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

Modernisation du Programme des dépenses des ménages

Christiane Laperrière, Denis Malo et Johanne Tremblay[1]

Résumé

L'Enquête sur les dépenses des ménages recueille des informations importantes pour la mise à jour du panier de l'Indice des prix à la consommation et pour le Système de comptabilité nationale au Canada. Ces données servent également à une large communauté d'utilisateurs généralement intéressés à l'analyse des dépenses en fonction de caractéristiques socio-économiques des ménages. Le contenu détaillé et les pratiques traditionnelles de collecte de données axées sur une entrevue personnelle et un journal de dépenses imposent un lourd fardeau aux répondants et engendrent des coûts de collecte élevés. Le programme des dépenses des ménages explore donc le potentiel de nouvelles sources de données, ce qui s'inscrit dans les objectifs de modernisation de Statistique Canada. Dans cet article, certaines des sources de données alternatives à l'étude seront présentées, et le potentiel et les limites de celles-ci seront discutés dans le contexte du programme. L'article présentera également les défis rencontrés dans l'exploration de ces nouvelles sources de données, les idées novatrices envisagées pour leur classification et leur intégration et les résultats des évaluations en cours.

Mots-clés : Modernisation; programme des dépenses des ménages; sources alternatives; intégration de données.

1. Introduction

Face aux défis auxquels ils sont confrontés, les programmes d'enquêtes sur les dépenses des ménages cherchent à se moderniser et ce, à l'échelle internationale (Eurostat, 2017). Cette réalité n'échappe pas au contexte canadien dont le programme, même s'il a adopté le modèle international de collecte, veut se positionner pour mieux réagir aux défis actuels et à venir liés à la collecte de ces données. Ainsi, la recherche de sources de données alternatives et leur intégration, tout en protégeant les renseignements personnels, est toute naturelle pour ce programme. Le présent article décrit les études en cours en ce qui a trait à l'exploration et l'exploitation de ces dernières. Une description du programme des dépenses des ménages (PDM) actuel est donnée à la section 2 et les objectifs du projet de modernisation sont détaillés à la section 3. Les sections 4 et 5 décrivent les enjeux concernant les sources de données portant respectivement sur les dépenses liées aux logements et sur les données de transaction. La section 6 conclut par un sommaire des enjeux rencontrés et par une description des travaux futurs.

2. Programme des dépenses des ménages

Le PDM actuel repose principalement sur l'Enquête sur les dépenses des ménages (EDM) qui est une enquête annuelle volontaire dont la taille d'échantillon est d'environ 17 500 ménages dans les 10 provinces (Statistique Canada, 2018). L'EDM combine deux modes de collecte : une entrevue et un journal de dépenses. L'entrevue personnelle assistée par ordinateur est utilisée principalement pour recueillir des dépenses plus importantes et moins fréquentes. Les périodes de référence sont établies selon le type de dépenses (1 mois, 3 mois, dernier paiement, 4 semaines). Le journal, sous format papier, est utilisé pour recueillir des dépenses fréquentes et plus petites sur une période de deux semaines. Ces dépenses seraient plus difficiles à se rappeler dans le cadre d'une entrevue rétrospective. Les données portant sur le revenu des ménages proviennent de sources administratives et sont ajoutées aux données des répondants suite au

[1]Christiane Laperrière, Statistique Canada, 100 promenade du pré Tunney, Ottawa (Ontario), K1A 0T6 (christiane.laperriere@canada.ca); Denis Malo, Statistique Canada, 100 promenade du pré Tunney, Ottawa (Ontario), K1A 0T6 (denis.malo@canada.ca) et Johanne Tremblay, Statistique Canada, 100 promenade du pré Tunney, Ottawa (Ontario), K1A 0T6 (johanne.tremblay@canada.ca)

processus de collecte. Le fardeau de réponse de l'enquête est important puisqu'en plus d'avoir à compléter le journal, la durée moyenne de l'entrevue est d'environ 60 minutes. Les taux de réponse à l'entrevue sont autour de 65 % et ceux du journal entre 40 % et 45 %.

