Recueil des articles du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Ajustement des estimations de l'enquête sur les transports par des capteurs routiers installés en permanence à l'aide de techniques de capture-recapture

Jonas Klingwort, Bart Buelens, and Rainer Schnell[1]

Résumé

L'intégration des données de capteurs en statistique officielle est particulièrement utile si elles peuvent être couplées aux données d'enquête et aux données administratives. Dans cette application, de tels jeux de données des Pays Bas sont couplés un à un avec un identificateur unique pour quantifier et ajuster la sous-déclaration dans les estimations ponctuelles d'enquête. L'échantillon probabiliste comprend des propriétaires de camions enregistrés qui déclarent les déplacements et le poids de la cargaison. Les capteurs mesurent en continu chaque camion qui passe sur certaines stations routières. Les méthodes de capture-recapture sont utilisées pour estimer la sous-déclaration dans l'enquête. L'hétérogénéité des probabilités de capture et de recapture est modélisée au moyen d'une régression logistique et de modèles log-linéaires. Les résultats montrent que l'approche est prometteuse en ce qui concerne la validation et l'ajustement des données d'enquête à l'aide de données de capteurs externes.

Mots clés : données volumineuses, couplage d'enregistrements, validation des données, sous-déclaration, estimation multisource, pesée en mouvement

1. Introduction

En statistique officielle, produire des statistiques non-biaisées à partir des données d'enquête devient plus difficile et cher. Par conséquent, la recherche s'accroît en ce moment en ce qui concerne l'usage des données volumineuses pour la production de statistiques officielles (Daas et coll. 2015). À ce jour, les données volumineuses sont rarement employées dans la production statistique à cause du mécanisme de génération des données (Buelens et coll. 2014). Toutefois, à long terme, l'usage des données volumineuses en statistique officielle est inévitable (Lohr et Raghunathan, 2017). Alors, au lieu d'utiliser une seule source de données volumineuses, la recherche portant sur la combinaison de différents jeux de données probabilistes et non-probabilistes est une façon prometteuse d'utiliser les données volumineuses en statistique officielle (Shlomo and Goldstein 2015). Plus particulièrement, les différents problèmes des enquêtes et des données volumineuses pourraient être minimisés si une enquête et un capteur (collectant les données volumineuses) mesurent la même variable cible de sorte qu'il est possible de coupler les microdonnées résultantes avec un identificateur unique. Partant de ce principe, nous couplons les données d'enquêtes, de capteurs et administratives pour les statistiques de transport. En utilisant le jeu de données couplées, nous appliquons des techniques de capture-recapture (CRC) pour valider, estimer et ajuster le biais des estimations ponctuelles dû à la sous-déclaration des variables cibles de l'enquête.

2. Contexte de recherche

Le nombre d'enquêtes réalisées a augmenté ces dernières décennies (Singer, 2016), mais au même moment les taux de non-réponse se sont aussi accrus (Meyer et coll., 2015). En particulier, les enquêtes par journal imposent un lourd fardeau de réponse et donnent des taux de réponse très bas (Krishnamurty, 2008). Dans le passé, les enquêtes journalières sur la mobilité et les transports ont été validées et ajustées à partir de données GPS.

Il a été démontré que ces enquêtes ont souvent un biais négatif à cause de la sous-déclaration, qui varie entre 2.6% (Hassounah et coll. 1993) et 81% (Bricka et Bhat 2006). Ces études ont utilisé des appareils mobiles munis de GPS,

[1] Jonas Klingwort, University of Duisburg-Essen & Statistics Netherlands, Forsthausweg 2, Germany, 47057 Duisburg & CBS-weg 11, Netherlands, 6412 EX Heerlen, jonas.klingwort@uni-due.de; Bart Buelens, VITO NV, Boeretang 200, Belgium, 2400 MOL; Rainer Schnell, University of Duisburg-Essen, Forsthausweg 2, Germany, 47057 Duisburg

qui sont transportés par les véhicules ou les répondants. En pratique, les appareils GPS causent des problèmes à cause des coupures plannifiées ou imprévues, des délais dûs à la mise en veille, des problèmes de batterie, ou au fait que l'appareil n'est pas transporté (Bricka, Sen et coll. 2012; Shen et Stopher 2014). Au lieu d'utiliser des appareils GPS mobiles, nous utilisons des capteurs routiers installés de façon permanente pour valider et ajuster les estimations.

