Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

ÉVALUATION DE L'EFFET DE DIVERS PARAMÈTRES DE VÉRIFICATION SUR LA QUALITÉ DES DONNÉES

Katherine Jenny Thompson et Samson Adeshiyan[1]

RÉSUMÉ

Le présent article décrit l'évaluation de deux ensembles de procédures de vérification par ratio et d'imputation, fondés l'un et l'autre sur l'utilisation du sous-système généralisé de vérification et d'imputation (« Plain Vanilla ») du US Census Bureau au moyen des données de l'Economic Census de 1997. Nous comparons, après vérification et imputation, la qualité des macro et des microdonnées produites par les deux ensembles de procédures et exposons comment nos méthodes quantitatives nous ont permis de recommander certaines modifications des procédures courantes.

MOTS CLÉS : Vérification par ratio; totalisation; test aveugle.

1. INTRODUCTION

Le U.S. Census Bureau réalise un recensement économique les années se terminant par le chiffre 2 ou le chiffre 7 en envoyant par la poste plus de 4 millions de questionnaires aux établissements commerciaux qui fournissent des services aux membres du public et aux autres entreprises. Pour l'Economic Census de 1997, le Census Bureau a développé et utilisé un sous-système généralisé de vérification et d'imputation, qu'il a appelé « Plain Vanilla » (PV). Le sous-système de vérification PV comprend trois programmes distincts de vérification et d'imputation (rédigés de façon générique) : un module de vérification par ratio, un module de vérification de solde et un module de vérification. Les divisions spécialisées personnalisent ces programmes grâce au développement de scénarios de vérification qui décrivent comment le sous-système PV traite les vérifications d'un programme particulier. À titre d'exemple de fonctions du scénario, mentionnons l'énumération des items à vérifier par ratio et les poids de fiabilité connexes, la fourniture d'un modèle d'imputation assurant la mise en séquence par ordre précisé de préférence et la description des vérifications de solde.

Pour la partie de l'Economic Census réservée aux secteurs des services, l'utilisation du sous-système PV en 1997 représentait un changement important dans les méthodes de vérification et d'imputation par rapport à celles utilisées lors des recensements précédents. La méthodologie du module de vérification par ratio était notamment assez différente (voir la section 2). Par conséquent, à la fin des opérations de traitement du cycle de production de 1997, nous avons procédé à un contrôle de la qualité de la mise en œuvre du système PV pour chaque secteur de services. À la suite de ce contrôle, nous avons recommandé que plusieurs modifications soient apportées aux procédures de vérification par ratio utilisée en production. En outre, nous avons recommandé que l'on élabore des scénarios de vérification distincts pour chaque secteur commercial. Pour 1997, un seul scénario de production a été utilisé pour traiter les 5 secteurs commerciaux du secteur des services.

Pour évaluer l'effet de l'utilisation de ces deux implémentations distinctes des vérifications sur la qualité des données, nous avons réalisé un test sur un sous-ensemble de branches d'activité et d'éléments de données de

[1]Economic Programs Directorate, U.S. Bureau of the Census, Washington, D.C, U.S.A, 20233. Les auteurs remercient Scot Dahl, Ruth Detlefsen, Nash Monsour, Kenneth Sausman, Richard Sigman et Michael Walkup de leurs commentaires constructifs au sujet de versions antérieures du manuscrit. Le présent article décrit les résultats de travaux de recherche et d'analyse entrepris par les employés du Census Bureau. Il a fait l'objet d'un examen de portée plus limitée que celui auquel sont soumises les publications officielles. Le présent rapport est diffusé en vue de tenir les parties intéressées au courant des travaux de recherche en cours et de favoriser les discussions.

base en nous servant de données provenant du Recensement de 1997. Pour ce test, nous avons développé de nouveaux scénarios de vérification pour chaque secteur commercial en utilisant les méthodes de vérification par ratio et d'imputation par item recommandées à la suite du contrôle de la qualité susmentionné (Thompson et coll., 2001), puis nous avons soumis des jeux de données d'essai pour ces branches d'activités aux scénarios révisés (nouveaux) et aux scénarios originaux (production) (nommés « nouveaux » et « anciens » scénarios, respectivement, dans la suite du document). Après le traitement, nous disposions de trois valeurs concurrentes pour chaque élément de données vérifié pour chaque branche d'activité, à savoir la valeur définitive publiée, de l'élément de données figurant dans la base de données de production (supposée « correcte »), la valeur imputée selon l'ancien scénario/ancien paramètre et la valeur imputée selon le nouveau scénario/nouveau paramètre.

Le présent article décrit comment nous avons comparé, aux macro et micro-niveaux B, la qualité des données traitées durant le cycle de production de 1997 à celles obtenues en appliquant les procédures modifiées recommandées à la suite du contrôle de la qualité. La section 3 fournit des renseignements généraux sur l'étude d'évaluation. La section 4 décrit les méthodes utilisées pour comparer les macrodonnées (totalisations) et les résultats de ces comparaisons. La section 5 présente la méthode d'examen des microdonnées et les résultats connexes. La section 6 présente une discussion de ces résultats et la section 7, les conclusions.

