Discrete Mathematics for Data Imputation*

Renato Bruni
University of Roma “La Sapienza” - D.I.S.
Via M. Buonarroti 12, Roma, Italy, 00185.
E-mail: bruni@dis.uniroma1.it

April 16, 2003

Abstract
The paper is concerned with the problem of automatic detection and correction of inconsistent or out
of range data in a general process of statistical data collecting. The proposed approach is able to deal
with both qualitative and quantitative values. As customary, erroneous data records are detected by
formulating a set of rules. Erroneous records should then be corrected, by modifying as less as possible
the erroneous data, while causing minimum perturbation to the original frequency distributions of the
data. Such process is called imputation. By encoding the rules with linear inequalities, we convert
imputation problems into integer programming problems. The proposed procedure is tested on a real-
world case of census. Results are extremely encouraging both from the computational and from the data
quality point of view.

1 Introduction
When dealing with a large amount of collected information, a well-known relevant problem arises: perform
the requested elaboration without being misled by erroneous data. Data correctness is a crucial aspect of
data quality, and, in practical cases, it has always been a very computationally demanding problem. This
paper is concerned with the problem of automatic detection and correction of inconsistent or out of range
data in a general process of statistical data collecting. Examples of data collecting are cases of statistical
investigations, marketing analysis, experimental measures, etc. Without loss of generality, our attention will
be focused on the problem of a census of population carried out by collecting questionnaires. A census is
a particularly relevant process and actually constitutes the most fundamental source of information about
a country [13]. Errors, or, more precisely, inconsistencies between answers or out of range answers, can be
due to the original compilation of the questionnaire, or introduced during any later phase of information
conversion or processing.

As customary for structured information, data are organized into records. A record has the formal
structure of a set of fields. Giving each field a value, we obtain a record instance, or, simply, a record [12].
The problem of error detection is generally approached by formulating a set of rules that the records must
respect in order to be declared correct. Records not respecting all the rules are declared erroneous. In the
field of database theory, rules are also called integrity constraints [12]. Integrity constraints are verified by
correct records, and are generally checked before inserting a record into the database. In the field of statistics,
rules are often called edits [5]. Edits express the error condition, being verified by erroneous records. In
order to simplify our exposition, we consider here rules that are verified by correct questionnaires. Clearly,
rules can easily be converted from one representation to the other.

Given an erroneous questionnaire, the problem of error correction is usually tackled by changing some of
its values, obtaining a corrected questionnaire which satisfies the above rules and is as close as possible to the
(unknown) original questionnaire (the one we would have if we had no errors). Such process is called data

---

*Work developed during the research collaboration between the Italian Statistic Office (Istat) and the University of Roma
“La Sapienza”.

1
imputation. Many software systems deal with the problem of questionnaires correction, by using a variety of different edits encoding and solution algorithm (e.g. [1, 11, 15]). A very well-known approach to the problem, which implies the generation of all the rules logically implied by the initial set of rules, is due to Fellegi and Holt [5]. In practical case, however, such methods suffer from severe computational limitations [11, 15]. Computational efficiency could sometimes be obtained only by sacrificing the data quality issue. Another serious drawback is that simultaneous processing of quantitative and qualitative fields is seldom allowed. Recently, a declarative semantics for the imputation problem has been proposed in [6], as an attempt to give an unambiguous formalization of the meaning of imputation and of the behavior of the various imputation systems. Another logic-based approach to the problem is in [7].

An automatic procedure for data imputation by using a discrete mathematical model is here presented. In an earlier paper, an imputation procedure for the case when all the rules are expressed by using propositional logic is already developed [4]. That would not suffice when dealing with rules containing also mathematical operators. The effectiveness of a discrete approach is also showed, for a similar problem, by the theory of Logical Analysis of Data ([2] among other papers).

By encoding the rules into linear inequalities, an integer programming model of the problem is proposed, as explained in Sect. 2. Note that, since a very precise syntax for writing the rules was developed, such encoding could be automatically performed. A sequence of integer programming problems is therefore solved by means of a state-of-the-art integer programming solver (ILOG Cplex), as clarified in Sect. 3. Moreover, due to the peculiar problem’s structure, the efficient use of a separation routine for set covering problems was possible [10]. The proposed procedure is tested by executing the process of error detection and correction in the case of real world census data, as shown in Sect. 4. The practical behavior of the proposed procedure is evaluated both from the computational and from the data quality point of view. The latter analysis is carried out by means of recognized statistical indicators [8]. The overall software system developed for data imputation, called DIESIS (Data Imputation Editing System - Italian Software) is also described in [3].

