ANALYSIS OF DEPENDENCE OF DECISION QUALITY ON DATA QUALITY

Note: Figures will be improved for publication V28

Andrew U. Frank
Institute for Geoinformation and Cartography
Technical University Vienna, Austria
Gusshausstrasse 27-29/127, A-1040 Vienna

Abstract

GIS professionals seem to assume that better data leads to better decisions. Is this true? An analysis to determine the effects of data quality on the quality of decisions reveals relations that are useful to consider before blindly investing in data quality improvement.

This article analyzes data quality and how it influences the quality of a decision with an example of environmental engineering decisions. It shows that the uncertainty in aspects, which are poorly known, e.g., the necessary security levels, dominate the uncertainty of the decision. Efforts to collect more or better date to improve the data quality of the data stored in the GIS would not reduce uncertainty in the decision significantly. This result seems to be consistent with results from other studies for this very large class of decisions and it appears to be further generalizable and useful to investigated other decision situations.

1 Introduction

It is difficult to answer questions about the accuracy of results from spatial analysis. For scientific papers, it is sufficient to indicate the strength of the correlation to demonstrate which factors influence which outcomes, but if the results of spatial analysis are used to make political decisions or to design constructions, more pressing demands for the quality of the results are coming forward and the GI professional must, sometimes publicly, explain the certainty of her results. It is equally difficult, in a time of tight budgets, to justify the high cost of accurate data collection with the need in decision situations.

A critique of GIS could argue that all observations and other data in a GIS must necessarily have some error (Frank 2007b) and lead in combination with other erroneous data to unreliable results. The negative statement “Garbage in, garbage out” is often used to demean the speed of electronic calculation; it assumes a contrapositive of “good data, good results” based on a commonsense belief that the quality of the input data influences positively the quality of the result. This positive folk belief is used to justify projects to improve the quality of some data, e.g., higher precision for cadastral boundary points, assuming that such

efforts will improve the decision following from the data. Is this hypothesis confirmed by observations? There is contrasting evidence that even bad data are useful to help with decisions. Experience demonstrates daily that all sorts of spatial information systems with imprecise and incomplete information are beneficial. We are currently confronted with the public success of Google Earth, which is *not* diminished by the unknown and varying quality of the data. I have argued before (Frank 2007b) that (good) data quality means 'good' enough for this decision; this will be substantiated in this article.

A quest for a description on how input data quality affects the error and accuracy of results from GIS was an important goal already in the original NCGIA program (NCGIA 1989) and its first research initiative (Goodchild et al. 1989). Much detailed, but not conclusive research followed (Goodchild et al. 1998; Heuvelink 1998a; Heuvelink 1998b; Shi et al. 2002; Shi et al. 2003; Frank et al. 2004b; Frank et al. 2004a; Karssenberg et al. 2005; Wu et al. 2005). The equally important research in ontology for geographic information, primarily focusing on the ontology of space, later on geographic space-time, progressed in the philosophical tradition of perfect knowledge of the world (Frank et al. 1991) but not on ontological foundations for data quality. In related articles (Frank 2007b; Frank to appear 2007b) an ontology of *perfect knowledge* is contrasted with a novel ontology of *imperfect knowledge*. The set of realistic commitments for imperfect knowledge of the physical reality give the rational methodology to assess the quality of data and their influence on qualities of decisions used here. The present article demonstrates that the novel combination of analyzing at the same time ontological and data quality issues is producing practical advice for designing GIS, an argument initially made in a conference contribution (Frank 2007a). The example used is the design of a bridge over a small stream, specifically the clearance under the bridge to assure that all the rainwater falling upstream can flow through and no flooding will occur. For engineering design, levels of acceptable risk and who bears it, are fixed by established practice and standards; Agumya and Hunter provided previously a valuable discussion of risk resulting from uncertainty in geographic data and how to deal with (2002).

The article uses the result from the articles on tiered ontology (Frank 2001b; Frank 2003a) and on ontologies for imperfect data (Frank 2007b). In particular:

- point observations are *properties* observable at a point in space-time,
- *objects* are formed as regions with uniform values for some property, and
- objects have *attributes* that are summary descriptions, typically integrals of some property over the region of the object.

The treatment concentrates—similar to the referenced paper (Frank 2007b)—on physical properties and attributes; an extension to socially constructed reality must wait till the corresponding extension of the ontology of imperfection has been achieved.

Boin and Hunter's observation that GIS data is often used to produce other data, which makes assessment of the effects of data quality difficult (Boin et al. 2007) is a partial explanation why methods for error propagation (Heuvelink et al. 2006) are not widely used. Goodchild (2006) points to the academic nature of data quality discussions. Perhaps Frank's ontological clarification (Frank 2007b) defines their "elusive" end user as the user of GIS data who makes a decision that leads to material action in reality. This definition of use and end user of GIS data is used here and allows to overcome the empirically observed difficulty. The next section gives some background to the concept of information quality. Section 3 then gives a detailed description how an engineering design decision is made and introduces an example problem, which is used in section 4 to analyze the effects of data quality on the decision. Section 5 concludes with an argument for an improved statistical approach to decision making and some suggestions for more research.

2 Information Quality

The goal of human activities is to improve ones situation and—following the Golden Rule—to improve the 'condition humaine' in general. This is part of a Greco-Judaic tradition to control the world and use it (Genesis 1, 28). Information became central for the development of economy in the past few centuries. The industrial revolution in the 18th and 19th century improved the efficiency of the production of goods for human consumption and allowed an unprecedented increase in population; it combined improvements in government, taxation, and markets together with technical improvements in manufacturing (North 1981). North identifies a second economic revolution when scientific methods are used to produce systematically new knowledge to further advance technology and management (North 2005). This is evident in the current debate on directing universities to produce 'socially useful and responsible knowledge' combined with high levels of funding for universities but it is equally true for the new internet businesses. Information is today a production factor, comparable to the classical production factors of land, capital, and labor (Ricardo 1817; reprint 1996; Marx 1867; translated reprint 1992; Frank to appear 2007a).

