Edits – Data Cleansing at the Data Entry to assert semantic Consistency of metric Data

Hans-J. Lenz, Veit Köppen, Roland M. Müller, FU Berlin
{hjlenz, koeppen}@wiwiss.fu-berlin.de, roland@knowledge-commerce.org

Abstract

It is a matter of fact that the input of numeric data into databases needs careful screening to avoid semantic incoherency with respect to the knowledge at hand. In nearly all real applications such knowledge exists as models, i.e. as balance equations, behavioral equations or simply as definitions. The representation of those objects is possible by validation rules ("edits"), which are roughly speaking specially tailored tests. The methodology is presented, recent work in progress is shown, and a business application is presented.

Keywords: Edits, Data Cleansing, Validation Rules, Semantic Consistency, Semantic Integrity Constraints

1. Introduction

There exist two slogans which characterize the process of entering numeric or metric data into an operative (OLTP) or decision-making based (OLAP) database: "Garbage in – Garbage out" and "Numbers don't mind where they come from"(Lord), cf. [H93], [L93]. While the integrity problem as such is evident it is by no means clear how to solve it. If a corresponding model is available, which is believed to be true or useful, then validation rules called edits can be used. Such rules will be discussed in the sequel.

As an illustrating example, assume we have to insert into a database the tupel t = (sales, quantity_sold, price_per_unit) = (500, 100, 5). Evidently, the balance equation *M*: sales = quantity_sold * price_per_unit is fulfilled. Now, assume that each variable is measured with an identical measurement error of ε%=5%, and that t' = (490, 105) is observed instead of t. The question arises whether t' can pass the data entry or not, i.e. is t' coherent with the model *M* and the errors in the variables equal to ε%.

While simple, logical, and probabilistic edits are world-wide used in the front-ends of commercial and statistical information systems, edits based on statistics and fuzzy logic are not applied on large scale, although their special feature is that they allow for errors in the variables. While statistical edits use probability theory, cf. [K31], fuzzy edits make use of possibility theory to represent impreciseness of data, cf. [Z65], [KGK95].

2. Edits

The class of edits (or validation rules) is characterized by a set $\boldsymbol{F}$ of formulas or a rule base $\boldsymbol{B}$, which is linked to a given frame of discernment (view of a real system) and corresponds to a test function as a mapping T: $\boldsymbol{F} / \boldsymbol{B} \rightarrow \boldsymbol{D} \subseteq \boldsymbol{R}$. In the most simple case $\boldsymbol{D} = \{0, 1\}$, which means that tuple t is either accepted and fulfills the related integrity constraints embedded into $\boldsymbol{F} / \boldsymbol{B}$ or it is rejected, i.e. it does not satisfy the corresponding validation rules.

First of all we remind the reader of deterministic rules, cf. [WG86], and then turn to statistical and fuzzy edits.

2.1 Edits for deterministic Data

The general underlying assumption of deterministic edits is that all the data of concern are crisp, i.e. are measured or observed with no measurement errors.

This is a quite unrealistic assumption because it means that all variables are error-free. This assumption is mostly made only for the sake of mathematical convenience, and is quite crucial, especially if aggregated data are entered, typically generated in data warehouses or statistical information systems. Note, that if one of those variables is not error-free, the rest will be contaminated due to the relationships between variables.

Conceptually, the simplest edits are those applied to a single field or attribute with respect to: Data Type, Length, Subset Constraints, Scale, and Dimension. They are called "simple edits", cf. [WG86]. Using the syntax (<attribute> <predicate> <value>), the typical simple edits are: (age type integer), (code length 4), (date between [01.00.00-13.08.00]), (size scale cardinal), (cost unit €/year).

Another class of edits is given by logical rules. One way to define them is to use a Fellegi and Holt rule form, i.e. if <premises> then <conclusion>, cf. [FH76] and [MA75]. Let $(x_1,\ldots,x_p)$ and y be a numerical or symbolic vector, $(A_i)_{i=1,\ldots,p}$ and B corresponding sets, and (x_i *is* A_i) and (y *is* B) for i=1,…,p Boolean terms (mostly Horn clauses). Then a logic edit is defined as (*if* x_1 *is* A_1 *and* x_2 *is* A_2 *and*...x_p *is* A_p *then* y *is* B).

