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

TECHNIQUES D'OPTIMISATION POUR LA VALIDATION DES RÈGLES DE VÉRIFICATION ET POUR L'IMPUTATION DES DONNÉES

Renato Bruni[1], Alessandra Reale[2], Renato Torelli[2]

RÉSUMÉ

Le présent article traite de la détection et de la correction automatique des données incohérentes ou en dehors des limites permises dans un processus général de collecte de données statistiques. La méthode proposée s'applique aussi bien aux données qualitatives que quantitatives. Notre objectif est de surmonter les contraintes de calcul de la méthode de Fellegi-Holt, tout en retenant ses aspects positifs. Comme à l'accoutumée, les enregistrements de données doivent satisfaire un ensemble de règles afin d'être déclarés corrects. Grâce au codage de règles sous forme d'inégalités linéaires, nous créons des modèles mathématiques pour les problèmes étudiés. Le premier point pertinent est que l'ensemble de règles proprement dit est vérifié en vue de déceler les incohérences ou les redondances par résolution d'une série de problèmes de *faisabilité*. Le deuxième point pertinent est que l'imputation est réalisée par résolution d'une série de problèmes de *couverture d'ensemble*.

MOTS CLÉS : Imputation de données, modèle mathématique, optimisation.

1. INTRODUCTION

1.1 Description

Lorsqu'on traite une grande quantité de renseignements recueillis, le problème pertinent bien connu qui se pose est celui de procéder aux élaborations requises en ne tenant compte que des données correctes. Les erreurs ou, plus précisément, les incohérences entre réponses ou les réponses tombant en dehors des limites permises peuvent être observées au moment du dépouillement des questionnaires ou à une étape ultérieure du traitement des données. Le présent article traite du problème de la détection et de la correction automatique des incohérences ou des données tombant en dehors des limites permises dans un processus général de collecte des données statistiques. Nous nous concentrerons sur le problème du traitement automatisé de données démographiques hiérarchiques dans le cas d'un recensement de population.

Comme il est coutume de le faire pour des renseignements structurés, les données ont été classées dans des *enregistrements*. Un enregistrement à la structure formelle d'un ensemble de *zones*. En attribuant une *valeur* à chaque zone, nous obtenons un cas d'enregistrement ou, simplement, un enregistrement (Ramakrishnan, Gehrke, 2000). En général on résout le problème de la *détection des erreurs* en formulant un ensemble de règles que les enregistrements doivent respecter afin d'être déclarés *corrects*. Les enregistrements qui ne respectent pas ces règles sont déclarés *incorrects.* Les règles sont souvent rédigées sous forme de *règles de vérification* qui expriment les conditions de l'erreur. Néanmoins, elles peuvent être facilement converties d'une représentation à l'autre. Afin de simplifier notre exposé, nous considérons ici que les règles sont satisfaites par les questionnaires corrects. Étant donné un questionnaire erroné, le problème de la *détection des erreurs* est habituellement résolu en modifiant certaines valeurs et en obtenant ainsi un *questionnaire*

1 Université de Rome "La Sapienza", DIS, Via M. Buonarroti 12, Rome, Italie, 00185.

2 Istat, DISS, Via A. Ravà 150, Rome, Italie, 00100.

corrigé qui répond aux règles susmentionnées et qui s'approche autant que possible du *questionnaire original* (inconnu), c'est-à-dire celui que nous aurions si aucune erreur n'avait été commise.

De nombreux systèmes logiciels traitent le problème de la correction des questionnaires. Notre but est de surmonter les contraintes de calcul (Poirier, 1999; Winkler, 1999) de la méthode de Fellegi-Holt (1976) tout en retenant ses aspects positifs et en profitant de certaines caractéristiques de la nouvelle méthode d'imputation (NIM pour New imputation method) (Bankier, 1999). Nous construisons un modèle mathématique du problème et appliquons les méthodes d'optimisation les plus pointues élaborées par les spécialistes de la recherche opérationnelle.

