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

Regrouper et explorer des combinaisons de données de transactions électroniques à la recherche de plans de rechange pour l'EBM

Anders Holmberg[1]

Résumé

Lorsque Statistics Norway a annulé son enquête sur le budget des ménages de 2018, c'était en raison de préoccupations liées au compromis entre les coûts et la qualité et parce que l'organisme doutait si une approche essentiellement traditionnelle fondée sur une enquête et l'utilisation d'un journal serait satisfaisante. La décision d'annuler l'enquête a déclenché une intensification des recherches en vue d'acquérir d'autres sources de données sur la consommation des ménages. Par conséquent, le paysage national des données sur les transactions peut maintenant être bien décrit, et trois sources différentes de données sur les transactions électroniques et la façon de les combiner sont examinées à titre expérimental. L'une des sources comprend des données sur les transactions provenant d'un important fournisseur de services de paiement de la Norvège (les transactions par carte et autres transactions électroniques, ainsi que les transactions interentreprises). Le taux de couverture est assez élevé puisque les données englobent la plupart des transactions par carte effectuées en Norvège. Nous examinons également les données des caisses enregistreuses des chaînes de magasins de détail et des échantillons de données sur les membres des chaînes de magasins de détail (cartes de fidélité). Toutes ces sources de données sont intéressantes en soi, mais ce qui ajoute vraiment de la valeur, du moins du point de vue de la consommation des ménages, c'est de trouver des façons de combiner les différentes sources. Il peut s'agir par exemple de méthodes de couplage des enregistrements de transaction par carte aux données des caisses enregistreuses pour combiner la dimension démographique et la dimension de la consommation selon le niveau détaillé de la classification des fonctions de la consommation individuelle (COICOP). Le présent document traite des possibilités et des expériences méthodologiques et techniques découlant de ces travaux jusqu'à ce jour (milieu de l'année 2018).

Mots-clés : Combinaison de sources de données; consommation des ménages; transactions de paiement électroniques.

1. Introduction

1.1 Contexte

Les enquêtes dont la collecte de données repose sur la tenue d'un journal ont toujours visé à motiver et maintenir la participation des répondants (voir Edgar et coll. [2013]). Si on les compare aux enquêtes qui utilisent un instrument de mesure classique, le fardeau de réponse est généralement plus lourd (sur le plan du volume et de la fréquence). Le répondant doit non seulement retenir, consigner et déclarer différents faits concernant divers sujets, événements et activités, mais il doit également le faire assidûment sur une plus longue période. Dans ce contexte, il est normal de croire qu'il y a un risque accru de non-réponse et plus d'erreurs de mesure. Par ricochet, cela fait augmenter les coûts, car des activités d'assurance de qualité sont nécessaires.

Un groupe d'enquêtes sociales importantes qui présentent généralement ces caractéristiques et pour lesquelles ce dilemme se pose sont les enquêtes auprès des ménages qui recueillent des données sur le revenu et la consommation. Dans certains pays, on les appelle enquêtes sur le budget des ménages (EBM). Dans d'autres pays, seulement la composante consommation fait l'objet d'une collecte directe, d'où le nom d'enquête sur les dépenses des ménages EDM). C'est sur cette dernière que nous nous penchons ici. Je vais décrire les travaux exploratoires réalisés par Statistics Norway pour identifier, acquérir et évaluer des données sur les dépenses des ménages qui sont comptabilisées par les transactions électroniques ainsi que la façon dont elles peuvent être utilisées à des fins statistiques.

1.2 Enquête sur les dépenses de consommation de Statistics Norway

[1]Anders Holmberg, Statistics Norway, Postboks: 2633 St. Hanshaugen, Norway, 0131 Oslo.

Depuis 1958, en Norvège, Statistics Norway (SSB) produit des statistiques sur la consommation des ménages en utilisant une enquête par sondage appelée FBU (abréviation de Forbruksundersøkelsen). Les premières enquêtes réalisées en 1958, en 1967 et en 1973 visaient à obtenir des estimations détaillées de la consommation privée pour mettre à jour les pondérations de l'Indice des prix à la consommation. Entre 1974 et 2009, la FBU a été réalisée chaque année en y ajoutant un autre objectif : surveiller les habitudes de consommation de différentes catégories de ménages. En 2010, il a été décidé de faire une pause de deux ans jusqu'à 2012, et aujourd'hui, la FBU de 2012 est la plus récente publiée par SSB.

