

DÉTECTION DE VALEURS ABERRANTES LORS DU TRAITEMENT DES DONNÉES DE LA TAXE SUR LES PRODUITS ET SERVICES

Nelson Émond et Guylaine Dubreuil¹

RÉSUMÉ

La détection des valeurs aberrantes est essentielle à la production de résultats de qualité et il existe plusieurs méthodes qui ont été développées pour les détecter. La taxe sur les produits et services (TPS) est une source de données administratives. La TPS est également associée au secteur économique car elle recueille, entre autres, des informations sur le revenu et sur le montant de taxe versé pour le gouvernement fédéral.

Une méthode souvent utilisée pour la détection des données aberrantes dans le secteur économique a été élaborée par Hidioglou et Berthelot (1986). Il s'agit d'une des méthodes utilisées dans le cadre du traitement des données de la TPS. Le présent article décrira cette méthode et une variante de celle-ci qui fait intervenir la médiane des effets correspondant à une notion de distance. L'intérêt de cette variante est de réduire le nombre de paramètres nécessaires. Une nouvelle méthode, basée sur les travaux de Robert Philips (2001), faisant appel à un modèle de régression linéaire, sera aussi abordée. L'originalité de cette dernière est d'affecter à chaque donnée une probabilité d'être aberrante en tenant compte de l'influence de chacune des données par rapport aux autres dans une classe. Le temps de traitement est également un facteur important à prendre en considération si on veut modifier ou remplacer la méthode Hidioglou-Berthelot déjà en place pour la TPS. Divers résultats et un sommaire comparatif de ces méthodes seront présentés.

1. INTRODUCTION

1.1 Contexte

La taxe sur les produits et services (TPS) est entrée en vigueur en 1991. Par une entente signée entre Statistique Canada et l'Agence du revenu du Canada (ARC) en 1997, il est maintenant possible pour Statistique Canada d'avoir accès à ces informations. Sur une base mensuelle, l'ARC envoie sur ruban magnétique à Statistique Canada les informations relatives à la TPS.

Après traitements, ces informations constitueront une base de données auxiliaires complète permettant d'alléger le fardeau de réponse et de diminuer les coûts des enquêtes. Elles peuvent servir à remplacer ou à ajuster des données déjà existantes dans les enquêtes ou à en fournir là où il n'y en a pas. La TPS est une source auxiliaire d'information qui est d'autant plus intéressante qu'elle donne un reflet de l'ensemble de l'activité économique du pays.

1.2 Détection de valeurs aberrantes dans les fichiers de la TPS

Les données de TPS reçues de l'ARC à la Division des données fiscales de Statistique Canada ne sont pas utilisables directement par les enquêtes. Plusieurs étapes sont nécessaires avant qu'elles puissent être utilisées. Il faut corriger les incohérences, détecter les valeurs aberrantes, imputer les valeurs manquantes ou aberrantes, etc., avant de rendre les données disponibles aux utilisateurs. On entend par valeur aberrante toute valeur qui diffère de façon significative de la tendance globale des autres observations se rattachant à un ensemble de données ayant des caractéristiques communes.

¹ Nelson Émond, Statistique Canada, 120 avenue Parkdale Ottawa ON, Canada, K1A 0T6,
Nelson.Emond@statcan.ca

Guylaine Dubreuil, Statistique Canada, 120 avenue Parkdale Ottawa ON, Canada, K1A 0T6,
Guylaine.Dubreuil@statcan.ca

Dans le module de la détection des valeurs aberrantes, la méthode d'Hidiroglou-Berthelot (HB) (voir article Hidiroglou et Berthelot, 1986) est combinée avec d'autres méthodes avant de déterminer si une donnée est aberrante ou non. La méthode HB détecte des variations jugées anormales ou excessives en fonction du niveau de la variable d'intérêt. Avant de recommander une modification ou un remplacement de la méthode HB existante, il faut s'assurer que: les délais de traitement sont acceptables, l'efficacité de la méthode est stable dans le temps, la qualité de la détection des valeurs aberrantes est comparable, le pourcentage des données aberrantes est similaire et le nombre de paramètres est moindre.