2. MÉTHODOLOGIE DU MODULE PV DE VÉRIFICATION PAR RATIO

Une vérification par ratio consiste à comparer le ratio de deux éléments de données fortement corrélés à une borne supérieure et une borne inférieure, appelées tolérances. Les éléments de données déclarés qui tombent en dehors des tolérances sont considérés comme rejetés à la vérification et les valeurs de l'un ou des deux éléments de données faisant partie d'un ratio rejeté à la vérification sont imputées ou marquées d'un repère aux fins de leur examen par un analyste. Le module de vérification par ratio du sous-système PV utilise le modèle de vérification de Fellegi-Holt, donc, exécute l'ensemble complet de vérifications par ratio **simultanément**. Le programme détermine le nombre minimal de zones contenant des données déclarées qui doivent être modifiées pour satisfaire l'ensemble complet de vérifications (Greenberg, 1986). L'ensemble complet de vérifications s'entend des vérifications précisées par l'utilisateur et fournies dans le scénario (vérifications explicites) ainsi que les autres tests de ratio **impliqués** par l'ensemble explicite. [Nota : toute paire de vérifications par ratio possédant un élément de données commun implique une autre vérification par ratio]. Cette méthode est utilisée avec succès au Census Bureau par d'autres programmes économiques depuis les années 1980 (Greenberg et coll., 1990).

Par exemple, pour chaque secteur des services, des données sont recueillies sur la paie annuelle (APR), la paie au premier trimestre (QPR) et l'emploi (EMP). Pour s'assurer que la valeur imputée de APR ne soit jamais inférieure à la valeur imputée de QPR et que la valeur du ratio de APR à QPR ne soit jamais « très éloignée » de la moyenne de quatre généralement observée pour la branche d'activité, le développeur des règles de vérification spécifie que $1 \leq APR/QPR \leq 6$. Puisque l'emploi est habituellement un bon prédicteur de la paie annuelle, le développeur des règles de vérification définit un test explicite de correspondance entre ces deux variables en imposant des limites de tolérance propres à la branche d'activité ($BI \leq APR/EMP \leq BS$). Ces deux tests impliquent une troisième relation, à savoir le ratio de la paie du premier trimestre à l'emploi testé au moyen de $BI/6 \leq QPR/EMP \leq BS$. Après vérification et imputation, l'enregistrement doit satisfaire **les trois** règles de vérification.

Pour chaque élément de données, on peut utiliser des poids de fiabilité pour influer sur la probabilité que les données d'une zone particulière soient supprimées, la fiabilité des données étant d'autant **plus élevée** que le poids est faible. Le nombre de rejets pour chaque élément de données rejeté à la vérification est multiplié par le poids de fiabilité, si bien que réduire au minimum le nombre d'éléments de données à supprimer équivaut à maximiser le nombre pondéré de rejets.

Le module PV de vérification par ratio impute automatiquement des données pour remplacer les éléments de données déclarés manquants. Si un seul élément de données est déclaré, celui-ci n'est pas soumis à la

vérification (deux éléments de données non manquants sont nécessaires pour exécuter une vérification par ratio). Toutefois, le système impute un enregistrement complet à partir de l'élément de données unique (non vérifié) déclaré.

3. CONTEXTE

La partie de l'Economic Census réservée aux services est un recensement avec envoi et retour des questionnaires par la poste qui porte sur cinq secteurs commerciaux, à savoir le commerce de détail, le commerce de gros, les industries de services, les branches du transport, des communications et des services publics et les branches de la finance, des assurances et de l'immobilier (FAI). Les données sont recueillies au moyen d'environ 150 questionnaires différents, propres aux branches d'activité. Pour certains secteurs commerciaux, les établissements d'une branche d'activité particulière sont également classés selon la forme juridique de l'organisation, le genre d'opération et la situation fiscale. Nous utilisons ces classification de vérification-traitement pour notre évaluation, mais considérons chaque classe comme une branche d'activité.

Les spécialistes des divers secteurs commerciaux ont défini les branches d'activité utilisées pour le test. Elles ont été choisies parce qu'elles posaient des problèmes particuliers en 1997 et n'étaient pas destinées à représenter le secteur commercial dans son ensemble. Un « effet secondaire » de ce critère est qu'il pourrait être très difficile de développer des paramètres de vérification et d'imputation pour ces branches d'activité. Nous disposions d'un petit nombre de branches d'activité par secteur commercial : quatre pour le commerce de détail, 14 pour le commerce de gros, sept pour le secteur des services, quatre pour les services publics et quatre pour le secteur FAI. Nous avons réalisé notre évaluation par branche d'activité dans les secteurs commerciaux et utilisé uniquement les dossiers des établissements déclarants exploités toute l'année. Les éléments de données utilisés pour l'étude varient légèrement selon le secteur commercial. Outre la paie annuelle, la paie du premier trimestre et le nombre d'employés, tous les secteurs commerciaux recueillent des données sur les ventes et les revenus (VR). En outre, le secteur du commerce de gros recueille des données sur les dépenses d'exploitation, les achats, les stocks d'ouverture et les stocks de clôture et celui des services recueillent des données sur les dépenses d'exploitation des branches d'activité exemptes d'impôt.

