

# LA VÉRIFICATION ET L'IMPUTATION DES DONNÉES DE L'ENQUÊTE UNIFIÉE SUR LES ENTREPRISES

Marie-Claude Duval et Julie Bernier<sup>1</sup>

## RÉSUMÉ

Le système de vérification et d'imputation pour l'enquête du programme unifié des statistiques sur les entreprises (PUSE) permet la vérification et l'imputation de données quantitatives. Le système doit produire un fichier de micro-données complet de sorte que les données soient cohérentes au niveau de l'enregistrement. L'enregistrement est défini ici par l'unité de collecte (UC), représentant un établissement ou un groupe d'établissements, pour laquelle les données d'enquêtes sont recueillies. Le système généralisé de vérification et d'imputation produit par Statistique Canada, le SGVI, est utilisé. Faute de données historiques, l'imputation par donneur a été préconisée; spécifiquement la méthode du plus proche voisin. Les variables importantes et communes aux sept enquêtes du PUSE, telles le revenu total, les dépenses totales, les salaires et le nombre d'employés, ainsi que des données fiscales servent à choisir le plus proche voisin. Ce dernier est sélectionné à l'intérieur d'un groupe d'unités défini généralement par l'activité industrielle et par région géographique. Certaines contraintes ont dû être prises en considération lors de la mise en oeuvre du système et sont élaborées pour expliquer la stratégie utilisée pour la vérification et l'imputation des données.

MOTS-CLÉS : Imputation par donneur; imputation massive; donneur; receveur.

## ABSTRACT

The edit and imputation system for the Unified Enterprise Statistics Program (UESP) survey allows the edit and imputation of quantitative data. The system must produce a complete microdata file such that the data are coherent at the record level. The record is defined here by the collection entity (CE), representing an establishment or a group of establishments from which the survey data was gathered. The Generalized Edit and Imputation System (GEIS) produced by Statistics Canada is used. Due to the lack of historical data, donor imputation was recommended; specifically, the nearest neighbour method. The mandatory variables common to the seven pilot surveys of the UESP (total revenue, total expenses, salary and wages, and the number of employees), as well as tax data will be used to choose the nearest neighbour. It is selected within a donor group defined generally by industrial activity and by geographical region. Certain constraints have to be considered at the time of the implementation of the system. The impact of these constraints on the strategy used for the edit and imputation of data will be explained.

KEY WORDS : Donor imputation; Massive imputation; Donor; recipient.

## 1. INTRODUCTION

Dans le cadre de l'enquête du programme unifié des statistiques sur les entreprises (PUSE), un système automatisé de vérification et d'imputation (V&I) a été implanté. Ce système consiste, à partir du fichier de la collecte des données, à repérer et corriger les erreurs d'une part, et d'autre part, à remplacer les données non disponibles par des valeurs plausibles (Kalton & Kasprzyk, 1986). Une vérification préliminaire a été effectuée à la collecte et, au besoin, tous les suivis auprès des répondants ont pris place. Ainsi, un grand

nombre de problèmes ont été corrigés à la collecte et seuls les cas non résolus ont à être traités dans le système de V&I.

Dans le cadre du PUSE, quatre principaux objectifs devaient être pris en considération lors de l'élaboration du système de V&I, soient :

- i) La création d'un fichier de micro données complet ;
- ii) La cohérence et la validation des données au niveau de l'enregistrement ;

<sup>1</sup> Marie-Claude Duval, Julie Bernier, Division des Méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6

- iii) L'uniformité de la méthode entre les industries ;
- iv) Modifier le moins possible les valeurs rapportées par les répondants (Fellegi & Holt, 1976).

Pour atteindre les objectifs i) et ii), la vérification et l'imputation sont faites au niveau micro i.e. sur toutes les unités, une à une. Pour atteindre l'objectif iii), les mêmes types de règles de vérification et la même méthode d'imputation sont utilisées pour les sept industries. Pour atteindre l'objectif iv), un système minimisant le nombre de champs à imputer a été utilisé.