1.3 Formation des classes de détection de données aberrantes

Le revenu des entreprises et le montant de TPS qu'elles versent sont les 2 variables principales des fichiers de TPS. En outre, leur revenu provient de la vente de produits et services taxables et non taxables. Le fichier provenant de l'ARC contient l'ensemble de toutes les transactions des entreprises ayant un compte de TPS. Chaque enregistrement correspond à une transaction. On assigne une classe de détection de données aberrantes (h), pour chaque enregistrement, qui dépend de la fréquence (fq) de déclaration de la TPS (mensuelle, trimestrielle ou annuelle), de sa classification industrielle provenant du Système de classification des industries de l'Amérique du Nord (SCIAN) et de son revenu annuel estimé. Les entreprises qui remettent un montant de TPS correspondant à un taux très près de 15 % ou 7 % (TPS versée/revenu), selon qu'elles font partie d'une province harmonisée ou non, sont acceptées automatiquement comme valeurs non aberrantes et elles sont incluses dans la formation des classes.

2. MÉTHODES DE DÉTECTION DES VALEURS ABERRANTES

La méthode HB possède plusieurs avantages lorsqu'on la compare à d'autres. Tout d'abord, elle n'exige pas l'hypothèse de la distribution normale et ne requiert pas de table de décision. Il a fallu vérifier que cette méthode continue d'être efficace et coïncide bien avec les besoins des données de la TPS. Nous l'avons comparée avec la méthode HB modifiée qui offre la possibilité d'avoir moins de paramètres prédéfinis ainsi qu'avec une autre méthode faisant intervenir un modèle linéaire à une variable qui utilise une approche complètement différente (Philips, 2001).

L'étude a porté sur différents mois et nous en avons retenu 2 ayant des caractéristiques spécifiques: décembre 2003, car c'est le mois le plus actif, le plus complet et le plus diversifié et le mois de mars 2004, car il contient en proportion beaucoup de données trimestrielles. Les autres mois ne sont pas présentés puisqu'ils ne faisaient que confirmer les conclusions des 2 mois retenus.

2.1 Méthode Hidiroglou-Berthelot standard

Dans le cadre du traitement des données de la TPS, la méthode HB est utilisée afin de permettre aux petites entreprises d'avoir un taux de variation dans le temps de leur revenu ou de la TPS versée beaucoup plus grand sans être nécessairement aberrantes ou suspectes et de restreindre ce taux de variation pour les grandes entreprises.

Lorsqu'on utilise des tendances historiques, l'application de la méthode HB exige que pour une unité i de la classe h , la variable d'intérêt x soit positive pour la période courante t et la période précédente $t-1$: soit $x_{hi(t)}$ et $x_{hi(t-1)} > 0$. La variable d'intérêt x représente soit le revenu, soit la valeur de la TPS versée. La valeur courante sera aberrante ou suspecte si elle a subi une trop forte diminution (*inf*) ou une trop forte hausse (*sup*) par rapport à la valeur précédente. Une donnée est jugée critique (*crit*) si elle est à l'extérieur des bornes définies par nos paramètres et suspecte (*susp*) s'il y a un doute sur sa valeur auquel cas elle ne sera pas imputée. Les données critiques et suspectes sont exclues du calcul des moyennes et des ratios servant à l'imputation.

Étapes de la méthode pour une classe donnée « h »:

1. Calcul du ratio de croissance : $r_{hi} = \frac{x_{hi(t)}}{x_{hi(t-1)}}$

2. Transformation : $s_{hi} = \begin{cases} 1 - \frac{r_{hM}}{r_{hi}} & \text{si } 0 < r_{hi} < r_{hM} \\ \frac{r_{hi}}{r_{hM}} - 1 & \text{si } r_{hi} \geq r_{hM} \end{cases}$ où r_{hM} est la médiane des r_{hi}
3. Calcul des effets : $e_{hi} = s_{hi} [\text{Max}(x_{hi(t)}, x_{hi(t-1)})]^U$ où $0 \leq U \leq 1$, U : facteur de courbure
4. Calcul de la valeur du premier quartile (e_{hq1}), de la médiane (e_{hM}) et du troisième quartile (e_{hq3}) des effets (e_{hi})
5. Distance : $d_{hq1} = \text{Max}(e_{hM} - e_{hq1}, |A \cdot e_{hM}|)$
 $d_{hq3} = \text{Max}(e_{hq3} - e_{hM}, |A \cdot e_{hM}|)$
 où A est déterminé par l'utilisateur et assure une valeur minimale de la distance.