Deux différences importantes caractérisent les ensembles de vérifications par ratio prévus par les nouveaux et les anciens scénarios. En premier lieu, les anciens scénarios contiennent un plus grand nombre de variables vérifiées par ratio, y compris des tests sur des données complémentaires et des données administratives. En deuxième lieu, les anciens scénarios fournissaient des tolérances pour les ensembles complets de vérifications par ratio (explicites et implicites), ce qui se traduisait par des fourchettes d'acceptation très serrées. Les nouveaux scénarios ne contiennent plus aucun test sur les données complémentaires (souvent, ces éléments de données étaient faiblement corrélés aux éléments de données de base, ce qui provoquait des rejets à la vérification/imputations fondés sur des liens ténus) et ne spécifient qu'un ensemble très limité de vérifications par ratio explicites (APR/QPR, APR/EMP et SLS/APR pour tous les secteurs commerciaux, ainsi que quelques tests supplémentaires pour une branche d'activité du secteur des services et pour le secteur du commerce de gros). Bien que nous ayons recommandé d'inclure des tests de ratio portant sur les données administratives dans les nouveaux scénarios, nous ne les avons pas inclus dans nos scénarios d'essai, à cause de préoccupations opérationnelles. Nous avons appliqué divers poids de fiabilité (aux mêmes éléments de données) dans chaque scénario. Enfin, les paramètres liée à la vérification et à l'imputation n'étaient pas tous de même qualité. Les paramètres des anciens scénarios avaient subi plusieurs révisions alors que ceux des nouveaux scénarios n'avaient été révisés qu'une seule fois. Par conséquent, nous avons considéré l'obtention de résultats d'aussi bonne qualité au moyen des anciens que des nouveaux scénarios comme une preuve de l'amélioration de la méthodologie dans le cas des nouveaux scénarios.

4. MACROÉVALUATION (COMPARAISON DE TOTALISATIONS)

Notre premier ensemble d'analyses vise à comparer des totalisations d'éléments de données provenant des résultats des anciennes et des nouvelles vérifications aux totalisations fondées sur les données définitives publiées pour 1997 (notre « norme d'excellence »). Pour cela, nous avons produit trois totalisations d'éléments

de données par catégorie branche d'activité/taille de l'établissement (petite, grande, totale), à savoir une totalisation par scénario (nouveau/ancien) et la totalisation des données définitives vérifiées. Puis, nous avons calculé les ratios des totalisations pour les anciens et les nouveaux scénarios aux totalisations des données définitives, en considérant comme étant la « meilleure » totalisation celle pour laquelle le ratio s'approchait le plus de l'unité. Si **les deux** ratios étaient compris dans un intervalle de plus ou moins 5 % par rapport à la valeur finale, nous avons considéré que les deux scénarios étaient équivalents. Le tableau 1 résume nos comparaisons des éléments de données pour le **total** des établissements sur les diverses branches d'activités d'un secteur commercial. Les profils de totalisation obtenu pour les petits et les grands établissements sont comparables.

Sauf dans le cas du commerce de détail, pour tous les éléments de données, les totalisations des données fondées sur les nouveaux scénarios de vérification sont généralement plus proches des valeurs définitives publiées que les totalisations correspondantes fondées sur les anciens scénarios, les premières étant souvent de loin meilleures. Les deux éléments de données sur les stocks pour le commerce de gros sont des exceptions notables : les deux catégories de scénarios donnent d'aussi mauvais résultats. Pour les six branches d'activités qui recueillent des données sur les stocks, le ratio moyen de la valeur des stocks d'ouverture selon le nouveau scénario à la valeur définitive des stocks d'ouverture et celui de la valeur des stocks de clôture selon le nouveau scénario à la valeur définitive des stocks de clôture était égal à 13,3 (min = 1,8, max = 35,8) et 15,4 (min = 1,9, max = 54,9), respectivement et le ratio moyen de la valeur des stocks d'ouverture selon l'ancien scénario à la valeur définitive des stocks d'ouverture et celui de la valeur des stocks de clôture selon l'ancien scénario à la valeur définitive des stocks de clôture était égal à 12,03 (min = 2,03, max = 35,9) et à 13,4 (min = 1,9, max = 54,9), respectivement. La vérification par ratio des éléments de données sur les stocks est difficile. Bien que la corrélation de ces éléments de données par paire soit forte, leur corrélation à d'autres éléments de données de base est mauvaise.

Tableau 1 : Comparaisons des éléments de données

Secteur commercial	Éléments de données	Nouveau meilleur	Ancien meilleur	Égalité
FAI	Ventes	3	1	0
	Paie annuelle	3	1	0
	Paie du 1er trimestre	3	1	0
	Emploi	4	0	0
Commerce de détail	Ventes	2	2	0
	Paie annuelle	2	2	0
	Paie du 1er trimestre	2	2	0
	Emploi	2	2	0
Services	Ventes	2	2	3
	Paie annuelle	2	1	4
	Paie du 1er trimestre	3	1	3
	Emploi	3	1	3
	Dépenses d'exploitation	0	1	0
Services publics	Ventes	2	1	1
	Paie annuelle	1	1	2
	Paie du 1er trimestre	1	1	2
	Emploi	2	0	2
Commerce de gros	Ventes	9	5	0
	Paie annuelle	8	4	2
	Paie du 1er trimestre	7	2	5
	Emploi	9	1	4
	Dépenses d'exploitation	6	5	3
	Achats	6	0	0
	Stocks d'ouverture	2	4	0
	Stocks de clôture	3	3	0

Ensuite, nous avons comparé les résultats totaux selon la branche d'activité dans chaque secteur commercial. Pour chaque branche d'activité, nous avons totalisé nos classifications des ratios (nouveau meilleur/ancien meilleur/égalité) pour chaque élément de données pour obtenir une valeur au niveau de la branche d'activité.