Au fil du temps, la conception de base de l'enquête n'a pour ainsi dire pas changé. Les données d'entrée de la FBU de 2012 provenaient d'un questionnaire, d'un journal (papier ou électronique) et de reçus de caisse. La taille de départ de l'échantillon était de 7 000 ménages. Deux entrevues étaient réalisées : une entrevue initiale et une entrevue finale après la période de 14 jours pendant laquelle les répondants consignaient des données dans leur journal. Des entrevues téléphoniques et en personne (au moyen de visites) étaient effectuées pendant la période de collecte sur le terrain qui durait 15 mois. Le taux de non-réponse s'établissait à 51 %, ce qui traduit une tendance croissante. En 1983, le taux de non-réponse de la FBU était de 33 % (Holmøy et Lillegård, 2014).

Après la FBU 2012, SSB a entrepris de développer et de moderniser l'enquête. L'objectif était de lancer une FBU modernisée en 2017, mais le lancement a par la suite été repoussé à 2018. L'amélioration de la conception de la collecte des données est l'un des principaux résultats des travaux de modernisation de l'enquête. Un outil numérique à remplir soi-même pour la consignation des données dans le journal ainsi que le codage automatique accru des reçus de caisse ont été proposés pour assurer une qualité suffisante des données et limiter les coûts. Mais ces suggestions n'ont pas été suffisamment convaincantes. À la fin de 2016, le conseil des directeurs de SSB a décidé d'annuler la FBU 2018 et de réaffecter les fonds à d'autres projets de l'organisation. La prochaine FBU est provisoirement fixée pour 2022.

1.3 À la recherche de sources de données alternatives

Pour le projet de modernisation de la FBU, on a également cherché d'autres sources de données pour évaluer les dépenses des ménages. SSB a tenté d'utiliser son autorité législative pour acquérir les *données sur les membres des programmes de fidélisation* des chaînes de détaillants. Sa tentative a été bloquée lors des négociations avec les fournisseurs. Ce n'est que trois ans plus tard, soit en 2017, qu'un ensemble de données d'essai ont été transmises à SSB à des fins d'exploration statistique, mais il était trop tard pour avoir une incidence sur la décision d'annuler l'enquête. Qui plus est, l'évaluation et l'analyse des possibilités d'utilisation de ces données pour obtenir des statistiques représentatives des dépenses de consommation n'ont pas fournis des résultats prometteurs. SSB s'est plutôt tourné vers d'autres sources de données; pour les ventes au détail quotidiennes, il a misé sur les données des *caisses enregistreuses et des lecteurs optiques* (souvent désignées par données scanner) provenant des transactions de vente numérisées.

Les données transactionnelles des caisses enregistreuses contiennent des détails pour chaque vente. Elles indiquent les produits vendus, les prix, le volume, le point de vente, l'heure, etc., mais ne nous renseignent pas sur le consommateur et le ménage. SSB a toutefois examiné d'autres sources et un autre moyen de combler cette lacune.

En étudiant le contexte des transactions financières en Norvège, on constate que les *transactions de paiement électroniques* des banques, des institutions financières et d'autres entreprises offrant des services de paiement peuvent être accessibles pour établir des statistiques, car il existe un dépositaire central de données en Norvège. Même si la consultation des données n'est pas simple, le fait qu'il existe un dépositaire facilite grandement l'accessibilité des données. Les données concernent les transactions électroniques des particuliers et des entreprises, et indiquent le montant, le type de transaction, la date et les parties au niveau de la transaction. Les données des caisses enregistreuses renferment beaucoup de détails et leur couverture est plus ou moins adéquate, alors que les données transactionnelles de paiement sont moins détaillées, mais couvrent (en théorie) toutes les transactions de paiement électronique effectuées en Norvège. Dans les pages qui suivent, nous allons examiner l'idée selon laquelle ces sources peuvent se compléter pour obtenir des statistiques sur les dépenses de consommation dans le domaine de la vente au détail.