2. ENCODAGE DES RÈGLES DANS DES INÉGALITÉS LINÉAIRES

2.1 Enregistrements de données

Dans le cas d'un recensement, chaque enregistrement contient les réponses données sur un questionnaire par une famille entière. Une famille comprend un ensemble d'individus $I = \{1,\ldots, l\}$. Nous considérons généralement le même ensemble de zones $F = \{f_1,\ldots, f_m\}$ pour chaque individu. Si nous considérons les zones susmentionnées pour chaque individu, nous obtenons le genre de structure d'enregistrement qui suit, que nous appellerons aussi *structure du questionnaire* Q.

$$Q = \{f^1_1,\ldots, f^1_m, \ldots, f^l_1,\ldots, f^l_m\}$$

Un *cas de questionnaire* q ou, simplement, un questionnaire est donc :

$$q = \{v^1_1,\ldots, v^1_m, \ldots, v^l_1,\ldots, v^l_m\}$$

Chaque zone f^i_j, avec $i = 1,\ldots, l$, $j = 1,\ldots, m$, possède un *domaine* D^i_j, qui est l'ensemble de toutes les valeurs possible pour cette zone. Puisque nous avons affaire à des erreurs, le domaine comprend toutes les valeurs que l'on peut trouver sur les questionnaires, même celles qui sont incorrectes. Les zones sont habituellement catégorisées comme étant quantitatives ou qualitatives.

2.2 Règles

Un cas de questionnaire q est déclaré correct si, et uniquement si, il respecte l'ensemble de règles $R = \{r_1,\ldots, r_p\}$. Chaque règle est habituellement exprimée comme une disjonction ($\vee$) de conditions également appelées propositions (p_t). Les propositions peuvent aussi être inversées ($\neg p_t$). Par conséquent, les règles ont la structure de clauses (c.-à-d. une disjonction de propositions éventuellement inversées).

$$(p_1 \vee \ldots \vee p_u \vee \neg p_{u+1} \vee \ldots \vee \neg p_v)$$

Puisque toutes les règles doivent être respectées, une conjonction ($\wedge$) de propositions est simplement exprimée au moyen d'un ensemble de règles différentes, chacune faite d'une seule proposition. Toutes les autres relations logiques entre les propositions (implication, etc.) peuvent être exprimées en utilisant uniquement les opérateurs susmentionnés ($\vee$, $\wedge$, $\neg$). Nous appelons ici une proposition ayant trait aux valeurs d'une zone unique une *proposition logique*. Nous appelons en outre ici une proposition comportant des opérations mathématiques entre les valeurs des zones une *proposition mathématique*. Une proposition logique est, par exemple, (*âge* < 14), voire même (*état matrimonial* = *marié*). Une proposition mathématique est, par exemple, (*âge – nombre d'années de mariage* $\geq$ 14). Nous appelons *règles logiques* les règles exprimées uniquement au moyen de propositions logiques, *règles mathématiques* celles exprimées uniquement au moyen de propositions mathématiques et *règles logiques-mathématiques* celles exprimées en se servant des deux catégories de propositions. Un cas spécial de règles logiques est celui des règles qui délimitent le *domaine de faisabilité* $\check{D}^i_j \subseteq D^i_j$ de chaque zone. En fait, très souvent, certaines

valeurs du domaine sont inacceptables, quelles que soient les valeurs de toutes les autres zones. Ces valeurs sont appelées valeurs *hors des limites.*

Un exemple de règle logique exprimant que toutes les personnes qui déclarent être mariées devraient avoir au moins 14 ans est : ¬(*état matrimonial* = *mariée*) ∨ ¬(*âge* < 14). Les règles définissant le domaine de faisabilité de la zone *âge* sont par exemple : (*âge* ≥ 0), (*âge* ≤ 110).

2.3 Divisions des domaines en sous-ensembles

Nous disons que deux valeurs v'^{i}_{j} et v''^{i}_{j} sont *équivalentes* du point de vue des règles lorsque, pour chaque paire de questionnaires $q' = \{v_1, \ldots, v'^{i}_{j}, \ldots, v_l\}$ et $q'' = \{v_1, \ldots, v''^{i}_{j}, \ldots, v_l\}$, pour lesquels toutes les valeurs sont identiques sauf celles de la zone f^{i}_{j}, q' et q'' sont soit corrects soit incorrects.

Un point important est que nous pouvons toujours diviser chaque domaine D^{i}_{j} en n_j sous-ensembles

$$D^{i}_{j} = S^{i}_{j\,1} \cup \ldots \cup S^{i}_{j\,nj}$$

de façon telle que toutes les valeurs appartenant aux mêmes $S^{i}_{j\,k}$ soient équivalentes du point de vue des règles logiques. Il existe systématiquement un sous-ensemble de valeurs *hors des limites.* De surcroît, la valeur de certaines zones peut manquer. Ce genre de valeur est qualifiée de *blanc* et, selon la zone, peut ou non faire partie du domaine de faisabilité. Si la réponse en blanc fait partie du domaine de faisabilité, le sous-ensemble en *blanc* est également présent. Sinon, il fait partie du sous-ensemble *hors des limites*.