As an example, we consider the rule that a pupil $u \in \boldsymbol{U}$ from an elementary school population $\boldsymbol{U}$ should not be married and not be head of a household, i.e. *for all* $u \in \boldsymbol{U}$: *if* le (Age,15) *or* school (elementary) *then not* household_status (head) *and* marital_status (single).

Proceedings of the 18th International Conference on Scientific and Statistical Database Management (SSDBM'06)
0-7695-2590-3/06 $20.00 © 2006 IEEE

The underlying theorem of rule checking is given by the well-known Fellegi-Holt Theorem on "Normal Form Edits", cf. [FH76]. During the last decades a lot of progress was achieved to improve the performance of the Fellegi-Holt rule checking, either using heuristic constraint programming or integer programming techniques, cf. [BGH03], [WC02], [W99], and [GS84].

The third form of edits are numerical edits. They are restricted to numerical data type variables. Consider the following fact: an employee is now twenty-nine years of age (x_1), stayed six years at an elementary school (x_2), stayed seven years at a high-school (x_3), studied five years at a university (x_4), and is employed since two years (x_5).

The corresponding variables have to fulfil the inequality $\underline{a}'\underline{x} \geq \underline{b}$ with $\underline{a}' = (1,-1,-1,-1,-1)$, $\underline{x}' = (x_1,x_2,x_3,x_4,x_5) = (29, 6, 7, 5, 2)$, and $b = 6$.

The general approach is represented by $A\ x \geq b$ (linear numerical constraints), $x \geq 0$ (non negativity), $x \in X = \Pi$ range(xi) for all attribute vectors x.

We close the class of deterministic edits by referring to probabilistic edits, which generalise the former ones just by switching from a Boolean logic defined for an edit to a probability statement for each rule. The syntax is given by f(A,B,C,...)= *false* with Prob(edit) $\geq 1-\alpha$, where $0 < \alpha < 0.5$.

2.2 Edits for non-deterministic Data

Statistical Edits. Now we turn to statistical and fuzzy edits. To the best of the authors' knowledge the idea to embed numerical edits into a Gaussian framework is proposed by [S79], and was fully worked out in [LR91].

As an example we take six business indicators as part of a model *M*:

Capital = 60 ± 1, Profit = 10 ± 2, Sales = 55 ± 20, Expenditures = 45 ± 20, ROI = 0.1 ± 0.05, Margin = 0.5 ± 0.05.

The notation "Capital = 60 ± 1" means, that for 'Capital' a value of 60 is measured, and the measurement error is 1. Evidently, the six variables span the six-dimensional, real data space $\boldsymbol{R}^6$. On this space one linear and three non-linear constraints are defined, which stem from true assumptions on the model *M*:

Profit = Sales - Expenditures, ROI = Profit / Capital, Margin = Profit / Sales, CapitalTurnover = Sales / Capital.

Note, that in the following we assume the correctness of those models.

There is evidence that the data set does not fulfil the given constraints. For example, observe the semantic conflict Margin = 0.5 > Profit / Sales = 0.18. The question arises, whether this conflict can be resolved.

Our modelling is based upon a state space approach, cf. [LR91]. Let be $\underline{x}$ an observed state vector, $\underline{\xi}$ an unobservable (error-free) state vector, $\underline{\zeta}$ a dependent vector $\underline{\zeta} = \underline{H}\ \underline{\xi}$, $\underline{z}$ an observed vector of $\underline{\zeta}$, and $\underline{v}$, $\underline{w}$ vectors of measurement errors.

The state space equation is $\underline{x} = \underline{\xi} + \underline{v}$, and observational equation $\underline{z} = \underline{H}\ \underline{\xi} + \underline{w}$.

