

ÉVALUATION, AMÉLIORATION DE LA QUALITÉ DU REGISTRE DES FERMES ET LES BÉNÉFICES RETIRÉS

Martin Lessard¹

RÉSUMÉ

Le Registre des fermes est employé lors de la création des bases de sondage de la majorité des enquêtes agricoles canadiennes. Il est également la principale source utilisée pour améliorer et évaluer la couverture du Recensement de l'agriculture. Le recensement et les enquêtes permettent la mise à jour du registre. Suite à la réconciliation du registre avec les données du Recensement de l'agriculture de 2001, des mesures d'amélioration et d'évaluation de la qualité ont été mises en place. Bien que ces mesures soient pour la plupart automatisées, un certain travail manuel est nécessaire. Grâce à ces améliorations, une intéressante stratégie d'identification des fermes potentiellement en double a pu être développée, ouvrant la porte à de multiples analyses. Enfin, les activités d'assurance de la qualité seront abordées, suivies de recommandations concernant la maintenance du registre.

MOTS CLÉS : Qualité; base de sondage; erreurs non dues à l'échantillonnage; élimination de doubles.

1. INTRODUCTION

1.1 Description

La grande majorité des gestionnaires d'enquêtes agricoles construisent et mettent à jour leurs bases de sondage à partir du Registre des fermes. Les gestionnaires du Recensement de l'agriculture, quant à eux, l'utilisent afin de réduire et évaluer la sous-couverture (voir Lachance 2005) et d'étudier les caractéristiques propres aux fermes sous-dénombrées. Bien que ces gestionnaires aient des objectifs différents, la couverture et la qualité de l'information du registre sont deux éléments essentiels à la réussite de leurs activités.

Cet article présente les procédures manuelles et automatisées utilisées afin d'améliorer la qualité du Registre des fermes, ainsi que les bénéfices retirés de cet exercice. Afin de bien cibler l'importance de chaque procédure, les particularités entourant le Registre des fermes sont décrites, de même que plusieurs composantes du Recensement de l'agriculture, la plus importante source de mise à jour du registre.

1.2 Registre des fermes

Le Registre des fermes est en quelque sorte l'équivalent d'un bottin téléphonique agricole. Il contient l'information permettant de contacter près de 304 000 fermes canadiennes. Le statut (en affaire, hors des affaires ou vendue), nom et type d'activité agricole de la ferme, ainsi que la superficie totale des cultures font partie intégrante du registre. La majeure partie de ces informations provient du dernier Recensement de l'agriculture. Le registre compte également plus de 415 000 exploitants agricoles, au plus trois par ferme et dont certains exploitent plus d'une ferme. Selon le registre, un total de 12 763 exploitants travailleraient sur deux fermes « en affaire » et un nombre additionnel de 741 exploitants travailleraient sur plus de deux ferme « en affaire », comparativement à respectivement 3942 et 122 exploitants au recensement. Il y a présentement un peu plus de 264 000 fermes « en affaire » exploitées par près de 378 000 exploitants sur le registre.

1.3 Enquêtes agricoles

Chaque année, une soixantaine d'enquêtes agricoles permettent de recueillir de l'information sur les produits agricoles (bétail, culture, fruits et légumes, etc). Ces enquêtes permettent également la mise à jour du registre. Les principales variables mises à jour sont le statut de la ferme, son type d'activité agricole et l'information de contact.

¹ Martin Lessard, Statistique Canada, Immeuble R.H. Coats, 17^e étage, Ottawa(Ontario), Canada, K1A 0T6, lessmar@statcan.ca.

1.4 Recensement de l'agriculture

Tous les cinq ans, le Recensement de l'agriculture, mené conjointement avec le Recensement de la population, dresse le profil des fermes au Canada (voir Statistique Canada 2001), recueillant de l'information sur leurs exploitants, leurs activités agricoles et certains aspects financiers de l'exploitation (dépenses, ventes et valeur de la machinerie agricole). L'information de contact est recueillie pour un maximum de trois exploitants agricoles.

Des erreurs non dues à l'échantillonnage surviennent à presque toutes les étapes du recensement. Dès la collecte, le peu d'information recueilli pour les refus occasionne des erreurs. L'image instantanée du registre créée la journée du recensement afin d'améliorer et d'évaluer la couverture au Recensement de l'agriculture peut également être source de multiples erreurs, de par sa sous-couverture des fermes, sa sur-couverture et l'information administrative manquante ou inexacte. Parmi les erreurs possibles, il y a les erreurs d'appariement, lorsqu'une ferme du recensement est liée à la mauvaise ferme du registre ou encore lorsqu'un appariement est manqué. Les appariements manqués peuvent ensuite générer des fermes en double au recensement, puisque les fermes du registre les plus importantes et non appariées au recensement font l'objet de suivis afin de réduire la sous-couverture du recensement. Qui plus est, la présence de fermes initialement en double sur l'image instantanée du registre pourrait également créer des doubles au recensement via ce même type de suivi. Enfin, le processus d'élimination de doubles qui consiste à identifier et corriger la sur-couverture au Recensement de l'agriculture pourrait également occasionner des erreurs de couverture. Des erreurs surviennent lorsque deux fermes sont faussement identifiées comme des doubles, une d'elle étant éliminée par erreur, ou lorsque des fermes en double ne sont pas identifiées ou identifiées mais résolues comme non-doubles.

