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

Système de traitement de données transactionnelles

Agnes Waye, Serge Godbout et Nathalie Hamel[1]

Résumé

Les données transactionnelles sont de plus en plus utilisées, à la fois comme données administratives et dans les enquêtes. La richesse et le volume des données permettent d'obtenir des renseignements précieux et d'analyser les tendances de manière plus approfondie. Toutefois, de tels ensembles de données de grande taille et de structures complexes posent des défis uniques en matière de traitement et d'estimation de données, et les méthodes classiques de traitement de données nécessitent des solutions adaptées. À Statistique Canada, l'infrastructure statistique de traitement des données transactionnelles présente des lacunes. En raison du degré élevé de souplesse nécessaire, nous avons cerné le besoin d'élaborer un système plus robuste pour le traitement des données transactionnelles. On a élaboré un système de traitement des données transactionnelles pour les enquêtes sur les transports, qui comprennent de nombreuses enquêtes comportant des données transactionnelles. Jusqu'à maintenant, une enquête a été intégrée à ce système (Enquête sur la base tarifaire) et, progressivement, d'autres enquêtes provenant des programmes de statistiques sur l'aviation, le transport ferroviaire et le camionnage y seront également intégrées. Ce système met en œuvre les étapes de la phase du processus, telles que définies dans le Modèle générique du processus de production statistique (GSBPM), y compris des caractéristiques comme l'importation de données, la vérification et l'imputation, l'intégration de données, l'équilibrage et l'estimation. Le présent article présentera la définition et les caractéristiques particulières des données transactionnelles, la façon dont elles sont traitées, les leçons tirées, les défis auxquels nous nous sommes heurtés ainsi que les problèmes qu'il faudra résoudre à l'avenir dans le système de données transactionnelles.

Mots-clés : Traitement des données; traitement; vérification et imputation; estimation; données transactionnelles; intégration des données.

1. Introduction

1.1 Contexte

Les données transactionnelles sont de plus en plus utilisées, à la fois comme données administratives et dans les enquêtes. La richesse et le volume des données permettent d'obtenir des renseignements précieux et d'analyser les tendances de manière plus approfondie. Toutefois, de tels ensembles de données de grande taille et de structures complexes posent des défis uniques en matière de traitement et d'estimation de données, et les méthodes classiques de traitement de données nécessitent des solutions adaptées. À Statistique Canada, l'infrastructure statistique de traitement des données transactionnelles présente des lacunes. Plus particulièrement, le Programme de la statistique des transports de Statistique Canada utilise des données transactionnelles pour les statistiques officielles et les produits analytiques. Citons trois exemples d'enquêtes sur le transport utilisant des données transactionnelles : l'Enquête sur la base tarifaire, les Statistiques relatives aux mouvements d'aéronefs et l'Enquête sur l'origine et la destination des marchandises transportées par camion. L'Enquête sur la base tarifaire constitue une source complète et régulière de données sur les passagers, les recettes et les tarifs moyens aériens d'après les différents genres de tarifs (Statistique Canada, 2018a). L'enquête sur les Statistiques relatives aux mouvements d'aéronefs fournit des estimations sur le mouvement des aéronefs au Canada. Transports Canada et NAV CANADA utilisent cette information pour évaluer la charge de travail des contrôleurs de la circulation aérienne, l'activité des aéronefs et l'utilisation des pistes d'atterrissage (Statistique Canada, 2018b). L'Enquête sur l'origine et la destination des marchandises transportées par

[1]Agnes Waye, Statistique Canada, 100 promenade Pré Tunney, Ottawa, Canada, K1A 0T6 (agnes.waye@canada.ca); Serge Godbout, Statistique Canada, 100 promenade Pré Tunney, Ottawa, Canada, K1A 0T6 (serge.godbout@canada.ca); Nathalie Hamel, Statistique Canada, 100 promenade Pré Tunney, Ottawa, Canada, K1A 0T6 (nathalie.hamel@canada.ca)

camion sert à mesurer les mouvements de marchandises et la production de l'industrie du camionnage au Canada (Statistique Canada, 2017). En raison du manque d'outils de traitement des données transactionnelles des enquêtes sur les transports, un nouveau système de traitement des données transactionnelles a été élaboré. Jusqu'à maintenant, une enquête a été intégrée à ce système (Enquête sur la base tarifaire) et, progressivement, d'autres enquêtes provenant des programmes de statistiques sur l'aviation, le transport ferroviaire et le camionnage y seront également intégrées. Ce système met en œuvre les étapes de la phase du processus, telles que définies dans le Modèle statistique général du processus opérationnel (MSGPO), y compris des caractéristiques comme l'importation, la vérification et l'imputation, l'intégration, l'équilibrage et l'estimation de données. Le présent article présentera la définition et les caractéristiques particulières des données transactionnelles, la façon dont elles sont traitées, les leçons tirées, les défis auxquels nous nous sommes heurtés ainsi que les problèmes futurs qu'il faudra résoudre dans le système de données transactionnelles.