Because information is now a production factor, efforts to include "knowledge" in the accounting of large companies are under way (Schneider 1999). O'Hara and Shadbolt discuss methods to value information in a business administration tradition (O'Hara et al. 2001). The problem of measuring quantity and quality of information has not been solved in general yet. Easily observable and countable substitutes (number of patents, number of scientific publications, etc.), which are expected to be proportional to the actual knowledge, are widely

used and the usefulness of an ontological foundation demonstrated (O'Hara et al. 2001). I have suggested a method to measure the quantity of pragmatic (useful) information (Frank 2001a; Frank 2003b), but the approach is currently viable on a micro level only.

What do we mean when we say that information is of high quality? Before the computer age, one would have said 'the information is from reliable sources', qualifying the information only indirectly by its source. *Quality of information* is a concept only emerging in the 80's. Scientists—especially astronomers and surveyors—commented on the quality of *observations* in the 18th century; surveyors have generalized this approach to evaluate the precision of observations and led to the data quality discussion in GIS (Chrisman 1985; Robinson et al. 1985; Frank 1990). In today's information economy, quality of information becomes important for business and production. Engineers need to know the sources for data they can trust. This trust in the quality of data used is traditionally established through accumulating experience over extended periods. In today's global economy, experience has to be replaced by rules (Dueck 2006) to allow faster evolution. Business processes using data go astray if the inputs are wrong and this gives an alternative approach to the question of data quality (Wand et al. 1996). The loss for U.S. businesses due to data quality problems is estimated as $600 billion for 2002, which is about 5% of GDP (Eckerson 2006).

3 Reasoning in Environmental Decision Making

This article uses a case as a running example to demonstrate the approach before progressing to a generalization. I investigate the arguments related to a decision, because the importance of data quality for decisions can only be understood in an 'end-to-end' analysis. This design of a bridge over a small stream provides a realistic, but necessarily much simplified, example from an environmental project. Essentially the same example can be found in many text books, e.g., by Keller (2007), which demonstrates that this is indeed a prototypical design decision.

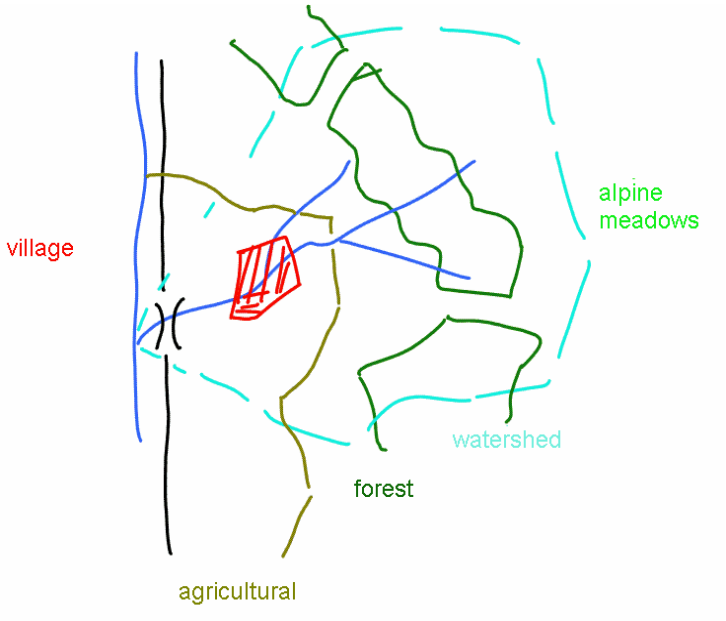

Figure 1: The situation: a bridge over a small stream

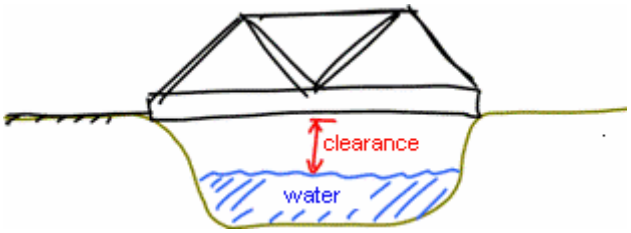

Figure 2: Cross section of opening under bridge

3.1 Example design problem

Protection against flooding has become increasingly important; an engineer must check that bridges crossing a river leaving enough clearance for the water to flow under the bridge even after heavy rainfall. If the clearance under the bridge does not leave a sufficient cross section for water to flow, water is backed up and leads to flooding upstream. The engineer is required to compare the maximum flow *D* possible to pass under the bridge against the assumed maximum quantity *Q* of water flowing in the stream after a heavy rainfall. If the maximum flow *D* under the bridge is more than the quantity *Q* of water following a heavy rainfall, then the area upstream is safe of floods from water backing up upstream from the bridge.

Engineers compute the two quantities *D* and *Q* separately, combining three different models:

- to estimate the maximum intensity of rainfall to be expected,
- to compute runoff after such a rainfall, and
- to calculate the quantity of water flowing through a channel.

> "The standard engineering practice for quantifying the risk of flooding requires that a design storm be selected, that a hydrologic model be used

> to calculate the peak flow runoff generated by the design storm, and that a hydraulic model be used to calculate the maximum height of water at a particular location." (Topanga 2006)

3.2 Assessment of Situation and Selection of Model

It has been observed that most engineering decisions can be brought to the form $D\text{-}Q>O$, (Schneider 1999; Frank 2007a), which leads to a statistical test for $D\text{-}Q$. Crucial for the design and the use of data is the appropriate formalized model—the translation of reality in a model. What are the objects interacting? What are the relevant properties and how do they interact? What is the event that endangers lives and could damage properties? What is the mechanism that causes it? In this case, the risk of flooding caused by heavy rain is identified and the water quantity after rainfall and the maximum flow under the bridge are the relevant quantities to compare. Formulae are derived from the model and link these quantities to observable values, which are stored in a GIS and used in the decision.