3. Données

La population cible de l'enquête des Pays Bas (2015) est la flotte commerciale hollandaise, excluant l'armée, les véhicules agricoles et commerciaux vieux de plus de 25 ans (avec un poids ≥ 3.5t et une capacité de chargement ≥ 2t). L'échantillon comprend 33 817 camions tirés du registre national des véhicules. Un objectif central de l'enquête journalière obligatoire est de recueillir les données sur les poids de cargaison transportés par les camions. Par conséquent, les propriétaires de camions doivent déclarer le nombre de jours où le camion a été utilisé. 3 597 cas sont classés comme de la non-réponse. Les catégories de réponse concernant les activités associées aux camions sont: camion utilisé (22 454), camion non-utilisé (5 304), et camion loué (2 462). Ce dernier cas est défini comme de la non-réponse technique et est exclu de l'analyse puisque la validité de la réponse ne peut être vérifiée. La sous-déclaration est attendue à cause de la non-réponse et les réponses invalides en répondant par erreur qu'un camion n'est pas utilisé.
Les données des capteurs sont recueillies par le réseau de capteurs routiers pour la pesée en mouvement (WIM, Weigh-in-Motion) géré par l'administration routière nationale des Pays Bas comprenant 18 stations de mesure. Au passage, le poids du véhicule est mesuré.

Bien que les capteurs ne couvrent pas toutes les autoroutes aux Pays Bas, ils sont installés aux emplacements ayant un grand volume de circulation et aux carrefours routiers. En 2015, il y avait 35 669 347 camions enregistrés dont, en utilisant la combinaison unique de la plaque d'immatriculation et du jour, 44 011 pouvaient être couplés un-à-un à des trajets déclarés dans l'enquête. Des contrôles de qualité et du nettoyage des données ont été appliqués en suivant les lignes directrices développées par Enright et OBrien (2011). Des corrections d'erreur de mesure ont été appliquées aux poids des essieux en utilisant l'imputation conditionnelle basée sur la moyenne. En utilisant une règle déterministe de correction d'erreur, le poids d'un essieu est imputé par le poids moyen des essieux restants si le poids mesuré est plus grand que 20t. Si plusieurs essieux ont un poids excédant 20t, la valeur moyenne des essieux restants avec un poids plus petit que 20t est utilisée ici aussi. La modélisation prédictive basée sur la régression linéaire ($r^2 = adj.r^2 = 0.54$) a été appliquée pour corriger les poids des camions roulant en dehors de l'intervalle [60;120] km/h de vitesse recommandé. Pour 17 321 des 44 011 camions appariés, aucun trajet ne pouvait être couplé au camion. Pour 11 341 cas, la reconnaissance optique de caractères a failli et pour 5 980 cas le trajet n'était pas listé dans le registre. Les poids manquants ont été imputés par le poids moyen d'un trajet vide, conditionnel à la classification automatique du camion et sa capacité de chargement. Le registre hollandais des véhicules et le registre des entreprises sont couplés aux micro-données en utilisant la plaque d'immatriculation et le trimestre annuel comme variables de couplage. Puisque les capteurs mesurent le poids de toute l'unité (camion, remorque et cargaison), les poids du camion et de la remorque ont été soustraits en utilisant l'information du registre des véhicules. La valeur résultante est le poids de la cargaison transportée, qui correspond à la définition du poids déclaré dans l'enquête. Dans 3 945 cas, les poids de cargaison furent négatifs et mis à zéro. Enfin, une correction proportionnelle du biais a été appliquée, en calant les poids de cargaison mesurés par le capteur à ceux déclarés dans l'enquête. Le facteur de correction a été obtenu du sous-ensemble de véhicules qui ont été observés à la fois par l'enquête et par les capteurs. Cela a produit une réduction approximative de 14% pour les poids de cargaison mesurés par les capteurs. Les observations avec des données manquantes dans les registres ont été exclues de l'analyse (ce qui explique la différence entre les 44 011 appariements et les 43 775 appariements dans la Table 4.2-1).