2. Données transactionnelles

2.1 Caractéristiques des données transactionnelles

On peut considérer que les données transactionnelles appartiennent à la famille des mégadonnées. Elles ont en commun avec ces dernières les 4 v : volume, vélocité, véracité et variété. En général, les données transactionnelles proviennent de très grands ensembles de données en raison de la fréquence élevée à laquelle les données sont rapportées et du grand nombre de transactions. La vélocité des données en est également une caractéristique particulière, car les données sont souvent transmises très rapidement. Elles diffèrent en cela des données traditionnelles (comme les données d'enquête) pour lesquelles un long délai peut séparer la création des données de leur collecte. La véracité des données renvoie à la façon dont la qualité des données peut varier, selon le fournisseur ou dans la durée. La variété décrit la diversité des formats que peuvent avoir les données transactionnelles entrantes.

Elles sont habituellement collectées à une fréquence prédéfinie (par exemple, quotidienne ou hebdomadaire). Comme exemples de données transactionnelles, citons les données financières, par exemple les factures, ou les données logistiques, par exemple les enregistrements de voyage. Les variables peuvent être liées au temps (p. ex., la date), à la classification (p. ex., les types de produits) ou à de l'information numérique (p. ex., les recettes). Les variables d'intérêt pour les données transactionnelles sont souvent agrégées afin que soient produites les informations statistiques ciblées. On peut par exemple calculer le total des ventes d'un mois en additionnant les ventes de toutes les transactions du mois.

2.2 Comparaisons entre données transactionnelles et données traditionnelles

Observons ici ce à quoi peuvent ressembler des données transactionnelles, à partir de l'exemple de l'Enquête sur la base tarifaire. Le tableau 2.2-1 montre le nombre total de passagers de chaque transporteur aérien. Dans l'Enquête sur la base tarifaire, chaque transporteur aérien nous fournit tous les trimestres un fichier contenant les coupons de vol, semblable au tableau 2.2-2.

Tableau 2.2-1
Enregistrements des transporteurs

Transporteur	Total des passagers
A	100
B	200

Tableau 2.2-2
Enregistrements des transactions

Transporteur	Paire de villes	Passagers
A	AB	40
A	AC	60
B	AD	100
B	AE	100

Les données transactionnelles peuvent être plus souples que les données traditionnelles, c'est-à-dire plus souples que les données recueillies à partir d'enquêtes traditionnelles. Le tableau 2.2-3 montre à quoi ressembleraient les données traditionnelles si nous voulions ajouter des passagers pour chaque secteur (international, intérieur et transfrontalier). Comme le montre le tableau, nous avons ajouté trois variables distinctes pour le nombre de passagers. Au tableau 2.2-4, nous constatons que les données transactionnelles présentent la même information que les données traditionnelles sans qu'il soit nécessaire d'ajouter trois nouvelles variables. Seule une nouvelle variable, « secteur », a été ajoutée aux enregistrements transactionnels. Cet exemple illustre le fait que les données transactionnelles nécessitent moins de nouvelles variables pour montrer l'information au niveau du domaine comparativement aux données traditionnelles.

Tableau 2.2-3
Données traditionnelles

Transporteur	Total des passagers	Passagers – International	Passagers – Intérieur	Passagers – Transfrontalier
A	400	100	200	100
B	400	300	40	60

Tableau 2.2-4
Enregistrements des transactions

Transporteur	Paire de villes	Passagers	Secteur
A	AB	100	International
A	AC	200	Intérieur
A	AZ	100	Transborder
B	AD	300	International
B	AE	40	Intérieur
B	AF	60	Transborder

Observons maintenant un scénario où nous souhaitons ajouter un nouveau secteur, « étranger ». En cas de données traditionnelles (tableau 2.2-5), nous devons ajouter une nouvelle variable qui montre le nombre de passagers étrangers. En revanche, en cas de données transactionnelles (tableau 2.2-6), il est inutile d'ajouter de nouvelles variables. Il suffit d'ajouter directement des enregistrements aux données transactionnelles avec une valeur « étranger » pour la variable de secteur.