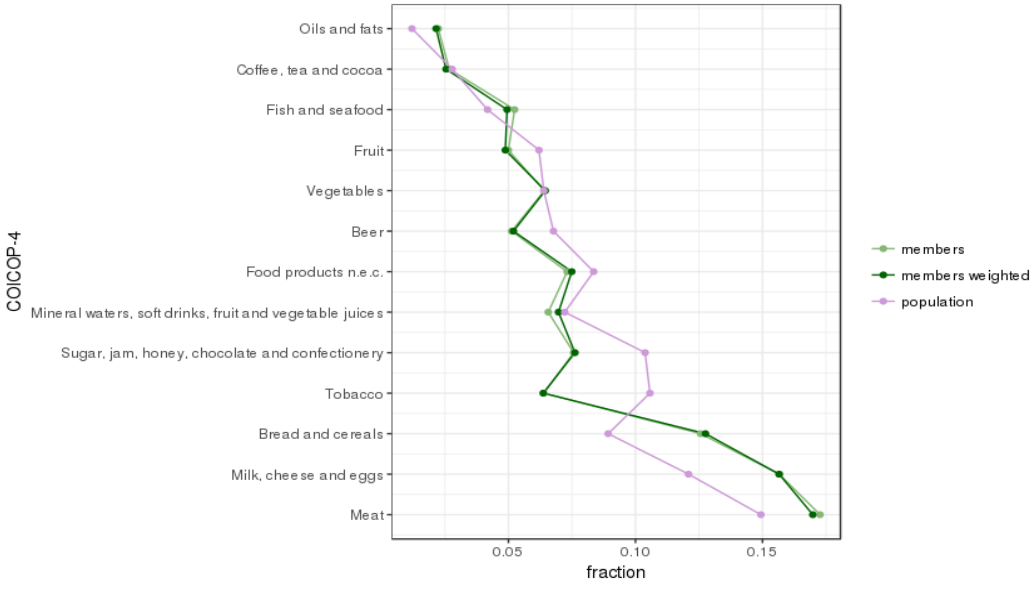
2. Exploration des données de transaction

2.1 Expériences relativement à l'acquisition de données provenant des cartes de fidélité et des caisses enregistreuses

En Norvège, trois acteurs dominent le marché des biens de consommation courante. SSB a déployé des efforts considérables pour instaurer un dialogue avec ces acteurs afin d'avoir accès à leurs données. C'est ce qui fait, entre autres, que l'on a accèes depuis longtemps à des données agrégées des caisses enregistreuses pour établir l'indice des prix à la consommation. Ces dernières années, les détaillants ont mis en place des programmes de fidélité incluant des cartes électroniques et des bases de données sur les consommateurs qui enregistrent l'utilisation des offres de produits par les membres et leurs habitudes d'achat. Idéalement, ces bases de données seraient le reflet des profils de dépenses des membres et des consommateurs. Voilà donc pourquoi les données pourraient être utiles pour produire des statistiques sur les dépenses des ménages.

Statistics Norway a analysé les données de l'une de ces bases. Tous les achats enregistrés par les cartes de fidélité des membres pendant un mois en 2016 ont été étudiés pour évaluer la répartition démographique et les habitudes d'achat des membres par rapport à celles des non-membres (Buelens et coll. 2018). Les auteurs de l'étude ont relevé des problèmes importants liés à la représentativité. En effet, la répartition démographique et les habitudes de dépenses des membres s'écartent nettement d'autres statistiques de référence. Des tentatives en vue de corriger ces écarts à l'aide de données auxiliaires du registre de la population de Norvège sont demeurées infructueuses, et l'étude a conclu que les différences ne pouvaient pas s'expliquer par l'âge, le sexe et le lieu. Par ailleurs, certains produits entrant dans la classification des fonctions de consommation individuelle (COICOP) ne sont pas uniquement vendus dans les magasins des chaînes de détaillants; on les retrouve dans plus de points de vente. Pensons au tabac dont la faible proportion des ventes est illustrée dans la figure 2.1-1.

Figure 2.1-1
Distribution des dépenses par COICOP pour les membres des programmes de fidélité et pour la population selon l'information utilisée dans l'Indice des prix à la consommation



En raison de l'important biais de sélection observé dans les données des membres des programmes de fidélité et des autres investissements considérables nécessaires pour acquérir des données adaptées à la production, SSB a suspendu toute nouvelle recherche de données de ce type. Il a plutôt été décidé de concentrer les efforts sur une analyse plus approfondie des données des caisses enregistreuses.