3.3 Accepted Risk

Engineering design decisions, like other decisions in life, are never 100% free of risk; engineers can construct systems with an arbitrary small risk of failure at a corresponding price. The lower the accepted residual risk the higher the price of the construction. The difficulty in the public debate is (a) to quantify the risk and the expected damage and (b) the non-correspondence between the persons benefiting from the reduced cost of the construction—typically the owners—and the person, suffering the eventual damages—typically the public. In a political process the lawmaker fixes a required level of security against damage the construction must fulfill. An economic optimum is achieved if the different system with which we interact provides the same marginal security level, i.e., the cost of reducing the risk is for every subsystem similar (Schneider 2000)

The risk of flooding is expressed as a recurrence interval. A probability that a flood occurs is fixed as the accepted risk, for example, 1% but traditionally expressed as 100 year recurrence interval. If a flood endangers very sensitive areas where human life may be lost a smaller risk is ordered by law or building codes, if the area that would be flooded contains only minor, or no buildings, a higher risk, e.g., a 30 years recurrence interval (~3% probability) is acceptable. It is worth noting that a 100 year recurrence interval does not mean that such an event will occur every 100 years, or only once in 100 years.

Engineers do not attempt to predict precisely the future: "Our interest is not in the impossible task of predicting the date of a storm event that exceeds the design capacity of a practice or structure. Rather, we need to know the likelihood of occurrence (probability) of an event with a specified intensity and duration. This is called its return period or frequency of occurrence" (Huggins 2006).

3.4 Design storm

The largest storm that occurs at most once in the selected recurrence interval is obtained from tabulations of past events. This gives the "design storm" for which the bridge and other constructions are designed. The design storm gives the rainfall intensity with which the designed clearance of the bridge must cope.

The rain intensity i_{50} for a recurrence period of 50 years at a given location is found in a table as, for example, 5mm/hour. This means that the probability of a rainfall of a fixed minimal duration more intense then 5mm/hour will occur in any year with the probability 1/50 (USDAForestService 2006). It does not mean that it will occur only in 50 years. The probability that such an event occurs before is 64% (Widmoser 1976). (For the sake of brevity, I have left out a discussion of the dependency of the rainfall intensity on the length of a rain and the determination of the relevant duration). Rainfall intensity for different recurrence intervals are tabulated. Statistics of observations at a number of points over a long period of time are condensed to the tables engineers use. The recurrence interval is a way to describe the probability for a rainfall of a certain intensity to occur, the longer the interval, the larger is the exceptional rainfall.

3.5 Runoff calculation

With the determined rainfall intensity, the peak runoff is calculated using the so-called *rational formula* (ems-i 2006):

$$Q = \frac{1}{3600} CiA \qquad \textit{(Eq. 1)}$$

where:

Q - peak flow (m^3/s).
C - dimensionless runoff coefficient.
i - rainfall intensity (m^3/s, mm/hr).
A - catchment area (m^3/s, ha).

The runoff coefficient models the infiltration into the ground, which reduces the overland flow. Infiltration rates are highly correlated with the land use and are tabulated for engineering use as runoff coefficients (McCuen 1989). Todini (1988) gives a survey of the history and more details about rainfall runoff modeling.

The rainfall intensity is the value for the design storm (numerous simplifications are again necessary to keep the example short). For the runoff calculation the watershed must be identified as an object and its area computed; the watershed boundary is derived from terrain height observations, from which terrain aspect (i.e., the direction of maximal descent) follow. The determination of the watershed is performed in a raster or TIN model of the terrain, interpolated from available observations (Frank et al. 1986); Fisher and Tate (2006) discuss the effects of errors in elevation data on derived data, e.g., watersheds.

3.6 Maximum flow under bridge

An approach to calculate the maximum flow under a bridge uses Manning's formula:

*Discharge = Area * Velocity.*

The velocity is calculated as

$V = 1/n\ (R^{2/3})(S^{1/2})$

and thus

$$D = A \times V = 1/n\ A^{5/3}\ p^{-2/3}\ S^{1/2} \qquad \text{(Eq. 2)}$$

where:

A = channel cross-sectional area
D = Discharge
V = average flow velocity (meters/second)
n = roughness coefficient
S = channel slope
R = hydraulic radius (meters) = A/P
P = wetted perimeter

(Keller et al. 2007)

The factors in this formula are obtained from observation, e.g., terrain height differences gives the channel slope, the hydraulic radius is obtained from measurements of the channel cross section, the roughness coefficient is estimated from observation of the channel properties and then looked up in tabulation of data obtained from experiments.

3.7 Decision whether design is acceptable

A design of a stream cross section under a bridge is acceptable if $D > Q$; for the assumed situation, the engineer calculated D as $D = 90$ m^3/s and $Q = 80$ m^3/s; this design satisfies $D > Q$. According to standard practice, this design is considered safe at the assumed level (in this case a 50 year recurrence interval for the design storm). The next section discusses how much this decision is influenced by imperfection in the data.

4 Influence of Data Quality

The engineering decisions depend on the quality of the input data. Engineering practice uses security factors to increase loads and to reduce bearing capacity to account for imperfections in the model and the data (Schneider 2000). These factors accumulate experience with the practice that results from a technology level and the available data sources.

The GIS contributes data to engineering decisions and the attention in the GIS community has been mostly on the quality of environmental data, assuming that better GIS data would lead to better decisions. The detailed assessment of this assumption by de Bruin (2001) showed that improvements in the measurements contribute little to less expensive designs and are therefore not economically justified.