The next step is to estimate the unknown parameters $\underline{\xi}$ and $\underline{\zeta}$. This can be simply achieved by a generalized least squares (GLS) estimation approach. Let $\underline{u} = (\underline{v}, \underline{w})$, $E(\underline{u}) = \underline{0}$, and $E(\underline{v}\ \underline{w}) = \underline{0}$ due to partial information from independent data sources. It follows

$$G = E(uu') = \begin{bmatrix} P & 0 \\ 0 & Q \end{bmatrix},$$

where $\underline{P}$, $\underline{Q}$ are covariance matrices of the measurement-error vectors $\underline{v}$ and $\underline{w}$. Let $\underline{y}' = (\underline{x}', \underline{z}')$ and $\underline{J} = (\underline{\boldsymbol{I}}, \underline{H})$, then we estimate $\underline{\xi}$ by solving GLS: min $\| \underline{y} - \underline{J}\ \underline{\xi} \|_{\underline{G}}$ subject to the linear constraint $\underline{\zeta} = \underline{H}\underline{\xi}$. If the variables are connected by multiplication or division then the constraint $\underline{\zeta} = \underline{H}(\ \underline{\xi}\ )$ is to be used, cf. [LR00] and [R01].

We close with an example:

$x_1 = 30 \pm 20$, $x_2 = 30 \pm 10$, $z = 50 \pm 10$.

We assume that all variables have a Gaussian distribution, are jointly independent, and a confidence level of α=5% is used. We get the following estimates:

$$\hat{\xi}_1 = 23 \pm 12, \quad \hat{\xi}_2 = 28 \pm 9, \text{ and } \quad \hat{\xi} = 28 \pm 9.$$

There are several effects of the estimation procedure which are of interest to data cleansing. First, the constraints lead to smaller (or equal) error intervals of the estimates compared with the original measurements. The estimates are shifted into the "right" direction, if a constraint is not fulfilled. The shifts are proportional to the standard deviation of the errors. Estimates are equal to their observations if a measurement is error-free. Finally, the estimates fulfill the system of equations up to numerical errors.

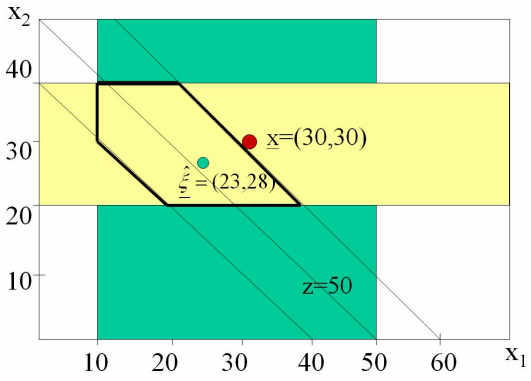

Fig. 1: Constrained Data Space

In Fig. 1 we observe that the boldly bounded area for the intervals of ξ_1, ξ_2 and ζ is non-empty. We call

such a data set 'weak inconsistent'. Strong inconsistency of data occurs, if the area is empty. In such cases no reasonable estimates exist.

Fuzzy Edits. Now we look at Fuzzy Edits. We remember that the basic assumptions of the probabilistic approach presented above were *joint Gaussian distribution* and *linearity of relationships*. It is a disadvantage of statistical edits, that the Gaussian distribution is not closed under non-linear transformations induced by products and ratios of variables. As there exists no finite-parametric distribution which is closed under all four arithmetic operations, an alternative is to use Fuzzy Edits.
The *Extension Principle* of Zadeh, cf. [KGK95], allows all four arithmetic operations on fuzzy variables.
The modeling is performed as follows: Each variable $x_1, x_2, \ldots, x_p$ is treated as a *fuzzy set*, i.e. its degree of membership to a fuzzy set is described by a membership function μ_i: $\boldsymbol{R} \rightarrow \boldsymbol{R}_{[0,1]}$. The impreciseness of measurements is mapped as the length of the support of each fuzzy variable (set), defined as support$(x) = \{x \in \boldsymbol{R} \mid \mu(x)>0\}$, cf. Fig.2.
We assume that all equations are "separable", i.e. each equation can be uniquely solved for each variable existing in this equation.

