WP. 20

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

WORK SESSION ON STATISTICAL DATA EDITING
(Madrid, Spain, 20-22 October 2003)

Topic (ii): Developments related to new methods and techniques

DATA VALIDATION THROUGH MEASUREMENTS ON CONDITIONAL DISTRIBUTIONS OF A LOGICALLY RELATED GROUP OF VARIABLES

Supporting Paper

Submitted by Liaison Systems S.A., Greece[1]

[1] Prepared by George A. Petrakos (george.petrakos@liaison.gr), Gregory E. Farmakis (gregory.farmakis@liaison.gr) and Photis Stavropoulos (photis.stavropoulos@liaison.gr).

Data Validation through Measurements on Conditional Distributions of a Logically Related Group of Variables

George A. Petrakos, Gregory E. Farmakis, Photis Stavropoulos

Liaison Systems S.A.
Research and Development of Information Systems
Acadimias 77, Athens, 106 78 GR, http://www.liaison.gr

E-Mail: George.Petrakos@liaison.gr
Gregory.Farmakis@liaison.gr, Photis.Stavropoulos@liaison.gr

Abstract

Data validation, i.e. the process of detecting erroneous, or probably erroneous, values within statistical data sets, is normally done through the application of validation rules, such as ranges or deterministic correlations among variables. The presented research work on the contrary, is based on the concept of probabilistic validation treatment, based on the conditional probability distribution of multi-dimensional random variables. After the probability space is defined, the conditional probabilities are calculated from historical data and/or prior information, and an estimation of the underlying probability structure is derived. Thus, a generic and consistent validation treatment is feasible, which can then be applied without the overhead of expressing and formally declaring variable-specific validation rules. Obviously, due to the significant number of dimensions as well as the important volume of statistical data, this process can demanding in terms of both computing and data management, therefore requiring a specifically optimised software and data warehouse architecture.

The methodological and software concepts presented here are being implemented and validated within the INSPECTOR project, funded by the EC's 5th Framework Programme.

Keywords: Data validation, probability measure, conditional probability, Borel sets, statistical inference

1.Introduction

Statistical agencies collect large amounts of data, which they process and analyse. The results, in the form of tabulations and estimates, are reported to a variety of users ranging from international organisations and national governments to the general public. Many important decisions are based on statistical information and therefore the quality of the data on which this information is based is of utmost importance. This explains the interest on editing and imputation methods.

In the domain of official statistics, information is delivered to the final information consumer at the end of a several-stage life cycle starting at data collection from various possibly non-homogeneous sources (i.e. questionnaires of various forms, administrative sources etc.). No matter what the information quality assurance processes are at the downstream stages (such as aggregation, compilation of metadata or publication and delivery), it is the quality of these collected primary data, which inherently defines the quality of the information to be delivered.

Before being loaded to the data warehouse, data are either available in the form of files or stored in databases with the observations from each responding unit taking the form of a record. Editing involves specifying a set of validation rules (the edits), checking each record against them, flagging records that fail one or more edits and imputing some fields of the flagged records to make them satisfy all edits. The most important problems from a theoretical point of view are establishing internal consistency of the edits (i.e. that the set of declared edits can be satisfied by some possible records) and error localisation (i.e. determining which fields of a flagged record need to be imputed). A very important problem from an implementation point of view is the speed of execution of edit and imputation software.

The aim of this introduction is to give a brief overview of editing and then specify where this paper fits into the general framework. Since we only deal with data validation, we will not give emphasis to advances in imputation. Editing attracted the attention of statisticians since at least the 1960s; Nordbotten (1963, 1965), Pritzker et al (1965), Freund and Hartley (1967). An early attempt at error localization based on probabilistic arguments was presented in Naus et al (1972). The landmark work however was Fellegi and Holt (1976). Their method permitted both the checking for internal consistency of edits and error localization. The latter takes the form of identifying the minimum number of fields which, when imputed will make the record pass all edits. It has now been generalized to permit the assignment of confidence weights to the fields. The method is very demanding in computer power and time and has stimulated research into methods for speeding it up. For advances up to 1999 see Winkler (1999) and the references therein. Other recent techniques include the New Imputation Methodology of Statistics Canada (Bankier (1999), Bankier et al (2000)), which first finds possible donor records for imputation and then chooses the fields to impute and a method based on the Fourier elimination method for solving systems of linear inequalities, developed at Statistics Netherlands (see de Waal (2000) and the references therein).

