Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2009 Proceedings

International Conference on Information Systems (ICIS)

2009

Simulations of Error Propagation for Prioritizing Data Accuracy Improvements in Multi-Criteria Satisficing Decision Making Scenarios

Irit Askira Gelman
Tucson, AZ, askira@cox.net

Simulations of Error Propagation for Prioritizing Data Accuracy Improvements in Multi-Criteria Satisficing Decision Making Scenarios

Completed Research Paper

Irit Askira Gelman
3565 E. Calle Alarcon, Tucson, AZ
Askira@cox.net

Abstract

This research addresses the need for models that guide data quality design and resource allocation decisions. Broadly, our research problem is: Given an information system that utilizes a set of data sources for producing required information, how can we determine the gain in information accuracy and, subsequently, the economic return if the accuracy of a chosen data source is improved? An earlier paper by the author approaches this problem through a construct and a model. The construct, named damage*, is defined as the change in information accuracy that results from improving the accuracy of a chosen data source. The model that is provided together with this construct enables its quantification as well as a simple ranking of inputs according to the damage that errors in each inflict. The model suits environments in which the data are applied mostly by satisficing, multi-criteria decisions, such as databases. This paper reports on a series of Monte Carlo Simulations that validate the ranking component of the model under conjunctive decisions, and, in addition, explore and characterize special conditions in which a predicted ranking is not assured to be correct.*

Keywords Information quality management, information accuracy, multi-criteria decisions, conjunctive decision rules, satisficing decisions, ranking, Monte Carlo simulation.

Introduction

The overall annual cost of poor data quality to businesses in the US has been estimated in the hundreds of billions of dollars (Eckerson 2002) and the overall cost to individual organizations is believed to be 10%-20% of their revenues (Redman 2004). However, these estimates are not impressive enough, apparently, to drive organizations to action. For instance, most organizations have no plans for improving data quality in the future (Eckerson 2002). In the face of this neglect, there is a mounting conviction among both practitioners and researchers that an understanding of the economic aspect of data quality can be crucial for convincing organizations to increase their efforts. An understanding of the economics of data quality can guide decisions on how much to invest in the quality of their information and how to allocate limited organizational resources (Wang and Strong 1996).

This paper is a product of a research project that addresses the need for models that support resource allocation and design decisions that center on information quality. In particular, this research considers the accuracy dimension of information quality (Wang and Strong 1996). Broadly, the questions that are of interest in this research project include, for instance:

(1) Assuming an information system that utilizes a specified set of input sources for producing required information, how can we identify the input sources that would yield the highest gain in information accuracy if their accuracy is improved? How can we identify the input sources that would offer the highest economic return if their accuracy is improved?

(2) How can we quantify the gain in information accuracy that would result from improving the accuracy of a

chosen data source, and the subsequent economic return?

A model that provides answers to these questions can support design decisions on the accuracy of different data sources as well as resource allocation among the sources. Ultimately, both users and providers of the information system can benefit from such a model.

An earlier paper (Askira Gelman, forthcoming) approached these questions by developing a construct and a model for estimating that construct (Hevner et al. 2004; March and Smith 1995). The perception that underlies the proposed construct, named *damage*, is that, all other things being equal, it would be beneficial to assign priority to the elimination of input errors that have a higher negative effect on output accuracy (i.e., generate more output errors) over input errors that have a lower negative effect. In practical settings where, often, not all other things are equal, an estimate of the damage should be weighed by, or combined with, values of other relevant factors, in order to yield a more comprehensive evaluation of an investment in the accuracy of a chosen input.

Our model quantifies the damage in a popular class of applications and, in addition, enables a simple ranking of inputs of such applications according to the damage that errors in each inflict. Specifically, the model assumes an information system that employs dichotomous, multi-criteria satisficing decision rules. Instances of such systems include databases and expert systems that utilize domain knowledge in the form of multi-criteria satisficing decision rules. The term "satisficing," has been coined by Herbert Simon to denote problem-solving and decision-making that aims at satisfying a chosen aspiration level instead of an optimal solution (Simon 1957). Research indicates that satisficing rules agree with human choices and inferences in diverse situations involving complex problems, severe time constraints, or lack of information (e.g., Einhorn 1970; Einhorn 1971; Einhorn 1972; Lussier and Olshavsky 1979; Mintz 2004; Park 1976; Payne 1976; Phipps 1983).