Par exemple, pour le domaine composé de nombres entiers $D^{i}_{âge}$, les valeurs inférieures à 0 ou supérieures à 110 sont *hors des limites*. La réponse en blanc ne fait pas partie du domaine de faisabilité, donc appartient au sous-ensemble *hors des limites*. Nous obtenons les sous-ensembles suivants :

$$S^{i}_{âge\,1} = \{0,\ldots, 13\},\ S^{i}_{âge\,2} = \{14,\ldots,17\},\ S^{i}_{âge\,3} = \{18,\ldots,25\},$$
$$S^{i}_{âge\,4} = \{26,\ldots,110\},\ S^{i}_{âge\,5} = \{\ldots, -1\} \cup \{111, \ldots\} \cup \{\textit{blanc}\}.$$

2.3 Variables

Nous définissons ici les variables de notre modèle. Nous avons un ensemble de *l*×m variables entières $z^{i}_{j} \in \{0,\ldots,U\}$, une pour chaque domaine D^{i}_{j}, un ensemble de $l(n_1 + \ldots + n_m)$ variables binaires $x^{i}_{j\,k} \in \{0,1\}$, une pour chaque sous-ensemble $S^{i}_{j\,k}$, et un ensemble de $l(n_1 + \ldots + n_m)$ variables binaires $\bar{x}^{i}_{j\,k} \in \{0,1\}$, qui sont les compléments des $x^{i}_{j\,k}$.

Nous représentons la valeur v^{i}_{j} du questionnaire par la variable entière z^{i}_{j} en définissant une correspondance entre les valeurs du domaine et les nombres entiers compris entre 0 et une valeur supérieure U. Cette valeur U est telle qu'aucun élément de tout domaine de faisabilité ne concorde avec elle.

$$\psi: D^{i}_{j} \rightarrow \{0, \ldots, U\}$$
$$v^{i}_{j} \rightarrow z^{i}_{j}$$

Les domaines qualitatifs sont également mis en concordance avec l'ensemble de nombres entiers en choisissant un ordonnancement. Toutes les valeurs *hors des limites* concordent avec le plus grand nombre utilisé U. Si elle fait partie du domaine de faisabilité, la valeur *en blanc* est codée au moyen de la valeur entière qui suit immédiatement la plus grande valeur entière du domaine de faisabilité, mais qui est plus petite que U.

Nous codons l'appartenance d'une valeur à un sous-ensemble en utilisant la variable binaire $x^{i}_{j\,k}$.

$$x^{i}_{j\,k} = \begin{cases} 1 & \text{si } v^{i}_{j} \in S^{i}_{j\,k} \\ 0 & \text{si } v^{i}_{j} \notin S^{i}_{j\,k} \end{cases}$$

Enfin, les variables binaires complémentaires sont liées aux premières au moyen de ce que nous appelons des *contraintes de couplage*.

$$x^{i}_{j\,k} + \overline{x}^{i}_{j\,k} = 1$$

Nous lions les variables entières et binaires au moyen d'un ensemble d'inégalités linéaires appelées *contraintes de pont*. Elles imposent que, lorsque z^{i}_{j} a une valeur telle que $v^{i}_{j} \in S^{i}_{j\,k}$, la valeur de la variable $x^{i}_{j\,k}$ correspondante soit 1 et celle de toutes les autres variables binaires $\{x^{i}_{j\,1}, \ldots, x^{i}_{j\,k-1}, x^{i}_{j\,k+1}, \ldots, x^{i}_{j\,nj}\}$ soit 0.

2.4 Encodage des règles

Les propositions logiques sont exprimées au moyen de variables binaires et les propositions mathématiques, au moyen de variables entières. Les règles ayant trait à plus d'un individu sont exprimées au moyen des variables pertinentes pour les divers individus. Chaque règle logique ayant une structure de clause

$$(p_1 \vee \ldots \vee p_u \vee \neg p_{u+1} \vee \ldots \vee \neg p_v)$$

peut être convertie à l'inégalité linéaire qui suit. Nous représentons par $x(p_t)$ et $\overline{x}(p_t)$ la variable logique et la variable logique complémentaire correspondant à p_t, respectivement et, en sélectionnant l'ensemble $\{p_1,\ldots, p_u\}$ de propositions positives et l'ensemble $\{p_{u+1},\ldots, p_v\}$ de propositions négatives, nous définissons les vecteurs d'incidence correspondants a^{π} et a^{ν}.

