ASSESSING THE QUALITY OF DATA WITH A DECISION MODEL

Andrew U. Frank

Department of Geoinformation and Cartography, TU Vienna, Gusshausstrasse 27-29/E127, A-1040 Vienna, Austria

frank@geoinfo.tuwien.ac.at

ABSTRACT.

The quality of GIS data cannot be assessed independently of what it should be used for ("fitness for use"); this assessment requires currently human expert interaction. If the user describes how he uses the data to make a decision, a general method for assessing the quality with respect to their decision is possible. The decision process is divided into two phases, the first producing a decision model and the second uses this model to arrive at the decision. The same decision model is then used to assess the quality of the decision as it derives from the quality of data. Examples for engineering and commercial decisions are given.

1. INTRODUCTION

Research in data quality is hindered by a lack of understanding of what quality for data means. The slogan "data quality is 'fitness for use'" is not giving an answer because it leaves open the question to what use the data should be fit. Data, especially GIS data, can be used in many ways; remember that a precursor of GIS was called "multi-purpose cadastre" (Arentze et al. 1992; Harvey 1997)! Data is used to improve decisions; decisions can be made without pertinent information (case of 'hull' information, e.g., none, inappropriate) and decisions are not necessarily changed after data are needed—only confidence is increased (Frank to appear 2007). GIS data can be used to improve many decisions, from ordinary, everyday decisions in wayfinding (left or right here?) to complex decisions about the location of a new nuclear power plant or a new factory or the violation of an international treaty (Abushady 2005).

The quality of the information influences the decision—it must be assessed with respect to the decision making process: can it be used to make this decision, does the lack of quality influence the outcome?

The diversity of the decisions GIS data is used for makes it difficult to understand how the quality of the data affects the decision. This is further complicated by the psychological complexity of how people actually make decisions. A number of studies have shown how data quality propagates from the data stored to data derived from a GIS to help making decisions (Karssenberg et al. 2005). Bruin et al (2001; 2003) investigated whether acquiring better data for a particular decision is worthwhile.

Schneider (1999) and Frank (to appear 2007) have been able to reduce decisions as they are made by engineers when designing technical artifacts to a statistical test. Once the engineer has selected the model and parameters to include, the decision itself can be reduced to a comparison of two desired quantities. This approach is generalized here to as broad a range of decisions as possible.

This approach to data quality from the perspective of a user is different from describing data quality from the perspective of the data producer, working with a specification; typically precision of location (Timpf et al. 1996). Unfortunately, such quality descriptions from the producer perspective are seldom relevant for users of the data (Shyllon et al. 2004).

In this paper I review in section 2 briefly the model for engineering decisions as proposed before (Frank to appear 2007). In section 3 different types of decisions are analyzed. Achatschitz (Twaroch et al. 2005) investigates how the user's situation can be captured separately in an interactive process; the models her work produces can be used to assess the propagation of data quality to decision quality as described here. Ignoring the psychological complexity of decision, especially if made in a group, a similar reduction to a comparison of values devised from the data stored can be achieved. Section 4 then generalizes the model for random errors in the data and section 5 discusses propagation of different data quality aspects from stored data to desired quantities.

As a result, the paper shows a reduced model of decision making, which separates the psychological complexities of taking a decision into a first phase in which the "problem" is conceptualized into a decision test and a model selected. This process is in most decisions not consciously performed or verbalized. In the second phase the decision is computed according to the model selected. It is possible to construct the model used 'after the facts' when the decision is made and one can reconstruct the process. This reconstructed model can then be used to assess how data quality has influenced the decision; which make the method described not only of theoretical interest, but also practically applicable.

With this division of a complex decision into two steps the propagation of data quality can be computed, because error propagation affects only the second one and can be formalized. The paper identifies the processing steps for which propagation of imperfections are necessary and points to the research needed to give general rules for the ones not currently well understood.

A note on terminology: I prefer to speak of *imperfections* of the data (Frank to appear 2007) and to characterize these. This is focusing on the effects such imperfections have on the (imperfect) result, and I avoid statements like "low data quality" or "lack of data quality". All data contains imperfections and it seems conceptually simpler to address these imperfections, rather than talk about data quality, which describes the degree of absence of imperfections.

2. ENGINEERING DESIGN DECISIONS

Engineering design, for example for buildings, bridges, sewage systems, etc. is based on physical observations that are combined in formulae. The results are used to decide if a design satisfies the requirements and is acceptable or not. Error propagation is applicable here and one can ask how much every value computed is influenced by the error in the data. Schneider has analyzed the influence of assumptions about load, strength of materials or required safety levels (Schneider 1999).

