

LE SYSTÈME GÉNÉRALISÉ DE VÉRIFICATION ET D'IMPUTATION ET SES RÉCENTS DÉVELOPPEMENTS

Claude Poirier¹

RÉSUMÉ

Le Système généralisé de vérification et d'imputation (SGVI) a été développé à Statistique Canada au milieu des années 80 pour satisfaire les besoins en imputation des enquêtes économiques canadiennes. Étant entièrement fondé sur des règles de vérification linéaires, le SGVI traite des données numériques qui se doivent d'être continues et non négatives. Des techniques de programmation linéaire sont utilisées pour localiser les champs à être imputés, et des algorithmes de recherche sont utilisés pour traiter l'imputation de façon automatique. Parmi ses méthodes d'imputation, on retrouve l'imputation déterministe, l'approche par donneur, et l'imputation par estimateur. Des développements récents ont été initiés pour rendre le système plus pratique. Les modifications désirées ont été identifiées à partir d'expériences passées avec plus de 30 enquêtes économiques. Les modules sont en remaniement pour pouvoir être utilisés indépendamment les uns des autres, et le système sera modifié pour interagir avec des ensembles de données SAS en plus des bases de données Oracle. Ce document décrit les fonctions actuelles, les changements prévus et la nouvelle architecture qui procureront aux utilisateurs un produit plus flexible.

MOTS CLÉS : Vérification; imputation; restructuration; système généralisé.

ABSTRACT

The Generalized Edit and Imputation System (GEIS) was developed at Statistics Canada in the mid 1980's to meet the imputation requirements of the Canadian economic surveys. Being entirely driven by linear edit rules, GEIS processes numeric data that are assumed to be continuous and non-negative. Linear programming techniques are used to localize fields to be imputed, and search algorithms are used to perform automatic imputations. Among its imputation methods, there are the deterministic imputation, the donor approach, and the imputation by estimators. Recent developments were initiated to make the system more practical. The desired modifications were identified based on past experiences with up to 30 economic surveys. The modules are being redesigned to be used independently from each other, and the system will be modified to interact with SAS datasets in addition to Oracle databases. This paper describes the current functionality, the planned changes and the new architecture which will provide the users with a more flexible product.

KEY WORDS: Edit; Imputation; reengineering; generalized system.

1. LES FONCTIONS DU SGVI

Un système typique de vérification est un outil qui permet de détecter des incohérences dans des données, qui extrait les enregistrements en erreur en indiquant quels principes ou règles ont mené à l'erreur, qui permet à des experts du sujet de modifier manuellement les données à l'aide d'informations auxiliaires, puis qui peut vérifier si des erreurs sont encore présentes. Le tout se fait de façon itérative jusqu'à ce qu'il n'y ait plus d'erreur ou jusqu'à ce qu'on accepte qu'un processus automatique fasse les derniers correctifs. Idéalement, l'ensemble du processus vise des données de qualité acceptable avec un coût ou un délai de production raisonnable. La version actuelle du Système généralisé de vérification et d'imputation, le SGVI-6.5.0, n'offre pas vraiment de

fonction typique de vérification. Il constitue plutôt un produit automatisé qui identifie, pour chaque enregistrement, les variables qui devraient être modifiées puis qui les impute pour obtenir des données qui respectent des relations spécifiques. Il est habituellement utilisé après qu'un processus préliminaire de vérification ait été complété dans les phases de collecte, de saisie et de restructuration de données. Des techniques de programmation linéaire sont utilisées pour identifier les variables qui devraient être imputées et des algorithmes de recherche sont utilisés pour imputer automatiquement les données. Le processus est entièrement basé sur des règles linéaires de vérification et d'imputation.