$$\sum_{u=1,..,n} [a^{\pi}_{u}\, x(p_t) + a^{\nu}_{u}\, \overline{x}(p_t)] \geq 1$$

La seule différence due à l'utilisation de propositions mathématiques est que celles-ci ne correspondent pas à des variables binaires mais à des opérations entre variables entières. Nous limitons les règles mathématiques à celles qui sont linéaires ou linéarisables, car il existe des méthodes de résolution plus rapide pour ces dernières. En particulier, nous acceptons des règles composées par division ou par multiplication de deux variables. Pour une discussion des inégalités linéarisables, consulter, par exemple, Williams, (1993). Parfois, nous devons introduire d'autres variables binaires, par exemple pour encoder des disjonctions de propositions mathématiques. Notons, en outre, que nous développons une syntaxe très précise pour les règles. Par conséquent, l'encodage pourrait être réalisé par une procédure automatique.

À titre d'exemple, considérons la règle logique qui suit.

$$\neg(\textit{état matrimonial} = \textit{marié}) \vee \neg(\textit{âge} < 14)$$

En remplaçant les variables logiques et entières, nous obtenons $\overline{x}^{i}_{\text{état matrimonial\_marié}} \vee \overline{x}^{i}_{\text{âge }\{0,\ldots,13\}}$. Cette formule devient l'inégalité linéaire suivante :

$$\overline{x}^{i}_{\text{état matrimonial\_marié}} + \overline{x}^{i}_{\text{âge }\{0,\ldots,13\}} \geq 1$$

À titre d'autre exemple, considérons la règle logique-mathématique suivante :

$$\neg(\textit{état matrimonial} = \textit{marié}) \vee (\textit{âge} - \textit{années de mariage} \geq 14)$$

En remplaçant les variables logiques et entières, nous obtenons $\overline{x}^{i}_{\text{état\_matrimonial marié}} \vee (z^{i}_{\text{âge}} - z^{i}_{\text{années de mariage}} \geq 14)$. Cette formule devient l'inégalité suivante :

$$U\,\overline{x}^{i}_{\text{état\_matrimonial marié}} + z^{i}_{\text{âge}} - z^{i}_{\text{années de mariage}} \geq 14$$

En tout, à partir de l'ensemble de règles, nous obtenons un ensemble d'inégalités linéaires (auquel nous ajoutons les contraintes de couplage et les contraintes de pont) et, à partir de l'ensemble de réponses à un questionnaire, nous obtenons les valeurs pour les variables introduites. Par construction, toutes les valeurs attribuées aux variables tirées de questionnaires corrects, et uniquement ces valeurs, satisfont toutes les inégalités linéaires, donc le système linéaire

$$\begin{cases} A^{\pi} x + A^{v} \overline{x} \geq 1 \\ A^{\pi} x + A^{v} \overline{x} + B z \geq b \\ x^{i}_{jk} + \overline{x}^{i}_{jk} = 1 \\ z_j \in \{0,\ldots,U\},\ x^{i}_{jk},\ \overline{x}^{i}_{jk} \in \{0,1\} \end{cases} \qquad (1)$$

où les cœfficients matriciels A^{π}, A^{v} et B (respectivement pour les variables x, $\overline{x}$, et z) et les cœfficients b sont donnés, comme plus haut, par encodage de toutes les règles en inégalités. Brièvement, un questionnaire q doit satisfaire (1) pour être correct.

3. MODÈLES MATHÉMATIQUES DES PROBLÈMES

3.1 Validation de l'ensemble de règles comme problème de faisabilité

L'ensemble de règles ne doit présenter aucune *incohérence* (autrement dit, les règles ne doivent pas se contredire) et, de préférence, aucune *redondance* (autrement dit, les règles ne doivent pas être impliquées logiquement par d'autres règles). Dans le cas de questionnaires réels, les règles peuvent être fort nombreuses, puisqu'un nombre élevé de règles permet de mieux dépister les erreurs. Afin de nous assurer que l'ensemble de règles ne comporte pas d'incohérence ni de redondance, nous étudions les solutions du système d'inégalités linaires (1). Lorsque chaque cas possible de questionnaire q est déclaré incorrect, nous obtenons une situation qualifiée d'*incohérence complète* pour l'ensemble de règles. Lorsque l'incohérence des règles semble être causée uniquement par des valeurs particulières de certaines zones, nous avons la situation (encore plus insidieuse) d'*incohérence partielle* de l'ensemble de règles. Dans le cas d'un grand ensemble de règles, ou lors d'une phase de mise à jour des règles, des incohérences peuvent se produire facilement.