4.1 Selection of Model

Abstracting the complex reality to a formalizable model is the crucial step in engineering and other decisions. What is to be included as relevant influence? What can be left out? It seems that most errors in engineering leading to accidents are caused by selection of an inappropriate model and leaving out relevant factors. This step is, however, not directly influenced by data quality. An indirect influence may result from lack of data available and forcing the use of less appropriate data and models. This is why metadata, i.e., data describing the data, is so important. The use of data from the web, often optimistically described, and of which is effectively less known than about data the decision maker has collected herself increases the danger of errors in the data selection.

Several types of errors can be hidden in the model selected. In the example case, the maximum runoff could be due to snow melting in spring and not due to rainfall or a retention basin could have been constructed upstream. In both cases a different model is required, considering snow melting or retention effects. The classification of the situation gives the model and the formulae used, which in turn lists the objects to identify and the properties of them relevant for the decision.

4.2 Properties Observed

The attributes relevant for object formation (Frank 2003a; Frank 2007b) and the properties of objects necessary for the decision follow from the model and the selected formulae (Achatschitz 2006). If values for these properties are not available, engineers use observations that are available and strongly correlated with the quantities required. For example, the infiltration coefficient is derived from land use, the roughness coefficient is not measured, but the visual appearance of the channel is sufficient indication to give ranges for the value to use (Table 1).

	Values
Smooth, open stream channels with gravel bottoms	around 0.035-0.055.
Very winding, vegetated, or rocky channels	around 0.055 to 0.075
Smooth earth or rock channels	0.020 to 0.035

Table 1: Roughness coefficient *n* (Keller et al. 2007).

4.3 Incompleteness of the Data

Engineering decisions require data about a very specific part of the world; the lack of data for other areas does not affect the decision. Data collected about past states of the world are fine for engineering decisions, because engineering is based on (physics and chemical) natural laws, which are valid independent of time (Feynman 1998). Problems arise if social forces affect input factors; for example urbanization increases the runoff coefficient.

4.4 Level of Detail

The observations of continuous variables like terrain height and infiltration coefficient or rainfall intensity are available only at some sampling points; values between these points are interpolated. For example the values for a 50 year rain tabulated for the counties in South Carolina varies from 6.65 to 8.9 inches for 24 hours (Purvis et al. 1988), which indicates that the variations are smooth and interpolation justified and not much increasing the error already present in the data.

4.5 Object formation

The abstraction to a model cuts from the continuous world determined objects: rainfall events, watershed, areas of similar land use/land cover, a stream with a channel. Each of these objects is based on a property that changes rapidly at the boundary of the object and is uniform inside; the object formation is influenced by errors in these observations (Frank 2007b). The relevant object properties are then computed for the object region from other observed properties again influenced by error. The watershed is derived from height observations typically in a regular raster (Digital Elevation Model Digital Terrain Model), where the error in height results in errors of the height observation and the sampling density (Fisher et al. 2006).

After forming the watershed, the interesting quantities are the integral of the rainfall intensity multiplied by the runoff coefficient (rational formula):

$$q_R = \int_A i(a) \cdot c(a)\, da \qquad (Eq.\ 3)$$

This can be simplified for small areas where rain intensity is assumed constant and the runoff coefficients are assumed uniform for regions of uniform land cover. The runoff coefficient is tabulated for land cover classes; regions of uniform land cover are formed (for example as in Figure 1) and the product of area times local runoff coefficient is summed:

$$q_R = k \cdot i \cdot \sum_A c(a) \cdot a\,. \qquad (Eq.\ 4)$$

4.6 Observation Error

The observation error and its influence on the quantities computed for a function $r = f(u,v,w)$ is well known. Under the assumption that observation error for *u*, *v*, *w* is random distributed with the standard deviations σ_u, σ_v, σ_w respectively and uncorrelated, the standard deviation for the error on *r* is given by the formula for error propagation:

$$\sigma_r^2 = \sigma_u^2\left(\frac{df}{du}\right)^2 + \sigma_v^2\left(\frac{df}{dv}\right)^2 + \sigma_w^2\left(\frac{df}{dw}\right)^2 \qquad (Eq.\ 5)$$

If the error is described by a percent error (sigma/value) then error propagation for product formulae become (Physicslabs 2006):

$a = b \cdot c$ *with* σ_a, σ_b, *and* σ_c *gives*

$\sigma_a^{\,2} = c^2 \cdot \sigma_b^{\,2} + b^2 \cdot \sigma_c^{\,2}$ *which simplifies to*

$$\frac{\sigma_a^{\,2}}{a^2} = \frac{\sigma_b^{\,2}}{b^2} + \frac{\sigma_c^{\,2}}{c^2}.$$

4.6.1 Error on design storm

The design storm for a 50 year recurrence period is at this location i_{50} = 12 m^3/s. This converts to a percentage or (σ_i/i); the mean maximum annual flow for this river assumed to be 10 m^3/s with a σ of 2 m^3/s, which gives for a 2% probability (z=2.05)

$i_{50} = i + 2.05 \cdot \sigma_i$

$i = i_{50} - 2.05 \cdot \sigma_i = 12 - 2.05 \cdot 2 = 10\ m^2/sec.$

Note that the value for i to be used now is 8 m^3/s, i.e., less than the value used before; the difference is the security margin. The desired security level of 2% will be reintroduced later.