Une valeur sera considérée aberrante si : $\begin{cases} e_{hi} < e_{hM} - K_{inf,crit,fq} \cdot d_{hq1} \\ e_{hi} > e_{hM} + K_{sup,crit,fq} \cdot d_{hq3} \end{cases}$

et suspecte si : $\begin{cases} e_{hM} - K_{inf,susp,fq} \cdot d_{hq1} \leq e_{hi} < e_{hM} - K_{inf,crit,fq} \cdot d_{hq1} \\ e_{hM} + K_{sup,susp,fq} \cdot d_{hq3} < e_{hi} \leq e_{hM} + K_{sup,crit,fq} \cdot d_{hq3} \end{cases}$

où $K_{inf,susp,fq} \leq K_{inf,crit,fq}$, $K_{sup,susp,fq} \leq K_{sup,crit,fq}$ sont des paramètres déterminés par l'utilisateur et sont différents selon la fréquence de déclaration de la TPS.

Le nombre de paramètres fournis par l'utilisateur est de 14 (12 pour les différentes valeurs de K , 1 pour U et 1 pour A) pour chaque variable d'intérêt. Ces valeurs ne sont pas remises à jour à tous les mois. Une des études a été de vérifier que ces valeurs restent efficaces dans le temps.

L'impact d'un changement du facteur de courbure (U) est montré à la figure 1. Dans notre système, ce facteur de courbure est $U=0,33$. La figure 2 illustre un exemple des zones du statut des données (aberrantes, suspectes, correctes) pour un facteur de courbure spécifique ($U=1$) en relation avec la taille des entreprises.

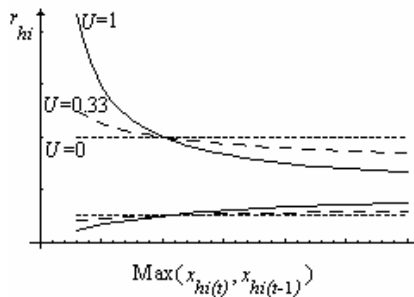


Figure 1 : Impact du facteur de courbure U sur les bornes supérieures et inférieures de la méthode HB.

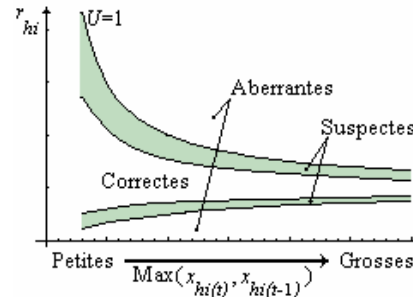


Figure 2 : Statut des données en fonction du ratio « r_{hi} » et de la taille des entreprises (de petites à grosses) pour $U=1$.

2.2 Méthode Hidirolou-Berthelot modifiée

Le but de la méthode HB modifiée est de diminuer le nombre de paramètres définis par l'utilisateur pour rendre leurs mises à jour plus faciles et plus rapides. En enlevant le facteur de la fréquence de déclaration, il y a seulement 6 paramètres par variable d'intérêt à définir au lieu de 14. Pour ce faire, il faut que la méthode prenne davantage en considération les données pour compenser l'effet de la perte du facteur de la fréquence de déclaration. La méthode HB modifiée s'ajuste plus naturellement aux données que la méthode HB standard en introduisant la médiane des effets. Naturellement, la méthode HB modifiée devra conserver la même qualité et minimiser les changements dans le système déjà en opération.

Cette méthode est la même que précédemment jusqu'au calcul des effets (étape 5 ci-dessus).

$$6. \text{ Calcul des écarts : } ecart_{hi} = \begin{cases} \frac{(e_{hi} - e_{hM})}{d_{hq1}} & \text{si } e_{hi} < e_{hM} \\ \frac{(e_{hi} - e_{hM})}{d_{hq3}} & \text{si } e_{hi} \geq e_{hM} \end{cases}$$