4. Méthodes

Soit $\delta_{i,j}^{svy}$ l'indicateur, qui est égal à 1 si le véhicule i était sur la route le jour j de sa période d'enquête selon la réponse à l'enquête, et qui est égal à 0 sinon. Soit $\delta_{i,j}^{wim}$ l'indicateur, qui est égal à 1 si le véhicule i est enregistré par une station de capteurs le jour j, et qui est égal à 0 sinon. $\Theta_{i,j}$ est défini comme le poids de cargaison transporté par le camion i le jour j. Lorsque $\delta_{i,j}^{svy} = 1$, la somme des poids de cargaison déclarés dans l'enquête est utilisée, sinon lorsque $\delta_{i,j}^{wim} = 1$ les mesures de la cargaison du capteur sont utilisées. Si le véhicule fut enregistré par les capteurs plusieurs fois par jour, le maximum des poids mesurés à ces occasions est pris. Deux variables cibles sont considérées:

le nombre total de camions-jours (D) et le poids total de cargaison transportée (W). Un camion-jour est défini comme un jour où le camion était sur la route aux Pays Bas. Les statistiques d'enquête habituelles sont des estimations par post-stratification, avec des poids calculés pour prendre en compte le plan de sondage et pour corriger la sélection dûe à la non-réponse. Les totaux pour D et W sont estimés par $\widehat{D}^{SURV} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy}\right)$ et $\widehat{W}^{SURV} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \Theta_{i,j}\right)$. Les observations des capteurs sont simplement ajoutées aux observations de l'enquête pour produire les estimations $\widehat{D}^{SURVX} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7} \delta_{i,j}^{svy} \vee \delta_{i,j}^{wim}\right)$ et $\widehat{W}^{SURVX} = \sum_{i=1}^{N}\left(w_i \sum_{j=1}^{7}\left(\delta_{i,j}^{svy} \vee \delta_{i,j}^{wim}\right) \Theta_{i,j}\right)$. C'est une façon simple d'inclure les données des capteurs et de fournir une borne inférieure pour les estimateurs CRC. Coupler les données d'enquête et de capteurs produit trois sous-ensembles d'unités : les éléments uniquement dans l'enquête, ceux uniquement dans les données des capteurs et ceux uniquement dans les deux jeux de données (Table 4.2-1). La case vide représente les camions et trajets qui n'ont pas été déclarés dans l'enquête et n'ont pas été enregistrés par un capteur, respectivement. Dans la présente étude, la première occasion de capture est l'enquête où les camions sont considérés comme saisis et inscrits à des journées particulières de la période d'enquête ($\sum_{i,j} \delta_{i,j}^{svy}$). La seconde occasion de capture est le capteur où $\left(\sum_{i,j} \delta_{i,j}^{wim}\right)$ sont capturés au total, dont $\left(\sum_{i,j} \delta_{i,j}^{svy} \wedge \delta_{i,j}^{wim}\right)$ sont recapturés. L'estimateur de Lincoln-Petersen (Lincoln 1935; Petersen 1893) utilise les totaux de ces sous-ensembles pour estimer les tailles de population (D) et (W) par $\widehat{D}^{LP} = \frac{n_1 n_2}{m_2}$ et $\widehat{W}^{LP} = \frac{\left(\sum_{i,j} \delta_{i,j}^{svy} \Theta_{i,j}\right)\left(\sum_{i,j} \delta_{i,j}^{wim} \Theta_{i,j}\right)}{\sum_{i,j}\left(\delta_{i,j}^{svy} \wedge \delta_{i,j}^{wim}\right)\Theta_{i,j}}$.