L'unité utilisée pour la vérification et l'imputation (V&I) est l'unité de collecte pour laquelle les données d'enquête sont recueillies. Généralement, l'unité de collecte correspond à l'établissement échantillonné à moins que l'unité répondante ait choisi de déclarer l'information pour plus d'un établissement sur le même questionnaire. L'unité de collecte est ici utilisée dans le système de V&I puisque c'est la seule information connue à l'étape de la V&I. L'étape de l'allocation au niveau de l'établissement a lieu après le processus de vérification et d'imputation.

Il existe, en général, un questionnaire pour chaque industrie de l'enquête pilote. Un système est mis sur pied pour chaque type de questionnaire. Cependant, pour l'industrie de la construction et l'industrie de la restauration, il existe deux types de questionnaires : les questionnaires plus détaillés (les questionnaires longs) et les questionnaires moins détaillés (les questionnaires courts). Les questionnaires longs sont envoyés aux établissements ayant un revenu de 150,000\$ et plus, et les questionnaires courts aux établissements de moins de 150,000\$. Un fichier de micro-données contenant toute l'information demandée sur les questionnaires longs doit être produit pour l'estimation. Le système de vérification et d'imputation est donc responsable de transformer les questionnaires courts en questionnaires longs.

## 2. SYSTÈME UTILISÉ

Le système choisi pour la vérification et l'imputation est le système généralisé de vérification et

d'imputation (SGVI), implanté à Statistique Canada (Cotton, 1993, Kovar, MacMillan & Whitridge, 1991).

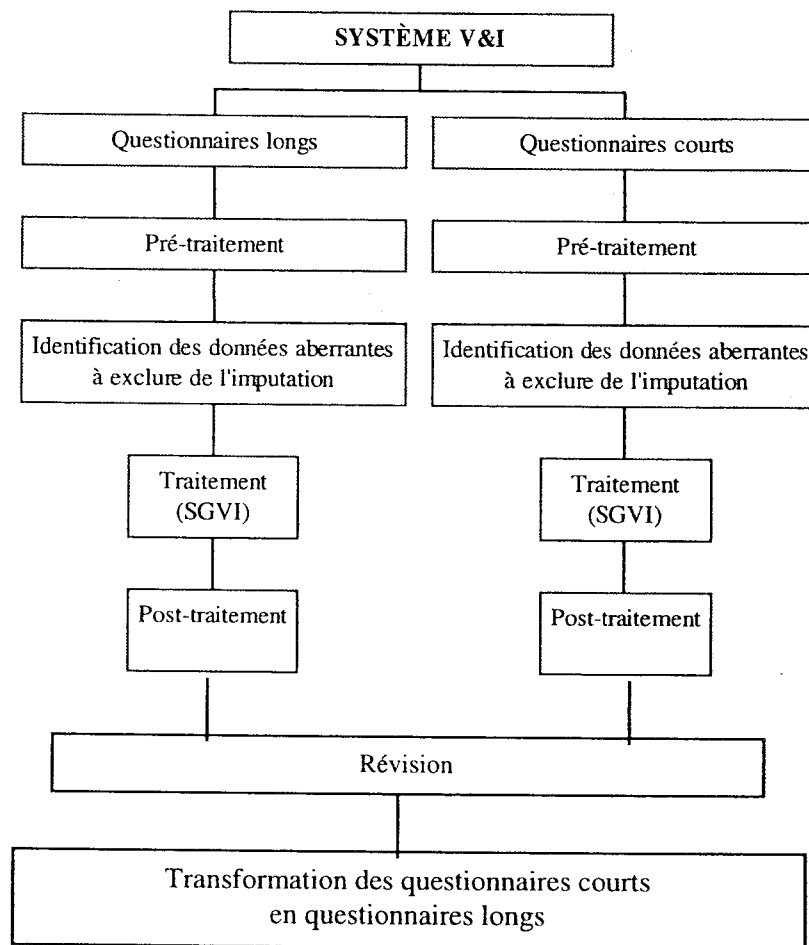
Les avantages de ce système sont :

- i) Il est déjà implanté, donc le temps requis pour rendre fonctionnel le système est plus court ;
- ii) La stratégie employée par le SGVI consiste à minimiser le nombre de champs devant faire l'objet d'une imputation, répondant ainsi à un des objectifs ;
- iii) Il permet l'imputation par donneur ;
- iv) Il travaille sous forme de groupes de règles de sorte que toutes les règles à l'intérieur d'un groupe doivent être satisfaites en même temps et non pas de façon séquentielle ;
- v) Il travaille avec des groupes d'unités de sorte qu'une unité pourra être imputée par des donneurs à l'intérieur d'un groupe défini par l'utilisateur.