Le tableau 2 résume les comparaisons au niveau de la branche d'activité pour le **total** des établissements. De nouveau, sauf pour le commerce de détail, les nouveaux scénarios produisent généralement des résultats plus proches des résultats définitifs totalisés que les anciens scénarios. Malheureusement, pour les deux branches d'activité du secteur du commerce de détail pour lesquelles les anciens scénarios de vérification étaient meilleurs, la valeur des totalisations des données imputées selon les nouveaux scénarios de vérification étaient de deux à trois fois plus grandes que celles des totalisations définitives. Au départ, nous pensions que la variation des résultats selon le scénario était due à des fourchettes de tolérance trop larges si bien que nous

avons resserré les bornes pour les vérifications par ratios pour les branches d'activité en question et avons procédé à une nouvelle vérification. Le deuxième ensemble de totalisations des données imputées selon les nouveaux scénarios ne différait pas de façon notable de l'ensemble précédent. Manifestement, il ne s'agissait pas d'un problème de paramètres.

Tableau 2 : Comparaisons au niveau de la branche d'activité

Secteur commercial	Nouveau scénario meilleur	Ancien scénario meilleur	Égalité
FAI	3	1	0
C. détail	2	2	0
Services	3	2	2
Services publics	2	1	1
C. gros	8	4	2

Pour caractériser les cas mal vérifiés au moyen des nouveaux scénarios, nous avons examiné, pour chaque élément de données, les enregistrements pour lesquels les valeurs imputées selon le nouveau scénario étaient les plus importantes. Pour les deux branches d'activité, l'écart entre les totalisations avec imputations selon l'ancien scénario ou le nouveau scénario était causé par quelques établissements posant les problèmes de déclaration suivants : toutes les valeurs exprimées en dollars étaient déclarées en utilisant une mauvaise unité (valeurs déclarées en unités au lieu de milliers), ou bien **un seul** élément de données de base était déclaré, et la valeur était manifestement incorrecte (anormalement grande ou petite). En général, les vérifications par ratio ne donnent pas de bons résultats si l'on essaye de corriger la première catégorie d'erreur (« arrondissement »). Les nouveaux scénarios visaient à repérer les erreurs d'arrondissement grâce à l'application d'un coefficient de fiabilité assez faible aux données sur l'emploi, seul élément de données de base non exprimé en dollars. Cette stratégie se soldait par un échec lorsque le nombre d'employés faisait défaut ou qu'il n'était pas déclaré correctement et avait également l'effet indésirable de ne presque jamais modifier la valeur déclarée de l'emploi. Pour les deux branches d'activité problématiques du secteur du commerce de détail, seuls les éléments de données exprimés en dollars étaient déclarés par les établissements commettant l'erreur d'arrondissement et les nouveaux scénarios ont produit une imputation de valeurs de l'emploi qui concordait avec les valeurs **non arrondies** (unités). La deuxième situation, c.-à-d. la déclaration d'un seul élément de données, ne peut être corrigée au moyen de la vérification par ratio. En lieu et place, le module impute un enregistrement complet à partir de l'unique élément de données déclaré. Pour les branches d'activité en question, l'élément de données déclaré était la paie annuelle ou les ventes, déclarées en unités au lieu de milliers.

Comme nous soupçonnions que ces deux problèmes de déclaration n'étaient pas uniques au secteur du commerce de détail, nous avons fait une macro-comparaison de tous les secteurs commerciaux. Partant de toutes les valeurs déclarées pour les quatre éléments de données de base communs (APR, QPR, SLS, EMP), nous avons reclassé les établissements dans les sept catégories suivantes : 1) imputation totale (délinquants), 2) un élément de données déclaré, valeur déclarée incorrecte, existence de données administratives, 3) un élément de données déclaré, valeur déclarée incorrecte, pas de données administratives, 4) un élément de données déclaré, valeur déclarée correcte, existence de données administratives, 5) un élément de données déclaré, valeur déclarée correcte, pas de données administratives, 6) au moins deux éléments de données déclarés, pas d'erreur d'arrondissement et 7) au moins deux éléments de données déclarés, au moins une erreur d'arrondissement.

La catégorie « un élément de données déclaré » a été divisée en sous-catégories « valeur déclarée correcte/incorrecte » pour examiner l'effet des nouveaux paramètres d'imputation : le scénario comportant les meilleurs paramètres d'imputation devrait produire systématiquement des totalisations plus proches des résultats définitifs totalisés. Nous avons en outre divisé la catégorie « un élément de données déclaré » selon l'existence de données administratives. Les anciens scénarios incluaient des tests concernant les données administratives, si bien que les cas pour lesquels existaient des données administratives n'étaient pas réellement des cas « un élément de données déclaré » pour les deux catégories de scénario. Après reclassification des établissements, nous avons produit les mêmes totalisations et ratios que ceux décrits plus haut et avons procédé à une classification croisée selon la catégorie de « caractéristique de l'enregistrement » (au lieu de la catégorie de taille d'établissement), puis nous avons réalisé des analyses comparables à celles présentées au tableau 1 et 2.

Les deux catégories de scénarios ont donné d'aussi bons résultats l'une que l'autre pour les établissements pour lesquels il a fallu procéder à une imputation totale. Comme prévu, les anciens scénarios donnaient lieu à une meilleure imputation lorsque un seul élément de données (correct ou incorrect) était déclaré, à condition qu'au moins un élément de données administratif soit disponible. Les anciens scénarios donnaient aussi de meilleurs résultats que les nouveaux lors de la correction des erreurs d'arrondissement (probablement parce qu'ils contenaient des tests portant sur les données administratives). Sinon, les nouveaux scénarios donnaient généralement lieu à une meilleure imputation. Les deux situations où les anciens scénarios étaient systématiquement meilleurs représentaient un très faible pourcentage des établissements testés (moins de 2 % des établissements de tout secteur commercial).