2 Encoding Rules into Linear Inequalities

In Database theory, a record schema \( R \) is a set of fields \( \{ f_1, \ldots, f_h \} \). A record instance \( r \) is a set of values \( \{ v_1, \ldots, v_h \} \), one for each of the above fields. In the case of a Census, each record contains the answers given in one questionnaire by an entire household. A household consists in a set of individuals \( I = \{ 1, \ldots, l \} \) living together in the same housing unit. We generally consider for every individual the same set of fields \( F = \{ f_1, \ldots, f_m \} \). Considering all such fields for all such individual, we have the following kind of record structure, that we will also call questionnaire structure \( Q \).

\[
Q = \{ f_1^1, \ldots, f_m^1, \ldots, f_1^l, \ldots, f_m^l \}
\]

A questionnaire instance \( q \), or, simply, a questionnaire, is therefore:

\[
q = \{ v_1^1, \ldots, v_m^1, \ldots, v_1^l, \ldots, v_m^l \}
\]

Example 2.1. In the case of a census, fields are for instance age or marital status, corresponding examples of values are 18 or single.

Each field \( f_j^i \), with \( i = 1 \ldots l, j = 1 \ldots m \), has a domain \( D_j^i \), which is the set of every possible value for that field. Since we are dealing with errors, the domains include all values that can be found on questionnaires, even the erroneous ones. Fields are usually distinguished in quantitative and qualitative ones. A quantitative field is a field on whose values are applied (at least some) mathematical operators (e.g. \( >, + \)), hence such operators should be defined on its domain. Examples of quantitative fields are numbers (real or integer), or even the elements of an ordered set. A qualitative field simply requires its domain to be a discrete set with finite number of elements. We are not interested here in considering fields ranging over domains having a non-finite number of non-ordered values. The proposed approach is able to deal with both qualitative and quantitative values.

\footnote{More informations available at www.cplex.com.}
Example 2.2. For the qualitative field marital status, answer can vary on a discrete set of possibilities in mutual exclusion, or, due to errors, be missing (blank) or not meaningful (n.m.).

\[ D^i_{\text{marital status}} = \{\text{single, married, separate, divorced, widow, blank, n.m.}\} \]

For the quantitative field age, due to errors, the domain is

\[ D^i_{\text{age}} = (-\infty, +\infty) \cup \{\text{blank}\} \]

A questionnaire instance \( q \) is declared correct if and only if it respects a set of rules \( R = \{r_1, \ldots, r_p\} \). Each rule can be seen as a mathematical function \( r_s \) from the Cartesian product of all the domains to the Boolean set \( \{0,1\} \).

\[ r_s : D^1_i \times \ldots \times D^1_m \times \ldots \times D^1_l \times \ldots \times D^1_m \rightarrow \{0,1\} \]

Rules are such that \( q \) is a correct questionnaire if and only if \( r_s(q) = 1 \) for all \( s = 1, \ldots, p \). Rules should be expressed according to some syntax. In our case, each rule is expressed as a disjunction (\( \lor \)) of conditions, also called propositions \( (p_v) \). Propositions can also be negated \( (\neg p_v) \). Therefore, rules have the structure of clauses (i.e. a disjunction of possibly negated propositions). By introducing, for each rule \( r_s \), the set \( \pi_s \) of the indices of the positive propositions and the set \( \nu_s \) of the indices of the negative propositions, \( r_s \) can be written as follows.

\[ \bigvee_{v \in \pi_s} p_v \lor \bigvee_{v \in \nu_s} \neg p_v \tag{1} \]

Since all rules must be respected, a conjunction (\( \land \)) of propositions is simply expressed using a set of different rules, each made of a single proposition. As known, all other logic relations between propositions (implication \( \Rightarrow \), etc.) can be expressed by using only the above operators (\( \lor, \land, \neg \)). Differently from the case of propositional logic, propositions have an internal structure. We need to distinguish between two different structures. A proposition involving values of a single field is here called a logical proposition. A proposition involving mathematical operations between values of fields is here called mathematical proposition.

Example 2.3. A logical proposition is, for instance, \( (\text{age} < 14) \), or \( (\text{marital status} = \text{married}) \). A mathematical propositions is, for instance: \( (\text{age} - \text{years married} \geq 14) \).

We call logical rules the rules expressed only with logical propositions, mathematical rules the rules expressed only with mathematical propositions, and logic-mathematical rules the rules expressed using both type of propositions.

A special case of logical rules are the ones delimitating the feasible domain \( D^1_i \subseteq D^1_i \) of every field. Very often, in fact, some values of the domain are not acceptable, regardless of values of all other fields. They are called out-of-range values. By removing the out-of-range values from the domain we have the feasible domain \( D^1_{\text{range}} \).

