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

VÉRIFICATION ET IMPUTATION DANS UN SYSTÈME STANDARD DE TRAITEMENT D'ENQUÊTES ÉCONOMIQUES

Richard Sigman[1]

RÉSUMÉ

Le Census Bureau des États-Unis a élaboré un logiciel appelé StEPS (Standard Economic Processing System ou Système standard de traitement d'enquêtes économiques) pour ainsi remplacer 16 systèmes servant au traitement des données de plus d'une centaine d'enquêtes économiques courantes. Le présent document décrit la méthodologie et la conception des modules StEPS de vérification et d'imputation et résume les réactions des utilisateurs à l'emploi de ces modules dans le traitement de leurs données d'enquête.

Mots clés : Traitement des données d'enquête; Enquêtes économiques; StEPS

1. INTRODUCTION

Le Census Bureau des États-Unis fait enquête auprès des ménages, des établissements et des entreprises. Il sera surtout question ici des enquêtes-entreprises. Le Census Bureau parle dans ce cas d'*enquêtes économiques*, parce qu'elles procurent aux économistes et aux autres analystes les estimations et les ensembles de données nécessaires aux analyses macro-économiques et micro-économiques. Ainsi, le Bureau of Economic Analysis se reporte aux estimations des enquêtes économiques pour sa comptabilité nationale des revenus et des dépenses. Ces enquêtes peuvent amplement varier pour ce qui est des caractéristiques des unités de déclaration et de la teneur des questionnaires. Elles se ressemblent souvent cependant dans leurs exigences de traitement de l'information, ce qui devait amener le Census Bureau à regrouper les systèmes de traitement des données d'un grand nombre de ses enquêtes économiques. C'est l'élaboration et l'utilisation d'un logiciel généralisé appelé Standard Economic Processing System (StEPS) qui ont permis ce regroupement.

Le présent document décrit les capacités de vérification et d'imputation de StEPS. La section 2 raconte comment les modules en question ont vu le jour et la section 3 dépeint tout le système StEPS dans ses grandes lignes. La section 4 décrit plus en détail les modules en question et la section 5 présente deux exemples de données stockées dans ce système qui ont servi à produire des données quantitatives de gestion d'enquête. Enfin, la section 6 résume la rétroaction reçue des utilisateurs au sujet de l'utilisation de StEPS en vérification et en imputation et la section 7 présente les conclusions et expose les activités futures.

2. ÉLABORATION DES MODULES DE VÉRIFICATION ET D'IMPUTATION DE StEPS

Le Census Bureau compte un certain nombre de directions qui réalisent des recensements et des enquêtes. La direction des programmes économiques fait des recensements économiques tous les cinq ans et mène des enquêtes économiques à intervalles mensuels, trimestriels et annuels dans les domaines de la fabrication, de la construction, des services commerciaux, des services gouvernementaux et du commerce extérieur. Les directions de l'organisme sont responsables de la conception des méthodes d'enquête, ainsi que des systèmes de traitement des données recueillies dans le cadre de leurs recensements et de leurs enquêtes.

[1]Richard Sigman, Census Bureau des États-Unis, ESMPD, pièce 3108-4, Washington, D.C. 20233, États-Unis, richard.s.sigman@census.gov . Le présent document expose les résultats de recherches et d'analyses effectuées par le personnel du Census Bureau. Il a été soumis à un examen plus restreint que celui que le Census Bureau destine à ses publications officielles. Sa diffusion vise à informer les intéressés et à favoriser les discussions.

La direction des programmes économiques a élaboré StEPS pour le traitement de ses enquêtes en cours. Dans quelques-unes de ces enquêtes, on permet aux répondants de fournir leurs données par Internet, et il s'agit principalement d'enquêtes postales avec suivi téléphonique. Dans nombre d'entre elles, des analystes des domaines effectuent ce suivi, mais dans quelques autres, il y a des commis qui, dans un centre d'appels, communiquent avec les non-répondants postaux. Les enquêtes économiques courantes sont diversifiées sur le plan de la taille, de la périodicité et de la nature des unités de déclaration. Cette périodicité peut être mensuelle, trimestrielle, annuelle ou quadriennale. Les tailles d'échantillons des enquêtes actuellement mises en traitement StEPS vont de 12 unités dans l'enquête sur les contenants de verre à 60 000 dans l'enquête annuelle sur les dépenses en immobilisations. Les unités de déclaration des enquêtes économiques sont notamment les établissements et les divisions des entreprises, les entreprises dans leur intégralité et les projets de construction.

On a commencé les travaux d'élaboration de StEPS en 1995 en se renseignant sur les besoins des utilisateurs et en étudiant une diversité de stratégies de conception. On a interviewé des responsables d'enquêtes, des spécialistes du traitement des données d'enquête et des concepteurs de systèmes. Les documents de planification rédigés par les concepteurs de StEPS ont été examinés par des conseillers extérieurs qui connaissaient bien les besoins de traitement des premières enquêtes où StEPS serait utilisé (Ahmed et Tasky, 2001).