An example that illustrates an information system in the class which is of interest here involves decisions on buying real estate properties. Nowadays, these decisions are often facilitated by an online real estate database like REALTOR.COM or REALESTATE.YAHOO.COM. Users formulate satisficing decision rules, primarily conjunctive rules, which express their needs. Suppose that, due to the high number of available properties, a conjunctive decision rule is employed for the initial screening of alternatives (e.g., Lussier and Olshavsky 1979; Payne 1976) or throughout the entire selection process. For instance, consider a decision maker that examines classic variables such as location, price, and number of bedrooms; let us say that he or she looks for a residential property in zip code 85719 that has 2 or 3 bedrooms and is priced at $375,000 or below. This decision-maker's preference is expressed by the conjunctive decision rule: zip code = 85719 *and* number of bedrooms = 2 or 3 *and* price $\leq$ $375,000. When the source of the data is a database, each decision variable corresponds to a suitable database attribute. A property is included in the resultant set of suitable properties if and only if the data about the property indicate that it satisfies the entire decision rule. Obviously, in agreement with the common experience, one may assume that the real estate data are not free of errors. These errors can lead to incorrect classifications of properties as fulfilling or not fulfilling a relevant decision criterion, which can result in incorrect property selection and exclusion decisions. A property that does not satisfy the specified criteria may thus be included in the short list of suitable properties (a false positive), while a property that has the desired attributes may be excluded from that list (a false negative).

To the extent that the decision maker actually has influence over the accuracy of the data that they are using, they can benefit from a model that quantifies or ranks the relevant database attributes in terms of the damage that errors in each attribute bring on the accuracy of the property selection decision. For instance, if they can choose among competing real estate databases, then this model can affect their choice.

In the majority of practical settings, however, data are used over and over again in different ways—data usage is rarely limited to one scenario. In the case of the real estate database, database attribute subsets are used in a variety of satisficing decision rules, mainly, perhaps, in conjunctive decisions. Assuming this type of multi-purpose database, its providers can benefit from our model if the *average damage* on a collection of decisions can be estimated, rather than estimating the damage that errors bring on a single decision (Askira Gelman 2008).

According to the emerging understanding about the fundamentals of good design research (Winter 2008; Hevner et al. 2004; March and Smith 1995; Nunamaker et al. 1991), the proposed construct and model must be evaluated. In agreement with that understanding, this paper reports on a set of empirical tests, implemented through Monte Carlo Simulations, of the model component that ranks decision inputs in terms of the damage that errors in each input inflict on a decision. Our simulations validate this model for conjunctive decisions, and, in addition, explore and characterize the special conditions in which the proposed model is not assured to predict the correct damage ranking.

The paper is organized as follows. The next section describes the primary assumptions of our model. A literature review is provided next, followed by an overview of the theory that serves as a foundation for the current, empirical study. A later section describes the research method. The results are detailed in a subsequent section. We conclude with a discussion of the implications of this work, its limitations, and future research directions.

Assumptions

This section exposes the major concepts and unique assumptions that underlie our work. To begin with, this paper typically uses the term *data* to describe the raw, unprocessed input of an information system; the term *information* mostly designates the output of an information system. *Accuracy* is defined as the degree to which the data or information are in conformance with the true values. On the output side, in particular, a decision error is registered whenever a decision based on the available inputs deviates from the outcome of the same decision based on error-free inputs. (Note that, from the perspective of the effect on decision accuracy, there is no need to differentiate between an incorrect input value and a value which is out-of-date or based on a nonstandard unit such as a price that is given in a nonstandard currency, etc.) Data and information accuracy are measured by the probability of error occurrence. Despite the fact that the implementation of this measure can be costly, studies that use error probability or error rate, error magnitude, or various fusions of the former to measure accuracy are common in the research literature. A possible explanation of that popularity is the relative ease with which such measures can often be examined through the prevailing, broad analytical frameworks. Furthermore, there is a growing literature that offers practical solutions for deriving the measures (e.g, Ballou et al 2006; Motro and Rakov 1997; Parssian 2006; Hipp et al. 2001). We will briefly discuss the important question of the implementation of the model in the final section of this paper.