Les données de l'EDM servent à mettre à jour les poids du panier de biens et services de l'Indice des prix à la consommation. Elles sont également utilisées par le Système de comptabilité nationale, en particulier comme intrant pour dériver le produit intérieur brut, de même que par plusieurs ministères fédéraux et provinciaux pour développer des politiques et des programmes sociaux et économiques. Finalement, divers groupes souhaitant mieux comprendre les enjeux reliés aux habitudes de dépenses des Canadiens utilisent les données de l'EDM. Il s'agit donc d'un vaste ensemble d'utilisateurs ayant des besoins variés, un élément important à considérer dans un plan de modernisation. Il existe également une forte demande pour des données sur les dépenses avec une plus grande valeur analytique. Ceci se traduit par des requêtes concernant l'ajout de contenu à l'enquête et l'amélioration de la capacité de mener des analyses pour des populations particulières. Dans le cadre du modèle actuel, le fait de répondre à ces demandes se traduirait respectivement par une augmentation du fardeau de réponse et par une augmentation importante de la taille d'échantillon et donc, des coûts de collecte. Pour ces raisons, et parce que le maintien des taux de réponse dans les dernières années a été obtenu au prix d'un effort accru, le programme explore présentement l'utilisation de données alternatives en remplacement ou en complément à une enquête.

3. Objectifs de la modernisation du PDM

Les grands objectifs du plan de modernisation du programme s'inscrivent dans les objectifs de modernisation de Statistique Canada. Le programme vise une utilisation accrue de données administratives et alternatives, de même qu'une amélioration de la valeur analytique des données pour les utilisateurs actuels et à venir. Lorsque la collecte des données s'avérera nécessaire, le programme vise également à diminuer le fardeau de réponse et à recueillir les données de façon moderne, par exemple, en utilisant des outils électroniques.

Les utilisateurs énumérés à la section précédente ont des besoins très variés, certains n'utilisant les données de l'EDM qu'à un niveau agrégé (avec ou sans information sociodémographique), d'autres devant mener leurs propres analyses à l'aide des microdonnées pour, par exemple, évaluer les facteurs reliés à l'insécurité alimentaire ou les dépenses nécessaires pour subvenir aux besoins des enfants. Certaines des sources alternatives disponibles pourraient ne pas être au niveau requis et donc, ne pas répondre à ces besoins. Il apparaît évident à ce stade que plusieurs modèles ou stratégies seront requis. Il faudra donc identifier des familles de besoins et déterminer quels agencements de données alternatives ou de données d'enquête et alternatives pourraient répondre à ces besoins. Évidemment, les enjeux liés à la cohérence des estimations issues de tels modèles devront être considérés.

Le processus d'acquisition des données alternatives au niveau de l'Agence est un processus qui comporte généralement trois phases :

- Dans un premier temps, le besoin et le potentiel d'utilisation sont identifiés pour une source donnée;
- Dans un deuxième temps, des données préliminaires sont disponibles pour en permettre l'évaluation et pour déterminer si le processus d'acquisition doit se poursuivre;
- Finalement, certaines sources ont été acquises officiellement et sont disponibles à des fins d'évaluations et pour leur utilisation en production.

Le PDM est présentement en mode exploratoire pour déterminer quelles sources pourraient lui être bénéfiques et il est dépendant du processus d'acquisition de l'Agence. Tout en appliquant des contrôles stricts quant à la confidentialité et la protection des données, le potentiel et les limites des sources disponibles sont évalués et de la rétroaction est fournie selon la phase d'acquisition de la source. Dans la suite de l'article, les résultats de ces évaluations seront présentés pour deux catégories de sources : des dépenses liées aux logements et les données de transaction.

4. Données alternatives liées aux logements

Les dépenses liées aux logements sont importantes puisqu'elles représentent environ 30 % des dépenses de consommation courante des ménages. On retrouve dans cette catégorie des dépenses telles que les paiements

hypothécaires, les paiements de loyers pour les locataires, les services publics et les taxes municipales. Ensemble, ces quatre composantes représentent 75 % des dépenses liées aux logements. Même si ces dépenses sont relativement régulières, il peut être nécessaire pour les répondants à l'entrevue de consulter des factures ou des relevés pour rapporter les montants exacts. L'attrait pour les sources alternatives de cette catégorie réside dans le fait qu'elles sont liées entre elles par le concept d'adresses.