Grâce à l'encodage de l'ensemble de règles en un système d'inégalités linéaires de la forme (1), l'incohérence complète n'a lieu que si, et uniquement si, (1) n'est pas faisable. Une incohérence partielle en rapport avec un sous-ensemble S^{i}_{jk} survient si, et uniquement si, le système devient infaisable lorsque l'on ajoute la contrainte $x^{i}_{jk} = 1$.

En outre, dans le cas de l'incohérence, nous souhaitons rétablir la cohérence. La méthode des règles de suppression correspondant aux inégalités que nous ne pouvons satisfaire n'est pas utile. En fait, chaque règle a sa fonction et ne peut être supprimée; elle peut uniquement être modifiée par l'utilisateur qui l'a rédigée. Par contre, la sélection d'un ensemble de règles contradictoires peut amener l'utilisateur à les modifier. Ceci revient à déterminer quelle partie du système cause l'infaisabilité, c'est-à-dire un

sous-système irréductible infaisable (IIS) (Amaldi, Pfetsch, Trotter, 1999). Par conséquent, nous utilisons un résolveur de faisabilité qui, en présence de cas infaisables, est capable de sélectionner un SII.

Certaines règles pourraient logiquement en impliquer d'autres, donc être redondantes. Il serait préférable de les supprimer, car diminuer le nombre de vérifications tout en maintenant le même pouvoir de dépistage des erreurs peut simplifier le processus complet et le rendre moins sujet aux erreurs. Le problème de l'implication logique (Loveland, 1978) peut être formulé sous forme de problème de faisabilité. Une règle r_s est impliquée par un ensemble de règles R si, et uniquement si, le système d'inégalités linéaires obtenu à partir de R, conjugué à l'inégalité linéaire obtenue par la négation logique de r_s, est infaisable. Il est possible de vérifier conséquemment si chaque règle r_s est redondante en testant la faisabilité du système obtenu à partir de l'ensemble de règles $\{(\{R\} \setminus r_s) \cup \neg r_s\}$. Nous pouvons vérifier la redondance de chaque règle en appliquant à chacune d'elle l'opération susmentionnée.

3.2 Problèmes d'imputation en tant que couverture d'ensemble

Après la phase de la *validation des règles*, nous sommes certains que le système (1) est faisable et a plus d'une solution. Le dépistage des questionnaires incorrects q_e se résume alors au problème consistant à vérifier si l'attribution des valeurs de variable correspondant à un cas de questionnaire q satisfait (1). Cette opération ne demande qu'un très petit effort de calcul.

Si un *questionnaire incorrect* q^e est décelé, le processus d'*imputation* consiste à modifier certaines des valeurs figurant sur ce questionnaire pour obtenir un *questionnaire corrigé* q^c qui satisfait le système (1) et est aussi proche que possible du *questionnaire original* q^o (inconnu), c'est-à-dire celui que l'on obtiendrait si aucune erreur n'était commise. Deux grands principes devraient être suivis durant le processus d'imputation : apporter le nombre minimal de modifications aux données erronées et modifier aussi peu que possible la distribution originale des fréquences des données (Fellegi, Holt, 1976).

En général, on donne le coût de la modification de chaque valeur de q^e en se fondant sur la fiabilité de la zone. Le questionnaire q^e correspond à une attribution de valeurs aux variables. En particulier, nous avons un ensemble de $l(n_1 + \ldots + n_m)$ valeurs binaires e^i_{jk} et un ensemble de $l \times m$ valeurs entières g^i_j.

Nous avons un coût $c^i_{jk} \in \Re_+$ pour la modification de chaque e^i_{jk}, et un coût $\hat{c}^i_j \in \Re_+$ pour la modification de chaque g^i_j.

Le questionnaire q^c que nous voulons obtenir correspond aux valeurs des variables $(x^i_{jk}, \bar{x}^i_{jk}, z^i_j)$ dans le cas de la solution optimale.

