

ANALYSE ET IMPACT DES UNITÉS PRÉSUMÉES INACTIVES SUR LA BASE DE DONNÉES DE LA TAXE SUR LES PRODUITS ET SERVICES

Guylaine Dubreuil, Louis Pierre, Sébastien Labelle-Blanchet et Karine Liu¹

RÉSUMÉ

Statistique Canada reçoit mensuellement des renseignements concernant la taxe sur les produits et services (TPS) auprès des entreprises constituées ou non en société et procède au traitement de la base de données de façon à ce qu'elle soit complète. Pour ce faire, un certain nombre d'unités est imputé pour cause de retard. Or, il y a lieu de croire que les unités en retard pour un certain nombre de périodes consécutives sont en fait décédées ou inactives.

Pendant cette présentation, nous expliquerons la stratégie adoptée lors du traitement des données de la TPS faisant en sorte que des unités soient identifiées comme étant inactives après un certain nombre de retards consécutifs. Nous poursuivrons avec une analyse du taux de succès, à savoir quel pourcentage des unités considérées inactives s'avère l'être réellement. Une analyse confrontant la surestimation due à l'imputation des unités qui sont en fait inactives, la sous-estimation des unités considérées inactives qui ne le sont pas et la sous-estimation des unités naissantes dont on ignore l'existence sera également présentée afin d'évaluer la pertinence de la stratégie mise en place. En conclusion, nous discuterons des mesures à prendre pour améliorer la stratégie d'identification des unités inactives, le cas échéant.

1. INTRODUCTION

1.1 Description de la base de données

L'Agence du revenu du Canada (ARC) recueille des renseignements concernant la taxe sur les produits et services (TPS) auprès des entreprises constituées en société et celles non constituées en société. Depuis 1997, l'ARC partage mensuellement ces renseignements avec Statistique Canada. Après traitement, les données de la TPS s'avèrent être une source de données administratives mensuelles offrant une bonne solution de rechange pour diminuer le coût et le fardeau de réponse associés aux activités d'enquête économique.

La base de données de la TPS est une base longitudinale où chaque unité (déclarant) a un ensemble de transactions. Pour chaque transaction, les ventes et la taxe perçue sont déclarées pour une période de référence donnée, tout dépendant du revenu annuel de l'entreprise. Les déclarants sont tenus de remettre leur déclaration selon une fréquence mensuelle s'ils ont un revenu annuel supérieur à 6 millions de dollars, selon une fréquence trimestrielle s'ils ont un revenu annuel entre 500 mille et 6 millions de dollars, et selon une fréquence annuelle s'ils ont un revenu annuel inférieur à 500 mille dollars. Les déclarants qui remettent mensuellement ou trimestriellement ont un délai d'un mois après la fin d'une transaction pour déclarer les montants perçus en TPS au gouvernement alors que les déclarants annuels ont un délai de trois mois. À cela s'ajoute un autre délai pour que l'ARC puisse effectuer la saisie et le traitement des données. Ce n'est qu'environ sept semaines après la fin d'un mois de référence donné que Statistique Canada reçoit les premières données disponibles. À ce moment, puisque des données se rapportant à ce mois restent encore manquantes, elles sont alors imputées. Si les données manquantes sont par la suite déclarées, elles apparaîtront sur le fichier en provenance de l'ARC les mois suivants et la base de données de la TPS sera mise à jour en conséquence, jusqu'à concurrence de 12 mois.

1.2 Traitement des données

Une fois les données reçues, Statistique Canada procède au traitement du fichier de données de la TPS mois après mois, sur une période de douze mois, en commençant par le mois *m-11*, puis *m-10* et ainsi de