Une première source à l'étude concerne les dettes des consommateurs. Cette source contient des données trimestrielles sur les paiements et soldes pour une grande variété de prêts, tels que les hypothèques, les lignes de crédit et les prêts automobiles et étudiants. Les données sont au niveau de chaque individu et agrégées par type de prêt, ce qui veut dire, par exemple, que pour une personne ayant plusieurs prêts d'un type donné, seuls les totaux pour les paiements et les soldes sont disponibles. Cette source est très intéressante pour le programme puisqu'elle permettrait de recueillir des données sur les paiements et les soldes hypothécaires. Bien que les soldes hypothécaires ne représentent pas des dépenses en tant que telles, certains utilisateurs s'y intéressent. Les évaluations préliminaires de cette source ont permis d'identifier certains défis. D'abord, les prêts conjoints (ce qui est souvent le cas pour les hypothèques) apparaissent avec la même information pour chacune des personnes impliquées. L'identification et la résolution de ces doublons devraient faire partie d'une stratégie de traitement des données. Pour certains utilisateurs, il est important de pouvoir distinguer l'information associée aux hypothèques contractées sur les résidences principales de celles des résidences secondaires. Comme les données sont agrégées au niveau des individus, cette différenciation représentera un défi important sans l'aide de sources auxiliaires. Finalement, certains paiements hypothécaires comprennent les taxes municipales ou des paiements reliés aux assurances habitation. Il faudrait développer une stratégie permettant d'identifier et de retirer ces montants des paiements réguliers.

Une seconde source concernant les prêts hypothécaires a été identifiée et pourrait s'avérer utile dans l'évaluation et le traitement de la source principale mentionnée au paragraphe précédent. Une agence du gouvernement fédéral fournit une assurance aux emprunteurs hypothécaires dont la mise de fonds est inférieure à un certain seuil. Les données de cette source concernent donc les hypothèques couvertes par cette assurance, ce qui représente environ le tiers des hypothèques au Canada. Ce produit n'a donc pas une couverture complète des ménages avec une hypothèque, mais il a une très bonne couverture des hypothèques assurées et donc, des institutions financières qui émettent des hypothèques. Il serait donc possible d'utiliser ce produit pour évaluer la couverture de notre source principale en termes d'institutions financières.

Toujours dans la famille des sources de données alternatives liées aux logements, une autre source à l'étude concerne les fournisseurs d'électricité. À l'EDM, on demande aux répondants de rapporter leur dernier paiement en ce qui a trait aux frais en électricité pour leur logement. L'Agence évalue les données de certains fournisseurs et des résultats préliminaires montrent que la qualité de l'information nécessaire pour l'intégration aux programmes est généralement bonne. Par exemple, la qualité des adresses est bonne en milieu urbain alors qu'elle peut poser certains défis en milieu rural en raison de l'utilisation d'adresses postales ou d'adresses non civiques telles que des coordonnées. Il est important de préciser que le marché de la production et de la distribution de l'électricité est relativement segmenté au Canada, car il existe quelques dizaines de distributeurs. Ceci est un facteur important pour l'acquisition et l'uniformisation des données. Pour ce qui est des frais en électricité, il existe peu de valeurs manquantes ou aberrantes dans les données évaluées. Les grandes valeurs observées sont souvent expliquées par la nature d'un logement qui serait de toute façon exclu de l'univers de l'enquête, tel qu'un logement collectif ou une entreprise.