L'approche basée sur la vraisemblance proposée par Huggins (1989) et Alho (1990) modélise l'hétérogénéité des probabilités de capture à partir des variables explicatives conditionnées aux éléments saisis. Un modèle logistique est utilisé pour modéliser les probabilités de capture de chaque élément à chaque occasion. Ainsi, les variables explicatives sont utilisées pour modéliser les probabilités de capture $\hat{P}_{ij}^{s}$ et $\hat{P}_{ij}^{w}$, qui sont les probabilités de capture pour l'enquête et le capteur, respectivement. L'estimateur de Horvitz-Thompson (Horvitz and Thompson 1952) est utilisé pour estimer D et W par $\widehat{D}^{HUG} = \sum_{i,j} \frac{1}{\hat{\psi}_{ij}}$ et $\widehat{W}^{HUG} = \sum_{i,j} \frac{\Theta_{i,j}}{\hat{\psi}_{ij}}$, avec $\hat{\psi}_{ij} = 1 - \left(1 - \hat{P}_{ij}^{s}\right)\left(1 - \hat{P}_{ij}^{w}\right)$. L'estimateur HUG_{int} est le modèle avec seulement le terme constant. Fienberg (1972) a présenté les modèles log-linéaires pour l'estimation de la taille d'une population fermée. Afin de modéliser l'hétérogénéité des probabilités de capture dans l'enquête (A) et le capteur (B), un nombre variable de variables explicatives disponibles peuvent être inclues dans le modèle. Étant donné la variable explicative X, la table de contingence 2x2 est développée en une table de contingence 4x4 $\log m_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_x^X + \lambda_{ax}^{AX} + \lambda_{bx}^{BX}$. Pour chaque niveau des variables explicatives inclues, une taille de sous-population est estimée, dont la somme donne la taille totale de la population. Cette méthode est utilisée pour estimer $\widehat{D}^{LL}$ et $\widehat{W}^{LL}$. À partir de techniques CRC, le nombre dans la case vide est estimé. Deux variables cibles de l'enquête sont estimées: le nombre de camions-jours (D) et le poids de cargaison transporté correspondant (W). En outre, tous les estimateurs sont appliqués de façon stratifiée. Puisque cette étude porte sur les soupçons à propos de la sous-déclaration causée par la non-réponse et les mauvaises déclarations dans le RTFS, les (non-) répondants du RTFS représentent la population étudiée. Le nombre de véhicules à l'étude est N. Par conséquent, les indicateurs $\delta_{i,j}^{svy}$ et $\delta_{i,j}^{wim}$ pour estimer D et W sont divisés en S strates, avec N_s unités d'échantillonnage dans la strate s. Dans chaque strate $\widehat{D}_s$ et $\widehat{W}_s$ sont estimés. Les strates sont basées sur les variables explicatives dans les modèles (voir la section 4.1). Dans chaque strate le niveau le plus probable de sous-déclaration est estimé.

4.1 Sélection du modèle et estimation de variance

Les variables explicatives des modèles linéaires et logistiques sont sélectionnées de façon itérative (à parti du BIC, Bayesian Information Criterion). Puisque le modèle log-linéaire permet seulement les variables catégorielles, et pour conserver l'information complète des variables explicatives, la sélection du modèle est basée sur un modèle logistique. Dans le modèle log-linéaire, les cinq variables avec la plus grande pusissance prédictive dans les deux modèles logistiques sont combinées. À cette fin, les variables explicatives continues ont été catégorisées à partir de leurs quantiles. Avec $\delta_{i,j}^{svy}$ comme variable dépendante du modèle logistique, les variables explicatives sélectionnées étaient: la classification de l'activité économique (NACE, classification of economic activity), la classification de la taille de l'entreprise, la capacité de chargement totale de la flotte, le nombre de roues, la puissance en chevaux, la masse maximale du camion, la masse du camion vide, la masse maximale de la remorque, le statut du propriétaire (personne ou entreprise), et la province où se situe le propriétaire. Avec $\delta_{i,j}^{wim}$ comme variable dépendante, les variables explicatives suivantes étaient sélectionnées: la classification de l'activité économique (NACE), transport commercial ou privé, la classification de la taille de l'entreprise, la taille de la flotte de véhicules, la capacité de

chargement totale de la flotte, la classe du camion, le type de carburant, la puissance en chevaux, la masse du camion vide, la masse maximale de la remorque, le nombre d'essieux, la largeur du camion, la longueur du camion, le statut du propriétaire (personne ou entreprise), la province où se situe le propriétaire, l'année de fabrication, et la classification du véhicule. Les variables sélectionnées pour le modèle log-linéaire étaient la classification de l'activité économique (NACE), transport commercial ou privé, la classification de la taille de l'entreprise, la taille de la flotte de véhicules, la capacité de chargement totale de la flotte, le nombre de roues et la puissance en chevaux. Puisque les camions sont les unités d'échantillonnage et non les nombres de camions-jours, le bootstrap a été utilisé pour prendre en compte cet effet de grappe dans les données (il y a plus de camions-jours que d'unités d'échantillonnage). De plus, les poids de cargaison sont regroupés dans les camions et ne sont pas i.i.d. Un échantillon aléatoire simple avec remise a été utilisé pour tirer les échantillons bootstrap. Aux fins d'estimation, un échantillon bootstrap comprend tous les éléments, à la fois de l'enquête et du capteur, qui sont disponibles pour les véhicules de l'échantillon bootstrap. La moyenne de la distribution bootstrap est calculée pour s'assurer que la procédure n'est pas biaisée. Les 2.5ème et 97.5ème centiles de la distribution bootstrap sont utilisés pour estimer les limites des intervalles de confiance de 95%.