Par contre, lorsqu'en cas d'erreur le SGVI doit choisir des champs à imputer, il ne permet pas de conserver prioritairement les données confirmées par le répondant lors de la collecte. Seules des règles de vérification essentielles et de base ont donc été utilisées dans le système pour minimiser l'imputation des données confirmées par le répondant. De plus, il n'est pas recommandé d'utiliser dans le SGVI plus de quarante variables à l'intérieur d'un groupe de règles. Or, les questionnaires des industries du PUSE contiennent plus d'une centaine de variables. Il a donc fallu diviser le questionnaire en parties pour faire l'imputation. Finalement, un traitement du fichier de données doit être fait avant et après l'utilisation du SGVI.

## 3. STRATÉGIE

La stratégie utilisée consiste tout d'abord à faire la vérification et l'imputation pour chaque type de questionnaire et ensuite, à transformer les questionnaires courts en questionnaires longs de sorte qu'un fichier de micro-données contenant toute l'information demandée sur les questionnaires longs soit produit. La stratégie peut être résumé par le diagramme suivant :



#### 4. PRÉ-TRAITEMENT

La pré-traitement consiste à traiter le fichier de collecte avant l'utilisation du SGVI. C'est à cette étape que certaines données manquantes sont remplacées par 0 lorsque cela s'applique. De plus, on identifie à cette étape les unités échantillonnées hors du champ de l'enquête (unités inactives, en double, plus en affaire, de mauvaise classification industrielle ou avec changement d'opérateur) qui seront exclues du système de V&I. Ces unités existent dû aux erreurs se trouvant sur la base de sondage ou de changements survenus depuis la dernière mise-à-jour de la base. On définit également à cette étape, le statut du questionnaire i.e. les questionnaires complétés, partiellement complétés et non-complétés, étant donné le traitement différent des questionnaires complétés et partiels vs les questionnaires non-complétés dans le SGVI.

#### 5. IDENTIFICATION DES DONNÉES ABERRANTES

La deuxième étape consiste à déterminer les unités aberrantes devant être exclues du groupe de donneurs lors de l'imputation. Les unités aberrantes sont les unités dont au moins une variable clé est éloignée des autres unités à l'intérieur du groupe de donneurs. Ces unités ne sont pas de bons donneurs car elles ont des caractéristiques trop différentes des autres unités à l'intérieur du groupe. La méthode utilisée est basée sur un calcul de distance entre les unités. Toutes les unités à partir duquel la distance, calculées sur les unités ordonnées, est supérieure à la médiane plus trois fois l'écart-type sont définies aberrantes (Nobrega, 1998). Les variables clés sont les variables de totaux communes aux sept enquêtes soient le revenu total, les dépenses totales, les salaires et le nombre d'employés.

## 6. TRAITEMENT (SGVI)

La traitement consiste à vérifier et imputer les données à l'aide du SGVI. La **vérification des données** n'a été basée que sur deux types de règles soient :

La règle arithmétique :  
 $W+X+Y=Z$

La règle de cohérence :  
Si  $X>0$  alors  $Y>0$

Aucune règle sur les valeurs ou ratios acceptables ( $a<X<b$ ,  $a<X/Y<b$ ) n'a été utilisée dans le système étant donné le peu de connaissance des industries nouvellement enquêtées et donc le peu de connaissance sur leurs valeurs limites et acceptables. Par contre, ces types de règles ont été mis sur pied à la collecte et en cas d'échec, le répondant peut changer les valeurs ou encore les confirmer. En omettant ce type de règles dans le système de V&I, on évite d'imputer des valeurs confirmées par le répondant, ce qui rejoint l'objectif de conserver autant que possible les valeurs rapportées par le répondant. De plus, en ne modifiant pas les caractéristiques distinctes d'une unité répondante, on évite de fausser les résultats lors d'analyses.