Les évaluations préliminaires ont également permis d'identifier certains défis potentiels dans l'utilisation de ces données. Premièrement, il est essentiel de bien comprendre les différences conceptuelles dans les données des différents fournisseurs. Par exemple, l'adresse de service, l'adresse de facturation et celle du compteur intelligent peuvent être disponibles, ou non, selon la source. Les variables transmises de même que leur définition peuvent donc être différentes et un traitement spécifique peut être nécessaire. Ceci est particulièrement important pour une géolocalisation précise du logement auquel se rapportent les frais en électricité. Le deuxième défi noté concerne le niveau de détail insuffisant des métadonnées associées à certaines sources, qui ne permet pas de répondre à tous les besoins du PDM. En effet, les coûts en électricité peuvent inclure, ou non, des frais uniques (branchement, arrérages, frais d'intérêts) et il a été observé que les métadonnées ne contiennent pas nécessairement ce genre de précision. Ceci représente un bel exemple de défi posé par l'utilisation de données recueillies ou générées par des sources externes et pour lesquelles l'agence statistique ne contrôle pas le concept. Finalement, le manque d'uniformisation des périodes de référence auxquelles se rapportent les paiements se traduirait par la nécessité de développer des stratégies de standardisation propres à chaque source. Ces ajustements permettraient de refléter les dépenses d'une année de

référence spécifique, ce qui est souvent nécessaire pour les programmes sur les dépenses. Somme toute, malgré les défis identifiés ici, le potentiel de remplacement de ce type de sources est intéressant pour les frais en électricité.

Les données sur les dettes des consommateurs et celles concernant les frais en électricité ne sont que deux exemples d'un ensemble plus vaste de données alternatives liées aux logements qui ont fait l'objet d'une évaluation. Ces deux sources ont été retenues pour des analyses plus poussées en raison du potentiel qu'elles avaient aux fins du remplacement de données d'enquête et de la diminution du fardeau de réponse qu'elles entraîneraient, surtout pour les données concernant les dettes. Malgré l'apparente facilité d'intégration a priori, les évaluations ont rapidement montré que des enjeux liés à la couverture, aux différences conceptuelles, au niveau de détail des métadonnées, à la multitude des sources et au prétraitement rendraient l'intégration moins directe que prévue. Ces obstacles sont clairement surmontables mais nécessiteront le développement de plusieurs méthodes d'ajustement.

5. Données de transaction

Dans cette section, deux autres types de données alternatives d'intérêt pour le PDM seront présentés: les données financières et les données de lecteurs optiques (désignées par données scanner dans la suite du document).

Les données financières sont définies comme étant des données sur les dépenses payées par divers modes de paiement tels que les cartes de crédit, cartes de débit, transferts électroniques, et chèques préautorisés. Ce type de sources de données présente un potentiel intéressant pour le PDM puisqu'il concerne directement les dépenses faites par des individus. La couverture de ces données peut varier d'une source à l'autre. Par exemple, les données financières provenant des institutions bancaires offrent une couverture très intéressante des modes de paiement (cartes de crédit, cartes de débit, transferts électroniques, lignes de crédit et chèques préautorisés). Les données financières provenant des compagnies de cartes de crédit, pour leur part, ne couvrent qu'un seul mode de paiement. Il faut noter que ces deux sources de données financières (c'est-à-dire provenant des institutions bancaires et des compagnies de cartes de crédit) ne couvriront pas les dépenses payées en argent comptant; des ajustements seront alors nécessaires pour corriger cette sous-couverture.

Les données scanner portent sur les transactions des ventes enregistrées aux caisses et faites en magasins. Dans la section 5.1, le contexte et les résultats d'une étude de faisabilité concernant les données financières seront présentés. Ensuite, dans la section 5.2, le potentiel et les limites des données scanner dans le contexte du PDM seront discutés. Finalement, la section 5.3 présentera une façon de combiner ces deux sources, les données financières et les données scanner, afin de tirer profit des avantages que les deux sources ont à offrir.