4.2 Couplage des données d'enquête et de capteurs

Dans la Table 4.2-1, les résultats du couplage entre les données d'enquête et de capteurs sont présentées. Le panneau gauche de la table montre 94 338 camions-jours déclarés dans l'enquête. Les capteurs ont enregistré 43 775 camions-jours dont 34 131 ont été déclarés dans l'enquête. 9 644 camions-jours ont été enregistrés par les capteurs, qui n'ont pas été déclarés dans l'enquête. Les capteurs n'ont pas enregistré 60 207 camions-jours, qui ont été déclarés dans l'enquête. Le panneau droit montre le poids de cargaison transporté en kilotonnes (kt) sur les camions-jours déclarés.

Tableau 4.2-1
Captures de camions-jours (D) et poids de cargaison transporté (W) dans l'enquête et selon les capteurs.

D	Enquête			W	Enquête		
Capteur	déclaré	non déclaré	∑	Capteur	déclaré	non déclaré	∑
enregistré	34 131	9 644	43 775	enregistré	376,83	99,13	475,96
non enregistré	60 207	–	60 207	non enregistré	576,88	–	576,88
∑	94 338	9 644	103 982	∑	953,71	99,13	1 052,84

953,71 kt ont été déclarés dans l'enquête. 475,96 kt ont été déclarés par les capteurs, dont 376,83 kt déclarés dans l'enquête. De plus 99,13 kt ont été enregistrés par les capteurs et n'ont pas été déclarés dans l'enquête. Les capteurs n'ont pas enregistrés 576,88t, qui ont été déclarés dans l'enquête.

5. Résultats

Table 5-1 montre les estimations de l'enquête et de CRC pour D et W. Selon $SURVX$, le niveau de sous-déclaration pour D et W est d'environ 6%. Les estimateurs HUG et HUG_{int} produisent environ 7% de sous-déclaration pour D et environ 13% pour W. L'estimateur LP produit environ 16% de sous-déclaration pour D et 19% pour W. Pour les deux variables cibles D et W le niveau de sous-déclaration le plus probable est d'environ 19% pour D et d'environ 20% pour W selon LL.

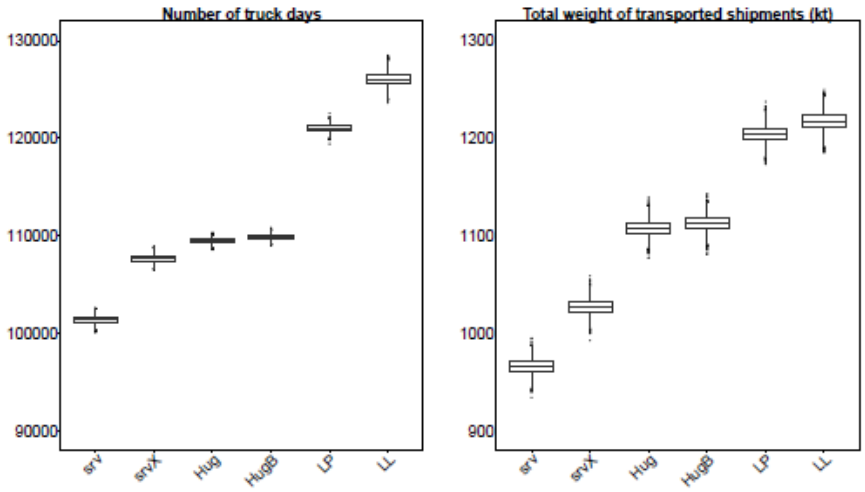
Tableau 5-1
Estimation d'enquête et de CRC pour D et W, variance bootstrap, erreur type et intervalle de confiance.