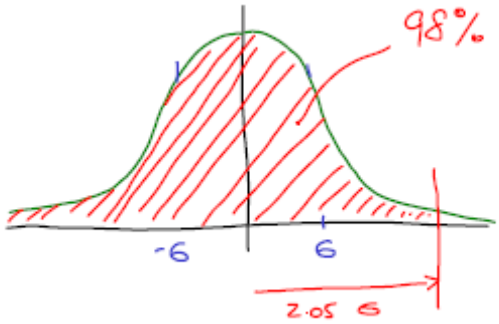

Figure 3: The 50 year recurrence means $v > 2.05 \cdot \sigma$

4.6.2 Error on maximum runoff

Runoff is calculated with the rational formula $Q = k \cdot C \cdot i \cdot A$. The rainfall intensity is taken from a table that is based on statistics from past rain events measured at locations in the region; the value for the area of interest is interpolated. The error in these tables is difficult to assess, but as observations increase, new tables are computed; comparing older tables with newer values gives some indication on the probable error (Purvis et al. (1988)give a short and drastic account) where an increase by a factor up to 3 for rainfall intensity is documented for new tables appearing 1953. I use here p_i = 20%.

The error on the runoff coefficient c is estimated as p_C = 25%, equivalent to 1/4 of the interval of the tabulated values. The error on the area—resulting from determination of the watershed—as is assumed as p_A = 5%. This gives the total error on the value for Q p_Q = 56%.

$$p_Q \approx \sqrt{p_C^2 + p_i^2 + p_A^2} \approx 32\%$$
$$p_C \sim 25\%$$
$$p_i \sim 20\% \qquad \text{(Eq. 6)}$$
$$p_A \sim 2\%$$

4.6.3 Error on maximum flow under bridge

The calculation for the maximum flow uses the formula:

$$D = 1/n\ (A^{5/3})\ (p^{-2/3})\ S^{1/2} \qquad \text{(Eq. 7)}$$

by error propagation we obtain:

$$p_D^2 = p_m^2 + \left(\frac{5}{3} p_A\right)^2 + \left(\frac{2}{3} p_p\right)^2 + \left(\frac{1}{2} p_s\right)^2 \qquad \text{(Eq. 8)}$$

The errors for the geometric term area *a*, wetted perimeter *p* and slope *s* are assessed to 2% each and uncorrelated. The roughness coefficient is less precisely determined and changes over time through growth of water plants; it is estimated as a quarter the tabulated interval, i.e., 10%.

$$p_D \sim \sqrt{p_n^{\,2} + \frac{25}{9} \cdot p_A^{\,2} + \frac{4}{9} \cdot p_P^{\,2} + \frac{1}{4} \cdot p_S^{\,2}} = 13\%$$
$$p_A \sim 2\%$$
$$p_n \sim 10\% \qquad \text{(Eq. 9)}$$
$$p_p \sim 2\%$$
$$p_s \sim 2\%$$

One observes that in both cases (Eq. 6 and Eq. 8) the error in the result is dominated by the largest error term, which is the error associated with the estimated rain intensity and the roughness of the channel. As a rule of thumb, an error that is less than one tenth of the largest error does not influence and can be ignored in the computation. As a consequence improvements in the quality of data have effects only for terms with large errors.

4.7 Influence of Imperfections in the Calculation

One must also check for imperfection introduced in the calculations. Engineers calculate exactly with the values obtained and the decision is based on a comparison desired vs. obtained and is again sharp. Typically 3 significant digits are used to make sure that rounding errors in the calculation do not influence the result (in these examples only 2 digits are given not to make the assumed values appear more meaningful than they are). This approach guarantees that for ordinary design errors, round off error does not influence the result. Calculating with more digits would —for ordinary cases—not improve the decision.

4.8 Error in the Decision

It is standard practice in engineering, to design structures such that the necessary resistance (strength) is more than the load $R > S$ or $R - S > 0$. In this case, this translates to a

comparison of the amount of water *D* that may flow under the bridge with the assumed maximal rainfall *Q*. If $D > Q$ the design is approved, if $D < Q$ then it is rejected.

If we want to analyze the influence of data quality on the decision, we must test $D - Q > 0$ with a statistical test. Selecting a level of probability of 98%, which corresponds to the 50 year recurrence interval, we obtain the test value:

$$\frac{D'-Q'}{\sqrt{\sigma_D^2 + \sigma_Q^2}} > 2.05 \qquad \text{(Eq. 10)}$$

The calculation uses the assumed values for *D* and *Q* and the standard deviations (a simplification using the error percentage as above is not possible). Standard engineering tests compare values, which include security margins and factors to allow for the recurrence period. The maximum flow under the bridge is underestimated to be safe and the design storm is for a 50 year recurrence period (2% probability). If we use a statistical test, then these security margins have to be removed first:

$$\begin{aligned} D &= D' + \sigma_D \cdot f \\ Q &= Q - \sigma_Q \cdot f \end{aligned} \qquad \text{(Eq. 11)}$$

The security factor is assumed to 2.05 for a 98% security level (for simplicity assumed normal distribution). The security factors reduced the maximum flow *D* passing under the bridge and increased the maximum flow after a rain storm. This gives

$$\begin{aligned} D' &= 90/(1-0.13) = 125\, m^3/s \\ \sigma_D &= 0.13 \cdot 125 = 17 m^3/s \\ Q' &= 80/(1+0.32) m^3/s = 48 m^3/s \qquad \text{(Eq. 12)} \\ \sigma_Q &= 0.32 \cdot 48 = 15 \\ t &= \frac{125-48}{\sqrt{17^2+15^2}} \cong \frac{77}{23} \sim 3.32 > 2.05 \end{aligned}$$

This design is safe at the desired level. The above calculations indicate, despite very substantial simplification that are necessary to keep the example short enough for a single journal article, that the safety margins are probably not equally distributed and a less expensive design with equal safety margins for each partial model might be achievable. Schneider investigated the design of a load bearing beam and comes to similar conclusions (Schneider 1999; Schneider 2000). This indicates an area for research for engineering science and statistics.