In engineering design, decisions can be abstracted to a comparison between the load on a system *S* compared with the resistance of the system *R* as designed. A design is acceptable if the resistance is larger than the load: *R > S resp. R – S > O.*

For a bridge, this means that the resistance *R* of the structure (i.e., maximum capacity) must be higher than the maximally expected load *S* (e.g., assumed maximum high water event). For a more environmental example: the opening under a bridge is sufficient and inundation upstream is avoided when the maximally possible flow *R* under the bridge is more than the maximal amount of water *S* expected from rainfall on the watershed above the bridge. To assess the influence of data quality on the decision, one computes the error on *(R – S)* using the law of error propagation and applies test statistics to conclude whether the value is lager than zero with probability *p* (e.g., 95%).

The law of error propagation for a formula
$r = f(a, b, \ldots)$

for random uncorrelated errors e_a, e_b, e_c on values *a, b, c, …* was given by C. F. Gauss as

$$e_r^2 = \left(\frac{\partial f}{\partial a}\right)^2 e_a^2 + \left(\frac{\partial f}{\partial b}\right)^2 e_b^2 + \ldots \qquad \textit{(Eq 1)}$$

where e_i is standard deviation of value *i*. If the observations are correlated, the correlation must be included (Ghilani et al. 2006). The test on $R – S > O$ is the

$$\frac{R-S}{\sqrt{\sigma_R^2 + \sigma_S^2}} > C \qquad \textit{(Eq 2)}$$

where *C* is determined by the desired significance, e.g., for 95% $C = 1.65$.

In such engineering design decisions a number of poorly known values must be used, e.g., the expected maximum rainfall in the next 50, 100, or 500 years, the maximum load on the bridge, the expected derivation from the plan in the building process, etc. and these may be correlated. The law or standards of engineering practice fix values for them. The accuracy of such general, fixed values to describe a concrete case is low and the effect of these uncertainties in a design decision high. This explains why more precision in observations is rarely warranted, because gains in a reduced construction are minimal. The uncertainties in the assumption about the load dominate the design decision. A rule of thumb for the law of error propagation engineers use is: Error terms that are one order of magnitude less than others have no influence on the result; this is the effect of squaring the standard deviations before adding them! For the formulae used to design an opening under a bridge to avoid inundations upstream, the comparison of the maximally possible flow with the largest flow expected in a period of 50 years gives for an example for $R = 200$ m^3/sec and for $S = 80$ m^3/sec, which satisfies $R > S$. Assuming error in the values used in the computation and propagating then to compute the standard derivation for *R* and *S*, we obtain, e.g., $\sigma_R = 60$ m^3/sec (30%) and $\sigma_S = 16$ m^3/sec (20%), a test at 95% level gives

$$\frac{200-80}{\sqrt{60^2+16^2}} = \frac{120}{62} = 1.93 > 1.65.$$

This design is therefore satisfactory.

Schneider (1999) discusses the selection of security levels, which are traditionally mandated as security factors, increasing the load and reducing the bearing capacity of a design. He shows that current values lead to designs that satisfy expectations, but a statistical viewpoint would result in similar levels of security for different subsystems and therefore a higher overall security level with less overall effort and for a better price.

3. OTHER DECISION SITUATIONS

Navratil has applied error propagation to simple derivations from observed values (Navratil et al. 2004). For example, the computation of the surface area of a parcel given the coordinates of the corners can be computed, if the standard derivations for the observations and their correlations are given. This uncertainty in the area is then sometimes multiplied by the going price per square meter and leads to critical comments by landowners about the quality of a land surveyor's work. The argument is false, because it does not consider a decision. In this section, some often occurring decisions are reformulated in the model proposed above and error propagation applied.

3.1 Decision to acquire a plot of land

The error in the computed area of a parcel (Figure 1) seems high, e.g., some square meters, when one considers the price per square meter one has to pay (i.e., € 550). Would more precise measurements be warranted?