All edit systems have as starting point a set of edits defined by subject matter experts. On the contrary, our approach presented in this paper is based on the concept

of probabilistic validation treatment, based on the conditional probability distribution of multi-dimensional random variables. After the probability space is defined, the conditional probabilities are estimated from historical data and/or prior information, and an estimation of the underlying probability structure is derived. Thus, a generic and consistent validation treatment is presented, which can be applied regardless of the semantics of the statistical variables, eliminating the overhead of identifying and formally defining variable-specific validation rules. These concepts, the suitable statistical methodology as well as the software architecture to implement it are presented in this paper.

Moreover, the application of the presented methodology has to be supported by intelligent data mining tools, capable of (i) calculating the probability distributions of the multi-dimensional variables based on "clean" data sets and (ii) processing the data sets under inspection against the derived distributions. Obviously, due to the significant number of dimensions as well as the important volume of statistical data, this process can be demanding in terms of both computing and data management, therefore requiring a specifically optimised software and data warehouse architecture.

The software architecture presented here concerns the design of data warehouse architecture, which is "aware" of the existence of multi-dimensional variables, inherently underlying the usual record based structure of the statistical data sets, and the design of a validation engine capable of applying the validation procedure.

2. Theoretical Aspects of Data Validation

According to the classification proposed by Petrakos & Farmakis, 2001, validation rules can run either horizontally or vertically in a statistical data array, apply to qualitative or quantitative variables, check the data type or domain and finally addressed to either a single variable or an entity or even an entire data schema. The application of such a classification to the definition and implementation of validation rules in an Official Statistics paradigm, the validation process in the new Integrated Information System of the National Statistical Services of Greece, is presented by Petrakos et all, 2001.

Trying to identify the theoretical frame of such a process, let us consider a set of observations called data and their associated measurable space (Ω, A, A ó-field of Ω), called sample space. Observations corresponding to different characteristics of population in Ω are considered as realizations of random variables defined as measurable transformations from (Ω, A) to B, where B the Borel sets of n-dim Euclidian space. The study of the distribution of different subsets of variables in a statistical data set or in a statistical database in general leads to certain methodologies aiming at the specification of data validation and even data editing rules. Parametric and non-parametric inference tools can be used to detect suspicious set of values in a holistic approach that takes into account the underline structure of the data under control.

Horizontally one can apply validation rules on a single variable by checking certain values against a predefined/estimated probability measure, or on an entity (a group of logically related variables).

In the later case certain observations are checked against conditional distributions in the data subset. An entity can be described as a group of variables

$Y=Y_i$, i=1,…,k, completed by another group $X=X_i$, i=1,…,m, defined on the same probability space (Ω, A), armed with a probability measure μ_θ, $\theta \in \Theta$. The distinction between the two groups is that Y consist of variables which their validity needed to be checked against the almost surely valid observations of variables in group X. Furthermore Y consist of transformation from to (Ω, A) to k-dim Euclidian space, while the X's are evaluated in any space X.

A logical rule which restricts the values of $Y=Y_i$, i=1,…,k, given a vector of observations for the other group $X=X_i$, i=1,…,m, can be embodied in the determination of the conditional probability distribution Y given X. The conditional probability measure $\mu_{m,\theta}$ in $A_m \subseteq A$, sub-field induced by X is given by

$$\mu_{m,\theta} = P\,(Y \in B,\ X \in A) / P\,(X \in A),$$

where $A, B \subseteq B$ and $P\,(X \in A) \neq 0$. If we note $\mu(A) = P\,(X \in A)$, and replace A with $A_h = [x, x+h]$ the function

$$q(x, B) = \lim_{h \to 0} \frac{P\,(X \in A_h, Y \in B)}{m(A_h)}$$

which is really equivalent to $P\,(Y \in B / X=x)$, the conditional probability of Y given certain values of X=x. For any vector of values x such that $\mu(A) \neq 0$ the set $C \subseteq B$, for which

$$\int_C q(x, B)\, m(dx) \leq c_0$$

where c_0 any selected small number, contains observations needed to be validated. The conditioning values of X, can be seen as a sufficient statistics T(X) in terms that the conditional probability measure $\mu_{m,\theta}$ in $A_m \subseteq A$ depends on X only through T(X).

Vertically let us consider $\hat{\imath} = (X_1, X_2, \ldots, X_n)$ a random sample on the probability space $(\grave{U}, A, \imath_{\grave{e}})$, $\grave{e} \in \Theta$ an open subset of R^k, $k \geq 1$, $A_n \subseteq A$ a ó-field induced by $\hat{\imath}$ $\therefore A_n = (X_1, X_2, \ldots, X_n)^{-1}(B)$. Under fairly general conditions, $P_{n,\grave{e}}$, the conditional probability measure $P_{\grave{e}}$ in A_n, $\forall$ $\grave{e} \in \Theta$ has a likelihood function asymptotically in exponential form. Therefore certain statistics, $T_i^*(\hat{\imath})$, $i \in I$, measuring distributional characteristics like location, spread, symmetry, etc., are asymptotically distributed according to known probability measures (N, ã, â, $\div^2$, etc.).