5.1 Données financières : Étude de faisabilité

Afin de confirmer le potentiel des données provenant des institutions bancaires, et d'en identifier les limites, une étude de faisabilité a été menée. Dans le but de recueillir des données à des fins d'exploration et avant de procéder à un projet pilote en collaboration avec les institutions financières, certains employés de Statistique Canada ont été invités à fournir sur une base volontaire et confidentielle leurs relevés bancaires et de cartes de crédit pour l'année de référence 2017. Au total, 52 employés ont accepté de participer et 42 d'entre eux ont fourni l'information complète pour la période de référence visée. Puisque les données financières ainsi obtenues provenaient de volontaires, l'objectif n'était pas de faire de l'inférence à la population, mais plutôt d'explorer le potentiel de ces données dans le contexte du PDM. Les données ont été recueillies de façon confidentielle et ne contenaient aucun identifiant personnel. Au total, le fichier combinant toutes les transactions recueillies contenait 31 000 transactions pour l'année de référence 2017. Le tableau 5.1-1 ci-dessous contient des exemples fictifs de transactions. L'information disponible pour ces données inclut la date de la transaction, des variables descriptives reliées à la transaction, le coût de la transaction et le type de compte bancaire utilisé. Le niveau de détail des variables descriptives varie d'une transaction à l'autre. Par exemple, tel qu'illustré dans le tableau 5.1-1, certaines transactions sont facilement identifiables (p. ex. « Hydro North » peut facilement être classifié comme une dépense en électricité). D'autres transactions sont difficiles, et parfois impossibles, à classifier (p. ex. « No de chèque » et « Email trfs » n'offrent aucun détail sur le type de dépense).

Tableau 5.1-1
Transactions fictives de l'étude de faisabilité sur les données financières

Date	Description 1	Description 2	Débit	Crédit	Type de compte
23/08/2017	BT FOOD #1254	PURCHASE 55695	58,87		Chèque
12/12/2017	NO DE CHEQUE	54	680,00		Chèque
30/10/2017	RED RIBBON PUB	RR PUB OTT ON	94,02		Carte crédit
06/09/2017	HYDRO NORTH		128,56		Chèque
26/10/2017	CANADA PAY/PAY			1 925,33	Chèque
01/02/2017	EMAIL TRFS	INTERAC E-TRF-58964	200,00		Épargnes
26/10/2017	MORTGAGE BANK	WEST RED BANK	789,63		Ligne de crédit

La classification des données financières en catégories de dépenses, selon un système prédéfini de classification (tel que celui utilisé pour l'EDM), est nécessaire dans le contexte du PDM. Cette étape peut représenter un grand défi, selon les types de dépenses, d'autant plus que les variables descriptives ne contiennent bien souvent que le nom du magasin et non pas les produits achetés. Pour les dépenses de type « vente au détail » achetés en magasin ou en ligne, il est parfois possible d'obtenir un lien direct entre la description et une catégorie de dépense. Par exemple, les restaurants, les taxis et les animaleries sont faciles à identifier étant donné que le nom du restaurant, de la compagnie de taxi ou du magasin, respectivement, apparaît dans la description; de plus, ces transactions sont associées directement à une seule catégorie de dépense. Les répondants à l'EDM sous-déclarent souvent les dépenses en restaurant et en taxi soit par simple oubli ou parce qu'ils ne gardent pas leurs reçus. Les données financières offrent donc un potentiel intéressant pour réduire l'effet de la sous-déclaration de certaines petites dépenses fréquentes sur les estimations. D'autres dépenses de type « vente au détail » sont plus ardues à classifier. Par exemple, des achats faits en épiceries ou dans des magasins à rayons sont identifiables par le nom du magasin qui apparaît dans la description de la transaction, mais il est impossible de connaître la liste de produits achetés. Pour ces cas, d'autres sources, telles que les données scanner, pourraient être utilisées pour obtenir le détail voulu. Ce point sera abordé plus en détail dans les sections 5.2 et 5.3.

En ce qui concerne les transactions liées au logement, certaines peuvent parfois être facilement associées à une catégorie unique de dépense; c'est le cas notamment des paiements d'hypothèque qui peuvent être identifiés à l'aide de mots-clés, tels que « hypothèque », apparaissant dans la description, ou des services publics ou en communication qui peuvent être identifiés par le nom du fournisseur. Cependant, le niveau de détail est moindre que celui fourni par l'EDM. En effet, les paiements d'hypothèque peuvent inclure des frais d'assurances ou des taxes municipales. En ce qui concerne les dépenses en communication, il est impossible de distinguer les paiements pour l'Internet, le téléphone et la télévision. Finalement, certaines dépenses sont généralement payées par chèque ou transfert électronique et la description ne fournit alors pas suffisamment de détails pour permettre une classification; c'est le cas, par exemple, pour le loyer et les frais de garde.