La détermination des besoins des usagers et des modes possibles de conception s'est trouvée grandement facilitée par trois activités antérieures. Avant l'élaboration de StEPS, la direction économique du Census Bureau avait conçu deux autres systèmes de traitement capables de prendre en charge les données de plusieurs enquêtes. On s'était doté dans les années 1980 du système * Current Industrial Reports + (CIR) pour le traitement des données de 75 enquêtes. Ce sont des enquêtes menées auprès des fabricants de produits industriels comme les peintures, les ampoules d'éclairage et les pièces moulées de fer et d'acier. Le programme CIR était écrit en FORTRAN et ses opérations de vérification et d'imputation obéissaient à des paramètres par enquête que pouvaient spécifier les utilisateurs expérimentés par la création de fichiers ASCII contenant des cartes paramètres à zones fixes. Au début des années 1990, on devait se doter du système * Generalized Annual Survey Processing + (GASP) pour le traitement des données recueillies auprès des entreprises de commerce et de services dans le cadre de huit enquêtes annuelles. Le programme GASP était écrit en COBOL et ses opérations de vérification et d'imputation étaient gérées par des fichiers paramètres par enquête qui venaient de programmeurs. Les fichiers paramètres de vérification contenaient du texte assimilable COBOL qui devait être converti en texte d'inclusion COBOL. Les fichiers paramètres d'imputation renfermaient, eux, des instructions dans un langage suffixé maison qui était interprété par le logiciel d'imputation. La force du système GASP résidait dans ses capacités d'estimation de paramètres de modèle pour des fonctions d'imputation par modèle.

Un autre exercice antérieur qui avait livré des indications utiles sur les besoins de vérification et d'imputation était l'inventaire fait par King et Kornbau (1994) des pratiques statistiques de la direction des programmes économiques. Les intéressés avaient constaté que les pratiques de vérification et d'imputation des enquêtes économiques courantes variaient considérablement en fonction du degré d'implication de commis, d'auxiliaires de traitement statistique et d'analystes de domaines. Dans quelques-unes de ces enquêtes, la vérification et la correction des erreurs étaient entièrement automatisées et comportaient dans certains cas un renvoi des unités suspectes aux analystes des domaines à des fins de révision manuelle. Dans d'autres enquêtes, seule la vérification était automatisée et les erreurs étaient corrigées par des commis ou des analystes des domaines. Dans d'autres encore, la vérification n'était que partiellement automatisée. King et Kornbau ont signalé que nombre des analystes d'enquête qu'ils avaient interviewés voulaient que l'on automatise davantage les opérations de vérification et d'imputation et que certains responsables d'enquêtes désiraient que l'on incorpore du graphisme à leurs systèmes de vérification (intégration, par exemple, de diagrammes de dispersion pour la constatation des valeurs aberrantes).

Une troisième activité antérieure qui a éclairé la conception d'un système de vérification et d'imputation a été la mise en place en 1994 d'un système de traitement des données de l'enquête sur l'irrigation des fermes de culture et d'élevage (Farm and Ranch Irrigation Survey ou FRIS). Ce programme écrit en SAS et un grand nombre de ses caractéristiques de spécification des vérifications et d'examen de leurs résultats en mode interactif s'est retrouvé dans StEPS. Ces caractéristiques comprenaient les fonctions suivantes (Monahan, 1994a, 1994b) :

- L'utilisateur définissait interactivement les vérifications par utilisation de menus de tests communs (tests de présence, d'étendue, de solde, de contrôle, etc.) et d'énoncés SAS pour la définition de tests plus complexes.
- Le système appliquait les tests ainsi définis aux données d'enquête et générait des listes d'enregistrements rejetés à la vérification avec des indications sur les éléments d'information liés à chaque rejet.
- Un module interactif d'examen et de correction permettait à l'utilisateur d'examiner les données des enregistrements rejetés. Celui-ci pouvait modifier les données et marquer des zones à des fins d'imputation. Le module servait aussi à imputer les valeurs de variables non déclarées antérieurement. Dans un fichier d'audit, on versait toutes les modifications apportées aux données et pouvait annuler des changements au besoin.

La direction des programmes économiques a mis au point un système de vérification et d'imputation appelé * Plain Vanilla + pour le traitement des données du recensement économique de 1997 pendant qu'elle élaborait le système StEPS. Plain Vanilla (PV) a ainsi été baptisé parce qu'il possède des capacités de vérification et d'imputation générales que l'on peut enrichir de codes informatiques par enquête (* garnitures +). Plus précisément, les divers secteurs spécialisés du recensement économique exécutent des fonctions supplémentaires de vérification et d'imputation qui ne font pas partie de PV. À cause de la masse de ces données, le recensement économique automatise plus ses fonctions que ne le font les enquêtes économiques courantes. Dans le système PV, les vérifications disponibles se limitent aux tests de rapports, de soldes et de contrôle des codes des enquêtes. On y repère les erreurs dans le cas des rejets aux tests de quotient. L'élaboration des modules de vérification et d'imputation de StEPS et du système PV a suivi une voie commune où on a déterminé les besoins des utilisateurs à partir de l'expérience qu'ils avaient vécue des systèmes antérieurs. L'objectif principal des concepteurs dans les deux cas était de remplacer une pluralité de systèmes de vérification et d'imputation par un système unique. Ces travaux d'élaboration ont toutefois différé par la nature des systèmes ainsi créés, c'est-à-dire un système hautement automatisé pour le recensement économique et un système très souple et facile à configurer pour les enquêtes économiques courantes. Il y a eu un avantageux transfert de connaissances de la conception du PV à celle de StEPS dans le domaine des méthodes d'imputation pour les rejets aux tests de soldes (Sigman et Wagner, 1997). Le gros du travail d'élaboration du système PV dans ce domaine a été transposé dans le système StEPS.