Le problème de la *localisation de l'erreur* consiste à trouver un ensemble V de zones pour lequel le coût total est minimal tel que q^c peut être obtenu à partir de q^e en modifiant (uniquement et complètement) les valeurs de V. Ceci revient à trouver le nombre minimal de modifications à apporter aux données erronées, mais respecte fort peu les distributions originales des fréquences.

Un *questionnaire donneur* q^d est un questionnaire correct qui devrait être aussi proche que possible de q^o. Le questionnaire q^d correspond à une attribution de valeurs aux variables. Plus précisément, nous avons un ensemble de valeurs binaires d^i_{jk} et un ensemble de valeurs entières f^i_j. Les donneurs sont sélectionnés conformément à une fonction de distance qui peut être complètement spécifiée par l'utilisateur.

$$\delta: (q^e, q^d) \rightarrow \Re_+$$

Le problème de l'*imputation par donneur* consiste à trouver un ensemble W de zones dont le coût total est minimal tel que q^c peut être obtenu à partir de q^e en copiant, à partir du donneur q^d (uniquement et entièrement), les valeurs de W. Il est généralement admis que cette opération altère peu les distributions

originales des fréquences; toutefois, les modifications apportées aux données erronées pourraient ne pas être minimales. Nous cherchons à résoudre les deux problèmes susmentionnés.

Introduisons les $\{l(n_1 + \ldots + n_m)\}$ variables binaires $y^i_{jk} \in \{0,1\}$ représentant les changements apportés à e^i_{jk}.

$$y^i_{jk} = \begin{cases} 1 & \text{si nous modifions } e^i_{jk} \\ 0 & \text{si nous gardons } e^i_{jk} \end{cases}$$

En outre, quand nous utilisons un donneur, introduisons $l \times m$ variables binaires $w^i_j \in \{0,1\}$ représentant les modifications que nous apportons à g^i_j.

$$w^i_j = \begin{cases} 1 & \text{si nous modifions } g^i_j \\ 0 & \text{si nous gardons } g^i_j \end{cases}$$

La minimisation du coût total des changements peut être exprimée par la fonction objective suivante (où les termes $\hat{c}^i_j w^i_j$ apparaissent uniquement dans le cas de l'imputation par donneur) :

$$\min \sum_{i=1,..,l} \sum_{j=1,..,m} \sum_{k=1,..,nj} c^i_{jk} y^i_{jk} + \sum_{i=1,..,l} \sum_{j=1,..,m} \hat{c}^i_j w^i_j \qquad (2)$$

Cependant, les contraintes (1) sont exprimées pour x^i_{jk}, $\bar{x}^i_{jk}$, z^i_j. Un problème important tient au fait qu'il existe un lien entre les variables qui figurent dans (1) et celles qui figurent dans (2).
En cas de localisation de l'erreur, la situation dépend des valeurs de e^i_{jk}, comme suit :

$$y^i_{jk} = \begin{cases} x^i_{jk} & \text{si } e^i_{jk} = 0 \\ 1 - x^i_{jk} & \text{si } e^i_{jk} = 1 \end{cases}$$

En fait, lorsque $e^i_{jk} = 0$, ne pas modifier ce terme revient à poser $x^i_{jk} = 0$. Puisque nous ne faisons aucune modification, $y^i_{jk} = 0$. En revanche, modifier ce terme revient à poser $x^i_{jk} = 1$. Puisque nous faisons une modification, $y^i_{jk} = 1$. Par conséquent, $y^i_{jk} = x^i_{jk}$.

Si, au contraire, $e^i_{jk} = 1$, ne pas le modifier revient à poser $x^i_{jk} = 1$. Puisque nous n'apportons aucune modification, $y^i_{jk} = 0$. En revanche, modifier le terme revient à poser $x^i_{jk} = 0$. Puisque nous apportons une modification, $y^i_{jk} = 1$. Donc, $y^i_{jk} = 1 - x^i_{jk}$.

Au moyen des résultats susmentionnés, nous pouvons réécrire la fonction objective (2).

Par conséquent, le problème de localisation de l'erreur peut être modélisé comme suit, où la fonction objective et un nombre cohérent de contraintes ont une structure de *couverture d'ensemble* (Garey, Johnson, 1976).

$$\min \sum_{i=1,..,l} \sum_{j=1,..,m} \sum_{k=1,..,nj} c^{i}_{jk} (1-e^{i}_{jk}) x^{i}_{jk} + \sum_{i=1,..,l} \sum_{j=1,..,m} \sum_{k=1,..,nj} c^{i}_{jk} e^{i}_{jk} \bar{x}^{i}_{jk}$$

subordonné à l'ensemble de contraintes (1).