Les données de l'étude de faisabilité ont permis de constater l'ampleur du défi que représente la classification en catégories de produits. Les variables descriptives de la transaction sont essentielles pour permettre une classification. Il est recommandé, pour ce type d'exercice de classification, de considérer l'utilisation d'algorithmes d'apprentissage automatique. En effet, des algorithmes de type supervisé permettraient une classification automatique de variables textuelles (la description de la transaction, dans ce cas-ci) en catégories prédéterminées de produits. Ces méthodes ne pourraient toutefois pas associer une catégorie de dépense dans les cas où la description serait trop vague. Dans certains cas, la description n'offre aucun détail (p. ex. « No de chèque »), mais la récurrence du paiement peut fournir de l'information importante. Par exemple, un chèque récurrent tous les mois, d'un montant assez élevé, pourrait possiblement être associé à un paiement de loyer. Des hypothèses seraient requises pour permettre de telles conclusions, mais celles-ci pourraient être validées par une étude plus globale de l'ensemble des transactions d'un ménage.

Les conclusions de l'étude de faisabilité démontrent le potentiel énorme de ce type de données pour le PDM. La majorité des modes de paiement sont inclus (cartes de crédit, cartes de débit, chèques et transferts électroniques) et

par conséquent, une grande portion des dépenses est couverte. Toutefois, les transactions payées en argent comptant ne sont pas couvertes. Il faudrait ajuster les données pour cette sous-couverture, tout en tenant compte que les transactions payées en argent comptant ne sont pas uniformes parmi les catégories de dépense. En effet, l'argent comptant est davantage utilisé pour des achats de plus faible valeur monétaire et pour certains types de dépense plus que d'autres (p.ex. : repas au restaurant, divertissement et stationnement) (Henry et coll., 2015). Un ajustement global ne serait donc pas suffisant; il faudrait considérer un ajustement spécifique à chaque catégorie de dépense. En ce qui concerne la classification des données financières, d'autres sources de données pourraient être nécessaires afin d'obtenir le détail des produits achetés. Les données scanner en sont un bon exemple.

5.2 Données scanner

Les données scanner portent sur les transactions des ventes enregistrées aux caisses et faites-en magasins. Ces données proviennent des détaillants et les fichiers présentement acquis par l'Agence sont à un niveau agrégé des produits. Les ventes totales et le nombre total d'unités vendues, pour un produit donné (dans un magasin et une semaine donnée) sont disponibles; l'information sur chaque transaction individuelle n'est pas fournie. Le tableau 5.2-1 donne un exemple fictif illustrant le format actuel des données scanner.

Tableau 5.2-1
Format actuel des données scanner

UPC	Description du produit	Emplacement dans le magasin	Semaine	Nom du magasin	Ventes ($)	Quantité vendue
2174060000	BOEUF HACHÉ EXTRA MAIGRE	Viande	4	Magasin #2	1 154,95	100
1122334455	PILULES POUR ALLERGIES	Méd. non prescrits	23	Magasin #6	83,88	12
1112223334	VERNIS À ONGLES, ROSE	Cosmétique	7	Magasin #3	1,99	1
6568400537	YOGOURT GREC 0%, VANILLE	Centre du magasin	15	Magasin #1	13,98	4
1020304050	BARRE SAVON, PEAU SENSIBLE	Santé et produits beauté	15	Magasin #3	19,90	10

Les variables du fichier fournissent l'information sur le code de produit universel (UPC) et une description très détaillée permettant une classification en catégories de produits. L'avantage intéressant de ce type de données, en comparaison aux données financières présentées à la section précédente, est justement le potentiel élevé de classifier les données en catégories de produits prédéterminées. Un désavantage, cependant, est que l'information est agrégée au niveau des produits et ne fournit aucune information sociodémographique. De plus, les données scanner ne couvrent pas seulement les dépenses faites par des ménages canadiens; l'impact de l'inclusion de dépenses d'entreprises et de voyageurs internationaux pourrait être non négligeable pour certains types de produits.