2. AMÉLIORATION DE LA QUALITÉ

2.1 Entités nuisibles

Une entité nuisible est une ferme qui autrefois était « en affaire » et qui maintenant, ne présente plus aucune activité agricole selon des critères bien établis. Plus précisément, c'est une ferme du registre « hors des affaires » qui ne se trouve sur aucune base de sondage, n'a pas été dénombrée au dernier Recensement de l'agriculture, ne peut être identifiée à l'aide de données de taxes et ne présente aucune activité agricole après la date du recensement.

Suite à la réconciliation du Recensement de l'agriculture avec le Registre des fermes, plus de 370 000 fermes étaient présentes sur le registre alors qu'un peu moins de 247 000 avaient été dénombrées au recensement. Une étude indépendante a estimé la sous-couverture du recensement à 5,6% au niveau canadien. Heureusement, la majorité des fermes manquées étant de moins grande envergure. Cette sous-couverture ne peut toutefois expliquer la grande différence entre le nombre de fermes des deux bases. Parmi les 123 000 fermes excédentaires, près de 69 000 correspondent à la définition d'entités nuisibles. Afin d'améliorer la qualité du registre, ces entités nuisibles ont été retirées du registre de façon automatisée, de même que 74 000 exploitants agricoles ne travaillant que sur ces fermes particulières.

2.2 Nouvelles fermes

Instaurée en 2003, l'Enquête annuelle pour le dépistage des fermes a pour but d'identifier des fermes « en affaire » parmi les fermes potentielles provenant de listes externes (les dossiers fiscaux, les listes obtenues de divers organismes, etc) et de les ajouter au registre. Également, elle a pour but de réactiver certaines fermes du registre dont le statut indique « hors des affaires ». Chaque année, cette enquête contacte 11 000 fermes potentielles. Au total, près de 1400 fermes ont ainsi été ajoutées au registre par les enquêtes de 2003 et 2004. Ces deux périodes de collecte ont également permis de changer le statut « hors des affaires » à « en affaire » pour plus de 3800 fermes du registre.

2.3 Ajouts et corrections d'informations administratives

De par sa nature atemporelle, la date de naissance est très utile pour les appariements avec le Registre des fermes. Le non respect du format, entre autre l'inversion fréquente de l'année, du mois et du jour, la présence de valeurs invalides et l'absence d'informations nuisent à sa qualité. Des améliorations à cet effet ont été apportées au registre. Près de 7000 dates de naissance ont été corrigées. De plus, l'utilisation de données administratives externes a également permis d'ajouter 10 900 nouvelles dates de naissance, après une évaluation soignée de la

qualité de l'information utilisée et après avoir contrôlé la qualité des liens établis entre le registre et les sources externes.

Lors de la création du Registre des fermes, il y a plus de 20 ans, tous les exploitants avaient été désignés comme étant des hommes. Bien que les valeurs de la variable « sexe » des exploitants aient été corrigées en grande partie depuis ce temps, la qualité de cette variable était jusqu'à tout récemment encore douteuse. Des efforts supplémentaires ont été déployés afin d'améliorer la qualité de cette variable dont l'inexactitude rendait l'information du registre moins crédible aux yeux des utilisateurs, en plus de nuire légèrement aux appariements. Ainsi, une table de référence a été créée à partir des fréquences « prénom-sexe » des exploitants agricoles du Recensement de l'agriculture de 2001. L'utilisation de cette table, combinée à des sources externes, a permis d'effectuer plus de 25 000 corrections sur les valeurs de la variable sexe du registre. La proportion d'exploitants de sexe masculin est ainsi passée de 78,6% à 74,1%, se rapprochant ainsi du pourcentage fourni par le Recensement de l'agriculture, soit 73,7%.

Les ajouts et corrections cités jusqu'à présent ont été réalisés de façon automatisée. Des améliorations ont également été apportées au Registre des fermes à l'aide de moyens assistés par ordinateur. En particulier, l'identification de fermes sans exploitant et la détection de noms ou prénoms invalides. Diverses techniques ont été employées afin de créer des listes de cas invalides. Jusqu'à présent, près de la moitié des 1000 cas soumis ont été manuellement résolus.