Example 2.4. A logical rule expressing that all people declaring to be married should be at least 14 years old is:

\[ \neg(\text{marital status} = \text{married}) \lor \neg(\text{age} < 14) \]

Rules delimiting the feasible domain for the field age are for instance:

\( (\text{age} \geq 0), (\text{age} \leq 110) \)

One can observe that, depending on the rules, some values (e.g. \( \text{age} \leq 32 \) or \( \leq 33 \)) appear to have essentially the same effect on the correctness of a questionnaire. Formally, we say that two values \( v_j^i \) and \( v_j^f \) are equivalent from the rules’ point of view when, for every couple of questionnaires \( q' = \{v_1^l, \ldots, v_j^l, \ldots, v_m^l\} \) and \( q'' = \{v_1^l, \ldots, v_j^l, \ldots, v_m^l\} \) having all values identical except for field \( f_j^i \), \( q' \) and \( q'' \) are either both correct or both erroneous:

\[ r_s(q') = r_s(q'') \quad \text{for all} \quad s = 1, \ldots, p \]
A key point is that we can always partition each domain $D_j^i$ into $n_j$ subsets

$$D_j^i = S_{j1}^i \cup \ldots \cup S_{jn_j}^i$$

in such a way that all values belonging to the same $S_{jk}^i$ are equivalent from the logical rules’ point of view (i.e., considering all and only the logical rules). A subset for the out-of-range values is always present. Moreover, the value for some field can be the missing value. Such value is described as blank, and, depending on the field, can belong or not to the feasible domain. If the blank answer belongs to the feasible domain, the subset blank is also present. Otherwise, it belongs to the out-of-range subset.

Example 2.5. Given the integer domain $D_{\text{age}}^i$, together with an hypothetic set of rules $R$ including those shown in Example 2.4, values below 0 or above 110 are out-of-range. The blank answer does not belong to the feasible domain, hence belongs to the out-of-range subset. We therefore have the following subsets.

$$S_{\text{age} \_ 1}^i = \{0, \ldots, 13\}, S_{\text{age} \_ 2}^i = \{14, \ldots, 17\},
S_{\text{age} \_ 3}^i = \{18, \ldots, 25\}, S_{\text{age} \_ 4}^i = \{26, \ldots, 110\}
S_{\text{age} \_ 5}^i = \{\ldots, 0\} \cup \{111, \ldots\} \cup \text{\{blank\}}$$

So far, the variables of our mathematical model can be defined. They are a set of $l \times m$ integer variables $z_j^i \in \{0, \ldots, U\}$, one for each domain $D_j^i$, a set of $(n_1 + \ldots + n_m)$ binary variables $x_{jk}^i \in \{0, 1\}$, one for each subset $S_{jk}^i$, and a set of $(l(n_1 + \ldots + n_m))$ binary variables $\bar{x}_{jk}^i \in \{0, 1\}$, which are the complements of the $x_{jk}^i$. We represent each value $v_j^i$ of the questionnaire with an integer variable $z_j^i$, by defining a mapping $\varphi_j^i$ (a different mapping for each field) between values of the domain and integer numbers between 0 and an upper value $U$. $U$ is such that no elements of any feasible domain maps to $U$.

$$\varphi_j^i : D_j^i \rightarrow \{0, \ldots, U\}
\forall v_j^i \mapsto z_j^i$$

Mapping for integer domains is straightforward. We approximate real domains with rational domains and then map them on the set of integer positive numbers. Qualitative domains also are mapped on the set of integer numbers by choosing an ordering. The integer variables are therefore:

$$z_j^i = \varphi_j^i(v_j^i)$$

Note that, in the case of the considered application, values were wanted to be integer. However, variables $z_j^i$ are not structurally bounded to be integer. All the out-of-range values map to the greater number used $U$. The blank value, when belonging to the feasible domain, is encoded with the integer value $b_j^i$ immediately consecutive to the greater value of the feasible domain $D_j^i$. Note that $b_j^i < U$ is always required. The membership of a value $v_j^i$ to the subset $S_{jk}^i$ is encoded by using the binary variables $x_{jk}^i$.

$$x_{jk}^i = \begin{cases}
1 & \text{when } v_j^i \in S_{jk}^i \\
0 & \text{when } v_j^i \not\in S_{jk}^i
\end{cases}$$

Finally, the complementary binary variables $\bar{x}_{jk}^i$ are bound the former ones by the following so-called coupling constraints.

$$\bar{x}_{jk}^i + x_{jk}^i = 1$$

The presence of the complementary variables is motivated by algorithmic issues. Integer and binary variables are linked by using a set of linear inequalities called bridge constraints. They impose that, when $z_j^i$ has a value such that $v_j^i$ belongs to subset $S_{jk}^i$, the corresponding $x_{jk}^i$ is 1 and all other binary variables \{\(x_j^1, \ldots, x_{jk-1}^i, x_{jk+1}^i, \ldots, x_{jn_j}^i\)\} are 0.