Les données scanner des détaillants présentement disponibles au niveau de l'Agence représentent environ 50 % de la part de marché des ventes de produits alimentaires au Canada. Le montant total dépensé en nourriture ne peut être dérivé puisque la couverture des détaillants est incomplète, mais la distribution en catégories de dépense pourrait être calculée, c'est-à-dire le total pourrait être alloué en grandes catégories telles que la viande, les fruits et légumes, les produits laitiers, etc. Une telle distribution serait agrégée, et non disponible par domaine sociodémographique, mais cette information pourrait toutefois être utile pour certains utilisateurs. Afin de confirmer le potentiel d'une telle distribution, il faudra en évaluer la qualité. La qualité de la distribution sera dépendante de la performance de l'algorithme de classification utilisé et de la couverture des données scanner présentement disponibles.

Pour l'instant, les données scanner d'une chaîne d'épicerie majeure au Canada ont été classifiées en catégories de produits alimentaires selon un algorithme d'apprentissage automatique. La distribution des ventes alimentaires totales

en catégories principales provenant des données scanner de cette chaîne d'épicerie a été comparée à celle provenant de l'EDM. Les résultats sont présentés au tableau 5.2-2 pour l'année de référence 2015.

Tableau 5.2-2
Distribution des données scanner d'une chaîne d'épicerie et de l'Enquête sur les dépenses des ménages (EDM) pour 2015

Catégorie	Données scanner (une chaîne d'épicerie)	EDM
VIANDE	18,1 %	19,5 %
POISSONS ET FRUITS DE MER	2,8 %	3,6 %
PRODUITS LAITIERS ET OEUFS	15,2 %	14,8 %
BOULANGERIE ET CÉRÉALES	14,5 %	14,7 %
FRUITS ET NOIX	12,7 %	12,3 %
LÉGUMES	11,8 %	11,1 %
AUTRES PRODUITS ALIMENTAIRES	24,2 %	23,9 %

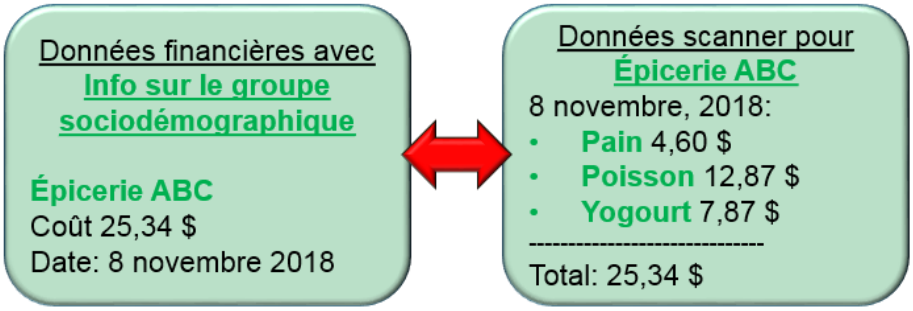
Il est intéressant de constater que les deux distributions sont très similaires, même si une seule chaîne d'épicerie n'est considérée (comparativement à l'EDM qui recueille les dépenses faites par les ménages dans tous les magasins). Des différences entre les deux distributions sont attendues, surtout à des niveaux plus détaillés de produits, puisque les données proviennent de deux sources distinctes ayant toutes deux des sources d'erreurs. Par exemple, les données scanner peuvent souffrir d'erreurs de classification et de couverture, et les données d'enquête peuvent souffrir d'erreurs d'échantillonnage et d'erreurs non dues à l'échantillonnage (p.ex. : erreur de rappel et sous-déclaration). Cependant, il est encourageant de constater que les deux distributions sont similaires, du moins pour les grandes catégories de produits alimentaires, et il est attendu que les distributions se rapprochent l'une de l'autre lorsque les données scanner d'autres détaillants deviendront disponibles.

