DIESIS: a New Software System for Editing and Imputation

DIESIS: *un Nuovo Software per il Controllo e la Correzione dei Dati*

Renato Bruni
Dipartimento di Informatica e Sistemistica dell'Università degli Studi di Roma "La Sapienza", Via M. Buonarroti 12, 00185 Roma, bruni@dis.uniroma1.it

Alessandra Reale
Istat – Direzione Censimento della Popolazione e Territorio, Via A. Ravà 150, 00142 Roma, reale@istat.it

Renato Torelli
Istat – Direzione Censimento della Popolazione e Territorio, Via A. Ravà 150, 00142 Roma, torelli@istat.it

Riassunto: In questo lavoro vengono presentati i risultati di una ricerca sul problema del controllo e della correzione dei dati. L'approccio proposto, implementato nel sistema software DIESIS, consente il trattamento congiunto di variabili quantitative e qualitative per dati con struttura gerarchica. Il controllo della consistenza e non ridondanza degli *edit* è affrontato risolvendo una sequenza di problemi di *feasibility*. Il problema dell'imputazione dei dati è affrontato risolvendo una sequenza di problemi di *set covering*. Questo approccio consente il superamento dei limiti computazionali presenti nei software che implementano la metodologia di Fellegi-Holt.

Keywords: Editing and Imputation, Hierarchical Data, Optimization.

1. Introduction

Preparing for the 2001 Population Census (PC), the Italian National Statistics Institute (Istat) planned research studies with the aim of improving the efficacy of the Editing and Imputation process. We describe here the new generalized software, called DIESIS (Data Imputation and Editing System - Italian Software), jointly developed by Istat and by the Department of Computer and Systems Science of the University of Roma "La Sapienza" (Bruni, Reale, Torelli, 2001).
We have tackled the problem of data completeness and consistency by means of an approach suitable to handle hierarchical data. Population Census data are collected at the household level (*unit*) with information for each person (*sub-units*) within the household. The problem of error detection is generally approached by formulating a set of edit rules, expressing the error condition. A unit is *failed* if it verifies at least one rule. A unit is *correct* if it does not verify any rule. Besides to preserving the distributions of individual variables (using *individual* edit rules), a crucial problem in

imputing PC data is also preserving the relationships among variables belonging to different persons within the household (using *between persons* edit rules). Usually, software systems based on the Fellegi-Holt (Fellegi and Holt, 1976) approach, developed by Statistical Offices, do not allow to define some kind of between persons edit rules (for example, edits concerning the difference between two ages). In fact, in such cases, mathematical relationships between numerical variables cannot be specified as edit rules in the editing and correction of both qualitative and numeric variables. Moreover, the number of edit rules, which is very large for real world problems, does not allow handling all the variables in a single step. This holds because Fellegi-Holt systems, requiring the generation of the complete set of edits, cannot afford such computational burden (Winkler, 1999).

DIESIS system is the results of a mathematical approach able to deal with both qualitative and quantitative variables. This approach, based on optimization techniques, overcomes the computational limits of Fellegi-Holt methodology, while maintaining its positive statistical features, and takes advantage of some useful characteristics of NIM (Bankier, 2000) and RIDA (Abbate, 1997). DIESIS performs detection and probabilistic correction of inconsistent or out of range data in a general process of statistical data collecting. In order to simplify our exposition, however, we will be focused on the case of a PC data editing and imputation.

The main characteristics of DIESIS are briefly described in section 2. The optimization techniques developed for edit validation and data imputation are resumed in section 3.

2. Main DIESIS System Characteristics

DIESIS system treats invalid or inconsistent responses for qualitative and numeric variables simultaneously. Edit rules must be defined by conjunction of terms. Terms can be either logical propositions, or linear inequalities, or inequalities containing the product of two variables.

Initially, the system locates *redundancies* among edit rules (i.e. edit rules which are logically implied by other rules) and checks for the presence of *inconsistencies* among edit rules (i.e. edit rules which are in contradiction with other ones) (Loveland, 1978). Moreover, if some deterministic imputation rules are used, it is possible to perform the check between all types of rules. DIESIS then divides units in correct ones and failed ones. Afterwards, the system performs imputation on all failed units, according to two different imputation strategies. It can *impute the minimum weighted number of variables given the available donors* (this approach will be called) or impute the *absolute minimum weighted change* (this will be called).