By using the above variables all the above mentioned rules can be expressed. Logic propositions $p_v$ are expressed by using the binary variables $x_{jk}^i$ or $\bar{x}_{jk}^i$, mathematical propositions $p_v$ are expressed by using the integer variables $z_j^i$. Rules involving more than one individual (called interpersonal rules) are expressed...
by using the opportune variables for the different individuals. By doing so, each logical rule $r_s$ having the structure (1) of a clause can be written as the following linear inequality

$$\sum_{i,j,k \in \pi_s} x_{jk}^i + \sum_{i,j,k \in \nu_s} \bar{x}_{jk}^i \geq 1$$

Moreover, with a commonly used slight abuse of notation, let $x, \bar{x}$ and $z$ be the vectors respectively made of all the components $x_{jk}^i, \bar{x}_{jk}^i$ and $z_j^i, i = 1, \ldots, l, j = 1, \ldots, m, k = 1, \ldots, n_j$. By introducing the incidence vectors $a_x^s$ and $a_{\bar{x}}^s$ respectively of the set of the positive propositions $\pi_s$ and of set of the negative propositions $\nu_s$, each logical rule can be expressed with the following vectorial notation.

$$a_x^s x + a_{\bar{x}}^s \bar{x} \geq 1$$

The only difference when mathematical propositions are present is that they do not correspond to binary variables but to operations between the integer variables. We limit mathematical rules to those which are linear or linearizable. In particular, we allow rules composed by a division or a multiplication of two variables. For a digression on linearizable inequalities, see for instance [14]. Occasionally, further binary variables are introduced, for instance to encode disjunctions of mathematical propositions. Note, moreover, that a very precise syntax for rules was developed. Therefore, encoding could be performed by means of an automatic procedure.

**Example 2.6.** Consider the following logical rule for all the individuals.

$$\neg (\text{marital status} = \text{married}) \lor \neg (\text{age} < 14)$$

By substituting the logical variables, we have the logic formula $\bar{x}_i^{\text{married status} (\text{married})} \lor \bar{x}_i^{\text{age} (0..13)}$, $i = 1, \ldots, l$. This becomes the following linear inequalities:

$$\bar{x}_i^{\text{married status} (\text{married})} + \bar{x}_i^{\text{age} (0..13)} \geq 1 \quad i = 1, \ldots, l$$

Consider the following logic-mathematical rule for all the individuals.

$$\neg (\text{marital status} = \text{married}) \lor (\text{age} - \text{years married} \geq 14)$$

By substituting the logical and integer variables, we have $\bar{x}_i^{\text{married status} (\text{married})} \lor (z_i^{\text{age}} - z_i^{\text{years married}} \geq 14)$, $i = 1, \ldots, l$. This becomes the following linear inequalities:

$$U \bar{x}_i^{\text{married status} (\text{married})} + z_i^{\text{age}} - z_i^{\text{years married}} \geq 14 \quad i = 1, \ldots, l$$

Finally, the following interpersonal mathematical rule between individual 1 and 2

$$\text{age (of 1)} - \text{age (of 2)} \geq 14$$

becomes the linear inequality

$$z_1^{\text{age}} - z_2^{\text{age}} \geq 14$$

Altogether, from the set of rules, a set of linear inequalities is obtained (to which the coupling constraints and the bridge constraints are added). From the set of answers to a questionnaire, values for the introduced variables are given. By construction, all and only the variable assignments corresponding to correct questionnaires satisfy all the linear inequalities, hence the linear system

$$\begin{cases} A^x x + A^\bar{x} \bar{x} & \geq 1 \\ B^x x + B^\bar{x} \bar{x} + B z & \geq b \\ x + \bar{x} & = 1 \\ z_j^i \in \{0,1\}, & i = 1\ldots l, j = 1\ldots m \\ x_{jk}^i, \bar{x}_{jk}^i & \in \{0,1\}, & i = 1\ldots l, j = 1\ldots m, k = 1\ldots n_j \end{cases}$$

(2)

The coefficient matrices $A^x, A^\bar{x}$ are given by encoding the logical rules, $B^x, B^\bar{x}, B$ and $b$ are given by encoding mathematical and logic-mathematical rules, and also by any other additional constraints such as the bridge constraints. Briefly, even if slightly improperly, a questionnaire $q$ must satisfy (2) to be correct.
3 Modeling the Problems

After a phase of rules validation, were the system (2) is checked to be feasible and to have more than one solution, detection of erroneous questionnaires \( q^e \) trivially becomes the problem of testing if the variable assignment corresponding to a questionnaire instance \( q \) satisfies (2).