3. APERÇU DE StEPS

StEPS est un système généralisé de traitement de données d'enquête que la direction des programmes économiques a conçu pour le remplacement de 16 systèmes hérités du passé. Le but était non seulement de réduire les ressources nécessaires à l'entretien des systèmes, mais aussi de procurer une meilleure capacité de gestion de traitement aux analystes et aux méthodologues des enquêtes. StEPS comprend des modules intégrés de soutien de collecte de données (impression d'étiquettes d'envoi postal, réception des questionnaires, etc.), de vérification, d'examen et de correction des données, d'imputation, de calcul d'estimations et de variances et d'administration de systèmes (spécification des paramètres et assignation et contrôle de travaux par lots). Au nombre des fonctions absentes de StEPS, on compte la constitution de bases de sondage, le prélèvement d'échantillons, la collecte effective de données et la diffusion de l'information. StEPS est un programme en SAS. On stocke données et paramètres dans des ensembles de données SAS. La direction des programmes économiques exécute StEPS surtout dans des machines Compaq Alpha avec UNIX comme système d'exploitation. La plupart des utilisateurs y ont accès par un progiciel de communication graphique (X-Windows) chargé dans leur micro-ordinateur de bureau.

Ahmed et Tasky (1999, 2000, 2001) nous renseignent davantage sur ce système. Tasky et coll. (1999) décrivent la stratégie de conception du système et les stratégies liées de programmation. Ils précisent que les concepteurs de StEPS ont retenu quatre grands concepts d'élaboration :

1. il faut concevoir un ensemble de structures types de données qui ne changent pas, quelles que soient les enquêtes et les données;
2. il faut se servir de paramètres (stockés dans des structures de données générales) pour déterminer les besoins de traitement par enquête;
3. il faut produire un ensemble riche d'enregistrements (toutes les données relatives à une unité de déclaration figurent dans un même enregistrement) * à la volée + pour certains modules;
4. il faut normaliser les désignations de zones et les valeurs possibles pour des concepts semblables.

StEPS se sert des noms de variables SAS pour dénommer les divers éléments d'information d'enquête. Dans les ensembles riches d'enregistrements, la convention de notation des désignations de variables est *trrrrrxx*, où *t* est le type de données (*t*=R pour les données déclarées, *t*=E pour les données vérifiées, *t*=A pour les données corrigées et *t*=W pour les données corrigées et pondérées), *rrrrr* est une désignation de base ou un code d'élément d'information (d'une longueur de jusqu'à cinq caractères) et *xx* est l'indicateur de période statistique relative (*xx*=00 pour la période statistique en cours, *xx*=01 pour la période statistique précédente, etc.). Comme StEPS se charge initialement des données des enquêtés tant dans *Rrrrrrxx* (données déclarées) que dans *Errrrrxx* (données vérifiées), nous employons dans les exemples qui suivent les désignations de données vérifiées lorsqu'il est question de données d'enquête stockées dans le système.

4. MODULES DE VÉRIFICATION ET D'IMPUTATION DE StEPS

StEPS compte un module de vérification et deux modules d'imputation de données. Ces derniers exécutent ce que le système appelle des imputations simples et des imputations générales. L'ordre habituel de déroulement est l'imputation simple, la vérification et l'imputation générale.

4.1. Module d'imputation simple

Le module d'imputation simple de StEPS impute des valeurs en équivalence de données déclarées. Les données ainsi obtenues sont marquées comme données déclarées. Une imputation simple fréquente consiste à faire du remplissage de zones, c'est-à-dire que StEPS introduit des données manquantes par inférence immédiate d'autres données. Ainsi, l'enquête annuelle sur le commerce de détail recueille les données suivantes :

etaxyn00 = perception de la taxe de vente par le détaillant (1 pour oui, 2 pour non);
ectax00 = valeur annuelle de la taxe de vente perçue;
ecsal00 = ventes annuelles sans la taxe de vente;
ectsal00 = ventes annuelles totales, taxe de vente comprise.

Si *etaxyn00* manque (ou est égal à l'unité pour oui), mais que *ectax00*=0 et *ecsal00=ectsal00*>0, une des règles d'imputation simple de StEPS fixe *etaxyn00* à 2 (pour non).

Les utilisateurs de StEPS peuvent définir en mode interactif deux types de règles d'imputation simple, à savoir des règles de traitement complexe de soldes et des règles de libre imputation. Dans ce dernier cas, l'utilisateur spécifie les énoncés SAS qui décrivent les conditions d'* erreur + et les interventions correspondantes lorsque ces conditions sont réunies. Les interventions en question peuvent être tout groupe d'instructions SAS, quelle qu'en soit la complexité. Dans le cas des règles de traitement complexe de soldes, on vise l'addition à un total dénoté *y* de valeurs partielles dénotées x_i, où i=1,2,...n. L'utilisateur spécifie une ou plusieurs corrections de *y* ou des x_i lorsque $y \ldots \Sigma x_i$, *y* et/ou un ou plusieurs des x_i manquent et que les données disponibles sont assez complètes pour que les données imputées soient jugées équivalentes aux données déclarées. On établit le caractère complet des données disponibles si le résidu absolu $|R|=|y-\Sigma x_i|$ ou ce même résidu exprimé en valeur

relative $|R|/y$ sont inférieurs à des valeurs de tolérance spécifiées par l'utilisateur. Celui-ci peut indiquer une ou plusieurs des corrections suivantes à apporter (Luery, 1999) :

ZERO-SET : assigne la valeur zéro aux x_i manquants;

YSUMX : assigne la valeur $\Sigma_{nm} x_i$ à yà , où $\Sigma_{nm} x_i$ est la somme des x_i non manquants;

RESIDUAL : si un seul x_i manque, fixation de celui-ci à $R=y-\Sigma_{nm} x_i$;

RAKE : calcul d'un x_i dénoté x_i' de sorte que $y=\Sigma x_i$'.