7. Calcul de la médiane des $ecart_{hi}$ pour chaque classe (h) : $ecart_{hM}$

8. Calcul de l'écart type des $ecart_{hi}$ pour chaque classe (h) en enlevant 5 % des plus petits et 5 % des plus grands $ecart_{hi}$: $ecart_std_h$

$$\text{Une valeur sera considérée aberrante si : } \begin{cases} ecart_{hi} < ecart_{hM} - K'_{inf,crit} \cdot ecart_std_h \\ ecart_{hi} > ecart_{hM} + K'_{sup,crit} \cdot ecart_std_h \end{cases}$$

$$\text{et suspecte si : } \begin{cases} ecart_{hM} - K'_{inf,crit} \cdot ecart_std_h \leq ecart_{hi} < ecart_{hM} - K'_{inf,susp} \cdot ecart_std_h \\ ecart_{hM} + K'_{sup,susp} \cdot ecart_std_h < ecart_{hi} \leq ecart_{hM} + K'_{sup,crit} \cdot ecart_std_h \end{cases}$$

Le nombre de paramètres fournis par l'utilisateur est de 6 (4 pour les différentes valeurs de K' , 1 pour U et 1 pour A) pour chaque variable d'intérêt.

2.3 Méthode utilisant un modèle linéaire à une variable²

Cette méthode suppose a priori qu'il existe une relation linéaire entre la valeur courante et la valeur précédente, ce qui est ici le cas, autant pour le revenu que pour la TPS versée. C'est un processus itératif qui enlève une ou plusieurs valeurs aberrantes à chaque itération. Au cours d'une itération, chaque enregistrement reçoit une probabilité d'être aberrant selon l'influence qu'il a sur la droite de régression. La recherche de valeurs aberrantes s'arrête lorsqu'on a atteint une corrélation acceptable entre les valeurs $x_{hi(t)}$ et $x_{hi(t-1)}$ restantes dans la classe ou lorsqu'il n'existe plus d'enregistrement avec une probabilité d'être aberrant suffisamment élevée. Les niveaux d'acceptation du coefficient de corrélation et du seuil de probabilité sont 2 paramètres qui sont fournis par l'utilisateur. Un des avantages de ce modèle est de réduire le nombre de paramètres requis.

3. COMPARAISON DES 3 MÉTHODES ET RÉSULTATS

On constate (Tableau 1) que les résultats sont similaires entre la méthode déjà instaurée dans le système (HB standard) et la version modifiée. Les cases ombragées représentent les cas équivoques où une méthode déclare ces enregistrements aberrants tandis que l'autre les juge corrects. On a soumis ces cas à un expert pour déterminer si une des méthodes avait tendance à mieux évaluer les valeurs aberrantes. En conclusion, il ne semble pas y avoir une méthode préférable par rapport à l'autre.

Tableau 1 : Fréquences des données aberrantes pour la variable revenu sur différents mois, toute fréquence de déclaration confondue.

Décembre 2003		HB standard				
		Critiques Inférieures	Suspectes Inférieures	Correctes	Suspectes Supérieures	Critiques Supérieures
HB modifiée	Critiques Inférieures	1879	221	29	0	0
	Suspectes Inférieures	219	589	286	0	0
	Correct	67	319	126916	623	76
	Suspectes Supérieures	0	0	653	1119	336
	Critiques Supérieures	0	0	49	377	3461

Mars 2004		HB standard				
		Critiques Inférieures	Suspectes Inférieures	Correctes	Suspectes Supérieures	Critiques Supérieures
HB modifiée	Critiques Inférieures	1875	254	27	0	0
	Suspectes Inférieures	183	529	261	0	0
	Correct	63	295	123399	564	47
	Suspectes Supérieures	0	0	662	985	290
	Critiques Supérieures	0	0	53	393	3353

² Étant donné que l'article qui sous-tend cette méthode n'est pas encore publié, les formules et le développement mathématique ne sont pas joints pour conserver l'exclusivité à son auteur. Pour plus de détails, voir Philips (2001).

De plus, l'analyse des tableaux montre que les paramètres n'ont pas besoin de changer selon le mois traité car le taux de données aberrantes et suspectes est à peu près constant dans le temps. Les mêmes conclusions sont applicables pour la variable de la TPS.