When detected an erroneous questionnaire \( q^e \), the imputation process consists in changing some of its values, obtaining a corrected questionnaire \( q^c \) which satisfies the system (2) and is as close as possible to the (unknown) original questionnaire \( q^o \) (the one we would have if we had no errors). In order to reach this purpose, two general principles should be followed during the imputation process: to apply the minimum changes to erroneous data, and to modify as less as possible the original frequency distribution of the data [5]. Generally, a cost for changing each value of \( q^e \) is given, based on the reliability of the field. It is assumed that, when error is something unintentional, the erroneous fields are the minimum-cost set of fields that, if changed, can restore consistency. Questionnaire \( q^c \) corresponds to a variable assignment. In particular, we have a set of \( l(n_1 + \ldots + n_m) \) binary values \( c^e_{jk} \) and a set of \( l \times m \) integer values \( g^i_j \). We have a cost \( c^e_{jk} \in \mathbb{R}_+ \) for changing each \( e^i_{jk} \), and a cost \( c^i_j \in \mathbb{R}_+ \) for changing each \( g^i_j \)

\[
\{ c^1_{11}, \ldots, c^1_{n_1}, \ldots, c^1_{m_1}, \ldots, c^l_{11}, \ldots, c^l_{n_1}, \ldots, c^l_{m_1}, \ldots, c^l_{m_1 n_1}, \ldots, c^l_{m_1}, \ldots, c^l_{m_1 n_1}, \ldots, c^l_{m_1} \}
\]

Questionnaire \( q^c \) that we want to find corresponds to the values of the variables (\( x^e_{jk}, \ x^c_{jk}, \) and \( z^e_j \)) at the optimal solution.

The problem of error localization is to find a set \( V \) of fields of minimum total cost such that \( q^c \) can be obtained from \( q^e \) by changing (only and all) the values of \( V \). Imputation of actual values of \( V \) can then be performed in a deterministic or probabilistic way. This cause the minimum changes to erroneous data, but has little respect for the original frequency distributions.

A donor questionnaire \( q^d \) is a correct questionnaire which should be as close as possible to \( q^c \). Questionnaire \( q^d \) corresponds to a variable assignment. In particular, we have a set of binary values \( d^e_{jk} \) and a set of integer values \( f^i_j \). Donors are selected according to an opportune distance function specified by the user.

\[
d : (q^e, q^d) \rightarrow \mathbb{R}_+
\]

The problem of imputation through a donor is to find a set \( W \) of fields of minimum total cost such that \( q^c \) can be obtained from \( q^e \) by copying from the donor \( q^d \) (only and all) the values of \( W \). This is generally recognized to cause low alteration of the original frequency distributions, although changes caused to erroneous data may be not minimum. We are interested in solving both of the above problems, and in choosing for each questionnaire the solution having the best quality.

Let us introduce \( l(n_1 + \ldots + n_m) \) binary variables \( y^i_{jk} \in \{0,1\} \) representing the changes we introduce in \( e^i_{jk} \).

\[
y^i_{jk} = \begin{cases} 1 & \text{if we change } e^i_{jk} \\ 0 & \text{if we keep } e^i_{jk} \end{cases}
\]

Furthermore, only in the case of imputation through a donor, let us introduce \( l \times m \) binary variables \( w^i_j \in \{0,1\} \) representing the changes we introduce in \( g^i_j \).

\[
w^i_j = \begin{cases} 1 & \text{if we change } g^i_j \\ 0 & \text{if we keep } g^i_j \end{cases}
\]

The minimization of the total cost of the changes can be expressed with the following objective function (where the terms \( \tilde{c}^i_j w^i_j \) appear only in the case of imputation through a donor).

\[
\min_{y^i_{jk}, w^i_j \in \{0,1\}} \sum_{i=1}^{l} \sum_{j=1}^{m} \sum_{k=1}^{n_j} c^e_{jk} y^i_{jk} + \sum_{i=1}^{l} \sum_{j=1}^{m} \tilde{c}^i_j w^i_j = (3)
\]
However, the constraints (2) are expressed by means of variables $x_{jk}^i$, $\bar{x}_{jk}^i$, and $z_j^i$. A key issue is that there is a relation between variables in (2) and variables in (3). In the case of error localization, this depends on the values of $e_{jk}^i$, as follows:

$$y_{jk}^i = \begin{cases} x_{jk}^i (1 - \bar{x}_{jk}^i) & \text{if } e_{jk}^i = 0 \\ 1 - x_{jk}^i (1 - \bar{x}_{jk}^i) & \text{if } e_{jk}^i = 1 \end{cases}$$

In fact, when $e_{jk}^i = 0$, to keep it unchanged means to put $x_{jk}^i = 0$. Since we do not change, $y_{jk}^i = 0$. On the contrary, to change it means to put $x_{jk}^i = 1$. Since we change, $y_{jk}^i = 1$. Altogether, $y_{jk}^i = x_{jk}^i$. When, instead, $e_{jk}^i = 1$, to keep it unchanged means to put $x_{jk}^i = 1$. Since we do not change, $y_{jk}^i = 0$. On the contrary, to change it means to put $x_{jk}^i = 0$. Since we change, $y_{jk}^i = 1$. Altogether, $y_{jk}^i = 1 - x_{jk}^i$.