Aucune règle comparant les données actuelles aux données historiques n'a été utilisée étant donné qu'ils s'agit d'industries nouvellement enquêtées et donc qu'aucune donnée historique n'est disponible.

La **méthode d'imputation** choisie est l'imputation par le plus proche voisin. Cette méthode a été préférée aux autres méthodes d'imputation pour les raisons suivantes:

- i) Il n'y a pas de donnée historique disponible pour imputer ;
- ii) Les données fiscales sont utilisées pour modéliser les données d'enquêtes lors de l'estimation. L'utilisation des données fiscales pour imputer les données d'enquêtes aurait pour effet de biaiser les modèles, ce qui n'est pas souhaitable ;
- iii) L'imputation par donneur conserve mieux la relation entre les variables que l'imputation par estimateur (imputation par la moyenne, par ratio, ...).
- iv) Pour la non-réponse totale, l'imputation par donneur a été préférée à la repondération pour être en mesure de fournir un fichier de micro-

données complet de l'échantillon initial aux utilisateurs. La repondération pourrait être envisagé dans le futur.

Les variables d'enquête importantes et communes aux sept industries du PUSE, telles le revenu total, les dépenses totales, les salaires et le nombre d'employés servent à choisir le plus proche voisin. Ce dernier est sélectionné à l'intérieur d'un groupe d'unités défini généralement par l'activité industrielle et par région géographique. En l'absence de données d'enquêtes, telle pour la non-réponse totale, le revenu total, les dépenses totales et les salaires déclarés dans les dossiers fiscaux sont utilisés pour trouver le plus proche voisin.

La stratégie à cette étape a été principalement de traiter le questionnaire section par section (exemples : section des revenus, section des dépenses) étant donné qu'il n'est pas recommandé d'utiliser plus de quarante variables par groupe de règles dans le SGVI. Le désavantage de cette méthode est que différents donneurs peuvent être utilisés d'une section à l'autre, ce qui n'est pas souhaitable si l'on veut garder autant que possible la relation entre les variables de différentes sections. Pour diminuer l'impact de ce problème, un groupe de règles initial a été défini permettant l'imputation des variables clés de différentes sections, soient le revenu total, les dépenses totales, les salaires et le nombre d'employés. Par la suite, la vérification et l'imputation des items par section sont effectuées. Ce problème ne s'applique pas pour la non-réponse totale. Une façon de faire dans le SGVI lors d'imputation massive est d'imputer un numéro de questionnaire au lieu de la centaine de variables du questionnaire, assurant ainsi le même donneur pour l'ensemble du questionnaire. Par la suite, dans un traitement à posteriori, il suffit de remplacer les données manquantes par les valeurs du donneur.

Certaines contraintes ont été spécifiées pour assurer une bonne qualité de l'imputation effectuée. L'imputation aura lieu si le nombre et le pourcentage de donneurs à l'intérieur d'un groupe sont suffisants (dans notre cas, au moins 30% de donneurs et au moins 30 donneurs) et si au moins un enregistrement donneur est satisfaisant. Un enregistrement donneur sera satisfaisant s'il permet à l'enregistrement receveur de satisfaire les règles de vérification post-imputation (dans notre cas, on exige que la somme des composantes soit différente du total d'au plus 10% une fois les champs appropriés imputés). Si certains enregistrements n'ont pu être imputés, de nouveaux groupes sont formés en agrégeant les groupes initiaux de même activité industrielle pour imputer ces enregistrements. Si certains enregistrements n'ont pu

être imputés avec les nouveaux groupes, de nouvelles règles post-imputation moins strictes sont définies (dans notre cas, la marge d'erreur passe de 10 à 30%). Enfin, les enregistrements restants non-imputés après toutes ces étapes seront imputés manuellement par les experts.

## 7. POST-TRAITEMENT

Le post-traitement consiste à traiter le fichier de données imputées de sorte qu'il soit utilisable pour l'estimation. Le post-traitement consiste principalement à ajuster au prorata les données imputées de sorte que la somme des composantes soit égale au total. La différence entre le total et la somme des composantes provient de la marge d'erreur acceptée dans les règles post-imputation à l'étape du traitement.