De même, dans le cas de l'imputation par donneur, la relation entre x^{i}_{jk} et y^{i}_{jk} dépend des valeurs de e^{i}_{jk} et d^{i}_{jk}.

$$y^{i}_{jk} = \begin{cases} x^{i}_{jk} & \text{si } e^{i}_{jk} = 0 \text{ et } d^{i}_{jk} = 1 \\ 1 - x^{i}_{jk} & \text{si } e^{i}_{jk} = 1 \text{ et } d^{i}_{jk} = 0 \\ 0 & \text{si } e^{i}_{jk} = d^{i}_{jk} \end{cases}$$

Notons que, même si nous ne modifions pas x^{i}_{jk} de e^{i}_{jk} à d^{i}_{jk}, et, conséquemment, z^{i}_{j} de g^{i}_{j} à f^{i}_{j}, nous pourrions encore devoir modifier z^{i}_{j} de g^{i}_{j} à f^{i}_{j} pour obtenir une meilleure solution. Afin de facililter le choix des valeurs de $z^{i}_{j,}$ nous utilisons l'information fournie par les variables x^{i}_{jk}. Nous prenons pour z^{i}_{j} la valeur du donneur lorsque nous devons procéder à des modifications des x^{i}_{jk} ou que, même si les x^{i}_{jk} ne changent pas, il est plus pratique de prendre la valeur de l'enregistrement donneur.

$$z^{i}_{j} = \begin{cases} g^{i}_{j} & \text{si } \forall\, k \; y^{i}_{jk} = 0 \\ f^{i}_{j} & \text{si } \exists\, k : e^{i}_{jk} = 1, \text{ ou que } f^{i}_{j} \text{ donne une meilleure solution} \end{cases}$$

En procédant de la même façon que pour le premier cas, le problème d'imputation par donneur peut être modélisé comme suit. De nouveau, la fonction objective et un nombre cohérent de contraintes ont une structure de couverture d'ensemble.

$$\min \sum_{i=1,..,l} \sum_{j=1,..,m} \sum_{k=1,..,nj} c^{i}_{jk} (1-e^{i}_{jk}) d^{i}_{jk} x^{i}_{jk} + \sum_{i=1,..,l} \sum_{j=1,..,m} \sum_{k=1,..,nj} c^{i}_{jk} e^{i}_{jk} (1-d^{i}_{jk}) \bar{x}^{i}_{jk} + + \sum_{i=1,..,l} \sum_{j=1,..,m} \hat{c}^{i}_{j} w^{i}_{j}$$

subordonné à l'ensemble de contraintes (1) et aux contraintes supplémentaires suivantes :

$$\begin{cases} z^{i}_{j} = f^{i}_{j} (w^{i}_{j} + \sum_{k=1,..,nj} y^{i}_{jk} /2) + g^{i}_{j} (1 - w^{i}_{j} - \sum_{k=1,..,nj} y^{i}_{jk} /2) \\ w^{i}_{j} \le 1 - \sum_{k=1,..,nj} y^{i}_{jk} /2 \\ w^{i}_{j} \in \{0,1\} \end{cases}$$

4. RÉSOLUTION DES PROBLÈMES

4.1 Problèmes de faisabilité

Étant donné le système d'inégalités linéaires résultant de l'encodage de l'ensemble de règles, nous résolvons au départ la série de problèmes de faisabilité découlant de leur validation. Ce genre de problèmes est résolu par une méthode énumérative fondée sur le branchement. Ce dernier est essentiellement une stratégie visant à *diviser pour conquérir*. L'idée consiste à diviser systématiquement la région de faisabilité de la programmation linéaire en subdivisions pratiques et à évaluer le problème de programmation en nombres entiers d'après ces subdivisions. En passant de la région à l'une de ces subdivisions, nous ajoutons une contrainte qui n'est pas satisfaite par la solution optimale de la programmation linéaire sur la région

mère. Donc, les programmes linéaires correspondant aux subdivisions peuvent être résolus efficacement. Consulter, par exemple, Nemhauser, Wolsey (1988) pour une explication complète. La durée chronométrée de la résolution de chaque problème de faisabilité est systématiquement inférieure à une seconde.