In the case of 'first donors then fields', the system selects for each failed unit a set of donors. Donors are correct units chosen among the nearest neighbors, that is, among units that resemble the failed unit. DIESIS finds a subset of potential donor units with the smallest distance from the failed unit. The distance function used is a weighted sum of the distance scores for each variable. This can be computed over all persons (this strategy will be called *all sub-units distance*) or over the subset of persons involved in verified edit rules (this strategy will be called *involved sub-units distance*). In both cases, all variables of persons are considered, not only those which enter verified edit rules. So far, DIESIS selects the imputation action to be used, by minimizing the

weighted number of *changes* performed in the failed unit, subject to the following constraints: imputed units should not verify any edit rule, not only those that were originally verified; imputed values must come from a single donor. For each donor selected, DIESIS finds the imputation action minimizing the following function:

$$\sum_k \sum_i c_{i,k}\, y_{i,k} \qquad (1)$$

where $y_{i,k}$ is a binary variable which value is 0 if the value of the *i*-th variable of the *k*-th person in the failed household is equal to the value of the *i*-th variable of the *k*-th person in the imputed household, and 1 otherwise. The $c_{i,k}$ is a real number representing the weight given for the *i*-th variable of the *k*-th person. Different weights can be assigned to different admissible values of the variables. In the above summation, *k* varies over the set of all persons (computation of *all sub-units objective function*) or only over the subset of persons involved in failed edit rules (computation of *involved sub-units objective function*). After computation of all the above minima, the imputation action corresponding to the minimum among them is chosen.
In the case of 'first fields then donors', instead, DIESIS localizes in the failed unit the minimum weight set of variables to be changed in order to obtain a correct unit. After this, the system imputes such values by performing a *joint* imputation (that is from a single donor) if possible, or a *sequential* imputation (that is from different donors) otherwise (Fellegi and Holt, 1976). Note that DIESIS overcomes computational limits of other systems implementing the Fellegi-Holt approach, since it does not need the generation of implied edits. The two above imputation algorithms ('first donors then fields' and 'first fields then donors') can also be used jointly. This means that the user can choose using the 'first fields then donors' when, for a given failed household, the number of changes proposed by the 'first donors then fields' algorithm is exceedingly higher. The system runs on MS Windows operating system, and a UNIX version is currently under development.

3. Optimization Techniques Developed for Edit Validation and Data Imputation

The set of edit rules is encoded by means of a system of linear inequalities (Bruni, Reale, Torelli, 2001). Inconsistencies and redundancies in the set of edits corresponds to particular structure of the obtained system of linear inequalities. In order to detect such situation, the *feasibility* of this system is analyzed (Bertsimas, Tsitsiklis, 1997). In particular, we need to solve a first sequence of feasibility problems arising from consistency checking, and a second sequence of new feasibility problems arising from redundancy checking. Such problems, which are well known in the field of Operations Research, are solved by means of an enumerative approach based on branching techniques (Bertsimas, Tsitsiklis, 1997). Elapsed times for solving each feasibility problem is always less than 1 second.
After detection of failed units, each failed unit imputation produces the sequence of minimization problems depicted in section 2. Such problems have an objective function expressing the aim of performing the minimum weighted changes, as roughly schematized in (1), and constraints imposing that edit rules must not be verified

anymore (the system of linear inequalities mentioned above). Other constraints arise from imposing the imputed values to be either the failed unit's ones or the donor unit's ones. Due to the peculiarities of the problem, most of these constraints are of a so-called *set covering* type (Nemhauser and Wolsey, 1988). After solving such sequence of *set covering problems*, the imputation giving the minimum value for the objective function is chosen. Such problems, again, are well known in the field of Operations Research. They are solved to optimality by means of branch and cut techniques (Nemhauser and Wolsey, 1988). Elapsed times for solving each set covering problem is again always less than 1 second.

4. Conclusions

The software system DIESIS, by treasuring optimization techniques, made a significant computation breakthrough. The sequence of arisen optimization problems can be solved to optimality by using state-of-the-art procedures. Each problem is solved to optimality in extremely short times. The statistical performances of the new software has been strictly evaluated and compared with the performance of the Canadian Nearest-neighbor Imputation Methodology (NIM) by a simulation study based on real data from the 1991 Italian PC (Manzari, Reale, 2001). NIM was selected for the comparative statistical evaluation because nowadays it is deemed to be the best methodology to automatically handle hierarchical demographic data. The results are very encouraging.

References

Abbate C. (1997) La completezza delle informazioni e imputazione da donatore con distanza mista minima, *Quaderni di Ricerca ISTAT n.4/1997.*

Bankier M. (2000) Canadian Census Minimum change Donor imputation methodology, *Proceedings of the Workshop on Data Editing*, UN/ECE, United Kingdom (Cardiff).

Bertsimas D., Tsitsiklis J.N. (1997) *Introduction to Linear Optimization*, Athena scientific, Belmont, Massachusetts.

Bruni R., Reale A., Torelli R. (2001) Optimization Techniques for Edit Validation and Data Imputation, in *Proceedings of Statistics Canada Symposiuum 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective"*, Canada (Ottawa).

Fellegi I. P., Holt D. (1976) A systematic approach to edit e imputation, *Journal of the American Statistical Association,* vol.71, pp. 17-35.

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

Manzari A., Reale A. (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,* Korea (Seoul).

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

Winkler W. E. (1999) State of Statistical Data Editing and current Research Problems, *Proceedings of the Workshop on Data Editing*, UN/ECE, Italy (Rome).