicatif de la fréquence des déplacemants ne permet pas de rendre compte des facteurs qui influencent la décision mobilité durant la période de référence. L'exemple le plus significatif est l'activité, qu'il s'agisse de l'échantillon web ou face-à-face. Dans le modèle qui distingue les individus mobiles des immobiles, le fait d'être inactif incite fortement les individus à ne pas se déplacer durant la période de référence. Une fois la décision de mobilité prise, cette variable a en revanche une influence positive sur le niveau de mobilité des répondants.

Si nous comparons les échantillons web et face-à-face, l'impact des variables sociodémographiques sur le nombre de déplacements déclaré diffère parfois :

- le rôle ambigu du statut (niveau d'activité) est similaire entre les deux échantillons ;
- dans l'échantillon face-à-face, le sexe a une influence complexe sur la mobilité (impacte différemment la décision de mobilité et le niveau de mobilité), alors que sur le web c'est l'influence du nombre d'enfants qui n'est pas identique entre les modèles ;
- certains effets sont stables et similaires entre les deux échantillons, comme celui de l'âge (négatif) et le fait de déclarer ses revenus (positif) ;
- d'autres variables, comme le nombre de personnes du ménage ont une influence stable sur la mobilité, mais opposée entre les deux échantillons (négative en face-à-face et positive sur le web) ;
- le fait de se déplacer le vendredi impacte uniquement la mobilité des répondants web (négativement), car les internautes avaient le choix du jour de connexion et donc implicitement du jour de référence pour la déclaration des déplacements (ce qui n'est pas le cas dans l'échantillon face-à-face) ;
- les variables relatives à la motorisation et à la possession d'un téléphone portable influencent positivement la mobilité dans les modèles appliqués à l'échantillon face-à-face mais ne sont pas significatives dans ceux menés sur l'échantillon web (ces variables ne sont pas discriminantes pous la population d'internautes).

L'immobilité est par définition une variable difficile à évaluer, puisqu'il est possible que des participants sous estiment volontairement leur niveau de mobilité. Le grand nombre d'immobiles de l'enquête web pourrait donc provenir d'une attitude de 'refus mou', qu'adopteraient certains répondants, ainsi soupçonnés de déclarer qu'ils n'ont pas bougé de chez eux plutôt que de se plier au cadre relativement complexe du questionnaire. Ces résultats laissent penser que l'immobilité ne peut être modélisée uniquement à l'aide de variables socio-économiques, bien qu'il ne soit pas possible, à ce stade de l'analyse, de diférencier les réels immobiles des 'refus mous'.

Conclusion de la partie III

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, au contraire, refus de l'interview à domicile et saisie des réponses en ligne). Les coefficients estimés des variables du modèle ne sont plus biaisés et reflètent l'impact des variables socio-économiques 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 qui impactent la mobilité des répondants web sont le genre, 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. L'intérêt de la méthode consiste également à fournir des clés pour éventuellement pouvoir redresser l'échantillon des répondants web, afin de rendre leur niveau de mobilité comparable à celui des répondants en face-à-face, bien que ce traitement soit risqué à un niveau agrégé.

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 concernait certains types de motifs (non contraints) et de mode de transports (marche à pied). 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. Il est intéressant de développer un modèle explicatif de l'immobilité selon le mode d'enquête afin de mieux appréhender les facteurs à l'origine de cette sous-mobilité. 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. C'est le cas du modèle Hurdle. Deux résultats importants ressortent de ces analyses. D'abord, certaines variables ont un effet complexe sur la mobilité, qu'une simple régression linéaire ne peux mettre en évidence. Ainsi, quel que soit le mode d'enquête, le fait d'être actif augmente la probabilité de se déplacer, mais une fois la décision de se déplacer prise, l'inactivité pèse sur le niveau de mobilité. Ensuite, des variables ont un effet similaire sur la décision de mobilité et son intensité, mais ne sont pas significative dans les deux échantillons. C'est le cas par exemple des variables liées à la motorisation ou à la possession d'un téléphone portable, qui jouent positivement sur la mo-

bilité des répondants en face-à-face. Les internautes étant fortement motorisés et équipés en moyens de communication, ces variables n'apparaissent pas significatives dans l'analyse appliquée à l'échantillon web. Inversement, le vendredi n'est pas un jour propice à la mobilité des internautes, alors que cette variable n'est pas significativement explicative de la mobilité dans l'échantillon face-à-face (les interviews sont réparties de manière uniforme entre les différents jours de la semaine).

Si le mode d'enquête semble influencer les facteurs à l'origine des comportements de mobilité et d'immobilité, il serait intéressant de développer un modèle capable de distinguer les réels immobiles des 'soft refusals', d'une part, de quantifier l'impact du mode sur l'immobilité, d'autre part.