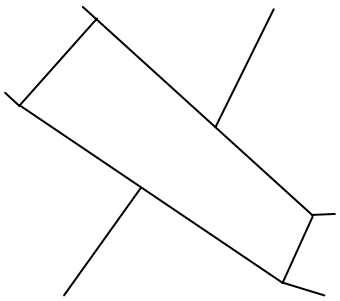

Figure 1: An example parcel

If one rephrases the question as a decision, e.g., whether one should buy the land? This can be seen as a test: are the benefits derived from the parcel larger than the cost? For simplicity

assume that we intend to develop the land and build an office building, where we earn 200 €/ m^2 when we sell it (cost of the construction deduced). The test whether this business opportunity is worthwhile is therefore benefit larger than price ($B > F$) or $B > F > O$ (i.e., anything left after the transaction?). Assuming the standard deviation on the benefit to be $\sigma_B = 0.3 \cdot 1'830'010 = 549'000$ we obtain

$$t = \frac{880'010}{\sqrt{550^2 + 549'000^2}} = \frac{880'010}{549'000} = 1.62,$$

which will occur with probability of ~ 94%. Note that for reason of constructing useful tables it is usual to fix the level of probability and then test, but it is also possible to ask what the probability to a given t value is. In this case, for an acceptable risk of 10%, the decision to buy is acceptable (significance 90%).

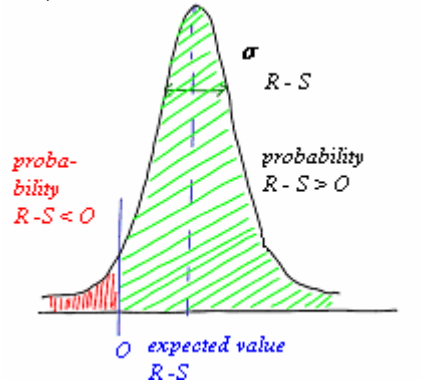

Figure 2: Statistical test for $R - S$

3.2 Find optimal choice

Many decision situations—especially personal decisions—consist of selecting the best choice from several variants. This can be seen as finding the variant with the highest benefit, computed with a formula including weights to indicate the importance of various aspects (Twaroch et al. 2005). For this formula the propagation of error for both data values and for the weights can be computed, using the methods described before. One can determine, with the method shown above, the probability that variant *1* with benefit v_1 and standard deviation σ_1 is indeed better than variant *2* (with v_2 and σ_2 respectively). Achatschitz has proposed to apply sensitivity analysis and inform the user how much his preferences (weights) had to change to make variant *2* to be the best.

3.3 Legal decisions

In a recent count case in Austria, the question was, whether a building was constructed too close to the parcel boundary or not. Abstracting from a number of technical issues of surveying engineering, the distance between boundary and building is established as 3.98 m with a standard deviation of 0.015 m. The law stipulates the required distance to 4.00 m. Is the building too close? A test, for $4.00 - 3.98 > O$ at 95% significance gives

$$\frac{4.00 - 3.98}{1.5} = \frac{2}{1.5} < 1.65.$$

The probability that the distance is shorter is ~ 91%. It depends on the particulars of the case and the judge, if this is considered sufficient evidence or not. I hope that if such cases are approached statistically, the courts will over time develop some standards.

4. OTHER DECISION SITUATIONS

A complex decision process can be split in a phase to select a model to use to make the decision and the phase of using the selected model to arrive at the decision. The discussion of examples in the previous section suggested that the influence of random errors can be computed with the regular error propagation formula if the decision is modeled formally. This section gives a generalized description.

4.1 Model of a decision

By model of a decision we mean the formal model of a particular decision; section 3 gave several examples. In general, a decision can be reduced to a test of a value being positive ($v > O$). The acceptance of an engineering design has immediately the form $R - S > O$, and other "yes/no", "go/no go" decision can be brought to this form. Selection of an optimal solution from a series of variants can be seen as the selection of the variant i with the highest value v_i. It seems easier to describe the two situations separately, but they can be merged into a single approach.

4.2 Binary decisions

A decision to do something or not has a decision model with the test $v > O$ (or can be rewritten to conform to this form). v is computed as a function

$$v = f(a_1, a_2, \ldots a_n, s_1, s_2, \ldots s_n) \qquad (Eq\ 3)$$

of input values $a_1, a_2, \ldots a_n$ describing the situation, which comes, for example, from the GIS, and values describing other factors $s_1, s_2, \ldots s_n$, for material constants, security factors, etc. If $v > O$ the action is carried out, the design built, etc. The influence of random errors in the data ($a_1, a_2, \ldots a_n$ and $s_1, s_2, \ldots s_n$) on the decision is computed by the law of error propagation (*Eq 1*) and a statistical test. From the standard deviations of the data ($\sigma_{a1}, \sigma_{a2}, \ldots \sigma_{an}$ and $\sigma_{s1}, \sigma_{s2}, \ldots \sigma_{sn}$) and the partial derivatives

$$\frac{\partial f}{\partial a_1}, \frac{\partial f}{\partial a_2} \ldots \textit{ of the equation 3}$$

the standard derivation σ_v of v is computed. From v/σ_v results a probability p that $v > 0$