Tableau 2.2-5
Données traditionnelles

Transporteur	Total des passagers	Passagers – International	Passagers – Intérieur	Passagers – Transfrontalier	**Passagers – Étranger**
A	400	100	200	100	**S.O.**
B	450	300	40	60	**50**

Tableau 2.2-6
Enregistrements des transactions

Transporteur	Paire de villes	Passagers	Secteur
A	AB	100	International
A	AC	200	Intérieur
A	AZ	100	Transfrontalier
B	AD	300	International
B	AE	40	Intérieur
B	AF	60	Transfrontalier
B	DF	50	**Étranger**

Dans le tableau 2.2-7, considérons un cas où il y a aussi une variable de recettes pour les transactions, ainsi qu'une nouvelle variable de domaine nommée Findesemaine (avec une valeur de 1 si le jour déclaré tombe en fin de semaine et de 0 autrement). Si le tableau 2.2-7 devait être transformé en tableau selon le format traditionnel comme le

tableau 2.2-5, il faudrait ajouter de nombreuses nouvelles colonnes pour chaque combinaison de Findesemaine et de Secteur pour les passagers comme pour les recettes.

Tableau 2.2-7
Enregistrements des transactions

Transporteur	Paire de villes	Passagers	Secteur	Recettes	Findesemaine
A	AB	100	International	100 000	1
A	AC	200	Intérieur	300 000	0
A	AZ	100	Transfrontalier	80 000	0

3. Détails du système

3.1 Cadre du système

Les fonctionnalités du système de traitement correspondent à la phase de « processus » du GSBPM. Ces étapes comprennent l'intégration des données, la validation, la vérification et l'imputation, la création de variables dérivées et le calcul d'estimations. Le système consiste en un ensemble d'outils qui doivent être intégrés aux nouvelles applications d'enquête. Le système est entièrement basé sur SAS et est composé d'un ensemble de macros SAS. Un processeur exécute toutes les étapes. Le système lit les paramètres et les étapes à partir des métadonnées, que l'utilisateur peut personnaliser. Si des paramètres doivent être modifiés, il suffit de mettre à jour la feuille de calcul contenant les métadonnées. Ce système est modulaire, souple et adaptable à différents modèles de traitement d'enquête.

Le système maximise l'utilisation des outils intégrés de Statistique Canada, comme BANFF pour la vérification et l'imputation et G-Est pour l'estimation (pour de plus amples renseignements sur ces deux outils intégrés, voir Statistique Canada, 2018c et 2018d). Il comprend également un guide de l'utilisateur.

En raison de la nature modulaire du système, toutes les étapes peuvent être réorganisées et répétées de toutes les façons possibles, tant que les fichiers d'entrée et de sortie sont correctement liés entre les étapes. Les tableaux ci-dessous présentent les feuilles de calcul de métadonnées décrivant les étapes de traitement. Le tableau 3.1-1 comprend les étapes qui doivent être exécutées et l'ordre dans lequel elles doivent l'être. Le tableau 3.1-2 présente les noms des paramètres et leurs valeurs pour chaque étape indiquée dans le tableau 3.1-1. Comme le montrent les tableaux, les étapes peuvent être réorganisées de toutes les façons qui soient, et les paramètres sont facilement modifiables au moyen de feuilles de calcul.

Chaque étape de traitement produit son propre ensemble de journaux et de données de sortie, ce qui facilite le débogage.