The calculations show clearly that the uncertainty in the data from the GIS does not significantly contribute to the uncertainty in the decision The area of the watershed has minimal influence; if the areas for different land covers are extracted to compute the runoff coefficient for the watershed with different runoff coefficient for different areas (Eq. 4) then another (usually small) improvement results. More influence results from the imperfection in

the rainfall intensity—which could also be stored in a GIS—and improvement of such data could improve the decision. The analysis has shown that a differentiated consideration is necessary and a general drive to improve all data is not justified.

4.9 Improve the Quality of the Decision

The calculation of error propagation shows which observation influences the quality of the result substantially; if a decision has important economic effects then one may decide to collect better data. If, for example, the runoff coefficient for the watershed seems high and a much less expensive bridge design would be feasible, if the runoff coefficient would be known with more precision, then actual rainfall events can be observed and the relation between rainfall intensity and runoff quantity is established and compared with the computed values. Such observations calibrate the model and give an empirical observation of the product of runoff coefficient times watershed area—errors in both of these terms are corrected. Such additional observations are usually costly, introduce their own data quality problems, and are only done if considerable cost savings in the construction may result from better data. The same argument applies also to efforts to improve the quality of the decision.

Above I assumed uncorrelated observations, which may be unjustified. For example, regarding the runoff coefficient, some observations indicate that it might be variable and approach for intense rain ~ 1.0, i.e., its value may be correlated with the rain intensity (Haupt 2000). A refinement with more precise data must assure that the selected model is detailed enough and does not ignore effects that are relevant if the data are more precise.

5 Conclusion

The example analyzed here is typical for engineering decision process; the design of a bridge, a building, a road all follow this pattern, leading to the comparison of two figures: expected event vs. resistance of the design. If the object designed is small or the decision of little consequence, then many shortcuts are taken, based on ‘rules of thumb’ and experience. Different kinds of engineering decisions require different statistical treatment, given that the distribution of the events considered are different; often occurring are Poisson or Gumbel distribution. The focus of this paper was on the generic principle and normal distribution was assumed for simplicity everywhere without falsifying the general message of the paper.

An engineering decision process is unlike a scientific argument or a legal argument (Lehmann et al. 2006) a probability argument: an engineer designs and builds structures that have the intended function and achieve the intended goals nearly always; it is technically impossible to build “100% safe” systems, because neither the actions of the human operators nor the environmental situation is completely predictable. Engineers strive for an economic optimum in the design: The system designed should work nearly always and the effect of

failures less costly than preventive measures to avoid them. The "state of the art" of engineering consists of rules of thumb, tables in which current knowledge derived from past experience is structured for use in design.

One could conclude that measurement precision is not important in engineering. The influence of errors in the measurement on the decision varies depending on the design problem. To assess the strength of a bridge under load, very accurate (better than 1/10 mm) observations of the deflection under known loads are necessary and useful because the models are very precise and the deflections small. For runoff calculation, accurate position information on the watershed boundary are not contributing much because the other factors (e.g., runoff coefficient) cannot be determined accurately everywhere and the model used is highly simplified. Experienced engineers know what observations with which precision are necessary. Problems arise when data collected for one application are used for another one, which is the typical GIS situation increasing further with interoperability.

The result of such engineering calculation include a number of parameters that are selected based on subjective judgments, experience, etc. Some engineer will include more factors, consider aspects others would not think of (e.g., the possibility of a fire occurring before a heavy rainfall and the influence of burned areas after a forest fire on the runoff coefficient). It is not surprising that different experts arrive at different predictions:

> "in a 1985 DPW .. estimated the peak flow at the mouth of Topanga Canyon at 15,200 cfs for a 100 year event, while only six years later they estimated the peak flow to be 20,600 cfs for a 50 year event." (Topanga 2006)

As far as possible, building codes fix judgments and engineers must use these assumptions and can replace them with their own better judgment only in very special cases. This is especially true for designs that could endanger the lives of others where the public fixes acceptable levels of risk (e.g., for fire protection). It is economically optimal to achieve everywhere similar marginal levels of protection compared to the cost of achieving this level of security; the building codes fix the security relevant design parameters (e.g., 50 year flood, assumed loads for bridges, etc.) and represent an effective assessment of a complex situation, developed over time. Even documented risk aversion (Wikipedia 2007) may be a rational development to compensate for the usually ignored high social cost of large accidents. The question how engineering and legal decisions combine remains open; this affects GIS as data quality decisions for GIS are often closely related to liability. From a legal perspective, security factors are sharp and have no error; if a design is checked in court, no tolerances apply. The example shows that a comprehensive assessment of the security level of the decision $D - Q > 0$ is underestimated if the quantities for D and Q are computed with individual security margins Schneider (2000); this is a challenge for engineering research.

GIS data is typically used to produce other data and it is difficult to observe directly the effects of data quality on decisions. The novel contribution in this article is to compare the effects of error and uncertainty in the inputs for a decision to construct a building not the production of data or a decision to acquire more data. de Bruin et al (2001) investigated whether efforts to improve the data would result in reduced cost to carry out a building project or not. Their result is comparable but their discussion is only indirectly related to the building decision. Improving the data qualities does not usually pay off.

It is sometimes argued by data providers that they have to provide the highest possible data quality because the data could be used for other purposes and they could be held liable. Cases tried in court are rare. In a 2006 decision the *House of Lords* (the highest court in the UK) argues very convincingly against such a liability (Lords 2006). In this case the provider of a chemical water analysis produced for agricultural use of the water did not observe variables that were years later determined crucial for the use of the same water for human consumption; this led to health problems and loss of life, but the court decided that the original data collector was not reliable for the later, not foreseen and not intended use of the water for human consumption (Attaran 2006).

Geographic data used for engineering or administrative decision making is usually collected with proper levels of quality to make the intended decisions. Over time and with experience an optimum is reached between the cost of improved data quality through more efforts when collecting the data and the cost of correcting errors in the decisions due to errors in the data. If geographic data is used for purposes it was not originally intended, for example using administrative data for environmental planning, the particulars of the quality of the data for this decision must be considered carefully.