Si y et tous les x_i sont des valeurs non négatives, l'option RAKE applique la formule $x_i' = x_i(y/\Sigma_{nm} x_j) = x_i(1+R/\Sigma_{nm} x_j)$, qui est reconnue par les analystes des enquêtes comme pratique acceptable lorsque $|R|/y$ est petit et inférieur à 0,05, par exemple. Cette pratique a de solides fondements statistiques là où l'erreur de déclaration d'une valeur partielle x_i se produit au hasard et que la variance de l'erreur aléatoire de cette déclaration, désignée par $var(x_i)$, est proportionnelle à x_i. Par la méthode des multiplicateurs de Lagrange, on peut démontrer que les chiffres partiels de cette méthode du quotient, x_i', minimisent la statistique chi-carré

$$P^2 = \sum_i \frac{(x_i' - x_i)^2}{var(x_i)}$$

sous réserve de la contrainte $y=\Sigma x_i$ (voir Deming, 1943, chapitre 5). Lorsque les x_i ne se limitent pas à des valeurs non négatives, Luery et Sigman (2000) montrent que, si $var(x_i)$ est proportionnel à $|x_i|$, la minimisation de la statistique chi-carré sous réserve de $y=\Sigma x_i$' donne la formule de correction $x_i' = x_i [1+\text{sign}(x_i) R/\Sigma|x_i|]$. Des écrans interactifs permettent à l'utilisateur de spécifier les règles d'imputation simple, lesquelles peuvent alors s'exécuter par lots, c'est-à-dire pour l'ensemble des cas, ou en mode interactif dans le module d'examen et de correction d'erreurs StEPS pour un seul cas.

4.2. Module de vérification

Le module de vérification de StEPS fait une détection automatisée d'erreurs de données, c'est-à-dire de valeurs qui, isolément ou par rapport à d'autres données, ne présentent pas le comportement prévu de déclaration. Il décèle seulement les erreurs sans changer les données. Il permet aussi à l'utilisateur de définir des vérifications interactivement et d'en examiner les résultats de diverses manières. Nous ne nous étendrons pas sur la question, mais précisons que ce module fait actuellement l'objet d'améliorations et qu'il comprendra désormais la fonction de vérification statistique décrite par Hidiroglou et Berthelot (1986) et évaluée par Hunt et coll. (1999) à l'aide des données de l'enquête mensuelle sur le commerce de détail.

L'utilisateur peut définir le type suivant de vérifications dans StEPS (Tasky, 2000a) :

Test d'élément d'information requis : on vérifie si la valeur d'un élément spécifié n'est pas égale à * manquant +.

Test d'étendue : on vérifie si la valeur de l'élément d'information se situe dans l'intervalle délimité par les valeurs minimale et maximale spécifiées.

Test de liste : on vérifie si la valeur de l'élément d'information spécifié est contenue dans une liste prédéfinie de valeurs.

Test de solde : on vérifie si l'addition d'éléments d'information partiels spécifiés donne le total indiqué.

Test de règle d'enquête : il s'agit d'un test libre qui valide des relations complexes entre éléments.

Test de valeur négative : on vérifie si la valeur de l'élément d'information spécifié n'est pas négative.

Tous ces tests, sauf le test de valeur négative, sont conditionnels, c'est-à-dire qu'ils doivent respecter une condition préalable avant que la vérification ne se fasse. On recourt fréquemment à des vérifications conditionnelles lorsque les unités d'un échantillon sont sélectionnées dans des secteurs économiques différents et que la vérification doit varier selon ces secteurs.

Voici des exemples de tests de règles de vérification pour l'enquête annuelle sur le commerce de détail avec les éléments d'information *ectax00* (perception de la taxe sur les ventes annuelles), *ecsal00* (ventes annuelles

sans la taxe de vente), *ectsal00* (ventes annuelles totales avec la taxe de vente) et le symbole **.** indiquant que des données manquent :

Taxe de vente uniquement déclarée : condition préalable : aucune; condition d'erreur : *ectax00*>0, *ecsal00*=. et *ectsal00*=.

Taxe de vente trop élevée : condition préalable : *ecsal00….* et ectsal00>0; condition d'erreur : *ectax00*>*ecsal00**0,15 ou *ectax00*>*ectsal00**0,15.

La définition de chaque vérification précise comment celle-ci s'exécutera. Les choix possibles sont une ou plusieurs des utilisations suivantes (appelées * événements + dans StEPS) :

Prévérification : il s'agit d'un sous-ensemble de vérifications à exécuter dans tous les cas avec examen des rejets;

Vérification complète : il s'agit d'un jeu complet de vérifications à exécuter dans tous les cas avec examen des rejets (il n'y a pas de tests de valeurs négatives dans une vérification complète);

Vérification d'imputation générale : il s'agit de vérifications à exécuter dans tous les cas détectés pour imputation;

Vérification individuelle : il s'agit de vérifications exécutées dans un cas particulier dans le module d'examen et de correction des erreurs.