At this point statistical inference tools can be used to check the validity of $\hat{\imath}$ through the calculation of the point and interval estimation of the corresponding parameters and its comparison with their analogous measurements in time and spatial data. These comparisons can be extended to the use of goodness of fit tests, which might show changes in distribution other than location parameters due to the presence of erroneous data.

3. The INSPECTOR Data Validation Architecture

Data validation software modules are fairly common in software packages meant to support extensive data "loading" processes, such as those encountered in data warehouse administration. These applications usually allow the declaration and storage of validation rules, using a formal notation language, which is then translated to the appropriate software procedures. The architecture proposed by the authors, and being implemented and validated within the framework of the INSPECTOR research project, allowed for the declaration of validation rules by means of algebraic operations on variable domains instead of logic operations on variable values.

The architecture includes:

(a) a repository holding both the information model of the survey (i.e. metadata on the multi-leveled hierarchical structure of the variables) and the domain definitions;

(b) a transient storage facility on which the data set under inspection can be stored and operated upon, while maintaining references to the variables structure metadata;

(c) a validation engine capable of generating the domain of the variable under inspection from the repository declarations and checking the corresponding values against these.

4.Requirements for the Validation Repository

We argued earlier that any existing horizontal validation rule, which restricts the values of $Y=Y_i$, i=1,…,k, given a vector of observations for another group $X=X_i$, i=1,…,m, can be can be discovered - and applied – based on the determination of the conditional probability distribution of Y given X.

This implies that for any given group of n atomic variables, for which the existence of rules – and therefore of the corresponding logically related variable sub-groups X and Y, are not a priori known, the probability distributions for all possible combinations of the n variables, r at a time, have to be calculated. Due to combinatorial explosion, even for moderate values of n, this leads to a high number of computationally intensive calculations, which is unacceptable for practical purposes.

On the other hand, this would ignore the knowledge we already possess concerning our multidimensional space, i.e. the fact that, since variables represent in fact information about a confined system of interrelated objects, we may define sub-groups of the variables where we expect rules to exist, based on the nature of the surveyed objects. Once conditional probability distributions – and thus validation rules - for these small sub-groups have been calculated, these groups can be treated as

variables themselves, combined to higher level sub-groups, based on the logical relationships between objects, and so forth. At each level of this hierarchical down – up regrouping, which will ultimately lead to the whole data record, a small number of distinct variables is treated, thus reducing considerably the computational load.

Consequently, a major requirement concerning the architecture of the validation repository, is that the information model has to be known to our validation system, or in other words, that the system must hold the appropriate meta-data on:

a) the inherent multi-leveled, hierarchical structure of the multi-dimensional variables;

b) the existence of logical relationships among variables;

in order to limit the scope of application of the probabilistic validation treatment to manageable groups of variables where the expectation of discovery of validation rules is high.

It must be noted here that this approach is quite different from the usual data mining concept, where no previous knowledge is usually exploited and the existence of patterns of interaction among any given variables is explored.

5.Repository Structure

A repository, i.e. a specifically designed metadata database, has been designed and implemented on a RDBMS, which is able to hold domain definitions for different kinds of atomic variables, as well as definitions of the logical structure of higher level multidimensional variables.

The repository can store and manage the definition of the hierarchical structure of the data models, consisting of:

a) Atomic data elements, i.e. simple variables, which represent either properties (or generally attributes of properties) of the objects under survey. Moreover, atomic data elements can be of different types, including numeric, ordinal or categorical etc.

b) A multi-leveled hierarchy of composite data elements, i.e. multidimensional variables, composed of either simple or other composite data elements and representing either complex properties of objects or entire objects;

c) Data records, i.e. composite data elements with the additional requirement of uniqueness within the data set, representing individual observations.

d) Data sets, consisting of data records of different types; and finally

e) Data schemata, consisting of different interrelated data sets.

Note that while levels (a), (b) and (c) are used to guide horizontal probabilistic validation treatment, levels (d) and (e) are used to guide the vertical one, while entities at each level are treated as single variables.

Moreover the system can also hold knowledge about the existence of logical relationships among variables. These relationships (of type "is dependent on") are clearly distinct from the structural ones (which are of type "is composed of").