¹guylaine.dubreuil@statcan.ca, louis.pierre@statcan.ca, sebastien.labelle-blanchet@statcan.ca, karine.liu@statcan.ca, Statistique Canada, 120 avenue Parkdale, Ottawa ON, Canada, K1A 0T6

suite, jusqu'au mois le plus récent pour lequel des données ont été reçues, que l'on dénote par le mois m . Le mois où une transaction est traitée est déterminé par le mois où celle-ci se termine. D'abord, on vérifie les données et on effectue une détection des données aberrantes. Ensuite, on détermine quelles sont les transactions manquantes parmi celles qui sont attendues en se basant sur la fréquence de remise des déclarants de la base de données. Les données erronées de même que les transactions manquantes sont ensuite imputées afin que l'on obtienne des données complètes pour toutes les transactions attendues. Étant donné que ces données ne se rapportent pas nécessairement à un mois de calendrier, un processus de calendarisation visant à transformer les données sur une base mensuelle en respectant la saisonnalité est appliqué. Enfin, comme un certain nombre de transactions trimestrielles et annuelles n'était pas attendu, une partie des données demeure manquante à ce stade. Ces données sont extrapolées jusqu'au mois traité le plus récent afin de fournir une base de données complète pour tous les mois de calendrier depuis 1998.

1.3 Stratégie d'identification des unités inactives

Étant donné que les déclarants ont un laps de temps plutôt limité pour remettre leur déclaration à l'ARC et qu'ils sont soumis à des pénalités en cas de retard, on pose l'hypothèse que les unités en retard pour un certain nombre de périodes consécutives sont en fait décédées, ou à tout le moins, temporairement inactives. Lors du traitement des données de la TPS, une stratégie a été développée en fonction du nombre de retards consécutifs de sorte que l'on cesse d'imputer les unités qui s'avèrent être inactives. La stratégie a été bâtie dans le but de réduire le biais dû à l'imputation d'un trop grand nombre d'unités inactives. Les unités sont dites inactives parce qu'elles ont cessé de remettre leur déclaration à l'ARC. La stratégie ne s'applique pas aux unités qui déclarent un revenu de zéro et qui pourraient aussi être perçues comme des unités inactives.

La stratégie est différente selon la fréquence de remise. Pour les unités mensuelles, un maximum de trois retards consécutifs est accepté avant qu'on cesse de les imputer. Si une unité n'a toujours pas remis sa déclaration pour un mois de référence donné au moment où celle-ci est traitée en tant que mois de traitement $m-3$, l'unité est alors considérée inactive depuis le moment où les retards ont débuté et les valeurs imputées durant la période d'attente sont supprimées de la base de données (voir figure 1).

Figure 1: Stratégie d'identification des unités inactives pour les déclarants mensuels n'ayant pas remis par exemple à partir de septembre 2003.

<u>Traitement effectué en vérification et imputation</u>				
Transaction en retard pour un 1 ^{er} mois	En retard en Sep (m)			
	Imputée			
Transaction en retard pour un 2 ^e mois consécutif	En retard en Sep ($m-1$)		En retard en Oct (m)	
	Imputée		Imputée	
Transaction en retard pour un 3 ^e mois consécutif	En retard en Sep ($m-2$)		En retard en Oct ($m-1$)	En retard en Nov (m)
	Imputée		Imputée	Imputée
Transaction en retard pour un 4 ^e mois consécutif	En retard en Sep ($m-3$)	En retard en Oct ($m-2$)	En retard en Nov ($m-1$)	En retard en Déc (m)
	Aucun (Inactive)	Aucun (Inactive)	Aucun (Inactive)	Aucun (Inactive)

Pour ce qui est des unités trimestrielles, un maximum de deux transactions en retard est accepté avant que l'on cesse de les imputer. Ce n'est donc qu'au moment du traitement du troisième trimestre en retard que l'unité sera considérée inactive. À titre d'exemple, si une unité doit remettre son rapport trimestriel couvrant la période du 1^{er} juillet au 30 septembre 2003 et qu'elle est en retard pour une première fois

lorsque septembre 2003 est traité en tant que mois m , le trimestre complet sera imputé. Il en sera de même pour un second trimestre en retard. Enfin, si un troisième trimestre consécutif en retard survient lorsqu'il est attendu au moment de traiter mars 2004 en tant que mois de traitement m , l'unité sera considérée comme étant inactive et la transaction ne sera pas imputée. D'autre part, on supprimera les valeurs imputées depuis le début des retards soient, la valeur du trimestre se terminant en septembre 2003, traité en tant que $m-6$, de même que la valeur du trimestre se terminant en décembre 2003, traité en tant que $m-3$. La période d'inactivité commence alors au début de la première transaction en retard, c'est-à-dire le 1^{er} juillet 2003. Ainsi, une unité trimestrielle en retard sera considérée inactive à la période de traitement $m-6$ si elle est toujours en retard à ce moment.