Lorsque StEPS exécute des vérifications par lots qui portent l'indicateur d'événement * prévérification + ou * vérification complète +, l'information sur les cas et les éléments d'information rejetés est versée dans un fichier de rejets au niveau de l'enquête. Si ces mêmes vérifications s'exécutent en mode interactif, l'information sur les cas et les éléments rejetés va à un fichier de rejets au niveau de l'utilisateur. On peut examiner ces fichiers de rejets de diverses façons. On peut les imprimer ou les visualiser en direct sous forme de listages de données. Ils se prêtent aussi à une vérification interactive dans le module d'examen et de correction d'erreurs de StEPS, où on peut visualiser toutes les données relatives à un cas et les rejets correspondants dans les fichiers de rejets, modifier des données et exécuter dans le cas en question des vérifications portant l'indicateur d'événement * vérification individuelle +. Il est également possible d'indiquer dans le module d'examen et de correction qu'un cas doit être laissé de côté dans les passages ultérieurs de vérification. L'utilisation de ce module en vérification interactive permet aux analystes des enquêtes de corriger rapidement les erreurs de données décelées. Willimack et coll. (2000) on eu recours à des groupes consultatifs d'analystes pour établir que les analystes expérimentés font leurs vérifications interactives de la manière suivante :

1. Ils examinent tous les messages de rejet à la vérification.
2. Ils relèvent les rejets * faciles + et les règlent.
3. Ils remettent le cas en vérification pour une constatation des erreurs restantes.
4. Ils caractérisent le cas et poussent la recherche au besoin.
5. Ils règlent tout rejet restant à la vérification.
6. Ils reprennent les étapes 3 à 5 au besoin.
7. Ils téléphonent à l'entreprise déclarante s'il y a des rejets à la vérification qui ne sont pas encore réglés.

4.3. Module d'imputation générale

Contrairement à ce qui se fait en imputation simple, les valeurs modifiées par le module d'imputation générale sont marquées comme valeurs imputées. Ce module fait ses imputations à l'aide d'estimateurs (Giles et Patrick, 1986) et corrige les soldes de manière que l'addition de valeurs partielles donne le total (voir Sigman et Wagner, 1997). Des écrans interactifs permettent à l'utilisateur de choisir à l'aide de menus des méthodes d'imputation pour des éléments d'information particuliers, ainsi que des interventions de correction de soldes. Le tableau 1 récapitule les modes d'imputation de valeurs individuelles d'éléments à l'aide de la notation suivante :

v = désignation d'élément pour la valeur à imputer;

v' = valeur imputée de *v*;

z_j = valeur de la variable auxiliaire j^{th} (une variable auxiliaire est une constante, une désignation d'élément autre que *v* ou une expression élément-constante liée au cas où il y a imputation de *v*);

S(f) = sommation de la désignation d'élément *f* sur un ensemble défini d'enregistrements;

$(S(f_1)/S(f_2))_I$ = rapport d'identiques entre la désignation d'élément f_1 et la désignation d'élément f_2, c'est-à-dire le rapport de deux sommes portant l'une et l'autre sur tous les enregistrements d'une cellule d'imputation liée où ni f_1 ni f_2 ne manquent et où ils satisfont à certains critères d'acceptation; un exemple en est $L \# f_1 / f_2 \# U$, où *L* et *U* sont spécifiés par l'utilisateur.

Tableau 1. Méthodes d'imputation individuelle d'éléments (Luery, 2001)

Désignation	Description	Formule
VALUE	Valeur d'une variable auxiliaire.	$v' = z_1$
SUM	Somme de variables auxiliaires.	$v' = z_1 + z_2 + \ldots + z_n$
PRODUCT	Produit de deux variables auxiliaires.	$v' = z_1 z_2$
RESIDUA	Variable auxiliaire, moins la somme d'autres variables auxiliaires.	$v' = z_1 - (z_2 + \ldots + z_n)$
ATREND	Variable auxiliaire multipliée par une tendance.	$v' = z_1(z_2 / z_3)$
MEAN	Moyenne d'une variable auxiliaire sur tous les enregistrements dans une cellule d'imputation où certains critères d'acceptation sont respectés.	$v' = \bar{z}$
RATIO	Prévision de rapport pour l'élément imputé.	$v' = (S(v) / S(z_1))_I z_1$
AUXRAT	Variable auxiliaire multipliée par un rapport d'identiques.	$v' = z_1(S(z_2) / S(z_3))_I$
SIMPREG	Variable auxiliaire multipliée par un coefficient de régression.	$v' = \beta_1 z_1$
MULTREG	Prévision en régression multiple pour l'élément imputé.	$v' = \beta_1 z_1 + \ldots + \beta_n z_n$

Lorsqu'on choisit plus d'une méthode d'imputation d'un élément, l'utilisateur spécifie une commande à donner à StEPS en application des méthodes retenues. Il peut assigner à chaque méthode une condition d'imputation qui doit être respectée pour que StEPS applique une méthode. Le tableau 2 présente les spécifications d'imputation de l'élément *ectax00* de l'enquête annuelle sur le commerce de détail (taxe perçue sur les ventes annuelles) avec les éléments suivants :

ecsal00 = ventes annuelles non pondérées sans la taxe de vente;

wcsal00 = ventes annuelles pondérées sans la taxe de vente;

etaxyn00 = indicateur de perception de taxe de vente avec les valeurs 1 pour oui et 2 pour non;

wctaxy00= élément recodé qui est égal à *wctax00* (taxe de vente annuelle pondérée) lorsque *etaxyn00*=1 et est autrement manquant;

wctaxb00 = élément recodé qui est égal à *wctax00* lorsque *etaxn00* est dans {1,2} et est autrement manquant.

Tableau 2. Spécifications d'imputation générale de l'élément *ectax00* de l'enquête annuelle sur le commerce de détail (Burton, 2000)

Condition	Méthode	Formule	Variables auxiliaires
etaxyn00=1	AUXRAT	$ecsal00*(S(wctaxy00)/S(wcsal00))_I$	z_1=*ecsal00*, z_2=*wctaxy00*, z_3=*wcsal00*
etaxyn00=.	AUXRAT	$ecsal00*(S(wctaxb00)/S(wcsal00))_I$	z_1=*ecsal00*, z_2=*wctaxb00*, z_3=*wcsal00*

Pour les enregistrements où *ectax00* est mis en imputation et *ectaxyn00*=1 (indication de perception de la taxe de vente), l'imputation de *ectax00* est fondée sur un rapport pondéré d'identiques calculé à partir d'autres enregistrements de la cellule d'imputation qui ont *ectaxyn00*=1. Pour les enregistrements où *ectax00* est mis en imputation et où *ectaxyn00* est manquant, l'imputation de *ectax00* est toutefois fondée sur un rapport pondéré d'identiques calculé à partir d'enregistrements où *ectaxyn00*=1 ou *ectaxyn00*=2.