Le SGVI est habituellement exécuté par étape, sa

¹ Claude Poirier, Statistique Canada, 120 ave. Parkdale, Ottawa, Canada, K1A 0T6, poircla@statcan.ca

structure ayant été pensée pour faciliter cette approche. Ses étapes peuvent être catégorisées en deux grandes fonctions : L'identification des erreurs, et leur imputation. Comme il est mentionné dans le document de Statistique Canada (1999), l'identification des erreurs inclut un module servant à spécifier des règles de vérification tout en assurant qu'elles soient cohérentes entre elles, un module de détection de données aberrantes, et un module qui permet de localiser les valeurs qui devront être imputées. De l'autre côté, l'imputation inclut une méthode déterministe, une méthode basée sur des modèles, une utilisant une approche par donneur, et une dernière basée sur le prorata.

1.1 Spécification et analyse des règles

La première étape, celle de spécification et d'analyse des règles, permet d'exprimer les critères et les relations entre les variables qui caractérisent les enregistrements acceptables. Les relations sont exprimées par un ensemble de n règles linéaires de vérification sous la forme

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1m}x_m &\leq c_1 \\ \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nm}x_m &\leq c_n \end{aligned}$$

où les a_{ij} et les c_i sont des constantes pré-définies par l'utilisateur, et où les x_j représentent m variables d'enquêtes. Les règles sont liées logiquement par des "et" ce qui signifie que chacune d'elles doit obligatoirement être respectée afin qu'un enregistrement soit qualifié d'acceptable. Les règles spécifiées par l'utilisateur sont analysées par le système pour assurer qu'elles soient cohérentes entre elles, et qu'elles ne cachent pas de redondances ni d'égalités. Les itérations de l'étape de spécification et d'analyse permettent à l'utilisateur d'identifier le meilleur ensemble de règles pour son application. Si elles ne sont pas spécifiées par l'utilisateur, le système ajoutera des règles de positivité pour les m variables d'enquêtes

$$\begin{aligned} x_1 &\geq 0 \\ \vdots & \\ x_m &\geq 0 \end{aligned}$$

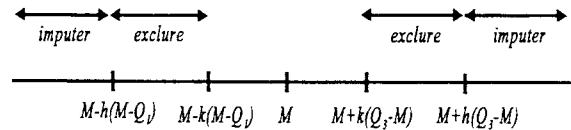
L'ajout de ses règles est essentiel parce qu'elles font partie des hypothèses de base des algorithmes de programmation linéaire. Bien que toutes les relations entre les variables doivent être exprimées de façon linéaire, quelques règles non-linéaires peuvent aussi être satisfaites à l'aide de transformations usuelles. C'est le

cas des règles conditionnelles du type "si $x_1 > 0$ alors $x_2 > 0$ " qui sont presque équivalentes à la forme linéaire " $\epsilon x_1 - x_2 \leq 0$ ", avec $\epsilon = 0,0000001$.

1.2 Détection des données aberrantes

La deuxième étape vise la détection des données aberrantes univariées en utilisant la méthode décrite par Hiridoglou et Berthelot (1986). La méthode identifie les observations aberrantes en les comparant à la médiane M et aux quartiles Q_1 et Q_3 , extraits de l'ensemble des enregistrements. Une valeur x sera déclarée aberrante si elle est hors de l'intervalle d'acceptation $[M-k(M-Q_1), M+k(Q_3-M)]$, où k est défini par l'utilisateur. Comme le montre la figure 1, cette méthode peut être utilisée pour identifier les valeurs qui devront être imputées ou qui devront être exclues des calculs subséquents.

Figure 1. Traitement des données aberrantes



1.3 Identification des valeurs à imputer

La troisième étape consiste à localiser les erreurs de façon à minimiser le nombre de valeurs à imputer. Ceci correspond au principe du changement minimal proposé par Fellegi et Holt (1976). L'idée est d'identifier les variables qui devront être imputées de façon à ce que l'enregistrement puisse respecter toutes les règles de vérification. Dans le SGVI, le problème d'optimisation est traité avec des algorithmes de programmation linéaire. Il s'exprime par des contraintes linéaires traitées par l'algorithme de Chernikova, comme l'ont détaillé Schioppa-Kratina et Kovar (1989). Le système permet à l'utilisateur d'utiliser des poids pour chacune des variables lorsqu'il désire exercer une influence sur l'identification des variables à imputer. Bien que l'algorithme soit coûteux à exécuter, il constitue une des principales qualités du SGVI.