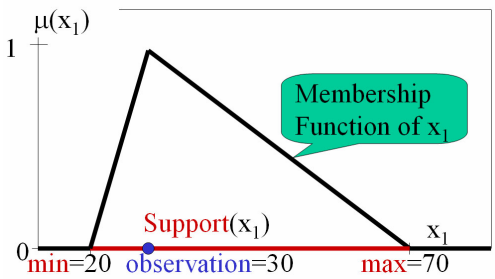

Fig.2: Support(x_1) of a fuzzy set x_1

Let x_1 be the vector of observations (one observation per variable), $\boldsymbol{X}$ the product-space spanned by the range of x_1, $\boldsymbol{F}$ a set of p fully specified fuzzy sets on $\boldsymbol{X}$ and $\boldsymbol{G}(x) = 0$ an algebraic system of non-linear equations $g_1, \ldots, g_q$.
The estimates ξ_{fuzzy} of unobservable variables ξ, which are called parameters, are computed by heuristically solving the following non-linear optimisation problem, cf. [LM00]:

$$\zeta_{fuzzy} = \arg\sup_{x \in X} \| \text{support}(x) \|$$

subject to $x \in \text{support}(x_0) \cap \text{support}(x|g_1) \cap \text{support}(x|g_q)$.
The overlap of supports is computed for the variable x together with its estimates from all equations where this variable shows up.

Comparative Study. In [DP04] the authors derived a theorem saying that the membership function gives an upper bound for a probability distribution. In the following we show the relevance of this theorem for edits. We use a specially designed benchmark data set linked to six business indicators.

Case	Sal	Cost	Cap	Prof	Mar	ROI	CTurn	Mod
1	100±5	80±4	80±4	20±9	0.2±0.1	0.3±0.1	1.3±0.1	FE
				20±6.4	0.2±0.1	0.3±0.1	1.3±0.1	SE
2	100±10	80±8	80±8	35±3	0.32±0.04	0.5±0.05	1.4±0.2/0.1	FE
				37±5.2	0.33±0.04	0.48±0.05	1.44±0.16	SE

Tab. 1: Comparative Study on Fuzzy (FE) and Statistical Edits (SE)

Note, that the intervals contract due to data cleansing, and that all balance equations are fulfilled. While the probabilistic approach (*SE*) produces symmetric confidence intervals, the corresponding intervals using the possibilistic approach (*FE*) are asymmetric.
We demonstrate the effects of varying sizes of the measurement error on the semantic consistency for another data set. While measurements for three variables *Sal*, *Cost*, and *Cap* are at hand, estimates for the fourth variable *Mar* must be imputated. The relation of error rates related to the probabilistic and possibilistic approach is about 1:3.

Case	Sal	Cost	Cap	Mar FE	Mar SE
1	100±5	80±4	80±4	0.2±[0.1, 0.1]	0.2±0.06
2	100±10	80±8	80±8	0.2±[0.18, 0.22]	0.2±0.11
3	100±50	80±40	80±4	0.2±[1.6, 2.0]	0.2±0.44

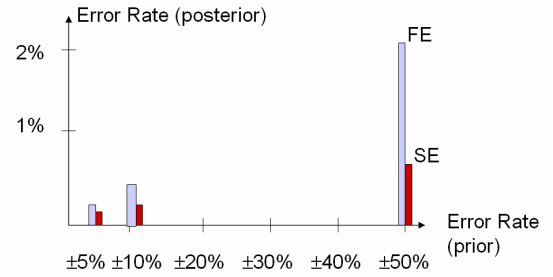

Fig.3: Comparison of Error Rates
– probabilistic and possibilistic approach

We close our discussion of both approaches with respect to strongly inconsistent data. Note, that under a Gaussian multivariate distribution defined on the 7-dimensional data space $\boldsymbol{R}^7$ there exists a solution, i.e. estimated or coherent values, while under the possibilistic approach with finite supports (intervals of finite length) no non-empty overlap exists.

Mod	Sal	Cost	Cap	Prof	Mar	ROI	CTurn
	100±5	80±4	80±4	30±1.5	0.2±0.01	0.4±0.02	?
SE	110±3	85±2.8	72±3	26±0.9	0.2±0.01	0.36±0.02	1.5±0.07
FE	-	-	-	-	-	-	-

3. Conclusion

Let us summarize this study as follows: There exist several classes of edits, which may be very useful. Leaving the deterministic rules aside, we put a stress on edits, which allow for errors in the variables.