Au nombre des interventions possibles de correction de soldes (l'addition de valeurs partielles x_i doit donner le total y), on compte les imputations simples (ZERO_SET, YSUMX, RESIDUAL et RAKE définies dans 4.1), plus les suivantes (Luery, 2001, section II) :

RAKEIMP : on soumet à la fonction RAKE toutes les valeurs partielles déjà imputées, c'est-à-dire que, si x_i est imputé, $x_i' = x_i(1+R/\Sigma_{imp}x_j)$, où $\Sigma_{imp}x_j$ est la somme des éléments antérieurement imputés.

ROUND : on divise les valeurs partielles par 1 000, puis applique la fonction RAKE (x_i / 1 000 remplace x_i dans la formule RAKE).

NSK : on fixe une variable non spécifiée par nature (Not-Specified-by-Kind ou NSK) comme égale au résidu $R= y-\Sigma x_i$ ou ajoute R à une valeur partielle spécifiée ou à la valeur partielle la plus élevée.

Le calcul de données d'imputation se fait par lots. En première étape, on crée un fichier riche contenant toutes les variables de l'enquête. Il y a trois passages dans ce fichier. Dans le premier, on met les données en imputation en exécutant des vérifications d'imputation générale et en vérifiant les soldes définis. On exerce un contrôle supplémentaire sur ce premier passage à l'aide d'indicateurs de contournement d'enregistrements et de mise d'éléments en imputation, ces deux fonctions pouvant être spécifiées dans le module d'examen et de correction d'erreurs. Dans un deuxième passage, on calcule les moyennes et les quotient d'identiques nécessaires. Comme dans le premier passage, il y a des indicateurs d'enregistrements et d'éléments que fixe l'utilisateur et qui permettent d'exercer un meilleur contrôle sur l'exclusion de données extrêmes ou suspectes de ces calculs. Enfin, dans le troisième passage, on impute les données mises en imputation, revérifie les soldes et procède à des interventions de correction de soldes. Comme aux premier et deuxième passages, il y a contrôle supplémentaire de l'opération par des indicateurs de contournement d'enregistrements et de mise d'éléments en imputation que fixe l'utilisateur (Tasky, 2000b).

5. DONNÉES DE GESTION D'ENQUÊTE DANS LES FICHIERS DE StEPS

À l'heure actuelle, 94 des enquêtes qu'effectue la direction des programmes économiques exploitent les modules de vérification et d'imputation de StEPS. Dans chaque enquête, on se sert des écrans interactifs du système pour créer des règles de vérification et d'imputation particulières, ce qui comprend des règles d'imputation simple, des tests de constatation des cas à examiner, des tests de mise de cas en imputation, des définitions de traitement complexe de soldes et des sélections de méthodes d'imputation générale d'éléments particuliers. Ces règles spécifiques aux enquêtes sont stockées dans des ensembles de données SAS. On fixe ainsi des règles sur mesure pour les éléments d'information et les relations de données qui sont propres aux diverses enquêtes. L'analyse (en dehors de StEPS) de ces fichiers de données SAS avec leurs règles par enquête peut nous renseigner sur la façon dont les enquêtes exploitent les modules de vérification et d'imputation.

Le tableau 3 recense les règles de vérification et d'imputation des 76 enquêtes où StEPS est actuellement utilisé dans une ventilation selon la périodicité de ces enquêtes et les types de règles. Il n'y a pas lieu ici d'examiner ce tableau au complet, mais il convient de noter que l'usage qui se fait actuellement de la fonction d'imputation générale vise principalement à l'imputation de données manquantes, c'est-à-dire au traitement des données mises en imputation par des tests d'éléments d'information requis par opposition à une correction des données rejetées par des tests autres. En fait, en poussant l'analyse des 173 tests de règles d'enquête pour la mise en imputation générale, on constate que ces 173 règles sont associées à cinq enquêtes seulement, d'où l'impression que, pour un grand nombre d'enquêtes où on exploite actuellement le système StEPS, un repérage

des erreurs n'est pas nécessaire, celui-ci demandant qu'au moins un des rejets vise plusieurs éléments d'information.

On se sert également des vérifications de StEPS pour constater les données à examiner dans le module d'examen et de correction d'erreurs du système. Dans ce cas, les données peuvent être modifiées en mode interactif par l'utilisateur, et ces modifications sont enregistrées dans un fichier de piste de vérification. Il s'agit d'un ensemble de données SAS que l'on peut analyser (en dehors de StEPS) pour obtenir des indications quantitatives sur les vérifications interactives. Farrar (2000) a analysé le fichier de piste de vérification de StEPS dans le cas des données de l'enquête annuelle sur le commerce de détail de 1998 pour ainsi étudier les modifications apportées par les analystes aux données sur les ventes annuelles de gros. Voici quelques-unes de ses constatations :

- Dans 8,5 % des cas, les données sur les ventes annuelles ont fait l'objet d'une vérification interactive.
- Dans une proportion de 65 %, il s'agissait de modifications cumulatives de plus de 100 millions de dollars (en valeur positive ou négative) apportées par les analystes. Dans certaines catégories de la classification type des industries, l'effet de ces vérifications sur les chiffres définitifs publiés était considérable, ce qui indique que les vérifications manuelles des analystes portent sur les erreurs les plus importantes et les plus significatives.