Tableau 3.1-1
Étapes de traitement

StepID [IDÉtape]	Module
1	Import [Importation]
2	Import [Importation]
3	stackFiles [Fichiersempilés]

Tableau 3.1-2
Paramètres de traitement

StepID [IDÉtape]	ParamName [NomParam]	ParamValue [ValeurParam]
1	inFileName [NomFichierentrée]	File1 [Fichier1]
1	outFileName [NomFichiersortie]	File1_out [Fichier1_sortie]
2	inFileName [NomFichierentrée]	File2 [Fichier2]
2	outFileName [NomFichiersortie]	File2_out [Fichier2_sortie]
3	listFilenames [listeNomsfichiers]	File1_out file2_out [Fichier1_sortie fichier2_sortie]
3	stackedFileName [NomFichierempilé]	File1File2Stacked [Fichier1Fichier2Empilé]

Les deux tableaux ci-dessus (tableaux 3.1-1 et 3.1-2) présentent un exemple élémentaire de mode possible d'exécution des étapes à l'aide d'un système de traitement composé de deux macros, l'une servant à importer et formater un fichier (appelée Import [Importation], avec deux paramètres) et l'autre à empiler plusieurs fichiers (appelés stackFiles [fichiers empilés], avec deux paramètres). Le tableau des étapes de traitement (tableau 3.1-1) indique la séquence des étapes à suivre, qui comprend l'étape d'importation de données (deux fois), puis l'empilement des fichiers. Le tableau des paramètres de traitement (tableau 3.1-2) donne les détails des étapes. Il faut d'abord importer un fichier nommé File1; le nom du fichier de sortie est File1_out. Ensuite, un fichier nommé file2 doit être importé; le nom du fichier de sortie est File2_out. Enfin, File1_out et File2_out sont regroupés dans un nouveau fichier appelé File1File2Stacked.

Comme nous l'avons indiqué plus haut, le système a été conçu pour le traitement des données transactionnelles sur le transport. A l'heure actuelle, une seule enquête a été traitée au moyen du système, l'Enquête sur la base tarifaire. Cette enquête collecte des données sur les recettes et le nombre de passagers auprès des transporteurs aériens pour mesurer les tarifs aériens moyens.

4. Défis

4.1 Volume

La grande taille des ensembles de données rend le traitement extrêmement lent. Il faut beaucoup de temps pour déceler les erreurs, car chaque programme est très long à exécuter. Nous recommandons de réduire le plus possible la taille des fichiers en éliminant les variables inutiles et de maximiser l'efficacité du système en diminuant autant que possible le nombre d'étapes. Le système actuel comprend des étapes juste avant celles de la vérification, de l'imputation et de l'estimation pour supprimer les variables inutiles. Il est important de procéder au début à des essais approfondis avec des fichiers très volumineux pour constater les limites du système.

4.2 Intégration des données

Quand on travaille avec des données transactionnelles, il faut souvent combiner des ensembles de données de différentes sources présentant des niveaux divers de détail, de fréquence et de qualité. Il est important de trouver d'abord une présentation, un format et des définitions communs avant de traiter les données. Nous avons retenu qu'il est important de créer le modèle de traitement en visant les fichiers d'estimation de sortie et non des fichiers d'entrée. Il faut commencer par les tableaux des fichiers de sortie, puis revenir en arrière pour déterminer les stratégies de traitement des fichiers d'entrée.

4.3 Conception

Étant donné que le système a priorisé les exigences de l'Enquête sur la base tarifaire, il a été élaboré pour un plan d'échantillonnage à deux degrés pour le moment. Dans l'Enquête sur la base tarifaire, le premier degré est un recensement des transporteurs et le deuxième degré est un échantillon de transactions de transporteurs pour certaines journées. Nous devions veiller à la cohérence entre les degrés et tenir compte des différents plans à chaque degré. Dans le futur, nous devrons s'atteler à l'élaboration d'autres caractéristiques pour d'autres modèles.

Il y a aussi eu des enjeux de couverture. Tout d'abord, il y a eu la question des transactions en double. Il est possible que différentes compagnies déclarent les mêmes transactions ou que des transactions s'annulent l'une l'autre. La question des transactions manquantes s'est également posée. Il est parfois difficile, voire impossible, de savoir s'il manque des transactions. On peut comparer les transactions aux données historiques, mais il reste difficile de déterminer avec certitude celles qui sont manquantes.

La leçon retenue ici est qu'il est important d'utiliser des méthodes d'étalonnage pour ajuster les données transactionnelles aux totaux des contrôles externes (autres enquêtes ou fichiers administratifs p. ex.). De plus, il est essentiel de collaborer étroitement avec des experts en la matière et les fournisseurs de données pour améliorer la qualité des données. Enfin, il est important de correctement définir les transactions qui relèvent de la portée des fournisseurs.