5. MICROÉVALUATION (TEST AVEUGLE)

5.1 Description du test

La comparaison de totalisations est un moyen assez objectif de déterminer s'il existe un écart systématique entre les deux ensembles de vérifications pour ce qui est de l'effet sur les totalisations. Cependant, l'influence des grands établissements est forte dans ce genre de comparaison. Bien que tous les questionnaires de l'Economic Census soient vérifiés par machine, les analystes limitent généralement le réexamen des cas rejetés à la vérification aux grands établissements vu les contraintes de temps. Les étapes subséquentes d'examen des données se concentrent habituellement sur les grands établissements parce que ce sont ceux qui ont le plus d'effets sur les totalisations. Donc, une des limites de la macroévaluation est que, si elle donne de bons résultats dans l'ensemble et pour les grands établissements en particulier, elle ne permet pas nécessairement de repérer les problèmes de vérification concernant les petits établissements.

Étant donné ces limites, nous ne voulions pas utiliser les microdonnées publiées pour notre comparaison de micro-niveau. En outre, les analystes des divisions spécialisées voulaient examiner les microdonnées produites selon les deux catégories de scénarios de vérification. Par conséquent, nous avons réalisé des tests aveugles pour toutes les branches d'activité étudiées. La raison qui a motivé au départ la réalisation de tests aveugles était de permettre aux analystes (du moins l'espérions-nous) de se familiariser avec les procédures révisées au cas par cas. Nous avions aussi prévu d'utiliser les résultats des tests aveugles pour confirmer les résultats de la macroévaluation présentés à la section 4 ou pour déceler des problèmes systématiques de vérification/imputation pour certaines catégories d'établissements (p. ex., les petits établissements).

Pour le test aveugle, nous avons fourni aux analystes de chaque secteur commercial des renseignements de base pour chaque élément de données[2] vérifié pour 200 cas sélectionnés au hasard (100 cas par catégorie de taille) par secteur commercial et nous leur avons demandé de choisir le résultat de vérification (vérification A ou vérification B) acceptable, s'il y en avait un. Les étiquettes pour les vérifications A et B ont été attribuées au hasard, pour que ni les analystes ni les vérificateurs ne sachent quel scénario avait été utilisé pour obtenir l'un et l'autre résultat. Pour éviter de biaiser éventuellement le résultat, nous nous sommes assurés que les analystes ne puissent **pas** identifier un établissement particulier. En outre, nous ne leur avons montré aucun indicateur de vérification. En fait, leurs outils d'examen des données étaient plus limités qu'ils ne le seraient dans un système de production. De surcroît, les résultats tabulés décrits à la section 4 ne leur ont été transmis que quand le test aveugle a été achevé. Nous avons demandé aux analystes d'examiner au moins 50 des 100 cas par catégorie de taille.

Nous voulions obtenir des éclaircissements au sujet des deux questions suivantes. En premier lieu, au niveau des microdonnées, les analystes préféraient-ils les résultats d'un scénario plutôt que ceux de l'autre? En deuxième lieu, les résultats de la micro-évaluation (préférence des analystes) concordaient-ils avec ceux de la macro-évaluation? Et, dans la négative, pouvions-nous expliquer les causes de ces différences et déterminer ce que nous pouvions faire pour les éliminer dans l'avenir? Pour répondre à la première question, nous avons

[2] Valeur déclarée, valeur administrative (si elle existe), valeur corrigée du Recensement de 1992 et indicateurs de vérification connexes, et résultats des deux scénarios de vérification distincts.

utilisé les analyses usuelles pour des variables catégoriques présentées à la section 5.2 et pour traiter le deuxième ensemble de questions, nous avons procédé à un examen exploratoire des enregistrements pour lesquels les analystes préféraient manifestement les résultats donnés par les anciens scénarios.

Notre capacité à analyser les résultats du test aveugle a été fortement limitée par le plan d'échantillonnage. Idéalement, nous aurions sélectionné un échantillon stratifié à l'intérieur des cellules branches d'activité - taille d'établissement selon l'« importance de l'écart entre les résultats des vérifications ». En outre, notre population aurait été l'ensemble de tous les établissements pour lesquels les résultats des vérifications (nouveau scénario/ancien scénario) contenaient au moins un élément de données de base modifié par l'un ou l'autre scénario. En raison du calendrier de traitement, nous n'avons pu participer à l'élaboration du plan de sondage. Par conséquent, nous avons fourni aux analystes un échantillon aléatoire stratifié d'établissements sélectionnés dans le secteur commercial (et non la branche d'activité) et la catégorie de taille (grand, petit) pour lesquels la valeur imputée en vertu de l'un ou l'autre scénario de vérification différait de la valeur publiée pour au moins un élément de données. Par conséquent, les données du test aveugle contenait une forte proportion de cas pour lesquels les résultats de la vérification étaient quasiment identiques.

5.2 Outils d'analyse statistique : Tests d'associations et de préférences des analystes

Test 1. Test d'associations

Nous avons commencé par déterminer si les analystes avaient une préférence pour l'un des deux scénarios de vérification au moyen du test de vérification d'hypothèses suivant :

H_0: Les choix des nouvelles ou des anciennes vérifications sont indépendants l'un de l'autre (autrement dit, aucune préférence entre les vérifications).
H_1: Les choix des nouvelles et des anciennes vérifications sont dépendants l'un de l'autre (autrement dit, un type de vérification est préféré par rapport à l'autre)

en utilisant le test type du chi-carré de Pearson (Agresti, 1990) et les dénombrements présentés au tableau 3.