Tableau 3. Dénombrement des règles de vérification et d'imputation

	Ensemble des enquêtes	Enquêtes sur la production industrielle			Enquêtes annuelles sur le secteur des services	Autres enquêtes
		Annuelles	Trimes-trielles	Men-suelles		
Nombre d'enquêtes	76	42	12	9	8	5
Nombre d'éléments des questionnaires	15472	10047	3006	766	682	971
Nombre de règles de vérification et d'imputation :						
Toutes catégories	56 820	38 959	12 081	3 168	1 954	658
Imputation simple :						
Libre	210	14	0	0	115	81
Soldes	112	1	0	4	91	16
Tests :						
pour examen	19 337	12 760	4 703	988	335	551
pour imputation générale	13 737	9 489	2 845	728	672	3
Règles d'imputation générale :						
Soldes	1 129	716	322	24	67	0
Modes d'imputation d'éléments	22 366	15 979	4 212	1 424	744	7
Ventilation du nombre de tests en vérification :						
Pour examen :						
Toutes catégories	19 337	12 760	4 703	988	335	551
Tests d'éléments requis	3	0	0	0	1	2
Tests de soldes	2 393	1 543	724	49	42	35
Tests de règles d'enquête	16 935	11 217	3 979	939	287	513
Autres types	6	0	0	0	5	1
Pour imputation générale						
Toutes catégories	13 737	9 489	2 845	728	672	3
Tests d'éléments requis	13 564	9 489	2 844	728	500	3
Tests de règles d'enquête	173	0	1	0	172	0
Ventilation du nombre de méthodes d'imputation générale pour imputation d'éléments						
Toutes les méthodes	22 366	15 979	4 212	1 424	744	7
RATIO	8 288	6 151	1 477	660	0	0
AUXRAT	2 675	1 748	491	36	398	2
VALUE	11 002	7 921	2 226	708	144	3
ATREND	390	159	18	20	192	1
SIMPREG, SUM, PRODUCT	11	0	0	0	10	1

6. RÉTROACTION DES UTILISATEURS SUR LES FONCTIONS DE VÉRIFICATION ET D'IMPUTATION DE StEPS

Dans les enquêtes de la direction des programmes économiques où on utilise actuellement le système StEPS, on se servait auparavant d'autres systèmes qui étaient très différents de StEPS : ils étaient spécifiques aux diverses enquêtes, ils exigeaient souvent des changements de code informatique si on voulait ajouter ou retrancher des éléments d'information et, souvent, seuls des programmeurs étaient capables de modifier les paramètres et de mettre les travaux en production. Les différences d'exploitation entre ces premiers systèmes et le système StEPS ont imposé des changements de paradigme aux gestionnaires, aux méthodologistes et aux

analystes. À la conférence méthodologique d'octobre 2000 de Statistique Canada et du Census Bureau des États-Unis, des utilisateurs de StEPS ont discuté en table ronde de ce qu'ils aimaient et de ce qu'ils n'aimaient pas des modules de vérification et d'imputation du système (Burton et Hanks, 2000). Voici les caractéristiques qu'ils prisaient :

- StEPS est un répertoire des méthodes statistiques reconnues.
- Il est facile d'y choisir des méthodes d'imputation générale.
- Les méthodes d'imputation et de traitement de soldes sont bien documentées.
- Les problèmes d'exécution sont faciles à résoudre pour ceux qui connaissent SAS.
- Les traitements par lots sont faciles à comprendre pour ceux qui connaissent SAS.
- Les tests de règles d'enquête sont des plus souples.
- Les divers types de vérifications peuvent s'exécuter séparément.
- Il peut y avoir vérification pour une seule observation ou un sous-ensemble d'observations.
- On peut regrouper dans un fichier les rejets à la vérification et ensuite les examiner.
- Des écrans de spécification permettent aux analystes de gérer le traitement des données d'enquête.
- Les analystes peuvent commander leurs propres travaux.
- On peut vérifier individuellement les éléments de la fonction d'imputation.
- Les résultats de StEPS sont semblables à ceux des systèmes antérieurs.

Voici des caractéristiques que les utilisateurs en table ronde ne prisaient pas dans les modules de vérification et d'imputation de StEPS :

- La durée de certains passages en lot est assez considérable, bien qu'on ait réussi à réduire les temps d'exécution.
- La pente est raide dans la courbe d'apprentissage pour les utilisateurs qui ne connaissent pas SAS.
- L'interaction entre les parties de StEPS et avec les besoins des enquêtes peut être complexe. Parfois, on a à écrire des programmes en SAS pour résoudre de subtils problèmes d'exécution.
- Il faut un grand nombre de paramètres dont la création et la mise à jour peuvent se révéler difficiles.
- Il a fallu spécifier des définitions de test de soldes dans les modules tant de vérification que d'imputation.
- Les méthodes d'imputation par lesquelles on calcule des rapports d'identiques ne sont pas d'une application facile pour les analystes.

Quelqu'un a mentionné en table ronde que les caractéristiques suivantes des modules de vérification et d'imputation de StEPS étaient à la fois positives et négatives :

- La souplesse de StEPS procure des moyens à l'utilisateur, mais il s'agit d'une utilisation intensive pour les usagers.
- Les fichiers et la syntaxe de SAS font de StEPS un système très souple pour ceux qui connaissent SAS, mais plus difficile d'utilisation pour ceux qui ne le connaissent pas.
- Des indicateurs détaillés fixés par l'utilisateur encadrent les fonctions de vérification et d'imputation de StEPS et, par conséquent, ces opérations peuvent être complexes.

