Data Cleansing at the Data Entry to assert Semantic Consistency

Hans-J. Lenz

Inst. of Statistics and Econometrics
Freie Univ. Berlin, Germany
Email: hans-j.lenz@fu-berlin.de

Abstract: We start with some illustrative examples of data consistency. Then we shortly discuss a set of **primitive edits**, and turn to practically useful, **logical edits** based on the well-known, but quite restrictive Boolean calculus, cf. Boskovits (2008). It is defined on a binary symbolic data space. We pass by **probabilistic edits** which are based on a measurable space (a set and a corresponding sigma algebra) allowing the application of the classical probability theory.

Next, we consider **arithmetic edits** as a (non-)linear equation system defined on a metric data space spanned by all given non-primary key attributes of data records. As an example think of a set of definitions or balance equations like S = Q * P with Sales S, Quantity Q, price per unit P) or F= M * B (force F, mass M and acceleration B). It is traditionally assumed that each data item is crisp.

A much more 'natural' class of validation rules (or semantic integrity constraints) is defined on a multivariate metric space, where we can allow for constraints and/or upper or lower bounds for specific attributes. The main idea is to substitute a crisp value of a datum by any kind of confidence interval characterized by an interval and a measure of confidence of covering the "true" value of an unknown parameter. We call them **statistical or fuzzy edits**. There exist at least two different approaches: Probability Theory mainly under a normal or Gaussian regime (Lenz and Rödel (1991)) or Possibility Theory (Fuzzy Set Theory), cf. Lenz and Müller (1998). There is empirical evidence that both approaches are practically equivalent iff all equations are linear and the cross-correlations between variables are ignorable. This is true because the embedded principles are the same: Folding in Probability Theory (Johnson and Kotz (1972)) and the "Extension Principle" in Fuzzy Set Theory, Zadeh (1965). Note, however, that the interpretation of such operations may strongly vary, cf. Dubois and Prade (2003).

Finally, as very restrictive assumptions are underlying both approaches, we apply MCMC-techniques with special emphasis on the Particle Filter Theory (Metropolis-Hastings simulation algorithm), cf. Chip (2004), Köppen and Lenz (2005) and Köppen (2008).

Keywords: Edits, validation rules, semantic consistency