Enfin, la recherche pour l'ajout et la mise à jour de numéros de téléphone sur le Registre des fermes est un processus manuel bien établi. Les nouveaux numéros sont essentiellement trouvés via le site *canada411.com* ou encore à l'aide de disques compacts vendus par les compagnies de services téléphoniques. Aucun chiffre n'est toutefois disponible concernant le nombre de mises à jour effectuées.

2.4 Élimination d'exploitants en double

Suite aux améliorations précédentes apportées aux variables individuelles, une stratégie d'identification automatique d'exploitants agricoles potentiellement en double sur le registre a été développée et mise en application. Un des outils servant à l'identification des doubles est le programme « maison » nommé *Vraisemblance*. Celui-ci compare deux chaînes de caractères et détermine la taille de la plus longue séquence d'éléments en commun. Basée sur la taille de cette séquence et la longueur totale des deux chaînes, une valeur de vraisemblance est calculée. Cette dernière est ensuite comparée à un seuil pré-établi par l'utilisateur et la similarité des deux chaînes est confirmée lorsque l'inégalité suivante est respectée:

$$\frac{2n}{l_1 + l_2} \geq \text{seuil} \quad (2.4)$$

où n est le nombre de caractères en commun, l_1 la longueur de la première chaîne de caractères et l_2 la longueur de la deuxième. Le seuil habituellement utilisé est 0,7. À titre d'exemple, si on compare « Martin Lessard » à « Martln Lessrad », la valeur de vraisemblance obtenue en excluant les espaces est 0,846 (soit 11/13) et les deux chaînes sont considérées similaires au seuil 0,7. Lachance (2004) fournit plus de détails.

L'identification des doubles potentiels d'exploitants agricoles s'est effectuée à l'intérieur des provinces. Les exploitants potentiellement en double ont été identifiés à l'aide d'appariements sur les valeurs de diverses combinaisons de variables administratives. Le tableau 1 qui suit présente les différentes combinaisons de variables utilisées, lesquelles incluent le nom et prénom de l'exploitant, sa date de naissance et ses numéros de téléphone. En plus du programme *Vraisemblance*, un dictionnaire de conversion des surnoms du type « Bob » et « Robert » a été employé. Les appariements partiels sur les dates de naissance et numéros de téléphone ont aussi été considérés. Les cas automatiquement identifiés ont été manuellement résolus, et ce, avec l'aide de variables auxiliaires. Le tableau 2 présente ensuite un exemple typique de résolution d'exploitants agricoles potentiellement en double. Plus de 8000 exploitants en double ont ainsi été éliminés. Il resterait un peu moins de 1500 exploitants potentiellement en double sur le registre.

Tableau 1
 Combinaisons¹ de variables administratives employées
 pour identifier des exploitants agricoles en double

Combinaisons	Nom ²	Prénom ²	Surnom	Date de naissance	Numéro(s) de téléphone
1	V(0,7)	V(0,7)	—	—	6/7 chiffres
2	V(0,7)	V(0,7)	—	7/8 chiffres	—
3	V(0,7)	—	exact	—	6/7 chiffres
4	V(0,7)	—	exact	7/8 chiffres	—
5	—	—	—	7/8 chiffres	6/7 chiffres

¹ Un exploitant agricole est potentiellement en double s'il satisfait au moins une des combinaisons listées.

² V(0,7) indique un appariement à l'aide du programme *Vraisemblance* pour un seuil 0,7.

Tableau 2
 Exemple de résolution manuelle d'une paire d'exploitants agricoles
 potentiellement en double obtenue à l'aide de la combinaison #3

Exploitants agricoles potentiellement en double					
Nom	Prénom	Date de naissance	Numéro de téléphone	Numéro de tél. alternatif	Adresse
Duplessis	Bob	1960-12-25	418-999-7999	—	RR1, St-Marie
Duplesea	Robert	0000-00-00	418-555-5555	418-999-1999	RR1, St-Marie
Résultat de la résolution manuelle					
Nom	Prénom	Date de naissance	Numéro de téléphone	Numéro de tél. alternatif	Adresse
Duplessis	Robert	1960-12-25	418-999-7999	418-555-5555	RR1, St-Marie

2.5 Élimination de fermes en double

Le nom de la ferme est sans doute la variable administrative la plus utile à l'élimination de fermes en double sur le registre. Comme ce nom n'est malheureusement disponible que dans moins de 30% des cas, l'identification de fermes potentiellement en double doit s'effectuer à l'aide d'une autre variable : l'identificateur unique de l'exploitant. Comme l'unicité des exploitants agricoles du registre a été renforcée par l'élimination des individus en double, il est maintenant possible d'identifier davantage de fermes ayant des exploitants en commun et de vérifier si ces fermes apparaissent bel et bien en double sur le registre, ou si, effectivement, certains exploitants exploitent plus d'une ferme.