Quant aux unités annuelles, une seule transaction en retard est acceptée, avec un sursis de 6 mois, avant qu'on considère une unité comme étant inactive. À titre d'exemple, si une unité doit remettre son rapport annuel couvrant la période du 1^{er} octobre 2002 au 30 septembre 2003 et qu'elle est en retard lorsque septembre 2003 est traité en tant que mois m , la transaction annuelle sera imputée au complet. Mais si au moment où on traite le mois de mars 2004 en tant que m , la transaction se terminant en septembre, traité en tant que $m-6$, est toujours manquante, la valeur imputée sera tout simplement supprimée et la période d'inactivité commencera au début de la transaction manquante, soit le 1^{er} octobre 2002.

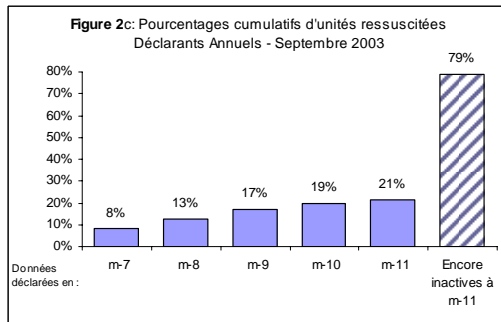
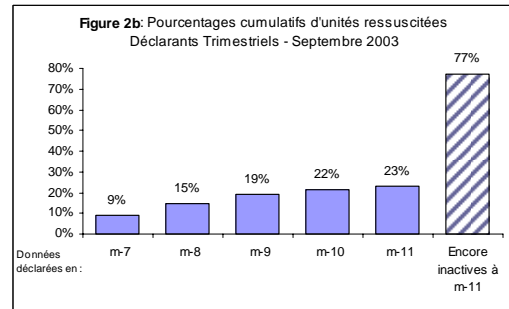
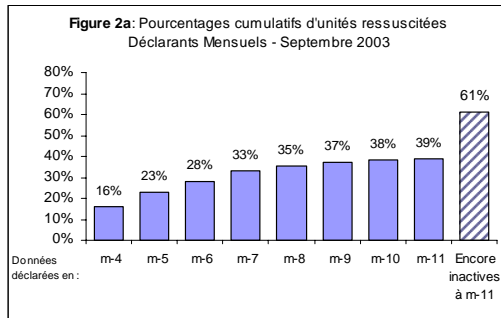
2. ANALYSE ET IMPACT DE LA STRATÉGIE

2.1 Taux de succès de la stratégie d'identification des unités inactives

Lorsqu'une unité est considérée inactive, elle demeure sur la base de données longitudinale de la TPS mais elle n'est plus imputée. S'il s'avère qu'elle n'est pas réellement inactive, elle devrait remettre sa déclaration durant les mois subséquents. Dans un premier cas, il peut arriver qu'une période de temps demeure manquante, mais que des données soient déclarées par la suite. La période manquante est alors tout simplement imputée par des valeurs de zéro et le processus de traitement habituel reprend pour les périodes déclarées qui suivent. Dans un second cas, il peut arriver qu'une unité ait tout simplement été très en retard et qu'elle décide de déclarer pour les périodes en retard après qu'elle ait été considéré comme étant inactive. Si elle déclare avant la dernière période de traitement, en $m-11$, la base de données sera tout simplement mise à jour en tenant compte des nouvelles données déclarées. Le calcul du taux de succès de la stratégie d'identification des unités inactives est basé uniquement sur ces dernières unités. Pour ce faire, nous évaluons le nombre d'unités considérées inactives pour une période de traitement donnée, puis nous suivons ces unités de mois en mois jusqu'à la dernière période de traitement en $m-11$ pour vérifier si elles ont finalement déclaré. On désigne les unités qui déclarent à nouveau par unités *ressuscitées*. Le taux de succès pour un mois de référence donné correspond au pourcentage d'unités encore inactives en $m-11$ sur le nombre total d'unités considérées inactives depuis le début de la transaction se terminant durant ce même mois de référence.