Our approach for identifying the data sources that yield the highest gain when their accuracy is improved, and for quantifying this gain, utilizes the concept of *damage*. The damage that errors in an input inflict on output accuracy is defined in this work as the *change in output error probability due to an increase in the error probability of that input*. The idea that motivates our focus on this construct is that, all other things being equal, it would be beneficial to assign priority to the elimination of errors that have a higher negative effect on output accuracy over errors that have a less negative effect. For instance, suppose that, by decreasing the error rate in one of the inputs by 1%, we decrease the decision error rate by 0.5%, while a decrease in the error rate of a second input by 1% decreases the decision error rate by 0.05%. Obviously, all other things being equal, it would be more effective to decrease the error rate of the first input than the second. In practical settings where, often, not all other things are equal, an estimate of the damage has to be weighed by, or combined with, values of other relevant factors (e.g., associated costs) in order to yield a comprehensive assessment of an investment in the accuracy of a chosen input. Technically, we use a *partial derivative* to implement the concept of damage. A derivative is a measure of the change in the output of a function when its input changes.[1] Therefore, by definition, it is consistent with the notion of damage as it is interpreted by this work.

This paper tests and explores a mathematical-statistical model that has been proposed in (Askira Gelman, formthcoming) for ranking the inputs of a conjunctive decision rule according to the damage that errors in each input creates. An important advantage of this model is its relative simplicity. Mainly, regardless of the number of decision variables that a decision utilizes, the ranking of two variables is based exclusively on parameters of the two variables—there is no need to account for characteristics of the other inputs of the decision. Furthermore, as will be demonstrated in this paper, the ranking of a given variable pair may often be obtained using rough estimates of the relevant parameters, alleviating the need for costly measurements. Evidently, however, a major disadvantage of a ranking relative to a full fledged quantitative measure is that a ranking may not be compatible with a broad quantitative assessment, i.e., an assessment that accounts for the damage as well as other relevant factors (such as, again, accuracy improvement costs).

[1] A partial derivative is the derivative of a function of multiple variables when all but one variable of interest are held fixed.

Literature Review

An implicit assumption of this research is that design decisions regarding the accuracy of data sources and resource allocation among these sources, take into account the intended use of the data. Contrary to an approach that does not differentiate between errors (e.g., Janson 1988; Parsaye and Chignell 1993), an approach that differentiates between errors based on the intended use of the data is compatible with the currently accepted definition of data quality as "fitness for use." The concept of fitness for use emphasizes the context of the data, mainly the uses, users, and suppliers of the data (Juran 1988). Current methodologies (e.g., Lee et al. 2002; Pipino et al. 2002; Wang 1998) apply diverse tools for capturing users' quality requirements and relating them to the actual state of data quality. Identified discrepancies between the actual and desired state assist in guiding the improvement efforts. Our research, however, introduces an additional, potentially relevant factor for design and resource allocation decisions, namely, the relationship between input quality and output quality.