1.4 Imputation des données

L'imputation des données constitue l'étape finale. Elle offre trois techniques bien connues : L'imputation déterministe, l'imputation par donneur, et l'imputation par estimateur, ou modèle.

Imputation déterministe

En se basant sur les règles de vérification, l'imputation déterministe identifie les cas pour lesquels il n'existe qu'une seule solution possible qui permette à l'enregistrement en erreur de satisfaire les règles. Par exemple, avec $x_1=10$, $x_2=?$, $x_3=?$, $x_4=20$ et $x_5=30$, et avec les règles de vérification

$$\begin{aligned}x_1 + x_2 - x_3 &\leq 0 \\x_3 + x_4 - x_5 &\leq 0\end{aligned}$$

on peut déduire que

$$10 + x_2 \leq x_3 \leq 10$$

Étant donné les règles de positivité, le SGVI détectera que $x_2=0$ et $x_3=10$ sont les seules valeurs possibles. Il les utilisera donc dans l'imputation déterministe.

Imputation par donneur

L'imputation par donneur remplace les valeurs de l'enregistrement en erreur en utilisant les valeurs de l'enregistrement valide qui lui ressemble le plus. Cette approche est aussi appelée imputation par plus proche voisin. Pour un enregistrement donné, un sous-ensemble de variables n'ayant pas besoin d'imputation sont automatiquement utilisées comme variables d'appariement. L'utilisateur peut aussi spécifier ses propres variables d'appariement qui s'ajouteront à celles identifiées par le système. Une différence normalisée est utilisée comme mesure de distance entre les donneurs potentiels et l'enregistrement à corriger. Plus de détails sont disponibles dans Statistique Canada (1999).

Par défaut, l'ensemble de donneurs inclut les enregistrements qui satisfont toutes les règles de vérification, mais l'utilisateur peut le réduire avec des

contraintes spécifiques. Il peut aussi spécifier des règles de post-imputation qui permettront d'assurer que le *plus proche voisin* s'apparente assez bien à l'enregistrement en erreur pour servir de source d'imputation. La recherche du donneur se fait par l'intermédiaire d'un arbre K-D (Statistique Canada, 1999) qui ajoute beaucoup aux performances du système. Après que le donneur soit identifié, ses valeurs sont transférées intégralement à l'enregistrement à imputer pour en corriger les erreurs.

Imputation par estimateur

L'imputation par estimateur, aussi appelée par modèle, fournit une grande variété de techniques utilisant autant des données historiques que des données actuelles. Des estimateurs pré-définis sont disponibles dans le SGVI : Valeurs précédentes, moyennes précédentes ou actuelles, tendances, ratios et régressions multiples. Si un estimateur non usuel est requis, alors le système permet à l'utilisateur de spécifier son propre modèle en utilisant les opérateurs de base : +, -, ×, ÷, exp. Un terme d'erreur avec une variance spécifique peut aussi être introduit dans ces modèles. Le tableau 1 donne quelques exemples d'estimateurs programmés dans le système.

Le SGVI permet d'utiliser différentes techniques d'imputation d'une section à l'autre d'un questionnaire, de même que d'une sous-population à l'autre. Une approche séquentielle où, à chaque étape, l'utilisateur inclut ou exclut des données imputées dans les étapes précédentes est aussi possible. À titre d'exemple, il peut être intéressant d'imputer par modèle ce qui n'a pu être imputé avec donneurs.