La dernière colonne des histogrammes des figures 2a, 2b et 2c représente respectivement le taux de succès des unités mensuelles, trimestrielles et annuelles pour les transactions se terminant en septembre 2003. Le rythme auquel les unités considérées inactives se remettent à déclarer y est également présenté.

On s'aperçoit que le taux de succès de la stratégie d'identification des unités inactives n'est que de 61% pour les unités mensuelles alors qu'il est respectivement de 77% et 79% pour les unités trimestrielles et annuelles. Il importe néanmoins de mentionner que ces taux de succès ont été calculés pour plusieurs mois de données, soient pour les transactions se terminant durant les mois de mai 2003 à décembre 2003.



Les taux de succès pour les unités mensuelles varient de 61% en septembre 2003 à 73% en novembre 2003. Il demeure néanmoins en deçà de 70% pour la plupart des mois à l'étude. Quant aux unités trimestrielles, les taux de succès varient de 73% en juin 2003 à 82,5% en décembre 2003 en restant généralement au-dessus de la barre des 75%. Enfin, le taux de succès des unités annuelles se maintient entre 73% en décembre 2003 et 80,5% en novembre 2003, tout en étant généralement au-dessus de 78%.

On remarque que la majorité des unités considérées inactives qui ressuscitent le font au cours du mois suivant, ce fait étant plus accentué pour les unités mensuelles. Pour ces unités où le taux de succès est moins élevé, il y a lieu de s'interroger à savoir si un mois d'attente supplémentaire n'optimiserait pas davantage la stratégie d'identification des unités inactives puisqu'entre 8% et 16% des unités considérées inactives en $m-3$ déclarent en $m-4$. Avec un seul mois supplémentaire d'attente, les taux de succès varieraient entre 68% et 80,5% avec un taux moyen de 75%, ce qui serait plus comparable aux taux de succès des unités trimestrielles et annuelles. Il faut néanmoins se rappeler que le fait d'attendre un mois supplémentaire implique qu'un nombre important d'unités inactives sera imputé un mois de plus, ce qui peut créer une surestimation. Une analyse plus approfondie s'avère donc essentielle avant de porter une conclusion.

2.2 Analyse de la surestimation et de la sous-estimation

Cette analyse a pour but de confronter la surestimation due à l'imputation d'unités qui sont en fait inactives et la sous-estimation des unités considérées inactives qui ne le sont pas. À cela s'ajoute une autre composante, soit la sous-estimation due aux unités naissantes dont on ignore l'existence jusqu'à ce qu'elles remettent leur déclaration une première fois. Pour effectuer une analyse précise sur un mois de référence donné, on procède à une comparaison entre les données calendarisées traitées en m et les données calendarisées du même mois, mais traitées en $m-11$. Le mois de traitement m correspond à la première version des données qui est disponible aux utilisateurs alors que le mois de traitement $m-11$ est la dernière version disponible. En période de traitement m , plusieurs unités sont imputées parce qu'elles sont manquantes et parmi celles-ci, certaines sont en fait inactives. Des unités en retard depuis déjà quelques mois ont aussi été considérées inactives alors que certaines ne le sont pas réellement. En période de traitement $m-11$, les données ont atteint leur maturité en ce sens que bien peu d'imputation y est effectuée, les unités qui avaient été considérées inactives à tort y sont ressuscitées et l'existence des unités naissantes y est alors connue pour la majorité des cas. La comparaison, portant sur le revenu, a été effectuée pour les mois de référence de septembre 2003 à novembre 2003. Les résultats du mois de septembre 2003 isolant les composantes de surestimation et de sous-estimation à l'étude pour chacun des types de fréquence de remise sont présentés au tableau 1.