Certaines fermes « en affaire » ayant au moins un exploitant en commun sont identifiées comme doubles potentiels. L'examen manuel de ces paires représente une tâche considérable et il est probable que certaines paires, plus que d'autres, soient de véritables doubles, d'où la nécessité de former des groupes basés sur ce degré de certitude. Dans le but d'effectuer ces regroupements, les fermes du registre ayant été dénombrées au recensement de 2001 sont d'abord identifiées. Le groupement des paires comporte ensuite plusieurs niveaux. Le premier niveau consiste à séparer les paires qui ont déjà été vérifiées lors du processus d'élimination de doubles du recensement de 2001 de celles qui ne l'ont pas été. Ensuite, les paires sont séparées selon le type de collecte des fermes recensées. Les questionnaires recueillis par les recenseurs, incluant ceux obtenus via les suivis pour la non-réponse, sont considérés comme obtenus via la collecte régulière. Tous les autres questionnaires sont obtenus via une méthode de collecte dite « non régulière », laquelle inclut les suivis pour l'identification des fermes manquées ainsi que les refus. Enfin, le dernier niveau de groupement est le ratio d'exploitants en commun dans la paire, où le dénominateur utilisé ici est le nombre d'exploitants de la ferme comportant le plus d'exploitants. Ainsi, pour une paire de fermes (A, B) exploitées respectivement par deux et trois exploitants dont deux sont en commun, ce ratio serait 2/3. Un ratio de 1 est considéré élevé; un ratio de 1/2 ou 2/3 est considéré modéré et, finalement, un ratio de 1/3 est considéré faible. Des échantillons sont ensuite tirés dans chaque groupe. Le pourcentage de doubles véritables identifiés déterminera le degré d'automatisation dans le cadre de la résolution des autres paires du groupe. De plus, ces pourcentages permettront de cibler les groupes où il faudra concentrer le travail manuel d'élimination de doubles. Une étude est actuellement en cours afin d'évaluer les doubles potentiels parmi les fermes jugées les plus à risque.

3. BÉNÉFICES RETIRÉS

L'analyse des fermes en double permettra de déterminer les activités principalement responsables de l'apparition de doubles. Ainsi, les gestionnaires du recensement pourront corriger et/ou améliorer les processus en cause. De plus, il sera possible d'évaluer la sur-couverture au Recensement de l'agriculture 2001, ainsi que le nombre de suivis inutiles effectués pour l'identification de fermes manquées. Plus généralement, les mesures établies pour améliorer la qualité du registre permettront de réduire les erreurs non dues à l'échantillonnage qui surviennent à presque toutes les étapes du recensement. Entre autres, les appariements au registre seront plus nombreux et de meilleure qualité, permettant ainsi une optimisation du budget alloué aux suivis et assurant du même coup une diminution du fardeau de réponse. Subséquemment, ces appariements amélioreront la couverture du recensement en plus d'augmenter l'efficacité de l'imputation et de faciliter la validation des données du recensement. Qui plus est, le traitement des questionnaires du Recensement de l'agriculture en sera accéléré.

Enfin, les enquêtes agricoles pourront également profiter des améliorations apportées au Registre des fermes, avant et après le Recensement de l'agriculture de 2006. Dès maintenant, la réduction des erreurs non dues à l'échantillonnage permettra d'améliorer la qualité des bases de sondage. À long terme, les enquêtes bénéficieront d'un recensement de meilleure qualité pour la création des bases de sondage.

4. CONCLUSION

Beaucoup de travail a été fait pour améliorer la qualité du Registre des fermes. Par contre, ces améliorations sont, pour la plupart, ponctuelles. Afin de maintenir la qualité, une partie des procédures employées pourrait maintenant être incorporée à une structure établie de mises à jour, et ce, sur une base régulière. Mais au-delà de la maintenance de la qualité, la prévention des erreurs constitue le véritable défi. À titre d'exemple, le simple fait de demander aux exploitants s'ils travaillent sur plus d'une ferme, au Recensement de l'agriculture, faciliterait l'élimination de fermes en double.

Que ce soient les gestionnaires du registre, du recensement ou des enquêtes, tous contribuent individuellement à la qualité du registre. Si un de ces intervenants augmente sa contribution, les autres en profitent.

RÉFÉRENCES

Lachance, M. (2005), « La couverture du Recensement de l'agriculture de 2006 », article présenté au Colloque francophone sur les sondages 2005, Québec, Canada.

Lachance, M. (2004), « Search Name Tool », rapport non publié, Ottawa, Canada: Statistique Canada.

Statistique Canada (2001), « Le recensement de 2001 en bref », publication n° 92-379-XIF au catalogue, Statistique Canada.