By using the above results, we can rewrite the objective function (3). Therefore, the problem of error localization can be modeled as follows, where the objective function and a consistent number of constraints have a set covering structure (see for instance [9]).

$$\min \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^{n_j} (1 - e_{jk}^i) c_{jk}^i x_{jk}^i + \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^{n_j} e_{jk}^i c_{jk}^i x_{jk}^i$$

Subject to

$$\begin{align*}
A^x x + A^e \bar{x} & \geq 1 \\
B^x x + B^e \bar{x} + B z & \geq b \\
x + \bar{x} & = 1 \\
z_j & \in \{0, \ldots, U\}, \quad i = 1 \ldots l, \quad j = 1 \ldots m \\
x_{jk}^i, \bar{x}_{jk}^i & \in \{0, 1\}, \quad i = 1 \ldots l, \quad j = 1 \ldots m, \quad k = 1 \ldots n_j
\end{align*}$$

Conversely, in the case of imputation through a donor, relation between $x_{jk}^i$ and $y_{jk}^i$ depends on the values of $e_{jk}^i$ and $d_{jk}^i$.

$$y_{jk}^i = \begin{cases} x_{jk}^i (1 - \bar{x}_{jk}^i) & \text{if } e_{jk}^i = 0 \text{ and } d_{jk}^i = 1 \\ 1 - x_{jk}^i (1 - \bar{x}_{jk}^i) & \text{if } e_{jk}^i = 1 \text{ and } d_{jk}^i = 0 \\ 0 & \text{if } e_{jk}^i = d_{jk}^i \end{cases}$$

In fact, when $e_{jk}^i = 0$ and $d_{jk}^i = 1$, not to copy the element means to put $x_{jk}^i = 0$. Since we do not change, $y_{jk}^i = 0$. On the contrary, to copy the element means to put $x_{jk}^i = 1$. Since we change, $y_{jk}^i = 1$. Altogether, $y_{jk}^i = x_{jk}^i$. When, instead, $e_{jk}^i = 1$ and $d_{jk}^i = 0$, not to copy the element means to put $x_{jk}^i = 1$. Since we do not change, $y_{jk}^i = 0$. On the contrary, to copy the element means to put $x_{jk}^i = 0$. Since we change, $y_{jk}^i = 1$. Altogether, $y_{jk}^i = 1 - x_{jk}^i$. Finally, when $e_{jk}^i = d_{jk}^i$, we cannot change $e_{jk}^i$, hence $y_{jk}^i = 0$.

Note, however, that even when we do not change $x_{jk}^i$ from $e_{jk}^i$ to $d_{jk}^i$, we still could need to change $z_j^i$ from $g_j^i$ to $f_j^i$ so that a better solution is achieved. For instance, this could help in satisfying some mathematical constraints without changing too many values. In order to guide the choice of values for $z_j^i$, information obtained by the $x_{jk}^i$ variables is used. We take for $z_j^i$ the value of the donor when a) changes on the $x_{jk}^i$ are made, or b) when, even if for all $k$ the $x_{jk}^i$ do not change, it is more convenient to take the donor value.

$$z_j^i = \begin{cases} g_j^i & \text{if } \forall k \in \{1, \ldots, n_j\} \ y_{jk}^i = 0 \\ f_j^i & \text{if } \exists k \in \{1, \ldots, n_j\} : y_{jk}^i = 1, \text{ or if } f_j^i \text{ is more convenient than } g_j^i \end{cases}$$

For each $z_j^i$, $n_j$ quantities $v_{jk}^i$ are defined. They are 0 or 1 when the corresponding $y_{jk}^i$ are 0 or 1.

$$v_{jk}^i = (x_{jk}^i(1 - e_{jk}^i) d_{jk}^i) + ((1 - x_{jk}^i) e_{jk}^i(1 - d_{jk}^i))$$

we have that the condition $\exists k \in \{1, \ldots, n_j\} : y_{jk}^i = 1$ becomes $\sum_k v_{jk}^i = 2$, and that the condition $\forall k \in \{1, \ldots, n_j\} : y_{jk}^i = 0$ becomes $\sum_k v_{jk}^i = 0$. Therefore, $z_j^i$ is $f_j^i$ when $\sum_k v_{jk}^i = 2$, and we need to choose between $f_j^i$ and $g_j^i$ otherwise.
By using the above, we can rewrite the objective function (3). Therefore, the problem of imputation through a donor can be modeled as follows. Again, the objective function and a consistent number of constraints have a set covering structure.