Tableau 1. Exemples d'estimateurs disponibles dans le SGVI

Estimateurs	Exemples
Historique	$\hat{y}_i = y_{i,Hist}$
Somme	$\hat{y}_i = u_i + v_i$
Moyenne	$\hat{y}_i = \bar{y}$
Ratio	$\hat{y}_i = (\bar{y}/\bar{x}) x_i$
Tendance	$\hat{y}_i = (\bar{y}/\bar{y}_{Hist}) y_{i,Hist}$
Régression	$\hat{y}_i = \beta_0 + \beta_1 u_i + \beta_2 v_i + \dots + \varepsilon_i$
Défini par l'utilisateur	$\hat{y}_i = (u_i + v_i) x_i / 1000$

2. LES FORCES ET LES FAIBLESSES

Du point de vue technique, le SGVI fonctionne sur tout ordinateur principal de type IBM de même que sur des plates-formes UNIX. Il a été développé en Langage C et interagit actuellement avec des bases de données Oracle. Il inclut une interface qui aide les usagers à spécifier leurs règles de vérification et leurs paramètres, mais cette interface ne peut être utilisée en période de production. Les fonctions décrites dans la section précédente sont adéquates pour la plupart des enquêtes économiques de Statistique Canada.

Le SGVI peut traiter les données par groupes, i.e. par sous-ensembles de variables reliées entre-elles, correspondant habituellement à des sections de questionnaire. Il peut aussi les traiter par classes d'imputation représentées par des sous-populations. Une des grandes forces du système est sa capacité à identifier les changements minimaux pour tout ensemble de règles définies avec des équations linéaires. Cette caractéristique accroît les chances de préserver une intégrité des données malgré les erreurs qui s'y trouvent. La fonction automatisée d'imputation par donneur basée sur les règles d'imputation est une autre force du système. Cette fonction s'exécute presque sans intervention de la part de l'utilisateur étant donné qu'elle identifie ses variables d'appariement elle-même, selon la qualité de chaque enregistrement qu'elle traite. Elle utilise simplement le patron de réponse, quel qu'il soit, pour chercher un donneur. La flexibilité du module d'imputation par donneur, les nombreux rapports de diagnostics, et un module d'initiation en direct, le tout joint au soutien permanent aux usagers constituent les autres aspects positifs du système.

La complexité d'Oracle en fait un outil assez lourd à utiliser et à maintenir du point de vue de l'utilisateur. De plus, les responsables d'enquêtes qui conçoivent leur propre système de vérification aimeraient avoir un accès direct aux fonctions d'imputation du SGVI. Malheureusement, le module actuel d'imputation ne peut fonctionner indépendamment de son module de vérification.

Étant donné que le SGVI ne traite que des données numériques, il ne peut être utilisé pour les enquêtes sociales typiques qui collectent plutôt des variables catégoriques. De plus, la fonction de vérification ne peut pas traiter des variables ayant des valeurs négatives. En pratique, si on prend l'exemple d'une enquête financière, la contrainte de positivité exige une transformation des données. Ainsi, des systèmes de pré-

traitement doivent être conçus.

3. LES DÉVELOPPEMENTS EN COURS

Comme l'explique Poirier (1999) dans son évaluation des systèmes d'imputation, il existe bien des qualités souhaitables à un système de traitement mais les objectifs et les priorités du développement font que, en pratique, elles ne peuvent toutes être présentes dans un même système. Chaque équipe de développement a ses propres clients à satisfaire, avec des techniques, des structures de données et des plates-formes particulières à soutenir. Dans le cas du SGVI, le but premier reste de fournir un outil flexible et convivial pour le traitement de données numériques.