## 8. RÉVISION

La cinquième étape est l'étape de la révision, une fois la vérification et l'imputation des données terminées. Il s'agit d'un système informatisé permettant de visualiser les données sous format questionnaire et permettant de modifier les données au besoin. À cette étape, les experts devront effectuer manuellement l'imputation si le système SGVI n'a pas été en mesure de le faire. L'imputation à l'intérieur du système SGVI pourrait ne pas avoir lieu si le nombre de donneurs disponibles à l'intérieur d'un groupe de données n'est pas suffisant, si aucun donneur n'est satisfaisant pour un enregistrement ou encore, si le SGVI n'a pas réussi à trouver de solution dans les délais permis. Également, les experts pourront visualiser et corriger les cas critiques i.e. les unités de grands impacts qui, soient ont été imputées, soient ont des relations aberrantes entre les variables. Les relations étudiées sont les ratios revenu-dépenses, salaires-dépenses et le salaire moyen par employé. Ce système contient une multitude d'information telle les champs imputés, les champs confirmés par le répondant, l'identification du donneur utilisé, les données provenant de la collecte et les unités critiques.

## 9. IMPUTATION MASSIVE DES QUESTIONNAIRES COURTS

Finalement, la dernière étape consiste à transformer les questionnaires courts en questionnaires longs. Trois méthodes ont été envisagées :

- i) L'imputation par donneur des questionnaires courts à partir des questionnaires longs. Cette

solution n'a pas été retenue puisque les questionnaires longs ne sont pas de bons donneurs pour les questionnaires courts. En effet, les questionnaires longs s'adressent aux établissements de plus grande taille alors que les questionnaires courts, aux établissements de plus petite taille.

- ii) La modélisation des données du questionnaires longs et l'application des modèles aux questionnaires courts. Cette solution n'a pas été retenue puisqu'il semble que la relation entre les variables est différente pour les établissements de petite et grande taille ;
- iii) L'utilisation des données fiscales. Cette solution a été retenue pour les variables disponibles dans les dossiers fiscaux.

Les données fiscales disponibles serviront à transformer les questionnaires courts en questionnaires longs, une fois que chaque type de questionnaires ait été imputé séparément. Un des problèmes est que certaines variables du questionnaire long ne se retrouvent pas dans les données fiscales. L'information ne sera donc pas disponible dans ces cas et cette contrainte devra être prise en considération lors de l'estimation. Un autre problème est que l'utilisation des données fiscales pour imputer les données d'enquête pourrait biaiser les modèles utilisant ces deux sources lors de l'estimation, et par conséquent introduire un biais dans les estimés. Par contre, on suppose ce biais minime puisque l'imputation par les données fiscales est appliquée uniquement sur des unités de petite taille.

## 10. CONCLUSION

La stratégie générale du système de vérification et d'imputation est basée sur un système informatisé dont les différentes étapes représentent un tout lors de la production. Il s'agit donc d'un système où toutes les alternatives ont été pensées et programmées pour pouvoir être exécutées automatiquement au besoin. Cette stratégie est importante étant donné les délais très courts alloués pour la vérification et l'imputation des unités de toutes les industries. Ce système est également basé sur une approche unifiée pour toutes les industries permettant d'atteindre un des objectifs de base. Finalement, le système de révision des données après imputation va permettre d'accroître la qualité des données.

## RÉFÉRENCES

- Nobrega Karla (1998), Outlier Detection in Asymmetric Samples : A Comparison of an Interquartile Range Method and a Variation of a Sigma Gap Method, *1998 Proceedings of the Survey Methods Section, SSC*.
- Cathy Cotton (1993), Description des fonctions du système généralisé de vérification et d'imputation, *Division des méthodes d'enquêtes entreprises, Statistique Canada*.
- J.G Kovar, J.H. MacMillan et Patricia Whitridge (1991), Système généralisé de vérification et d'imputation – Aperçu et stratégie, *Direction de la méthodologie, Statistique Canada*.
- Graham Kalton et Daniel Kasprzyk (1986), Le traitement des données manquantes, *Techniques d'enquête, juin 1986, Vol 12, n. 1*, pp. 1-17.
- Fellegi I.P. et Holt D. (1976), A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association* 71, pp. 17-35.