\[
\begin{align*}
\min_{x_{jk}, \bar{x}_{jk} \in \{0,1\}, \quad w_j^i \in \{0,1\}} & \quad \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^{n_j} (1 - e_{jk}^i) d_{jk}^i c_{jk}^i x_{jk}^i + \sum_{i=1}^l \sum_{j=1}^m \sum_{k=1}^{n_j} e_{jk}^i (1 - d_{jk}^i) c_{jk}^i \bar{x}_{jk}^i + \\
& \quad + \sum_{i=1}^l \sum_{j=1}^m e_{jk}^i w_j^i \\
\text{Subject to} & \quad \begin{cases} 
A^v x + A^v \bar{x} & \geq 1 \\
B^v x + B^v \bar{x} + B z & \geq b \\
x + \bar{x} & = 1 \\
z_j^i = f_j^i(w_j^i + \frac{1}{2} \sum_{k=1}^{n_j} v_{jk}^i) + g_j^i (1 - w_j^i - \frac{1}{2} \sum_{k=1}^{n_j} v_{jk}^i) & i = 1 \ldots l, j = 1 \ldots m \\
w_j^i \leq 1 - \sum_{k=1}^{n_j} v_{jk}^i & i = 1 \ldots l, j = 1 \ldots m \\
z_j^i \in \{0, \ldots, U\}, & i = 1 \ldots l, j = 1 \ldots m \\
x_{jk}^i, \bar{x}_{jk}^i \in \{0,1\}, & i = 1 \ldots l, j = 1 \ldots m, k = 1 \ldots n_j \\
w_j^i \in \{0,1\}, & i = 1 \ldots l, j = 1 \ldots m 
\end{cases}
\end{align*}
\]

The presence of the group of covering constraints and of that of equalities constraints allows the use of an additional separation routine during the solution of the above models.

4 Solving the Problems

The practical behavior of the proposed procedure is evaluated both from the computational and from the data quality points of view, as follows. Two large data sets representing correct questionnaires were initially perturbed by introducing errors. After this, detection of erroneous questionnaires was performed, as a trivial task. The proposed procedure is then used for the imputation of such erroneous questionnaires.

Data used for experimentation arise from the Italian Census of Population 1991. They consist in 45,716 four-person households and 20,306 six-person households (from a single region). Data perturbation consists in randomly introducing non responses (blank answers or out-of-range answers) or other valid responses (other values belonging to the feasible domain). Each data set was perturbed at four different increasing error levels, called 50, 100, 150, 200. Eight different data sets are therefore obtained:

\[(4_{0.50}, 4_{1.00}, 4_{1.50}, 4_{2.00}, 6_{0.50}, 6_{1.00}, 6_{1.50}, 6_{2.00})\]

The set of rules used for experimentations are real rules, developed under the supervision of the Italian Statistic Office. They consist in:

32 logic individual rules (to be repeated for each individual \(i \in I\));
35 logic interpersonal with 2 individuals rules (to be repeated for each couple of individual \((i,i') \in I\));
2 logic interpersonal with 3 individuals rules (to be repeated for each triple of individual \((i,i',i'') \in I\));
1 logic-mathematic individual rule (to be repeated for each individual \(i \in I\));
55 logic-mathematical interp. with 2 ind. rules (to be repeated for each couple of individual \((i,i') \in I\));
2 logic-mathematical interp. with 3 ind. rules (to be repeated for each triple of individual \((i,i',i'') \in I\));
1 logic-mathematical interp. with 4 ind. rule (to be repeated for each quadruple of ind. \((i,i',i'',i''') \in I\)).

For each erroneous questionnaire \(q^e\), the error localization problem (4) is solved at first. After this, a number \(b(q^e)\) of donor questionnaires for each \(q^e\) is selected. Such donors \(\{q_1^d, \ldots, q_{b_0}^d\}\) are selected in the above mentioned data sets, by choosing the nearest ones to \(q^e\), according to our distance function \(d\). The number \(b(q^e)\) increases when the imputation quality for \(q^e\) (given by the value of the cost function) is not satisfactory. Therefore, for each erroneous questionnaire \(q^e\), \(b(q^e)\) problems of imputation through a donor
(5) are solved. Altogether, for each erroneous questionnaire \( q^e \), \( b(q^e) + 1 \) imputation problems are solved. The corrected questionnaire \( q^c \) is finally obtained by choosing the best result among them.

Problems are solved by using a commercial implementation of a branch-and-bound routine for integer programming (ILOG Cplex 7.1). However, such solver allows the user to define specific separation subroutines to be used within its framework, obtaining therefore a branch-and-cut procedure. Since most of the constraints have a structure similar to those of set covering problems, a separation routine for the set covering polytope was used in order to generate valid cuts. Such separation routine, described in [10], is based on projection operations, which were possible thanks to the presence of the equality constraints above called coupling constraints. Since such cut generation is a relatively costing operation, it is preferable to perform it only at the very first levels of the branching tree, where its effect is greater.