Pour atteindre ce but, des priorités avaient été établies au milieu des années 80 alors que le système était en développement initial. On visait alors un outil centralisé qui fournirait des fonctions parmi les plus difficiles à développer soit, l'identification des changements minimaux et l'imputation par donneur. Le choix d'une base de données devait se faire en considérant les directions futures du monde des technologies. C'est ainsi que Oracle et C étaient choisis pour développer le système. Au cours des années qui ont suivi, alors que près de 30 enquêtes majeures utilisaient le SGVI, les modifications apportées au système visaient les besoins les plus pressants. À la fin des années 90, il s'était développé une expertise suffisante avec l'utilisation du SGVI pour identifier avec confiance la direction qu'on devait prendre. Une enquête auprès des usagers a aidé à bien identifier les priorités. L'expérience avec d'autres systèmes comme StEPS (Ahmed et Tasky, 2000), Solas (Statistical Solutions, 1997), NIM (Bankier et al., 2000), Plain Vanilla (Wagner, 2000), CherryPi (Van de Pol et al., 1997) et AGGIES (Perritt, 2000) a aussi été considérée dans la planification.

L'ajout de quelques fonctions était certes désiré mais on notait un besoin plus grand pour améliorer la flexibilité du système. En 1998, un projet majeur était alors initié pour développer une technique d'imputation par prorata, pour rendre les fonctions d'imputation directement accessibles, pour traiter les ensembles de données SAS en plus des bases de données Oracle, et pour briser le système en composantes pouvant être utilisées par d'autres applications. Le tout allait fournir indirectement une fonction d'imputation massive, technique qui ne nécessite pas de localisation d'erreurs.

3.1 Imputation par prorata

L'imputation par prorata sert à résoudre une inégalité entre une somme de parties et un total donné. Elle est particulièrement utile pour ajuster les valeurs provenant d'un donneur afin qu'elles satisfassent un total connu pour l'enregistrement en erreur. La méthode consiste à appliquer un facteur d'ajustement k aux parties de façon à ce que leur somme égale le total connu, i.e.

$$k x_1 + k x_2 + \dots + k x_m = y$$

Ici, le facteur de prorata k peut être dérivé et appliqué à toutes les variables x_i incluses dans la somme, ou seulement à celles ayant déjà été imputées au préalable. Cette dernière option est nécessaire pour corriger les inégalités résiduelles issues de l'imputation par donneur. Un contrôle sur la magnitude des transformations sera offert à l'utilisateur. Ainsi, il pourra spécifier des bornes minimale et maximale, k_{min} et k_{max} , pour le facteur de prorata.

3.2 Accès direct aux fonctions d'imputation

L'idée d'accéder directement les fonctions d'imputation vient des responsables d'enquêtes qui développent leur propre système de vérification et qui désirent utiliser le SGVI seulement pour ses fonctions d'imputation. Ceci exige la création de modules indépendants. La grande difficulté ici est de normaliser l'entrée des paramètres pour ces fonctions étant donné que les intrants proviendront directement de l'utilisateur, et non du module de vérification du SGVI.

Le développement a débuté avec les fonctions d'imputation par donneur et d'imputation par prorata qui devraient être disponibles dans leur forme "indépendante" dès l'automne 2000. L'imputation par estimateur telle que décrite à la section 1.4 suivra ensuite.

3.3 Oracle vs SAS

Depuis plusieurs années, il y a une utilisation grandissante de SAS à Statistique Canada. Ceci est appuyé par une directive récente de la haute gestion visant à accroître le développement de systèmes en SAS, étant donné le budget important dédié à une licence corporative. De plus, les divisions de la méthodologie, d'où vient la grande majorité des usagers du SGVI, font de SAS leur outil de travail quotidien pour concevoir et appliquer les méthodes d'enquêtes. Depuis plusieurs années déjà, on note un intérêt marqué pour une version du SGVI qui traiterait les fichiers SAS.

Le but du projet n'est pas de réécrire en SAS les 35000

lignes de la version actuelle en C, mais plutôt de modifier les étapes d'entrée et de sortie de données, et d'utiliser les capacités de SAS pour produire des procédures à partir de modules écrits en C. Notons que la conformité à Oracle sera préservée étant donné les ressources majeures déjà dépensées par les usagers pour mettre en place des applications actuelles.