6.Validation engine operation

Apart from holding the metadata necessary to guide the probabilistic validation treatment towards groups of logically related variables, the repository also holds the metadata required to execute the actual validation process. Here, all validation rules are translated into domains of corresponding multidimensional variables. Thus, a rule binding several atomic data elements is declared as a domain of a corresponding composite data element including the former, while rules binding different composite data elements are declared as domains of higher level elements, and so forth.

For atomic data elements, the process is straightforward, based on the declaration of domains, i.e. sets of allowable values. The later can be declared as finite sets or unions of ranges depending on the atomic data element type. For multidimensional variables (i.e. composite data elements and records) however, declaring a range, in a way suitable for an automated validation procedure, would be quite complicated.

Instead, the knowledge on the composition of the variable from lower level ones is once again exploited. The main concept here is that the system can generate the domain of a multidimensional variable if the domains of the component variables and a set of additional non-allowable values are known (Petrakos & Farmakis, 2001).

A validation engine can then generate the appropriate validation rules out of this metadata repository, acting in an iterative down-up way traversing the variables hierarchy. This is done at each consecutive level, by applying the algebraic operations implied by the multidimensional variable metadata (including the declaration of the non-allowable area) to the domains of the atomic variables to which the later can ultimately be decomposed, in order to generate the corresponding multidimensional variable's domain.

7.Discussion

The validation process described in this paper is based on the explicit knowledge of various distribution types (conditional, joint, marginal) within certain structures of the data set under control, which will initiate the use of statistical inference tools in order to uncover erroneous or suspicious data values. The use of the data set under control for the determination of these distributions is not the best practice since it will be impossible to discriminate between a group of erroneous values and a peculiar shape in the margins of the estimated distribution or to identify bias. The ideal scenario will be that the distributions are well defined from prior knowledge or a clean data set or a clean data subset of our data set. Furthermore, the declaration of stochastic domains for multidimensional variables, in the form of metadata in a relational database and in a way generalised enough to be exploited by a generic software application, may become tedious for some cases. These areas will be further investigated within the INSPECTOR project.

References

Bankier, M. (1999). Experience with the New Imputation Methodology used in the 1996 Canadian census with extensions for future censuses. Report presented at the *UN/ECE Work Session on Statistical Data Editing*, Rome, Italy, 2-4 June 1999.

Bankier, M, Lachance, M. and Poirier, P. (2000). 2001 Canadian census minimum change donor imputation methodology. Report presented at the *UN/ECE Work Session on Statistical Data Editing*, Cardiff, UK, 18-20 October 2000.

Fellegi, I. P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *Journal of the American Statistical Association*, 71, 353, 17-35.

Freund, R. J. and Hartley, H. O. (1967). A procedure for automatic data editing. *Journal of the American Statistical Association*, 62, 341-352.

Granquist, L. (1997). The new view on editing. *International Statistical Review*, 65, 381-387.

Granquist, L. and Kovar, J. G. (1997). Editing of Survey Data: How much is enough? In *Survey Measurement and Process Quality*, L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin (eds.), New York: Wiley, 415-435.

Naus, J. I., Johnson, T. G. and Montalvo, R. (1972). A probabilistic model for identifying errors in data editing. *Journal of the American Statistical Association*, 67, 340, 943-950.

Nordbotten, S. (1963). Automatic editing of individual statistical observations. *Conference of European Statisticians, Statistical Standards and Practice – No 2,* United Nations, New York.

Nordbotten, S. (1965). The efficiency of automatic detection and correction of errors in individual observations as compared with other means for improving the quality of statistics. *Bulletin of the International Statistical Institute*, Proceedings of the 35th session, Belgrade, 417-441.

Petrakos G., Farmakis G. A Declarative Approach to Data Validation of Statistical Data Sets, based on Metadata, to appear in *Statistica Applicata*, Vol.12.N.3, 2000.

Petrakos G., Kalogeropoulos K., Farmakis G., Stavropoulos Ph. A Classification Scheme of Validation Rules Applied to Statistical Data Bases accepted in *NTTS 2001 International Conference in Official Statistics*

Pritzker, L., Ogus, J. and Hansen, M. H. (1965). Computer editing methods – some applications and results. *Bulletin of the International Statistical Institute*, Proceedings of the 35th session, Belgrade, 442-465.

Waal, T. de (2000). New developments in automatic edit and imputation at Statistics Netherlands. *Report presented at the UN/ECE Work Session on Statistical Data Editing*, Cardiff, UK, 18-20 October 2000.

Winkler, W. E. (1999). State of statistical data editing and current research problems. Report presented at the *UN/ECE Work Session on Statistical Data Editing*, Rome, Italy, 2-4 June 1999.