Computational times in minutes for solving each data set (on a Pentium III 800MHz PC) are reported in Table 1. As observable, each single imputation problem is solved in extremely short times. Therefore, large data sets are imputed in very reasonable times.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Number of households</th>
<th># of problems solved</th>
<th>Total time</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_050</td>
<td>45,716</td>
<td>320,656</td>
<td>53.0</td>
</tr>
<tr>
<td>L_100</td>
<td>45,716</td>
<td>346,223</td>
<td>90.4</td>
</tr>
<tr>
<td>L_150</td>
<td>45,716</td>
<td>385,680</td>
<td>130.5</td>
</tr>
<tr>
<td>L_200</td>
<td>45,716</td>
<td>416,074</td>
<td>157.9</td>
</tr>
<tr>
<td>L_050</td>
<td>20,306</td>
<td>145,322</td>
<td>85.8</td>
</tr>
<tr>
<td>L_100</td>
<td>20,306</td>
<td>160,371</td>
<td>139.8</td>
</tr>
<tr>
<td>L_150</td>
<td>20,306</td>
<td>186,434</td>
<td>174.5</td>
</tr>
<tr>
<td>L_200</td>
<td>20,306</td>
<td>198,121</td>
<td>202.6</td>
</tr>
</tbody>
</table>

Table 1: Imputation times in minutes for 4 persons household and 6 persons households.

The statistical performances of the proposed methodology has also been strictly evaluated and compared with the performance of the Canadian Nearest-neighbour Imputation Methodology (CANCEIS) [1] by a simulation study based on real data from the 1991 Italian Population Census [8]. CANCEIS has been selected for the comparative statistical evaluation because nowadays it is deemed to be the best methodology to automatically handle hierarchical demographic data. Results of such comparison are very encouraging: the quality of the imputation performed by the proposed procedure is generally comparable, and sometimes better, than CANCEIS.

The quality of imputed data was evaluated by comparing the original questionnaires (here known) with the corrected ones. We report in Table 2 the value of some particularly meaningful statistical indicator: the percentage of not modified values erroneously imputed by the procedure \( (E_{true}) \); the percentage of modified values not imputed \( (E_{mod}) \); the percentage of imputed values for which imputation is a failure \( (I_{imp}) \). Therefore, lower values correspond to a better data quality. Reported value is computed as average on the demographic fields relation to head of the house, sex, marital status, age, years married.

<table>
<thead>
<tr>
<th>Data set</th>
<th>( E_{true} )</th>
<th>( E_{mod} )</th>
<th>( I_{imp} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>L_050</td>
<td>0.04</td>
<td>24.61</td>
<td>15.02</td>
</tr>
<tr>
<td>L_100</td>
<td>0.09</td>
<td>26.02</td>
<td>15.48</td>
</tr>
<tr>
<td>L_150</td>
<td>0.13</td>
<td>26.32</td>
<td>16.20</td>
</tr>
<tr>
<td>L_200</td>
<td>0.93</td>
<td>27.25</td>
<td>17.10</td>
</tr>
<tr>
<td>L_050</td>
<td>0.08</td>
<td>31.20</td>
<td>20.47</td>
</tr>
<tr>
<td>L_100</td>
<td>0.16</td>
<td>31.44</td>
<td>20.29</td>
</tr>
<tr>
<td>L_150</td>
<td>0.25</td>
<td>32.83</td>
<td>21.45</td>
</tr>
<tr>
<td>L_200</td>
<td>0.35</td>
<td>33.01</td>
<td>21.88</td>
</tr>
</tbody>
</table>

Table 2: Percentage of not modified values erroneously imputed \( (E_{true}) \), percentage of modified values not imputed \( (E_{mod}) \), percentage of imputed values for which imputation is a failure \( (I_{imp}) \).

The procedure introduces surprisingly few changes in fields that were not perturbed, is able to discover more then two times out of three the values which were modified, and imputes values which are generally correct. Note that, when randomly modifying values, the record can still appear correct, so detection of perturbed values inherently has no possibility of being always exact. Note, moreover, that for fields having many values, as the case of age, the correct imputation is extremely difficult.
5 Conclusions

Imputation problems are of great relevance in every process of data collecting. They also arise when cleaning databases which can contain errors. Imputation problems have been tackled in several different manners, but satisfactory data quality and computational efficiency appear to be at odds. A discrete mathematics model of the whole imputation process allows the implementation of an automatic procedure for data imputation. Such procedure repairs the data using donors, ensuring so that the marginal and joint distribution within the data are, as far as it is possible, preserved. The sequence of arisen integer programming problems can be solved to optimality by using state-of-the-art implementation of branch-and-cut procedures. Related computational problems are completely overcome. Each single imputation problem is solved to optimality in extremely short times (always less than 1 second).

The statistical performances of the proposed procedure has been strictly evaluated on real-world problems, and compared (in [8]) with the performance of the Canadian Nearest-neighbour Imputation Methodology, which is nowadays deemed to be the best methodology to automatically handle hierarchical demographic data. Results are very encouraging both from the computational and from the data quality point of view.

References