Le projet d'analyse et de développement en SAS a débuté avec le module de détection de données aberrantes décrit à la section 2. Maintenant que l'expertise avec les outils SAS est acquise, les travaux additionnels de restructuration sont prometteurs. Les fonctions de donneurs et de prorata devraient être disponibles dans leur version SAS dès l'an prochain.

3.4 Les algorithmes de base

Parallèlement aux travaux de conversion à SAS, l'équipe de développement s'attarde présentement à améliorer l'architecture du SGVI. Ainsi, les algorithmes de base gagneraient à être accessibles directement, au même titre que les fonctions d'imputation. La restructuration du SGVI vise donc l'extraction d'algorithmes tels que l'algorithme de Chernikova, la construction d'arbres K-D, l'algorithme de Hidioglou et Berthelot, la dérivation des règles implicites, etc. Dans un même temps, la restructuration permettra le soutien de la plate-forme Windows en plus de UNIX et de l'ordinateur principal.

4. CONCLUSION

La version actuelle du SGVI inclut les fonctions de base requises par la plupart des enquêtes-entreprises : Localisation des erreurs, détection de données aberrantes, imputation déterministe, par donneur et par estimateur. Les dix années d'utilisation du système ont permis de développer une expertise avec ses fonctions et d'en identifier les faiblesses. Suite à l'analyse des besoins, une restructuration du système s'imposait. Les travaux actuels s'orientent plutôt sur les aspects techniques que sur la méthodologie, malgré que les activités de développement visent une nouvelle fonction d'imputation, le prorata. Les priorités touchent la création d'algorithmes et de modules indépendants, de même qu'une utilisation dans l'environnement SAS.

RÉFÉRENCES

- Ahmed, S.A., et Tasky, D.L. (2000). "An Overview of the Standard Economic Processing System (StEPS)". International Conference on Establishment Surveys - II, Buffalo, États-unis.

- Bankier, M., Poirier, P., Lachance, M. et Mason P. (2000). "A Generic Implementation of the Nearest-Neighbour Imputation Methodology". Conférence 2000 de la Société statistique du Canada, Ottawa, Canada.
- Fellegi, I.P. et Holt, D. (1976). "A Systematic Approach to Automatic Edit and Imputation". *Journal of the American Statistical Association*, 71, 17-35.
- Hidioglou, M.A. et Berthelot, J.-M. (1986). "Contrôle statistique et imputation dans les enquêtes-entreprises périodiques". *Techniques d'enquête*, 12, 79-89.
- Perritt, K. (2000). "A Look into AGGIES, An Automated Edit and Imputation System". Conférence 2000 de la Société statistique du Canada, Ottawa, Canada.
- Poirier, C. (1999). "A Functional Evaluation of Edit and Imputation Tools". Atelier sur l'édition de données statistiques, Commission économique pour l'Europe, Nations unies, Italie (Rome).
- Schiopu-Kratina, I. et Kovar, J.G. (1989). "Use of Chernikova's algorithm in the Generalized Edit and Imputation System". Document de travail de la Direction de méthodologie, No. BSMD-89-001E, Statistique Canada.
- Statistical Solutions (1997). "Solus For Missing Data Analysis 1.0: User Reference". Cork, Ireland, Statistical Solutions Inc.
- Statistique Canada (1999). "Description des fonctions du système généralisé de vérification et d'imputation - V6.5.0". Rapport technique de Statistique Canada.
- Van de Pol, F., Buijs, A., van der Horst, G., et de Waal, T. (1997). "Towards Integrated Business Survey Processing". *Nouvelles directions pour les enquêtes et recensements, Recueil du Symposium international 1997*, Statistique Canada.
- Wagner, D. (2000). "Economic Census General Editing - Plain Vanilla". International Conference on ishment Surveys - II, Buffalo, États-unis.