Lorsqu'on a refait la série des tableaux pour vérifier si on pouvait enlever le facteur de la fréquence de déclaration dans l'évaluation des paramètres, on a observé une hausse allant jusqu'à 3 fois plus de cas équivoques pour la TPS versée et jusqu'à 2 fois plus pour le revenu en choisissant le pire des mois. Cela indique que les paramètres ne sont pas nécessairement indépendants de la fréquence de déclaration. Par contre, le nombre de cas équivoques reste faible (environ 0.5 % sur l'ensemble des données). Donc, il est difficile de conclure qu'il y a une diminution de l'efficacité de la détection des valeurs aberrantes en ne considérant pas la fréquence de déclaration pour la méthode HB d'autant plus que les paramètres utilisés pour l'étude doivent être optimisés.

On a ensuite comparé le modèle linéaire à une variable avec la méthode HB standard pour les valeurs aberrantes critiques seulement. Étant donné que le modèle linéaire à une variable demandait beaucoup trop de temps de traitement si on voulait traiter toute la base de la TPS, seules quelques classes ont été retenues et on s'est limité à l'étude des valeurs aberrantes critiques afin d'avoir une idée des possibilités du modèle. On a constaté une différence notable dans le nombre de cas équivoques entre le modèle linéaire à une variable comparativement à la méthode HB standard. Il y avait seulement la moitié des données aberrantes qui étaient identifiées simultanément par les 2 méthodes. La principale raison pour expliquer cette différence est que la méthode HB standard accepte une variation élevée du ratio entre $x_{hi(t)}$ et $x_{hi(t-1)}$ pour les petites entreprises et une variation faible du ratio pour les grosses entreprises. Le concept même de la méthode HB est de prendre en considération la taille de l'unité dans l'évaluation des données aberrantes. Par contre, le modèle linéaire à une variable se base sur l'impact d'une donnée sur la droite de régression. Il tient également compte de la taille mais ce facteur n'est pas aussi important. Le modèle de régression à une variable n'est pas retenu pour remplacer la méthode HB standard, car il ne satisfait pas un critère important qui était d'avoir une qualité de détection de valeurs aberrantes comparable.

4. CONCLUSION

Le principal avantage de la méthode HB modifiée par rapport à la méthode standard est la possibilité de diminuer le nombre de paramètres prédéfinis et leurs mises à jour. De plus, les paramètres K' de la méthode modifiée ont une signification et sont plus intuitifs car ils correspondent à un nombre d'écart types par rapport à la valeur médiane. On ne peut pas affirmer qu'il y ait une diminution de la qualité dans la détection de données aberrantes avec la méthode HB modifiée. La méthode modifiée demande cependant un temps de traitement légèrement plus long que la méthode standard. Il est à noter que les bornes de la méthode standard, qui délimitent les zones critiques, sont un peu moins affectées que les bornes de la méthode modifiée en présence de valeurs influentes mais non aberrantes telles que les valeurs acceptées automatiquement. Cela démontre que la méthode standard est un peu plus robuste que la méthode modifiée. Donc, on peut conclure qu'il est préférable de conserver la méthode actuellement en opération dans le système car les avantages de la version modifiée ne sont pas significatifs.

Le modèle linéaire à une variable demande très peu de paramètres à définir. Étant donné que c'est un processus itératif, cette méthode demande un temps de traitement beaucoup plus élevé que les méthodes HB. Il est très bien adapté lorsqu'on a peu de données par classe (<200). Dans le cas de la TPS, nous avons parfois plusieurs milliers d'enregistrements par classe. On ne peut pas vraiment conclure qu'il est meilleur ou pire que la méthode HB puisque les tests n'ont été faits que sur un sous-ensemble de classes. Tout semble indiquer que la méthode HB est mieux adaptée aux données de la TPS puisqu'on veut permettre un taux de changement plus significatif pour les petites entreprises et que le modèle linéaire ne le permet pas aussi facilement. Le principal handicap du modèle linéaire à une variable est un temps de traitement significativement plus élevé que les autres. Pour le moment, il n'est pas une solution envisageable pour le système de traitement des données de la TPS.

RÉFÉRENCES

Hidiroglou, M.A., et Berthelot, J. -M. (1986), « Contrôle statistique et imputation dans les enquêtes-entreprises périodiques », Techniques d'enquêtes, Vol.12, p.79-89.

Philips, Robert. (2001), « Outliers Detection Routine used in the Annual Survey of Manufactures (ASM)», publication interne, Ottawa, Canada: Statistique Canada.