5.3 Combiner les données financières aux données scanner

Les deux sections précédentes ont présenté deux sources de données intéressantes pour le PDM : les données financières et les données scanner. Ces sources de données alternatives ont des avantages et des inconvénients. Les données financières couvrent plusieurs modes de paiement et ont le potentiel de fournir des informations sur le domaine sociodémographique, mais ne fournissent pas suffisamment de détails par rapport aux produits achetés. Pour leur part, les données scanner fournissent de l'information très détaillée sur les produits, permettant une classification en catégories prédéterminées, mais n'offrent pas de détails sur les domaines sociodémographiques. Cette section portera sur l'idée de combiner ces deux sources de données, à l'aide d'un couplage d'enregistrements, afin de tirer profit des avantages respectifs des deux sources. Pour qu'un tel couplage soit possible, il est important de noter que les données scanner devraient alors être disponibles au niveau des transactions individuelles et non pas au niveau agrégé des produits. Par conséquent, l'acquisition des données scanner à ce niveau est d'intérêt pour le PDM.

À titre illustratif, l'exemple d'un achat fait dans une épicerie ABC peut être utilisé. La description de la transaction dans les données financières permettrait d'identifier l'épicerie ABC, le coût total et la date de la transaction, ainsi que le domaine sociodémographique. La description de la transaction dans les données scanner de l'épicerie ABC fournirait le détail des produits achetés, le coût individuel de chaque produit (totalisant le même coût fourni dans les données financières) et la date de l'achat. Un couplage d'enregistrements basé sur le nom du magasin, coût et date de l'achat permettrait ainsi d'obtenir la liste de produits achetés combinée à un domaine sociodémographique. L'exemple de la combinaison des deux sources est illustré à la figure 5.3-1 ci-dessous.

Figure 5.3-1
Combiner les données financières aux données scanner



Les données scanner, du fait qu'elles soient présentement disponibles à un niveau agrégé, peuvent servir à produire des distributions agrégées par catégorie de produits, tel que discuté à la section 5.2. Ces données ont également un potentiel additionnel lorsqu'elles sont disponibles au niveau des transactions. En effet, les données scanner pourraient alors être intégrées aux données financières afin de tirer profit des avantages des deux sources, et ainsi répondre aux besoins de certains utilisateurs. Il serait alors primordial que les deux sources de données contiennent des variables de couplage communes et de bonne qualité.

6. Conclusion

Dans cet article, diverses sources de données alternatives d'intérêt pour le PDM ont été présentées. Les données sur les dettes des consommateurs et celles concernant les frais en électricité offrent un potentiel intéressant pour le remplacement de données d'enquête, mais vont nécessiter un investissement important afin de mieux comprendre et mitiger les enjeux liés à la couverture, aux différences conceptuelles, au niveau de détail des métadonnées, à la multitude des sources et au prétraitement. Une étude de faisabilité sur les données financières a démontré l'énorme potentiel de ce type de données pour le PDM, surtout quant à la grande couverture des modes de paiement. Cependant, les dépenses en argent comptant ne sont pas couvertes, et les enjeux notés au niveau de la classification des transactions démontrent que des sources de données complémentaires, telles que les données scanner, seraient nécessaires pour obtenir le niveau de détail requis. Les données scanner actuellement disponibles peuvent fournir une distribution agrégée des ventes alimentaires totales en catégories principales de produits. Si ces données devenaient disponibles au niveau des transactions, un potentiel intéressant d'intégration de données pourrait être considéré en couplant aux données financières. Les données explorées jusqu'à maintenant ont un potentiel intéressant pour le PDM, mais il reste beaucoup de travail d'évaluation à effectuer avant de déterminer si elles pourraient remplacer ou être intégrées à des données d'enquête.

Bibliographie

Eurostat (2017), « Household Budget and Time Use Surveys – Launch of the Work of the Task Forces », Meeting of the European directors of social statistics, Luxembourg, mars 2017.

Henry, C., Huynh, K. P. et Shen, Q. R. (2015), « 2013 Methods-of-Payment Survey Results », Banque du Canada Document d'analyse No. 2015-4.

Statistique Canada (2018), *Guide de l'utilisateur, Enquête sur les dépenses des ménages, 2017*.