Les différences relatives de revenus sont ici calculées par rapport au revenu mature en $m-11$. De cette analyse, on remarque que pour les unités mensuelles, la composante de surestimation et les composantes de

sous-estimation s'annulent presque totalement, ce qui pourrait laisser croire que la stratégie d'identification des unités inactives est en fait très efficace dans une perspective d'analyse macroéconomique. Dans le cas des unités trimestrielles, on dénote une sous-estimation plus importante par rapport à la surestimation bien que celle-ci soit d'un ordre de grandeur raisonnable. Pour ce qui est des unités annuelles, il ressort une sous-estimation importante par rapport à la surestimation, ce qui est dû en grande partie au fait que la majorité des unités naissantes commencent à déclarer annuellement. Dans tous les cas, le nombre d'unités inactives en m puis ressuscitées est presque le double et même davantage par rapport au nombre d'unités imputées en m puis considérées inactives. Pourtant, le revenu total associé à ces premières unités est généralement inférieur ou presque équivalent à celui des unités imputées en m puis considérées inactives, ce qui signifie que les unités qui ressuscitent ont peut-être une caractéristique commune, celle d'être plutôt petites.

**Tableau 1 : Comparaison de la surestimation et de la sous-estimation par fréquence de remise
Données calendarisées de septembre 2003 traitées en m et $m-11$**

	Données traitées en m		Données traitées en $m-11$		Diff. rel. (%)
	#	Revenu total	#	Revenu total	
Unités mensuelles					
Imputées ou déclarées en m et $m-11$	162 638	184 955 580 830	162 638	184 997 338 564	-0,02
Imputées en m puis considérées inactives	4 649	923 737 834			
Inactives en m puis ressuscitées			8 346	581 055 508	
Naissances dont on ignore l'existence en m			2 051	350 444 821	
Total	167 287	185 879 318 664	173 035	185 928 838 893	-0,03
Unités trimestrielles					
Imputées ou déclarées en m et $m-11$	1 262 598	41 681 003 814	1 262 598	40 295 844 389	3,44
Imputées en m puis considérées inactives	53 042	1 018 623 122			
Inactives en m puis ressuscitées			118 866	866 506 675	
Naissances dont on ignore l'existence en m			28 029	517 614 991	
Total	1 315 640	42 699 626 937	1 409 493	41 679 966 055	2,45
Unités annuelles					
Imputées ou déclarées en m et $m-11$	440 089	6 410 893 657	440 089	5 372 533 486	19,33
Imputées en m puis considérées inactives	61 754	541 435 901			
Inactives en m puis ressuscitées			196 316	645 075 389	
Naissances dont on ignore l'existence en m			104 936	583 253 193	
Total	501 843	6 952 329 558	741 341	6 600 862 068	5,32

Cette constatation implique aussi que si on désire prolonger le temps d'attente avant de considérer une unité inactive, il vaut mieux agir avec parcimonie, car l'imputation d'unités inactives peut avoir de lourdes conséquences. Les mêmes tendances se reflètent pour les mois d'octobre et novembre 2003. Afin de pouvoir porter une meilleure conclusion de cette analyse, il serait important de la répéter par secteur industriel, car il est fort possible que les unités qui deviennent inactives ne soient pas nécessairement dans le même secteur industriel que les unités qui naissent.

3. CONCLUSION

Les deux analyses présentées dans cet article ont démontré des résultats assez opposés en ce sens que, le taux de succès de la stratégie d'identification des unités inactives semble nettement moins bon pour les unités mensuelles alors qu'il s'avère être meilleur pour les unités annuelles. À l'inverse, la comparaison entre les composantes de surestimation et de sous-estimation démontre des résultats satisfaisants pour les unités mensuelles, résultats qui se dégradent un peu pour les unités trimestrielles et encore plus pour les unités annuelles. Avant de décider s'il y a lieu de modifier la stratégie d'identification des unités inactives,

d'autres analyses telles qu'une analyse des caractéristiques des unités considérées inactives puis ressuscitées permettraient de faire un choix plus judicieux et même de raffiner la stratégie.