4.2 Problèmes de couverture d'ensemble

Après avoir décelé les questionnaires incorrects, pour chaque cas q^e, nous résolvons le problème de localisation de l'erreur. Par conséquent, nous commençons par choisir un nombre b de questionnaires donneurs, en sélectionnant ceux qui sont les moins éloignés d'après notre fonction de distance. Jusqu'ici, pour chaque questionnaire incorrect, on résout b problèmes d'imputation par donneur et on choisit l'imputation donnant la valeur minimale pour la fonction objective.

Il est parfois possible d'améliorer les méthodes de séparation et d'évaluation progressive (branch-and-bound) par ajout de coupures afin d'élaguer certaines branches de l'arbre de recherche. Les procédures de ce genre portent le nom d'algorithme de séparation et coupures (branch-and-cut). Une *coupure* est une inégalité satisfaite par toutes les solutions faisables du programme en nombres entiers. Les nouvelles contraintes réduisent successivement la région de faisabilité jusqu'à ce que l'on obtienne une solution entière optimale. La coupure représente un hyperplan qui passe entre la solution de relâchement de la programmation linéaire et le polytope entier, et enlève une partie du polytope relâché contenant la solution optimale de la programmation linéaire en n'excluant aucun point entier faisable. L'efficacité des méthodes de séparation et coupures devient évidente lorsque le nombre de cas augmente. Pour une discussion complète, consulter Nemhauser, Wolsey (1988). La durée chronométrée de la résolution de chaque problème de couverture d'ensemble est, elle aussi, inférieure à une seconde.

5. CONCLUSIONS

Un modèle mathématique du processus complet d'imputation permet d'appliquer une procédure automatisée d'imputation des données (DIESIS, Data Imputation and Editing System - Logiciel italien). Ce genre de procédure permet de réparer les données au moyen de donneurs en s'assurant que les distributions marginales et conjointes des données soient maintenues. La série de problèmes d'optimisation en nombres entiers qui se posent peuvent être résolus par mise en œuvre pointue de procédures de séparation et coupures. Chaque problème est résolu jusqu'à l'optimalité en un temps extrêmement court (toujours inférieur à une seconde).

Le rendement statistique du nouveau système a été évalué rigoureusement et comparé à celui de la méthode canadienne d'imputation par le voisin le plus proche (NIM pour Nearest-neighbour Imputation Methodology, Bankier, 1999) au moyen d'une étude par simulation fondée sur des données réelles provenant du recensement de la population de l'Italie de 1991 (Manzari, Reale, 2001). La NIM a été choisie comme base de comparaison pour l'évaluation statistique parce qu'à l'heure actuelle, elle est considérée comme étant la meilleure méthode de traitement automatique des données démographiques hiérarchiques. Les résultats des essais sont fort encourageants.

BIBLIOGRAPHIE

Amaldi, E., M.E. Pfetsch, et L. Trotter, Jr. (1999), "Some structural and algorithmic properties of the maximum feasible subsystem problem", *Proceedings. of 10th Integer Programming and Combinatorial Optimization conference, Lecture Notes in Computer Science 1610,* Springer-Verlag, pp. 45-59.

Bankier, M. (1999), "Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extensions for future Census" *Proceedings UN/ECE Work Session on Statistical Data Editing*, Working Paper n.24, Rome, Italie.

Fellegi, P. et D. Holt (1976), "A Systematic Approach to Automatic edit and Imputation" *Journal of the American Statistical Association*, 17, pp.35-71(353).

Garey, M. R., et D.S. Johnson (1979) *Computers and Intractability: A Guide to the Theory of NP--Completeness*, San Francisco: W.H. Freeman and Company.

Loveland, D. W. (1978), *Automated Theorem Proving: a Logical Basis*, North Holland.

Manzari A., et A. Reale (2001), "Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology", *Proceedings of the 53rd Session of the International Statistical Institute*, Séoul, août 2001.

Nemhauser, G. L., et L.A. Wolsey (1988). *Integer and Combinatorial Optimization*, New York: J. Wiley.

Poirier, C. (1999), "A Functional Evaluation of Edit and Imputation Tools" *Proceedings UN/ECE Work Session on Statistical Data Editing*, Working Paper n.12, Rome, Italie.

Ramakrishnan, R., et J. Gehrke (2000), *Database Management Systems*, McGraw Hill.

Williams, H. P. (1993), *Model Building in Mathematical Programming*, Chichester: J. Wiley.

Winkler, W. E. (1999), "State of Statistical Data Editing and current Research Problems" *Proceedings UN/ECE Work Session on Stat. Data Editing*, Working Paper n.29, Rome, Italie.