The article shows a method, with which one can, firstly, reconstruct the likely quality of the data collected considering the original decision process and then, secondly, compute the influence of imperfection in the data on the intended decision.

It may surprise, that data of unknown quality—even data with noticeable low quality—can be used to make decisions. Most decisions are very tolerant against error in the input and human decision makers are very experienced to cope with imperfections in the data. For example, Google Earth is widely used and has attracted attention in media beyond what was ever possible for GIS—despite a very limited offer of data and very variable quality of the data. One might conclude that availability of data is more important than quality of the data, which is not surprising, as it is a logical truism. The framework presented here may eventually be used to search automatically for data of suitable quality for a decision and advise or warn users about the uncertainties in the result on which they base their decision.

Future work should clarify of the risk remaining; especially the differentiation in first and second kind of errors (α- and β- error) in statistical hypothesis testing. Last, but not least, remains the extension to make the method displayed here to apply to decision in the social or legal realm. This will require methods to assess the quality of semantic classifications and how it is used.

Acknowledgement

I owe enormously to the engineering education I learned from Prof. Jörg Schneider at the ETH Zürich in courses on building statics and construction. This article 30 years later is a tribute to the quality of his teaching! I appreciate critical comments from my colleagues Gerhard Navratil and Claudia Achatschitz.

References

Achatschitz, C. (2006). *Preference Based Retrieval of Information Elements*. 12th International Symposium on Spatial Data Handling, Vienna, Springer.

Agumya, A. and G. J. Hunter (2002). "Responding to the Consequences of Uncertainty in Geographical Data." *International Journal Geographical Information Systems* **16**(5): 405-417.

Attaran, A. (2006). "Will Negligence Law Poison the Well of Foreign Aid? A Case Comment on: Binod Sutradhar v. Natural Environment Research Council." *Global Jurist Advances* **6**(1, Article 3.).

Boin, A. T. and G. J. Hunter (2007). *What Communicates Quality to the Spatial Data Consumer?* Proceedings of the 7th International Symposium on Spatial Data Quality (ISSDQ 2007), Enschede, The Netherlands.

Chrisman, N. (1985). An Interim Proposed Standard for Digital Cartographic Data Quality: Supporting Documentation. *Digital Cartographic Data Standards: An Interim Proposed Standard*. H. Moellering. Columbus OH, National Committee for Digtial Cartographic Data Standards. **6**.

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.

Dueck, G. (2006). *Lean Brain Management Erfolg und Effizienzsteigerung durch Null-Hirn*, Springer.

Eckerson, W. W. (2006). "Data Warehousing Special Report: Data quality and the bottom line." Retrieved 08.08.2006, 2006, from http://www.adtmag.com/article.aspx?id=6321&page.

ems-i. (2006). "Hydrologic Models - Basic Equation." Retrieved 09.21.06, 2006, from http://www.ems-i.com/wmshelp/Hydrologic_Models/Models/Rational/Equation/Basic_Equation.htm.

Feynman, R. (1998). *The Character of Physical Law*. Cambridge, Mass., The MIT Press.

Fisher, P. F. and N. J. Tate (2006). "Causes and Consequences of Error in Digital Elevation Models." *Progress in Physical Geography* **30**: 467-489.

Frank, A. U. (1990). Qualitative Spatial Reasoning about Cardinal Directions, University of Maine, NCGIA.

Frank, A. U. (2001a). The Rationality of Epistemology and the Rationality of Ontology. *Rationality and Irrrationality, Proceedings of the 23rd International Ludwig Wittgenstein Symposium, Kirchberg am Wechsel, August 2000*. B. Smith and B. Brogaard. Vienna, Hölder-Pichler-Tempsky. **29**.

Frank, A. U. (2001b). "Tiers of Ontology and Consistency Constraints in Geographic Information Systems." *International Journal of Geographical Information Science* **75**(5 (Special Issue on Ontology of Geographic Information)): 667-678.

Frank, A. U. (2003a). Ontology for Spatio-Temporal Databases. *Spatiotemporal Databases: The Chorochronos Approach*. M. Koubarakis, T. Sellis and e. al. Berlin, Springer-Verlag**:** 9-78.

Frank, A. U. (2003b). Pragmatic Information Content: How to Measure the Information in a Route Description. *Perspectives on Geographic Information Science*. M. Goodchild, M. Duckham and M. Worboys. London, Taylor and Francis**:** 47-68.

Frank, A. U. (2007a). *Assessing the Quality of Data with a Decision Model*. 5th International Syposium on Spatial Data Quality 2007, Enschede, NL.

Frank, A. U. (2007b). Ontologies for Imperfect Data in GIS. Vienna**:** 22.

Frank, A. U. (to appear 2007a). A Case for Simple Laws. *The Mystery of Capital and the Construction of Social Reality*. B. Smith, I. Ehrlich and D. Mark, Springer? 512.

Frank, A. U. (to appear 2007b). *Incompleteness, Error, Approximation, and Uncertainty: An Ontological Approach to Data Quality*. Geographic Uncertainty in Environmental Security. NATO Advanced Research Workshop, Kiev, Ukraine, Springer.

Frank, A. U. and E. Grum, Eds. (2004a). *Proceedings of the ISSDQ '04 Vol 1*. Geoinfo Series. Vienna, Austria, Institute for Geoinformation.