Nous avons ainsi obtenu toutes les ventes de l'une des principales chaînes de détaillants réalisées pendant un mois civil, c'est-à-dire quelque 220 millions d'enregistrements. Les reçus détaillés des transactions indiquent l'heure, le

nom du magasin, le numéro d'article commercial international (GTIN) de l'article vendu, le prix de l'article ainsi que le montant total de la vente. Ces informations sont utilisées pour l'examen du couplage décrit à la section 2.3.

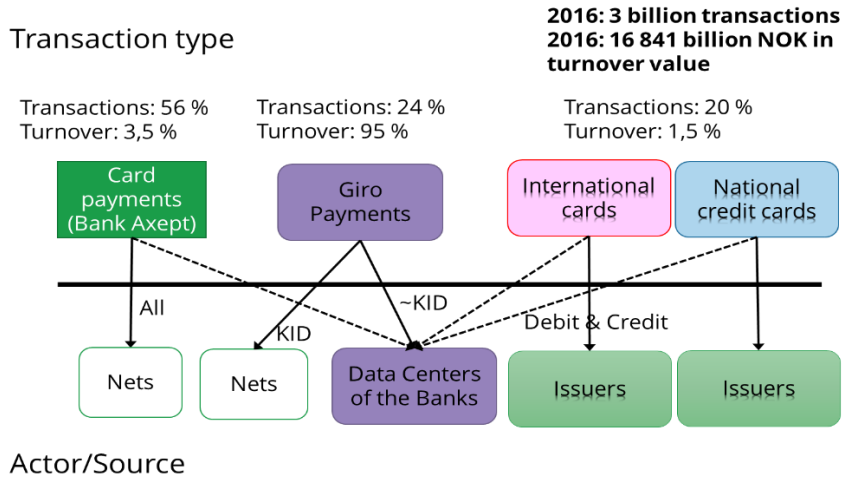
2.2 Transactions de paiement

Il existe différents types de transactions financières électroniques, mais celles qui nous intéressent sont les paiements effectués par les consommateurs. Les transactions entre entreprises, les virements de salaire d'une entreprise à un employé et les virements entre comptes bancaires sans achat en découlant sont exclus de l'étude. Ils revêtent cependant un intérêt plus large pour établir d'autres statistiques.

La figure 2.2-1 donne une vue d'ensemble des catégories de transactions et des acteurs en Norvège. On recense quatre grands types de transactions : (i) les transactions par carte de débit au moyen du système national BankAxept, (ii) les transactions par virement entre comptes bancaires et les transactions par carte de crédit, émises pour (iii) les Norvégiens ou (iv) consistant en des cartes de crédit étrangères. Tous les types de transactions passent par les centres de données des banques de Norvège et y laissent des traces. Les transactions par carte de crédit sont traitées par les émetteurs des cartes. Les paiements par virement peuvent être effectués avec ou sans code DIC. Le code DIC est un code d'identification du client qui simplifie le traitement des factures en repérant automatiquement la personne qui a fait le paiement.

Nets est un fournisseur de services de paiement qui reçoit toutes les transactions par virement avec un code DIC et toutes les transactions qui utilisent le système BankAxept, c.-à-d. les paiements au moyen de cartes de débit norvégiennes. De toutes les transactions financières électroniques, Nets gère près de 98,5 % de la valeur monétaire et 80 % des 3 milliards de transactions annuelles estimées.

Figure 2.2-1
Vue d'ensemble du système de transactions financières norvégien et de la répartition du nombre de transactions et des valeurs de transactions en 2016



SSB a pris contact avec Nets pour lui demander ses données d'essai couvrant une période d'un mois, ce qui a été possible uniquement grâce à la loi sur la statistique. La période de référence correspond à celle des données des cartes de fidélité et à celle des données des caisses enregistreuses obtenues des chaînes de détaillants. Les données de BankAxept de Nets sont celles qui présentent le plus d'intérêt pour les paiements des consommateurs aux entreprises. Les registres des transactions de paiement par carte sont horodatés (jusqu'à indiquer les secondes), indiquent à qui appartient le terminal BankAxept (le nom de l'entreprise) et où il se trouve, le compte à partir duquel est fait le paiement ainsi que le montant de la transaction.