Tableau 3 : Totalisation des données de l'essai aveugle

		Nouvelle vérification PV		
Ancienne vérification PV		Acceptable	Non acceptable	Total
	Acceptable	Toutes deux acceptables (N_{11})	Ancienne acceptable uniquement (N_{12})	N_{1+}
	Non acceptable	Nouvelle acceptable uniquement (N_{21})	Ni l'une ni l'autre acceptable (N_{22})	N_{2+}
	Total	N_{+1}	N_{+2}	N_{++}

Le rejet de l'hypothèse d'indépendance nous permet de conclure que les analystes ont tendance à préférer une vérification plutôt que l'autre, mais il ne nous indique pas **quelle** vérification est préférée. Pour le déterminer, nous nous concentrons sur les cellules ombrées dans le tableau 3, où l'analyste a fait un choix clair : « Nouvelle acceptable uniquement » (N_{12}) ou « Ancienne acceptable uniquement » (N_{21}). Formellement, nous avons procédé au test de vérification d'hypothèses qui suit.

Test 2. Test de préférences des analystes

H_0 : $p_{21} <= p_{12}$ (ou $p_{21}-p_{12}<=0$, autrement dit, la proportion de réponses « Nouvelle acceptable uniquement » est inférieure ou égale à la proportion de réponses « Ancienne acceptable uniquement ».)

H_1 : $p_{21} > p_{12}$ (ou $p_{21}-p_{12}>0$, autrement dit, la proportion de réponses « Nouvelle acceptable uniquement » est supérieure à la proportion de réponses « Ancienne acceptable uniquement ».)

où $p_{12} = N_{12}/N_{++}$ et $p_{21} = N_{21}/N_{++}$. Nous estimons la différence entre les proportions par $(p_{21} - p_{12})$, et l'erreur-type (se) par $se(p_{21} - p_{12}) = [(p_{21}(1-p_{21})/N_{++}) + (p_{12}(1-p_{12})/N_{++}) - (2p_{21}p_{12}/N_{++})]^2$, ce qui produit une statistique distribuée comme $t_{\alpha,N++-1}$. Le rejet de l'hypothèse H_0 donne la preuve que les analystes préfèrent le

nouveau scénario de vérification plutôt que l'ancien. Puisqu'il s'agit d'un test unilatéral, toute valeur négative de la statistique signifie que nous ne pouvons rejeter H_0.

5.3 Résultats

5.3.1. Données sommaires

Le tableau 4 présente les dénombrements pour les données du test aveugle selon le secteur commercial. Comme prévu, la réponse « Toutes deux acceptables » était le choix le plus courant. Pour le commerce de gros, la proportion de cas « Ni l'une ni l'autre acceptables » est forte, vraisemblablement à cause des résultats médiocres des vérifications des données sur les stocks (voir la Section 4).

Tableau 4 : Comptes pour l'examen des données par les analystes selon le secteur commercial

Choix de l'analyste	Secteur commercial				
	FAI	Commerce de détail	Services	Services publics	Commerce de gros
Toutes deux acceptables (N_{11})	137	294	509	176	765
Nouvelle acceptable uniquement (N_{21})	73	115	183	73	263
Ancienne acceptable uniquement(N_{12})	149	55	276	92	116
Ni l'une ni l'autre acceptable (N_{22})	45	41	82	13	744

Le tableau 5 présente les résultats de nos tests d'associations (Test 1) et de préférences des analystes (Test 2) appliqués aux dénombrements présentés au tableau 4. Sauf pour le secteur des services, le tableau 5 donne la preuve d'une association entre le choix des nouvelles et des anciennes vérifications par les analystes au niveau de signification de 5 %. Pour les secteurs des FAI et des services publics, les analystes préfèrent les anciennes vérifications aux nouvelles; pour les secteurs du commerce de gros et du commerce de détail, ils préfèrent les nouvelles vérifications aux anciennes; enfin, pour les services, nous n'avons pu tirer aucune conclusion quant au sens des préférences.

Tableau 5: Résultats des tests d'associations et de préférences des analystes

	Test 1	Test 2	
	Rejet de H_0	Statistique t	Rejet de H_0
FAI	Oui	-4,75	Non
C. détail	Oui	4,00	Oui
Services	Non	-4,50	--
Serv. publics	Oui	-1,25	Non
C. gros	Oui	7,00	Oui

Ces résultats sont fort intéressants, car ils semblent contredire ceux de la macro-évaluation exposés à la section 4. Cependant, avant de les interpréter, nous avons voulu confirmer qu'ils étaient réellement indicateurs des préférences des analystes plutôt que simplement fonction d'un mauvais échantillonnage. Par exemple, le tableau 5 donne la preuve que les analystes préfèrent le nouveau scénario pour le commerce de détail. Pour deux branches d'activité de ce secteur commercial, le nouveau scénario a donné des résultats nettement meilleurs que l'ancien au niveau macro (tableau 2). Si la majorité des cas « Nouvelle acceptable uniquement » (N_{21}) pour le commerce de détail étaient échantillonnés à partir de ces deux branches d'activité, alors les résultats du test aveugle concorderaient effectivement avec ceux de la macro-analyse. Un examen plus approfondi de la répartition des données d'échantillons selon la branche d'activité a été nécessaire.