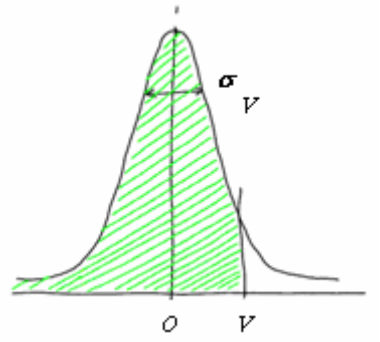

Figure 3: Probability of $v < V$

as the integral of the normal distribution curve with σ_v up to v (Figure 3). Usually a significance level is set first and only checked that v/σ_v is larger than the corresponding value $\Phi(p)$.

4.3 Selection

Selecting the optimal v_O from a set of variants v_i described with data values $a_i = (a_1, a_2, \ldots a_n)$ uses a valuation function f

$u_i = f(a_{i1}, a_{i2}, \ldots a_{in}, .s_1,. s_2, \ldots s_n, w_1, w_2, \ldots w_n)$ *(Eq 4)*

where the data describing the variant, constants and weights describing the importance of the criterion for the decision w_1, w_2, … w_n are combined. For each variant v_i we obtain a value u_i and select the variant v_O for which $u_O > u_i$ for any $i \neq O$.
The effect of random errors in the data on the values u_i is computed as the standard derivations of u_i using error propagation as above (Eq 1). The uncertainty of the selection of v_O compared to v_p (assuming to be the second best) can be tested as $v_O - v_p > O$ and the probability computed (Eq 2). From this, one must separate the test $v_O > O$, which gives the uncertainty that the best variant is an acceptable solution.

4.4 Assumption

Two assumptions must be stressed:

1. The data values are assumed to be influenced by random error, which follow a normal distribution and the errors are not correlated. If correlation is present, it can be taken into account by extensions of the formula for error propagation.
2. The decision model is fixed and not influenced by the imperfections in the data. The decision model includes all aspects of the decision, including subjective elements, which are assumed to be fixed for one decision. Note that for the computation of the effects of random errors in the data only the uncertainty of weights are required, not the exact weights, representing personal preferences.

5. GENERALIZATION: ERROR PROPAGATION IN DECISION FOR RANDOM ERRORS

The law of error propagation applies only to random errors in data values; in this section the approach is generalized to other types of imperfections:

5.1 Omissions and commissions

Omissions and commissions influence the computation of aggregates differently than the 'best solution'.

5.1.1 Aggregate values: If a number of values are summed (generally aggregated) to a single value, the effect of omission is a sum too small, the effect of commission is a sum too high. With given probability p_O and p_C for omissions and commissions, the effect on the sum is $s' = (\sum v_i) \cdot (p_O - p_C)$. This assumes that omissions and commissions are random, not systematic.

5.1.2 Selections: If v_O is the best variant then commissions could invalidate the choice. If p_{Oi} are the probabilities for the data a_i to be the result of commission errors then the probability is the sum of the p_{Oi}, because any single committed value invalidates the selection. For omission the effects are less devastating. A statistical test against a possible better variant not showing due to omissions can be constructed; intuitively, it seems only warranted if the probability of omissions is high.

5.2 Probability of normal and ordinal discrete values

Values not measured on a continuous scale do not have an error distribution, but only a probability to be in error, for example as a confusion matrix (Colwell 1983). For such values in a formula $v = f(a, \ldots)$ (Eq 3) the computation must bifurcate and compute with each possible value a_i (e.g., land use values 'forest', 'agricultural', …) the corresponding v_i and then compute the sum of the products of values times probability $v = \Sigma p_i \cdot f(\ldots v_i)$.

5.3 Fuzzy membership

For linguistic variables, e.g., 'large', sharp boundaries of applicability are impossible, and modeling with a fuzzy membership function is appropriate. Formulae for propagating imperfection modeled with fuzzy membership functions are known (Zadeh 1965). Schneider (1999) has shown that different distinctions for error do not influence the results. Further studies on effects of better models for fuzzy values seem warranted (Viertl 2006).