The literature on the relationship between input accuracy and output accuracy is vast. This relationship has been investigated in countless problem domains. Some of these problem domains are information pooling and group accuracy (e.g., Condorcet 1785; Grofman et al. 1983; Ladha 1995), propagation of measurement errors (e.g., Bevington 1969), feature selection (e.g., Elashoff et al. 1967; Toussaint 1971; Cover 1974; Fang 1979), expert resolution (e.g., Clemen and Winkler 1985), internal accounting control (e.g., Cushing 1974; Hamlen 1980; Stratton 1981; Yu and Neter 1973), ensemble learning (e.g., Ali and Pazzani 1996; Kuncheva et al. 2003), and multisensor fusion (Mitchell 2007). Naturally, the relationship between an information system's input accuracy and its output accuracy has also received significant attention in the Data and Information Quality (DIQ) literature in MIS (e.g., Ballou and Pazer 1985; Ballou and Pazer 1990; Ballou et al. 1998; Shankaranarayan et al. 2003; Wang et al. 2001; Motro and Rakov 1997; Parssian et al. 2004; Parssian 2006; Avenali et al. 2008). A distinctive characteristic of DIQ, which has influenced the type of questions that researchers in this community ask, is the assumption that the accuracy of a data source can be improved. This assumption is rare. For instance, unlike DIQ, the problem of feature selection, which is reflected through its name, is how to *select* the feature subset that would lead to the lowest classification error (Jain et al. 2000). In other words, the accuracy of the features is assumed to be fixed. Subsequently, the question of how to determine the gain in information accuracy that would result from improving the accuracy of a chosen input does not attract direct research in the feature selection and other relevant problem domains. Yet, from a DIQ perspective, this question is interesting.

In the DIQ literature, various frameworks for assessing the relationship between the quality of the raw data and the quality of query outputs have been proposed in the context of relational databases. These frameworks have sometimes been labeled *data quality algebra*. Reddy and Wang (Wang et al. 2001) assessed the relationship between the accuracy of the data and the accuracy of the output of a database query. Parssian et al. (2004) and Parssian (2006) investigated accuracy and completeness, and Ballou et al. (2006) targeted a broader set of data quality dimensions. Other relevant studies include (Motro and Rakov 1997) and (Avenali et al. 2008). A scenario that is partly similar to this work has also been addressed by Ballou and Pazer (1990), who proposed a framework for assessing the effect of input errors on the accuracy of dichotomous decisions. Ballou and Pazer considered decisions that are implemented by integrating multiple decision criteria through a conjunctive rule. Nonetheless, DIQ studies that investigated the relationship between input accuracy and output accuracy highlighted the *aggregate* effect of input errors, rather than the effect of errors in individual inputs, on the accuracy of the output of the information system.

An earlier paper by the author (Askira Gelman, forthcoming) approaches the problem by proposing the concept of damage and by developing a model that quantifies the damage in a popular class of applications and, in addition, enables a simple ranking of inputs of these applications according to the damage that errors in each inflict. A unique contribution of this paper that goes beyond the author's previous work on this topic is the empirical validation of the ranking component, and the clarification of specific conditions in which a prediction of the ranking model is not guaranteed to be correct.

Damage Ranking Model

In this section we introduce the mathematical-statistical damage ranking theory that underlies this paper. We describe the notation, model, and associated theory.

Variables and Notation

Consider a conjunctive decision rule that accounts for *N* decision variables, V_i, $i = 1, 2, .., N$ (e.g., V_1 represents the zip code of a property, V_2 represents the number of bedrooms, and V_3 represents the price). Specifically, V_i describes the ideal, error-free data. In this paper, the implementation of a conjunctive decision is assumed to be as follows. Initially, for every *i*, the value of V_i is tested against the matching decision criterion. The outcome of this test (zero for "false" or one for "true") is captured by a matching, dichotomous variable I_i. In the real estate decision, for instance, V_1 is tested against 85719 to derive the value of I_1 (e.g., if V_1=85718 then I_1=0, and if V_1=85719 then I_1=1); V_2 is tested against 2 and 3 to produce the value of I_2, and V_3 is tested against $375,000 to produce the value of I_3. The values of I_i, $i = 1, 2, .., N$ that are determined in this way are combined iteratively through a sequence of logical conjunction operations to generate the outcome of the decision. A decision can be either zero ("false" or "reject") or one ("true" or "accept"). We will use the symbol O_i to denote the outcome of applying the iterative process on $I_1, .., I_i$, ($O_1 \equiv I_1$). In the first iteration, the value of I_1 (or, equivalently, the value of O_1) is combined with the value of I_2 through logical conjunction, and the output is given by O_2. In the second iteration, the values of O_2 and I_3 are similarly combined, and the output is given by O_3, and so on. In the final iteration, the values of O_{N-1} and I_N are combined through logical conjunction, and the output is registered by O_N. It is easy to see that O_N registers the outcome of a conjunctive decision that accounts for all the decision variables.