2.3 Examen du couplage des sources de transaction et des données démographiques sur les ménages

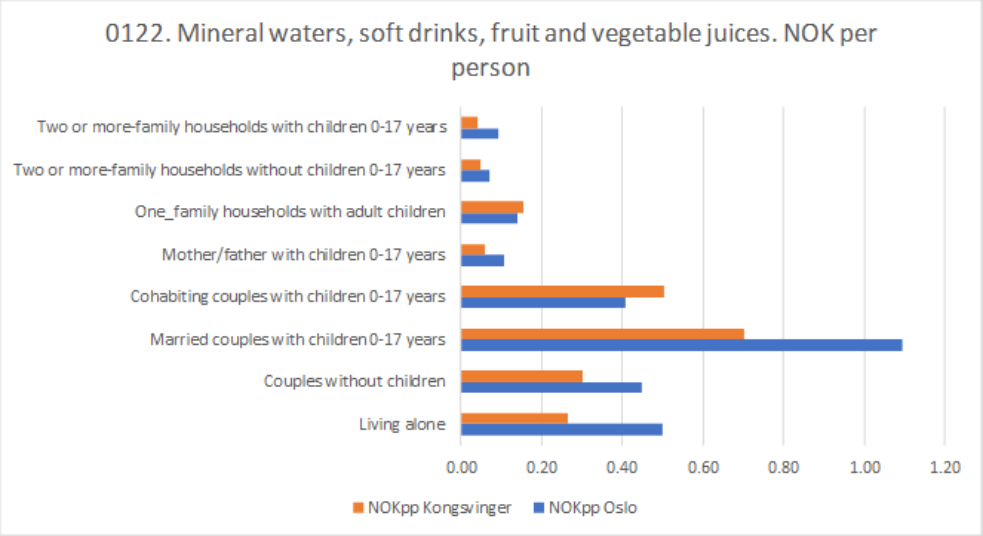
Fyrberg et coll. (2018) décrivent une étude de validation de concept visant à savoir si les données des caisses enregistreuses et des transactions de paiement combinées peuvent être utilisées pour établir des statistiques sur les dépenses des ménages. Pour ce faire, il faut effectuer une succession d'opérations de couplage et d'activités régulières de protection de la confidentialité.

La transaction monétaire est l'unité étudiée parmi les données de paiement et les données des caisses enregistreuses. Les renseignements sur l'émetteur et le récepteur constituent les attributs. Dans les statistiques sur les dépenses des ménages, l'unité étudiée est le ménage. D'où la nécessité d'établir un lien entre les ménages et les données transactionnelles, ce que SSB peut faire grâce à l'utilisation statistique du registre de la population, du registre des entreprises et du registre des comptes bancaires. Le registre des comptes bancaires est un répertoire des titulaires légitimes de comptes. L'utilisation de ces informations pour le couplage et la pseudonymisation des renseignements personnels et sur le ménage permet de lier les données de paiement de BankAxept au titulaire du compte qui effectue les paiements et de les classer dans une catégorie de ménage. Cette méthode n'a fait l'objet d'aucune analyse détaillée des erreurs, ni pour le couplage ni pour la catégorisation des ménages. Par contre, on fait le couplage à l'aide des numéros d'identification personnels, et la méthode utilisée pour classer les ménages est la même que celle utilisée pour le recensement norvégien fondé sur les registres. C'est donc dire qu'il s'agit d'une méthode fiable en ce qui concerne les statistiques sur les ménages. On ne sait toutefois pas dans quelle mesure les gens prêtent leur carte de débit à des membres ne vivant pas dans le ménage ou font des achats pour d'autres ménages, c'est une question qui devrait faire l'objet d'études plus poussées.

Cette méthode lie les catégories de ménages aux transactions de paiement pour ce qui est des dépenses. La prochaine étape consiste à ajouter des renseignements détaillés sur les dépenses. Lors de la validation de concept, les données des caisses enregistreuses des commerces vendant des biens de consommation courante ont été couplées associées aux données transactionnelles d'une seule journée. Pour créer les enregistrements couplés, on a utilisé les informations *heure*, *lieu de vente* et *somme totale de vente*. En procédant ainsi, 66 % de toutes les transactions des caisses enregistreuses ont pu être jumelées à la transaction d'un compte et classée dans une catégorie de ménage. Des études menées précédemment révèlent que près de 70 % de tous les achats en Norvège sont réglés par des cartes de débit utilisant le système BankAxept, on s'attend donc à ce taux de correspondance. Les autres transactions des caisses enregistreuses non appariées (34 %) consistent en des paiements par cartes de crédit, en espèces et par cartes émises par des magasins de la chaîne de détaillants. (On est à étudier les différences possibles entre l'ensemble de transactions appariées et celui des transactions non appariées.)