Le tableau 6 présente une stratification a posteriori des cas du test aveugle pour lesquels les analystes ont indiqué une nette préférence pour un scénario plutôt que pour l'autre (c.-à-d. les cas N_{12} et N_{21}) en fonction des catégories de branches d'activité du tableau 2. Une répartition comparable des dénombrements N_{12} et N_{21} dans les tableaux 4 et 6 donnerait la preuve que les analystes peuvent prédire les résultats de la macro-analyse d'après l'examen des microdonnées (autrement dit que le scénario préféré systématiquement produirait les meilleurs résultats de totalisation). Des résultats dissimilaires sont plus difficiles à interpréter. Par exemple, si la différence de qualité entre les deux ensembles de résultats de niveau macro était causée par quelques établissements (dont aucun n'a été inclus dans l'essai aveugle), alors les cas échantillonnés pour une branche

d'activité caractérisée par une réponse « Ancien scénario meilleur » à la section 4 pourraient effectivement être caractérisés par les mêmes résultats, voir même des résultats plus cohérents, lors de l'utilisation du nouveau scénario.

Tableau 6 : Répartition des cas N_{12} et N_{21} selon les catégories de branches d'activité utilisées pour la macro-évaluation

Secteur commercial	Nouveau scénario meilleur d'après la macro-évaluation	Ancien scénario meilleur d'après la macro-évaluation	Ancien et nouveau scénarios à égalité
FAI	66,22 %	0,00 %	33,78 %
Commerce de detail	39,41 %	61,59 %	0,00 %
Services	39,00 %	30,50 %	30,50 %
Services publics	23,64 %	17,58 %	58,79 %
Commerce de gros	58,84 %	20,58 %	20,58 %

Pour le secteur du commerce de détail, 39,41 % seulement des cas pour lesquels les analystes avaient une préférence nette pour l'un des scénario (N_{21} et N_{12}) ont été sélectionnés à partir des branches d'activité pour lesquelles le nouveau scénario donnait de meilleurs résultats globalement. Cependant, les analystes ont préféré le nouveau scénario comparativement à l'ancien dans environ 68 % des cas N_{12} et N_{21} (115/115+55). Dans les autres secteurs commerciaux, une forte proportion de ces cas N_{21} et N_{12} ont été échantillonnés à partir des branches d'activité où les nouveaux et anciens scénarios étaient à égalité. Dans ces cas, il est difficile d'établir un parallèle entre la préférence de l'analyste et les résultats de la macro-évaluation : par exemple, les analystes pourraient avoir préféré la vérification qui a préservé une plus grande quantité de données déclarées ou la modification de la valeur d'un élément de données plutôt que celle d'un autre. Pour les données sur le secteur des services publics, la majorité des cas N_{21} et N_{12} proviennent de branches d'activité où les deux scénarios sont jugés à égalité, ce qui rend impossible l'établissement de parallèles entre les résultats des macro et micro-évaluations. Pour le secteur FAI et celui du commerce de gros, la plupart des cas proviennent des branches d'activité où les nouveaux scénarios ont donné de meilleurs résultats. La majorité des cas testés pour le secteur FAI (66,22 %) ont été sélectionnés à partir de deux banches d'activité où les nouveaux scénarios donnaient de meilleurs résultats, mais où les analystes avaient tendance à préférer l'ancien scénario. La majorité des cas testés pour le commerce de gros (58,84 %) ont été sélectionnés à partir de branches d'activité où les nouveaux scénarios donnaient de meilleurs résultats et où les analystes avaient tendance à préférer le nouveau scénario.

5.3.2 Examen exploratoire des préférences des analystes

L'apparente contradiction entre les résultats des macro et micro-évaluations nous laissait perplexes. D'après les totalisations présentées à la section 4, les nouveaux scénarios produisaient des résultats de vérification nettement meilleurs, mais selon les résultats du test aveugle, les analystes avaient tendance à préférer les résultats produits par les anciens scénarios au niveau des microdonnées. Pour comprendre cette préférence, nous avons procédé à un examen exploratoire des enregistrements pour lesquels les analystes avaient manifesté une préférence nette pour les résultats produits par les anciens scénarios (cas N_{12}). Nous avons ainsi fait deux observations importantes (confirmées toutes les deux par les analystes). En premier lieu, les analystes préféraient habituellement le scénario donnant lieu au nombre le plus faible de modifications des valeurs déclarées ou s'appuyant sur les données administratives pour l'imputation, même si les données vérifiées définitives contenaient des ratios anormaux (par exemple, un ratio de la paie annuelle à la paie du premier trimestre de 12, valeur très éloignée de la moyenne de quatre observée pour la branche d'activité). En deuxième lieu, dans plusieurs cas, les analystes n'ont pu fournir suffisamment de règles pour établir les limites de tolérances pour le secteur commercial en question. Par exemple, d'après les spécifications des analystes, les tolérances du nouveau scénario assuraient que le chiffre des ventes soit égal ou supérieur à la paie annuelle. Lors de l'examen des résultats du test aveugle, nous avons appris que les analystes des données du secteur FAI préféraient en outre que le chiffre de vente ne soit pas plus de cinq fois plus élevé que la paie annuelle.