Frank, A. U. and E. Grum, Eds. (2004b). *Proceedings of the ISSDQ '04 Vol 2*. Geoinfo Series. Vienna, Austria, Institute for Geoinformation.
Frank, A. U. and D. M. Mark, Eds. (1991). *Cognitive and Linguistic Aspects of Geographic Space*. NATO ASI Series D. Dordrecht, The Netherlands, Kluwer Academic Publishers.
Frank, A. U., B. Palmer and V. Robinson (1986). *Formal Methods for Accurate Definition of Some Fundamental Terms in Physical Geography*. Second International Symposium on Spatial Data Handling, Seattle, Wash.
Goodchild, M. (2006). *Preface*. ISTE 2006.
Goodchild, M. and R. Jeansoulin, Eds. (1998). *Data Quality in Geographic Information - From Error to Uncertainty*. Paris, Hermes.
Goodchild, M. F. and S. Gopal, Eds. (1989). *The Accuracy of Spatial Databases*. Basingstoke, Taylor & Francis.
Haupt, R. (2000). *Regionalisierung von Hochwasserkennwerten in Mecklenburg-Vorpommern*. Rostock, Universität Rostock.
Heuvelink, G. B. M. (1998a). *Error Propagation in Environmental Modelling with GIS*. London, Taylor & Francis.
Heuvelink, G. B. M. (1998b). "Geographic Information Technologies in Society." 2007, from http://www.ncgia.ucsb.edu/giscc/units/u098/u098.html.
Heuvelink, G. B. M., P. A. Burrough and A. Stein (2006). Developments in Analysis of Spatial Uncertainty Since 1989. *Classics from IJGIS: 20 Years of the International Journal of Geographical Information Science and Systems*. P. F. Fisher. Boca Raton, CRC**:** 91-95.
Huggins, D. L. (2006). "Storm Rainfall Characterization." Retrieved 09.21, 2006, from http://pasture.ecn.purdue.edu/~engelb/abe526/Rain/.
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.
Keller, G. and J. Sherar. (2007). "Tools for Hydraulic and Road Design." *LOW-VOLUME ROADS ENGINEERING Best Management Practices Field Guide* Retrieved 04.11.2007, 2007, from http://ntl.bts.gov/lib/24000/24600/24650/Chapters/H_Ch6_Tools_for_Hydraulic_Design.pdf.
Lehmann, J., J. Breucker and B. Brouwer (2006). Causation in AI&Law. *ICAIL 2003 Workshop on Legal Ontologies & Web Based Legal Information Management*, University of Amsterdam, Faculty of Law**:** 34.
Lords, H. o. (2006). Binod Sutradhar v. Natural Environment Research Council. *House of Lords Session 2005–06; EWCA Civ 175; UKHL 33*. London, UK, House of Lords**:** 21.
Marx, K. (1867; translated reprint 1992). *Capital: Volume 1: A Critique of Political Economy*, Pengiun Classics.
McCuen, R. H. (1989). *Hydrologic Analysis and Design*. Prentice-Hall, Englewood Cliffs, NJ.
NCGIA (1989). "The U.S. National Center for Geographic Information and Analysis: An Overview of the Agenda for Research and Education." *IJGIS* **2**(3): 117-136.
North, D. C. (1981). *Structure and Change in Economic History*. New York, London, W W Norton &Company.
North, D. C. (2005). *Understanding the Process of Economic Change*. Princeton Oxford, Princeton University Press.
O'Hara, K. and N. Shadbolt (2001). *Issues for an Ontology for Knowledge Valuation*. Proceedings of Proceedings of the IJCAI-01 Workshop on E-Business and the Intelligent Web.
Physicslabs. (2006). "Uncer. & Error Propagation." Retrieved 09.21., 2006, from http://physicslabs.phys.cwru.edu/MECH/Manual/Appendix_V_Error%20Prop.pdf.
Purvis, J. C., W. Tyler and S. Sidlow (1988). Maximum Rainfall Intensity in South Carolina by County, State Climatology Office, South Carolina.
Ricardo, D. (1817; reprint 1996). *Principles of Political Economy and Taxation*, Prometheus Books.
Robinson, V. B. and A. U. Frank (1985). *About Different Kinds of Uncertainty in Collections of Spatial Data.* Seventh International Symposium on Computer-Assisted Cartography, Auto-Carto 7, Washington, D.C., ASP and ACSM.
Schneider, J. (1999). Zur Dominanz der Lastannahmen im Sicherheitsnachweis. *Festschrift zum 60. Geburtstag von Eduardo Anderheggen*, Institut für Baustatik und Konstruktion der ETH Zürich.
Schneider, J. (2000). "Safety - A Matter of Risk, Cost, and Consensus." *Structural Engineering International* **10**(4): 266-269.
Shi, W., P. F. Fisher and M. F. Goodchild (2002). *Spatial Data Quality*, Taylor & Francis.
Shi, W., M. F. Goodchild and P. F. Fisher, Eds. (2003). *Proceedings of The 2nd International Symposium on Spatial Data Quality '03*. Hong Kong, Hong Kong Polytechnic University.
Todini, E. (1988). "Rainfall-Runoff Modeling: Past, Present and Future." *Journal of Hydrology JHYDA7* **100**(1/3): 341-352.
Topanga. (2006). "Description of Hydrologic Models." Retrieved 09.21., 2006, from http://www.topangaonline.com/twc/water/APPM1.html.
USDAForestService. (2006). "Chapter 5 - Hydrology." *FSH 7709.56b -Transportation Structures Handbook* Retrieved 09.21., 2006, from http://www.fs.fed.us/im/directives/fsh/7709.56b/7709.56b,5.txt.

Wand, Y. and R. Y. Wang (1996). "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM* **39**(11): 86-95.

Widmoser, P. (1976). "Bestimmung von Bemessungswerten aus Hochwasserbeobachtungen." *Die Wasserwirtschaft* **66**(7/8): 199-203.

Wikipedia. (2007). "Risk Aversion." from http://en.wikipedia.org/wiki/Risk_aversion.

Wu, L., W. Shi, Y. Fang and Q. Tong, Eds. (2005). *Proceedings of the Fourth International Symposium on Spatial Data Quality (ISSDQ 05)*. Beijing, Peking University.