6. CONCLUSION

Separating the complex decision process in a step to establish a model for this decision and then to apply this case specific decision model to compute the outcome allows to assess the influence of imperfections in the data on the decision, and this achieves a realistic assessment of the quality of data as fitness for a particular use. For the assessment of data "fitness for use" the decision model need not be known with precision; it is only necessary to know the formulae used and plausible values for the weights. Detailed knowledge is not required and users may make decisions 'intuitively' and guided by subjective considerations—the assessment of the quality of the decision as a result of the imperfections in the data set remains valid if the used decision model approximates the personal decision process sufficiently. It is possible, and probably more realistic, to deduce the decision model after the decision has been taken and use it only for the assessment of the influence the data quality has on the decision. A detailed study how decisions in other application areas can be modeled with decision models of this type will show how easily the concept generalizes beyond what was discussed here. The examples studied in (Frank to appear 2007) and (Schneider 1999) give results that correspond to intuition, which is promising. Schneider (Petschacher 1996) has constructed software to treat engineering decisions and extensions to other cases seems possible and is planned as future work. It seems possible to include such tools in general GIS packages in the future.

ACKNOWLEDGEMENTS

Prof. J. Schneider at ETH Zürich taught me as a student about engineering decision; his life-long interest in risk of technical

system inspired the present approach to data quality. Thoughtful reviewers helped to improve the presentation.

REFERENCES

Abushady, A. (2005). *How Can Remote Sensing and GIS Help in the Verification of International Treaties?* RAST conference, Turkey.

Arentze, T., A. Borgers and H. Timmermans (1992). *Geographical Information Systems, Accessibility, and Multi-Purpose, Multi-Stop Travel: A New Measurement Approach*. Proceedings of EGIS '92, Munich, EGIS Foundation.

Colwell, R. N. (1983). *Manual of Remote Sensing*, ASPRS.

de Bruin, S., A. Bregt and M. van de Ven (2001). "Assessing Fitness for Use: The Expected Value of Spatial Data Sets." *Int. Journal of Geographical Information Science* **15**(5): 457-471.

de Bruin, S. and G. J. Hunter (2003). "Making the Trade-Off between Decision Quality and Information Cost." *Photogrammetric Engineering & Remote Sensing* **69**(1): 91-98.

Frank, A. (to appear 2007). *Incompleteness, Error, Approximation, and Uncertainty: An Ontological Approach to Data Quality*. NATO Advanced Research Workshop, Kiev, Ukraine.

Ghilani, C. D. and P. R. Wolf (2006). *Adjustment Computations Spatial Data Analysis*. Hoboken, New Jersey, John Wiley & Sons, Inc.

Harvey, F. (1997). Improving Multi-Purpose GIS Design: Participative Design. *Spatial Information Theory - A Theoretical Basis for GIS*. Berlin-Heidelberg, Springer-Verlag. **1329:** 313-328.

Karssenberg, D. and K. De Jong (2005). "Dynamic environmental modelling in GIS: 2. Modelling error propagation." *International Journal Geographical Information Systems* **19**(6): 623-637.

Navratil, G. and C. Achatschitz (2004). *Influence of Correlation on the Quality of Area Computation*. ISSDQ, Bruck a.d. Leitha, Austria, Department of Geoinformation and Cartography.

Petschacher, M. (1996). Programm VaP for Windows, IBK ETH Zürich.

Schneider, J. (1999). Zur Dominanz der Lastannahmen im Sicherheitsnachweis. *Festschrift zum 60. Geburtstag von Eduardo Anderheggen*, Institut für Baustatik und Konstruktion der ETH Zürich.

Shyllon, E. A. and G. J. Hunter (2004). *Handling Spatial Data Quality Semantics*. ISSDQ'04, Bruck a. d. Leitha, Austria, Department of Geoinformation and Cartography.

Timpf, S., M. Raubal and W. Kuhn (1996). *Experiences with Metadata*. 7th Int. Symposium on Spatial Data Handling, SDH'96, Delft, The Netherlands (August 12-16, 1996), IGU.

Twaroch, F. and C. Achatschitz (2005). *Conceptual Model for a Hotel Seeking Agent*. IWWPST '05 International Workshop on Web Portalbased Solutions for Tourism, Tampere, Finnland, Deaprtment of Geoinformation and Cartography.

Viertl, R. (2006). "Univariate Statistical Analysis with Fuzzy Data." *Computational Statistics & Data Analysis* **51**: 133-147.

Zadeh, L. A. (1965). "Fuzzy Sets." *Information and Control* **8**: 338-353.