L'ensemble de données appariées pour la seule journée de transactions choisie et une seule chaîne de détaillants contenait autour de 775 000 achats uniques avec une moyenne d'un peu plus de six articles par achat et une valeur totale de 155 millions de NOK. Tous les articles ont un code EAN qui permet de les associer à la classification COICOP. La figure 2.3-1 illustre le niveau de détail qu'on peut obtenir lorsqu'on traite l'ensemble de données qui en découle.

Figure 2.3-1
Estimation des dépenses quotidiennes d'articles de la COICOP 0122 par personne à Oslo et à Kongsvinger selon le type de ménage



La qualité des statistiques présentées ci-dessus n'a pas été étudiée lors de la validation de concept, car elle sortait du cadre de l'étude. Cette évaluation serait par ailleurs très difficile étant donné que ce niveau de détail n'a jamais été accessible auparavant. Soulignons l'estimation pour de petits domaines avec des statistiques tirées d'une petite commune comme Kongsvinger par catégorie de ménage, la taille d'échantillon et l'ancien design de la FBU sont bien loin de permettre d'obtenir des résultats d'une précision acceptable pour de tels estimés.

3. Résumé et travaux ultérieurs

3.1 Discussion

L'annulation de l'enquête a déclenché une intensification des efforts en vue de trouver d'autres sources de données et de les regrouper avec celles de Statistics Norway pour explorer d'autres façons de mener l'enquête. Jusqu'à présent, les expériences sont prometteuses. Nous avons découvert et examiné des sources de données nouvelles et inconnues jusqu'alors. L'idée d'utiliser les données sur les membres des programmes de fidélité a été mise de côté au profit d'une idée générale consistant à combiner différentes sources de transaction les unes aux autres et avec les données administratives disponibles. Une étude de validation de concept a incité SSB à poursuivre cette idée, surtout que les sources de données découvertes peuvent être utiles pour d'autres statistiques, non pas uniquement pour les dépenses des ménages. La possibilité d'avoir de nouvelles statistiques sur les nutriments, la santé et les prix ainsi que d'étudier des transactions interentreprises pour établir des statistiques structurelles sur les entreprises ne représente que quelques avantages qui découlent de la combinaison de sources de données.

Pour ce qui est des statistiques sur les dépenses des ménages, des travaux de développement sont encore nécessaires. Non pas uniquement pour mieux connaître les nouvelles sources de données, mais également pour convenir d'un design général pour faire des statistiques. Jusqu'ici, l'une des conclusions qu'on peut tirer est que l'ajout de données supplémentaires des autres grands acteurs du commerce au détail et de journées permettra probablement d'obtenir de meilleures statistiques sur les dépenses des ménages relatives aux produits de détail courants que celles de l'ancienne FBU. Les erreurs liées aux cas de non-réponse et aux mesures d'une FBU basée sur un journal seraient remplacées par des erreurs de couplage, d'unité (voir Zhang, 2012) et d'enjeux possibles concernant la représentativité, si les habitudes de paiement changent ou si de nouveaux grands acteurs s'ajoutent. L'exactitude, l'actualité et le potentiel analytique serait nettement amélioré grâce aux données transactionnelles. Toutefois, par rapport à l'objectif précédent de la FBU, la couverture des différents types de dépenses est faible. Par exemple, les acteurs du commerce de détail de biens de consommation courante ne couvrent pas très bien les produits comme l'alcool et les appareils

électroniques. Pour chacun des groupes de biens mal couverts, il pourrait y avoir de grands obstacles à franchir pour accéder aux données transactionnelles pertinentes. Il est peu probable que de tels investissements vaillent la peine seulement pour des statistiques sur les dépenses des ménages. Chaque nouvelle source de données doit être négociée, traitée et combinée. Il faut mettre en place et tenir à jour des solutions techniques et des modes de transmission de données sécurisée, car s'il n'y pas d'uniformité, ce qui arrive souvent, il pourrait être coûteux d'établir des modèles distincts de systèmes de production de données pour plusieurs fournisseurs de données. Même si les données transactionnelles norvégiennes peuvent également récupérer des signaux d'autres dépenses importantes, comme le coût du logement, il semble inconcevable et mal avisé qu'une enquête sur les dépenses des ménages soit uniquement fondée sur des transactions électroniques. Du moins, si les besoins en contenu sont similaires à ceux de l'ancienne FBU.