We presented two approaches, which are based on a sound methodology: Statistical Edits and Fuzzy Edits. Their assumptions and their interpretation are rather different. While the probabilistic approach needs a Gaussian regime for mathematical convenience, the possibilistic approach considers membership functions and ignores any assumption about dependencies between variables. Due to a theorem of Dubois and Prade [DP04], we illustrated with a benchmark study, that possibility measures are an upper bound of probability measures.

The question arises which effects are caused by the basic assumption of Gaussian distributions in data cleansing. This is investigated by using MCMC simulation based on the Metropolis-Hasting Algorithm in [KL05].

Acknowledgement

The first author thanks R. A. Müller and W. Schürer, DaimlerChrysler Research, for long-term research and development activities in planning, data validation, and controlling.

References:

[BGH03] Boskovitz, A. et al. A Logical Formalisation of the Fellegi-Holt method of Data Cleaning, ANU, TR-ARP-02-03, 10 p, 2003

[DP04] Dubois, D. and Prade, H. Possibilistic logic: a retrospective and prospective view. Fuzzy Sets and Systems 144(1): 3-23 (2004)

[FH76] Fellegi, I. P. and Holt, D. A systematic approach to automatic edit and imputation. JASA, 71, 17-35, 1976

[GS84]Greenberg, B.G. and Surdi, R. A Flexible and Interactive Edit and Imputation System for Ratio Edits. SRD report RR-84/18, US Bureau of Census, 1984

[H93] Hand, D. J., Deconstructing Statistical Questions, read before The Royal Statistical Society, London, 15th December, 1993

[K31] Kolmogoroff, A. N. Über die analytischen Methoden der Wahrscheinlichkeitsrechnung, Mathematische Annalen 104, 415, 1931

[KGK95] Kruse, R., Gebhardt, J. and Klawonn, J. Fuzzy-Systeme (in German), 2nd ed., Teubner, Stuttgart, 1995

[KL05] Köppen, V. and Lenz, H.-J., Simulation of non-linear stochastic equation systems. Proc. of the fifth St. Petersburg Workshop on Simulation, Ermakov, S. M. et al. (eds.), St. Petersburg, 373-378, 2005

[L93] Lenz, H.-J., Contribution to the discussion on "Deconstructing Statistical Questions" by David J. Hand, read before The Royal Statistical Society, London, December 15th, 1993

[LM00] Lenz, H.-J. and Müller, R. M.: On the Solution of Fuzzy Equation Systems, in: Della Riccia, G. et al. (eds.): Computational Intelligence in Data Mining, Springer, Vienna New York, 2000

[LR00] Lenz, H.-J. and Rödel, E., Controlling based on stochastic models, in: Computational Intelligence in Data Mining, Della Riccia, G. et al. (eds.), Springer, Vienna New York, 2000

[LR91] Lenz, H.-J. and Rödel, E., Data Quality Control, Proc. of the Annual Meeting of GÖR, Trier, 1991

[MA75] Mamdani, E.H. and Assilian, S., An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. Intl. Journal of Man Machine Studies, 7:1-13, 1975

[R01] Rödel, E.: Mixed linear regression with equi-cross-correlated errors. Statistical Papers, 2001

[S79] Schmid, B. Bilanzmodelle, ETH Zürich, Zürich, 1979

[W99] Winkler, W. E., State of Statistical Data Editing and Current Research Problems, UN/ECE Work Session on Statistical Data Editing, Working Paper No 29, Rome, 1999

[WC02] Winkler, W. E. and Chen, B.C. Extending the Fellegi-Holt model of statistical data editing. Research Reports Series, Statistics, #2002-02, US Bureau of Census, 2002

[WG86] Wetherill, G. B and Gerson, M. Quality assurance for data entry – an integrated approach, in: preprints of the 3rd Intl. Workshop on Statistical Quality Control, Technical Univ. of Denmark, Lynby, 1986

[Z65] Zadeh, L. A. Fuzzy Sets. Information and Control, 8:338-353, 1965