Estimateur	Estimation ponctuelle	Moyenne bootstrap	Erreur type bootstrap	Intervalle de confiance	Sous-déclaration estimée (en %)
$\hat{D}^{SURV}$	101 390	101,399	395,96	[100 643; 102 197]	–
$\hat{D}^{SURVX}$	107 666	107,672	380,66	[106 923; 108 441]	5,83
$\hat{D}^{HUG}$	109 439	109,440	244,73	[108 975; 109 926]	7,35
$\hat{D}^{HUG_{int}}$	109 882	109,885	246,86	[109 412; 110 376]	7,73
$\hat{D}^{LP}$	120 994	120,996	363,75	[120 304; 121 723]	16,2
$\hat{D}^{LL}$	125 954	126,034	737,46	[124 673; 127 577]	19,5
$\hat{W}^{SURV}$	965,3	965,23	8,20	[949,33; 981,40]	–
$\hat{W}^{SURVX}$	1026,83	1026,69	8,37	[1009,94; 1043,53]	5,99
$\hat{W}^{HUG}$	1108,58	1108,36	8,32	[1091,65; 1124,37]	12,92
$\hat{W}^{HUG_{int}}$	1112,59	1112,40	8,34	[1095,52; 1128,38]	13,24
$\hat{W}^{LP}$	1204,60	1204,38	9,14	[1185,83; 1221,89]	19,87
$\hat{W}^{LL}$	1216,85	1217,40	9,74	[1197,73; 1236,08]	20,67

La Figure 5-1 montre six estimateurs différents et la variance d'échantillonnage bootstrap (à partir des 3 000 échantillons bootstrap). Les six différentes estimations ponctuelles sont proches de la médiane et ne sont donc pas présentées.

Contrairement aux estimateurs basés sur la vraisemblance conditionnelle, la différence plus large entre les estimateurs basés sur la vraisemblance marginale montre un plus grand effet de la modélisation de l'hétérogénéité basée sur les variables explicatives. Il est recommandé de s'appuyer sur les estimations de LL puisqu'elles sont basées sur la vraisemblance complète et prennent en compte l'hétérogénéité des probabilités de capture.

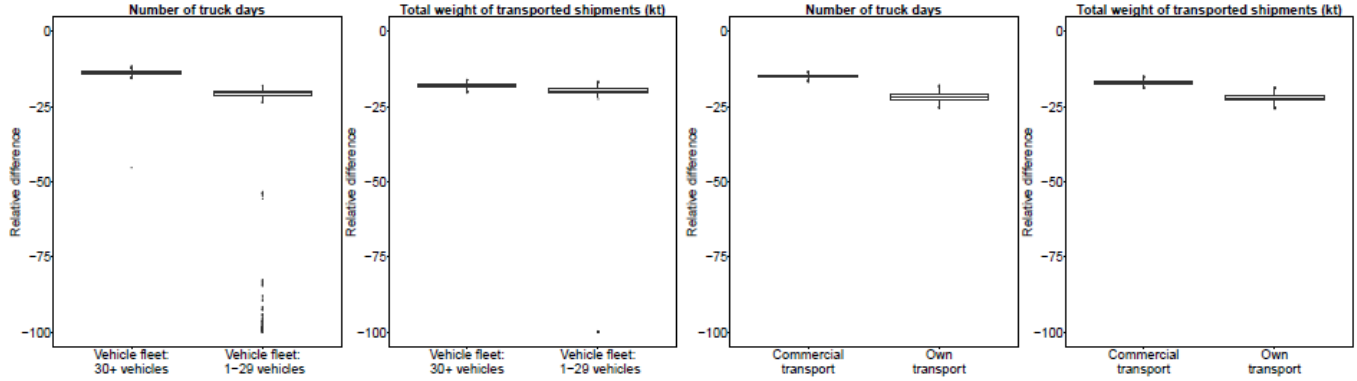
Figure 5-1
Effet d'estimateur sur l'estimation bootstrap des camions-jours et du poids de cargaison transporté.