Il serait plutôt intéressant de se pencher sur un modèle d'enquête consistant en un hybride entre une enquête traditionnelle et des données transactionnelles. Les données transactionnelles peuvent saisir des informations d'une grande précision et extrêmement détaillées concernant certains types de dépenses, p. ex., les articles de consommation courante et d'autres dépenses courantes identifiables et pouvant être attribuées aux ménages. On peut ensuite concevoir une enquête par échantillon complémentaire pour recueillir de l'information sur d'autres types de dépenses nécessaires qui ne sont pas couvertes. Pour réduire les coûts, le niveau de détail requis pour ces groupes de dépenses doit être inférieur à celui couvert par les données transactionnelles. Compte tenu de nos connaissances sur les erreurs de mesure lorsque les répondants tiennent des journaux ou répondent à des questions qui font appel à la mémoire et étant donné que l'un des facteurs qui influent sur les coûts est la collecte de données, il ne semble guère utile d'avoir un journal dans une telle enquête complémentaire, si les biens détaillés courants dépendent des transactions.

La personne qui communique avec le répondant devrait souligner qu'elle souhaite acquérir des informations qui ne peuvent être obtenues des données transactionnelles et vérifier si d'autres hypothèses au sujet de la répartition des dépenses des ménages sont valables. On peut notamment penser à la proportion des dépenses en activités de loisirs, et en voyage.

SSB poursuivra l'étude des données transactionnelles pour faire des statistiques sur les dépenses des ménages. Il est trop tôt pour dire si les prochaines statistiques publiées seront basées uniquement sur les données transactionnelles, sur un hybride entre une enquête traditionnelle et des données transactionnelles ou sur une FBU modernisée. Pour ma part, j'estime que si l'on choisit l'enquête, l'approche fondée sur un journal doit être éliminée ou sérieusement revue. Comme nous savons ce que nous pouvons obtenir des autres sources, le journal n'a pas sa place ou bien il devrait jouer un rôle différent de celui qu'il jouait auparavant, consistant à améliorer ou à assurer une des données détaillées rapportées en temps opportun.

Par ailleurs, si les transactions électroniques deviennent la principale source de données, il est alors nécessaire de «vérifier si la couverture est suffisante et de vérifier la représentativité en termes des ménages qu'on peut liés aux transactions et en termes de types de dépenses. Dans la négative, la complexité et les processus qui influent sur les coûts pour compléter les transactions par d'autres données pourraient très bien être une moins bonne solution que l'enquête traditionnelle. En Europe, on prévoit réglementer les statistiques sur le budget des ménages au sein du Système statistique européen. En fonction de la rédaction du règlement, il aura une incidence sur la façon de faire de SSB. Plus on exige que le niveau des dépenses autres que les biens de consommation courante soit détaillé, moins il est possible que SSB établisse un modèle statistique comprenant des sources combinées.

Bibliographie

Buelens, B., S. Amdam, et H. Holgersen (2018), « The use of loyalty card transactions data in Household Budget Statistics: an exploration », article présenté à la conférence européenne sur la qualité des statistiques officielles, du 26 au 29 juin 2018, Cracovie, Pologne.

Edgar, J., D. Nelson, L. Paszkiewicz, et A. Safir (2013), « The Gemini Project to Redesign the Consumer Expenditure Survey: Redesign Proposal », rapport de projet, 26 juin, Bureau of Labour Statistics.

Fyrberg, J., J. Zhiyang, J. Åmberg, A. Vestfossen, H. Grini, et A. Frøberg (2018), « Proof of concept – linking payment transaction data with purchase transaction data », rapport non publié, Oslo, Norway: Statistics Norway.

Holmøy, A., et M. Lillegård (1988), « Forbruksundersøkelsen 2012: Dokumentasjonsrapport », *Documents 2014/17*, Statistics Norway.

Zhang, L.-C. (2012), « Topics of statistical theory for register-based statistics and data integration », *Statistica Neerlandica*, Vol. 66, Nr. 1, p. 41-63.