7. CONCLUSIONS ET ACTIVITÉS FUTURES

StEPS est un système souple de vérification et d'imputation avec ses capacités de remplissage de zones, de vérification interactive (tests en vérification, plus fonction interactive de correction et de revérification en ligne), de vérification statistique et d'imputation machine. Les modules de vérification et d'imputation du système sont configurés selon les enquêtes à l'aide d'écrans interactifs permettant à l'utilisateur de définir des règles de vérification et d'imputation et de commander ses propres travaux en vue d'une évaluation des règles qu'il a spécifiées. StEPS stocke des données, des modifications de données et des paramètres de traitement dans des ensembles SAS, ce qui permet aux praticiens des enquêtes qui connaissent le SAS de créer des données quantitatives de gestion d'enquête. Les utilisateurs des modules de vérification et d'imputation de StEPS aiment la souplesse et les moyens que leur donne ce système, mais au nombre des inconvénients de systèmes généralisés, ils comptent les temps plus longs d'exécution par lots, la complexité accrue des relations

entre les différentes activités de traitement et la courbe d'apprentissage en pente raide que l'on doit parcourir pour apprendre à bien configurer le système en fonction des diverses situations d'enquête.

La direction des programmes économiques prévoit multiplier les enquêtes où StEPS est utilisé. Entre autres améliorations prévues des modules de vérification et d'imputation du système, on prévoit l'ajout de l'imputation * hot-deck + (ou la méthode par plus proche voisin), améliorer la vérification en mode graphique et ajouter des capacités de macro-examen de résultats totalisés (traitement en tableaux).

BIBLIOGRAPHIE

Ahmed, S. et Tasky, D. (1999), AThe Standard Economic Processing System: A Generalized Integrated System for Survey Processing, A *Proceedings of the Section on Government Statistics and Section on Social Statistics*, American Statistical Association, pp. 205-210.

_____ (2000), AStandardized Economic Processing System,@ *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp 633-642.

_____ (2001), AAre Generalized Systems the Way of the Furure: A Case Study on the Standard Economic Processing System (StEPS),@ *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, à paraître.

Burton, J. (2001). AGeneral Imputation Memo for StEPS: Supplement 8 (ARTS),@ document non publié, Washington DC: U.S. Census Bureau, Services Sector Statistics Division.

Burton, J. et R. Hanks (2000). Panel discussion, Session 8: Applications of Generalized Processing Systems, 2000 Statistics Canada/Census Bureau Methodological Interchange, Washington DC: U.S. Census Bureau, Methodology and Standards Directorate.

Deming, W.E. (1943). Statistical Adjustment of Data, New York: Wiley.

Farrar, R. (2000). AThe StEPS Audit Trail File as a Survey Management Tool@ ,rapport non publié, Washington DC: U.S. Census Bureau, Economic Planning and Coordination Division.

Giles, P. et C. Patrick (1986). AMéthodes d'imputation dans un système généralisé@, Tehcniques d'enquête, v. 12, pp. 53-65

Hidiroglou, M. et J. Berthelot (1986). AContrôle statistique et imputation dans les enquêtes-entreprises périodiques@, Techniques d'enquête, v 12, pp. 79-89.

Hunt, J; J. Johnson, et C. King (1999). ADetecting Outliers in the Monthly Retail Trade Survey Using the Hidiroglou-Bethelot Method, *Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 539-543.

King, C. et Kornbau, M. (1994), *Inventory of Economic Area Statistical Practice, Phase 2: Editing, Imputation, Estimation, and Variance Estimation*, Technical Report #ESMD-9401, Washington DC: Bureau of the Census, mars 1994.

Luery, D. (1999). ASimple Imputation for One Dimensional Balance Complexes,@ StEPS Decision Document #10, Washington D.C.: Bureau of the Census, Economic Statistical Methods and Programming Division.

_____ (2001). AGeneral Imputation,@ documentation non publiée, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

Luery, D. et R. Sigman (2000). ARaking When the Details are Positive and Negative,@ unpublished documentation, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

Monahan, J. (1994a). AFarm and Ranch Survey Data Processing Concepts,@ internal documentation, Washington, D.C.: U.S. Bureau of the Census, 24 octobre 1994.

_____ (1994b). AThe FRIS Processing System,@ document interne , Washington, D.C: U.S. Bureau of the Census, 20 novembre 1994.

Sigman, R. and D. Wagner (1997), AAlgorithms for Adjusting Survey Data That Fail Balance Edits,@*Proceedings of the Survey Research Methods Section*, Alexandria VA: American Statistical Association, pp. 576-581.

Tasky, D. (2000a). AStEPS Edit / Imputation Seminar - 2,@ documentation non publiée, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

_____ (2000b). AFlow of General Imputation,@ documentation non publiée, Washington DC: U.S. Census Bureau, Economic Statistical Methods and Programming Division.

Tasky, D.; Linonis, A.; Ankers, S; Hallam, D., Altmayer, L.; et Chew, D. (1999). AGet in Step with StEPS: Standard Economic Processing System,@ *Proceedings of the North East SAS Users Group*, pp. 167-178.

Willimack, D.; A.E. Anderson, et K.J. Thompson (2000). AUsing Focus Groups to Identify Analysts' Editing Strategies in an Economic Survey,@ *Proceedings of the International Conference on Establishment Surveys*, Alexandria, VA: American Statistical Association, pp. 1660-1665.

REMERCIEMENTS

L'auteur désire remercier Shirin Ahmed, Carol King, Don Luery, Deb Tasky et Jenny Thompson de leurs observations sur des versions antérieures du présent document.