4.4 Estimation

Comme on l'a vu à la section 4.2, le système a priorisé le plan de l'Enquête sur la base tarifaire. Dans le cas de recensements pour le premier degré, le système actuel est en mesure de calculer des estimations annuelles. Une des limites du système réside dans le fait qu'il peut traiter une estimation annuelle seulement si un plan identique est mis en œuvre pour chaque période de référence infra-annuelle (p. ex., trimestrielle). Il nous faudra donc élaborer des fonctionnalités supplémentaires pour tenir compte des différents plans pour les périodes de référence infra-annuelles.

L'estimation s'est heurtée à un autre défi : la question des unités inactives. Par exemple, dans l'Enquête sur la base tarifaire, nous pouvons avoir un transporteur qui ne fait pas partie de la portée de l'enquête pour un trimestre donné. Étant donné que les transactions sont utilisées dans l'estimation, toutes les unités inactives doivent être représentées dans l'ensemble des données de transaction afin que l'estimation de la variance soit calculée correctement.

4.5 Imputation

Pour ce qui est de l'imputation, il faut aussi tenir compte de certains éléments particuliers quand on travaille avec des données transactionnelles. Il faut accorder une attention particulière aux variables de domaine quand on utilise l'imputation historique. Si la variable de domaine est corrélée avec les variables d'intérêt, il est préférable de préserver la valeur historique de la variable de domaine. Par exemple, dans l'Enquête sur la base tarifaire, on a une variable nommée Findesemaine définie à la section 2.2. Quand on utilise une imputation historique, on ajoute la valeur 1 à l'année de la date de déclaration, mais on conserve la valeur historique de la variable Findesemaine. La raison en est que la variable des recettes, qui est une variable d'intérêt, est fortement corrélée à Findesemaine (les tarifs de fin de semaine tendent à être plus élevés). Par conséquent, il est important de préserver la valeur historique de la variable de domaine.

Dans l'exemple ci-dessous, les tableaux 4.5-1 et 4.5-2 montrent un calendrier partiel pour janvier 2017 et 2018, respectivement. Le 1er janvier était un dimanche en 2017, donc sa valeur pour Findesemaine est de 1. Toutefois, si on utilise les données de 2017 pour imputer les données de 2018, nous conserverions la valeur historique de Findesemaine pour le 1er janvier, même si, en 2018, cette date correspond à un jour de semaine.

Tableau 4.5-1
Calendrier de janvier 2017

Janvier 2017						
Dimanche	Lundi	Mardi	Mercredi	Jeudi	Vendredi	Samedi
1	2	3	4	5	6	7

Tableau 4.5-2
Calendrier de janvier 2018

Janvier 2018						
Dimanche	**Lundi**	Mardi	Mercredi	Jeudi	Vendredi	Samedi
	1	2	3	4	5	6

Il faut aussi prendre en compte les identificateurs uniques lors de l'imputation historique des données transactionnelles. Dans le système de vérification et d'imputation BANFF, il est nécessaire que les ensembles de données actuels et historiques aient des identificateurs uniques d'appariement afin que l'imputation historique soit correctement réalisée. Or, dans le cas des données transactionnelles, nous n'avons habituellement pas d'identificateurs d'appariement pour les deux ensembles de données. Une solution consiste à apparier les données par groupe de domaines plutôt que par identificateur unique au moment de l'imputation historique.

5. Conclusion

Le présent article a décrit les composantes de notre système de traitement de données transactionnelles, ainsi que les défis rencontrés et les leçons retenues lors de l'utilisation de données transactionnelles. Il reste encore beaucoup à faire pour perfectionner, améliorer et accroître les fonctionnalités du système**.** Différentes fonctionnalités devront être mises au point en fonction des exigences de l'enquête qui sera intégrée. Deux autres enquêtes seront bientôt traitées au moyen de notre système : les Statistiques relatives aux mouvements d'aéronefs et l'Enquête sur l'origine et la destination des marchandises transportées par camion.

Bibliographie

Statistique Canada (2017), *Enquête sur l'origine et la destination des marchandises transportées par camion.*

Statistique Canada (2018a), *Enquête sur la base tarifaire*.

Statistique Canada (2018b), *Statistiques relatives aux mouvements des aéronefs*.

Statistique Canada (2018c), *Banff (Vérification et imputation - Système Généralisé)***.**

Statistique Canada (2018d), *G-Est (Estimation - Système Généralisé).*

UNECE Statswiki (2018), *GSBPM v5.0*.