Par ailleurs, pour l'analyse stratifiée de D et W, la différence relative entre $SURV$ et LL est montrée à la Figure 5-2. Ici aussi, les estimations ponctuelles sont proches de la médiane et ne sont donc pas présentés. Pour les petites flottes de véhicules (1-29 véhicules) le niveau de sous-déclaration pour D est de 20% et pour W de 19%. Les flottes de véhicules plus larges (plus de 30 véhicules) montrent 13% de sous-déclaration pour D et de 18% pour W. Le

transport commercial montre 15% de sous-déclaration pour D et de 17% pour W. Pour le transport privé, le niveau de sous-déclaration le plus probable est de 22% à la fois pour D et pour W.

Figure 5-2
Stratification par grandeur de flotte de véhicules et type de transport, montrant l'effet de LL sur les estimations bootstrap et la différence relative entre $SURV$ et LL.



6. Conclusion

Nous avons démontré une utilisation particulière des données volumineuses en statistique officielle pour l'estimation et l'ajustement du biais de sous-déclaration. En utilisant des techniques CRC, des microdonnées provenant d'une enquête, de capteurs, et de donnés admnistratives ont été couplées. La combinaison proposée de sources de données et de méthodes semble produire des estimations raisonables selon la littérature. La méthode présentée ici s'applique à toute étude de validation où des données d'enquête, administratives et de capteur (ou tout autre source source externe de données volumineuses) peuvent être couplées au niveau des micro-données à l'aide d'un identificateur unique. Toutefois, puisque les capteurs ne sont répartis au hasard, les données des capteurs pourraient être biaisées. De plus, le logiciel de reconnaissance optique de caractères (OCR, Optical Character Recognition) ne reconnaît pas chaque plaque frontale ou arrière et les erreurs d'appariement pourraient influencer les résultats. Enfin, des méthodes d'imputations ont été utilisées pour estimer les mesures de capteurs manquantes. Une étude systématique est en cours concernant les effets de ces problèmes sur les résultats.

Bibliographie

Alho, J. M. (1990). Logistic regression in capture-recapture methods. Biometrics, 46, 623–635.

Bricka, S., & Bhat, C. (2006). Comparative analysis of global positioning system-based and travel survey-based data. Transportation Research Record: Journal of the Transportation Research Board, 1972, 9–20.

Bricka, S., Sen, S., Paleti, R., & Bhat, C. R. (2012). An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transportation Research Part C: Emerging Technologies, 21(1), 67–88.

Buelens, B., Daas, P., Burger, J., Puts, M., & van den Brakel, J. (2014). Selectivity of big data. CBS Discussion Paper, (2014–11).

Daas, P. J. H., Puts, M. J., Buelens, B., & van den Hurk, P. A. M. (2015). Big data as a source for official statistics. Journal of Official Statistics, 31(2), 249–262.

Enright, B., & OBrien, E. J. (2011). Cleaning weigh-in-motion data: Techniques and recommendations. Techinal Report, Dublin Institute of Technology & University College Dublin.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. Biometrika, 59(3), 591–603.

Hassounah, M. I., Cheah, L.-S., & Steuart, G. N. (1993). Underreporting of trips in telephone interview travel surveys. Transportation Research Record, 1412, 90–94.

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663–685.

Huggins, R. M. (1989). On the statistical analysis of capture experiments. Biometrika, 1(76), 133–140.

Krishnamurty, P. (2008). Diary. In P. J. Lavrakas (Editor), Encyclopedia of survey research methods (Volume 1, Pages 197–199). Thousand Oaks: Sage.

Lincoln, F. C. (1935). The waterfowl flyways of north america. Washington: United States Department of Agriculture.
Lohr, S. L., & Raghunathan, T. E. (2017). Combining survey data with other data sources. Statistical Science, 32(2), 293–312.

Meyer, B. D., Mok, W. K., & Sullivan, J. X. (2015). Household surveys in crisis. Journal of Economic Perspectives, 29(4), 199–226.

Petersen, C. G. J. (1893). On the biology of our flat-fishes. Kjøbenhaven: The Danish Biological Station.

Shen, L., & Stopher, P. (2014). Review of GPS travel survey and GPS data-processing methods. Transport Reviews, 34(3), 316–334.

Shlomo, N., & Goldstein, H. (2015). Editorial: Big data in social research. Journal of the Royal Statistical Society, Series A, 178(4), 787–790.

Singer, E. (2016). Reflections on surveys' past and future. Journal of Survey Statistics and Methodology, 4(4), 463–475.