6. DISCUSSION

Au départ, lorsque nous avons planifié cette étude d'évaluation, notre objectif était de prouver que, à condition d'être mis en œuvre correctement, le module de vérification par ratio du sous-système PV pouvait produire d'excellents résultats pour la partie du recensement économique couvrant les secteurs des services, moyennant une très faible intervention humaine, voir aucune. Nous avons en grande partie réussi. Toutefois, nous ne sommes pas arrivés au bout de la tâche. Cette évaluation a révélé deux problèmes systématiques importants – commun à tous les secteurs commerciaux – que posent les nouveaux scénarios de vérification : mauvais dépistage des erreurs d'arrondissement et augmentation (comparativement à l'ancien scénario) de la probabilité d'imputer un enregistrement complet à partir d'un élément de données unique non vérifié. Nous réduirons la fréquence de ces problèmes de vérification lors de l'Economic Census de 2002 grâce à l'intégration de tests de ratios aux données administratives dans nos scénarios de vérification. Nous veillerons aussi à éviter ces situations grâce au remplissage des éléments de données en blanc à l'aide d'autres sommes déclarées de données détaillées (et (ou) de données administratives) et à la correction des erreurs d'arrondissement avant la vérification par ratio.

Nous étions déçus au départ de l'apparente contradiction résultant de la macro-évaluation et du test aveugle. Nous savions que les nouvelles procédures amélioraient la qualité globale des données, mais que les analystes passant en revue les données résultant du test aveugle arrivaient à une conclusion différente. Les déficiences de l'échantillon proprement dit nous permettaient difficilement de comprendre les implications des résultats de l'essai aveugle. Malgré tout, ce dernier a été un outil analytique utile. D'abord, il a révélé une « discontinuité » entre la préférence des analystes et les exigences visant les ratios au niveau de la branche d'activité (p. ex., préférence pour les résultats des vérifications permettant de retenir un plus grand nombre de données déclarées ou de données administratives, même si ces résultats étaient en contradiction avec les tolérances établies au niveau de la branche d'activité). En outre, l'examen des cas pour lesquels les analystes acceptaient uniquement les résultats produits par l'ancien scénario (N_{12}) nous a permis de découvrir certains problèmes systématiques posés par les paramètres de vérification.

Nous considérons le test aveugle comme une « répétition générale » de l'examen des microdonnées de production. De toute évidence, les analystes connaissent leur domaine de spécialisation. Toutefois, ils n'en savent pas autant au sujet de la mise en œuvre des vérifications par ratio. Nous devons remédier à cette situation en offrant une formation qui permet de saisir le lien entre les tolérances des vérifications par ratio et les résultats des vérifications, de sorte que les analystes puissent intégrer toutes les exigences du programme dans leurs scénarios de vérification pour 2002.

Les analyses de macro-niveau décrites à la section 4 ont produit une assez bonne évaluation globale des résultats des vérifications et des problèmes systématiques de mise en œuvre des routines de vérification (surtout lorsqu'elles sont combinées à une micro-évaluation des enregistrements « problématiques »). Nous ne nous attendions pas vraiment à ce que la micro-évaluation des analystes soit aussi révélatrice. La micro-évaluation consiste à vérifier la cohérence interne des enregistrements individuels et non leur conformité aux normes établies pour la branche d'activité. Habituellement, il est impossible de déceler les valeurs aberrantes si l'on ne dispose pas de la distribution complète. Compte tenu de ceci, la mise en œuvre des routines de vérification PV pour 2002 devrait inclure une vérification sommaire continue des cas rejetés lors de la vérification par ratio pour une branche d'activité particulière afin de dépister les routines ou les tolérances de vérification par ratio problématiques au lieu de se fier à la capacité qu'ont les analystes de reconnaître et de décrire ce genre de problème.

7. CONCLUSION

Le présent article décrit l'évaluation de deux ensembles de procédures de vérification par ratio et d'imputation au moyen de données provenant de la partie réservée aux secteurs des services de l'Economic Census de 1997, à savoir les procédures de traitement utilisées pour la production des données et un ensemble modifié de procédures résultant des recommandations faites à la suite d'un contrôle de qualité. Nous avons montré que les

procédures modifiées amélioraient la qualité des données vérifiées. De surcroît, le processus d'évaluation a permis de déterminer d'autres améliorations qu'il faudra apporter aux procédures en prévision du Recensement de 2002.

Un produit moins mesurable, mais tout aussi important, de cette évaluation a été le dialogue établi entre les responsables de la mise en œuvre des vérification (méthodologistes) et les spécialistes des domaines. La discussion des résultats de la macro-évaluation et du test aveugle a été le point de départ d'un processus de formation depuis longtemps attendu concernant la mise en œuvre des vérifications. Grâce à la présente étude, nous sommes en train de mettre sur pied des groupes de travail comptant des méthodologistes, des spécialistes de la production et des spécialistes des domaines en vue d'élaborer des paramètres et des scénarios de vérification en prévision du Recensement de 2002. En outre, une fois que les groupes de travail seront établis, nous pourrons mettre au point à l'intention des analystes des cours de formation tenant compte des lacunes révélées par la présente évaluation.

BIBLIOGRAPHIE

Agresti, Alan (1990), *Categorical Data Analysis*, New York: Wiley.

Greenberg, B. (1986), "The Use of Implied Edits and Set Covering in Automated Data Editing," rapport non publié, Washington, DC: U.S. Bureau of the Census.

Greenberg, Brian; Draper, Lisa; Petkunas, Thomas (1990), "Possibilités interactives du SPEER (Programme structuré pour la vérification et l'étude de cas complexes dans les enquêtes économiques)," Recueil du Symposium 90 de Statistique Canada, *Statistique Canada*, pp. 259-269.

Thompson, K., Sausman, K., Walkup, M., Dahl, S., King, C., Adeshiyan, S. (2001), "Developing Ratio Edits And Imputation Parameters For the Services Sector Censuses (SSSD) Plain Vanilla Ratio Edit Module Test," rapport non publié, Washington, DC: U.S. Bureau of the Census.