Table 1. Notation

Symbol	Meaning
V_i	decision variable (random variable); describes the correct data
I_i	informs us whether V_i passes the decision criterion or not (dichotomous random variable)
O_i	the output of a decision based on $V_1, .., V_i$, (dichotomous random variable)
F_i^V	informs us whether the recorded value of V_i is correct or not (dichotomous random variable)
F_i^I	informs us whether the recorded value of I_i is correct or not (dichotomous random variable)
F_i^O	informs us whether the recorded value of O_i is correct or not (dichotomous random variable)
p_i^I, p_i^{FI}, p_i^{FV}, p_i^O, p_i^{FO}	expected (mean) values

While the former variables refer to error-free data, such data are rare in reality. The symbol F_i^V identifies a variable that informs us about the occurrence of a fault, or error, in the observed (recorded) value of V_i. Namely, F_i^V=1 if the recorded value of V_i is incorrect, and F_i^V=0 if that value is correct. Similarly, F_i^I refers to the occurrence of an error in the observed value of I_i, and F_i^O identifies the occurrence of an error in the value of O_i. Note that F_i^I=1 implies that F_i^V=1, i.e., an error in the recorded value of the dichotomous variable is always due to an error in the

corresponding decision variable. However, $F_i^V = 1$ does not necessarily imply $F_i^I = 1$, since not every error in the observed value of the decision variable generates an error in the dichotomous variable.

The variables in $\{ I_i, F_i^I, F_i^V, O_i, F_i^O : \ i = 1, 2, .., N\}$ are random variables that accept the values zero and one. For each of the variables in $\{ I_i, F_i^I, F_i^V, O_i, F_i^O : \ i = 1, 2, .., N\}$, we will mark the corresponding mean values with the symbol p and a combination of subscripts and superscripts that distinguishes the individual random variable, e.g., p_i^{FI} matches F_i^I (see Table 1). The mean of a random variable that informs us about the occurrence of an error is the same as the probability of error occurrence in that variable. This equivalence holds true since the value of such a variable is either zero or one.

Model

The damage ranking model that forms the foundation of this paper is introduced next (for more details, see Askira Gelman, forthcoming). The model and associated theory address a scenario in which one wants to rank the damage of errors in two decision variables that are combined by a conjunctive decision rule. Apart from these two variables, the conjunctive decision rule that combines them may join any number of decision variables.

Initially, we define the term damage:

Definition 1 (damage): Assume a decision O_N which is derived from $\{V_1, .., V_N\}$ through conjunction. Let $V_m \in \{V_1, .., V_N\}$. The damage that errors in the recorded values of V_m inflict on the recorded values of O_N is defined as the partial derivative $\partial p_N^{FO} / \partial p_m^{FV}$.

A second concept, damage type II, which focuses on $\{I_1, .., I_N\}$ rather than $\{V_1, .., V_N\}$, simplifies the ensuing presentation, and, as an example in the last section indicates, can sometimes be used independently of the concept of damage.

Definition 2 (damage type II): Assume a decision O_N which is derived from $\{I_1, .., I_N\}$ through conjunction. Let $I_m \in \{I_1, .., I_N\}$. The damage that errors in the recorded values of I_m inflict on the recorded values of O_N is defined as the partial derivative $\partial p_N^{FO} / \partial p_m^{FI}$.

Unlike Definition 1, Definition 2 interprets input errors as errors in the classification of input data values as satisfying or not satisfying the decision criterion (i.e., false negatives or false positives).

In this paper our concern is to verify and study a fundamental ranking model that presupposes a single decision or decision rule. The model is based on certain statistical independence assumptions. Fortunately, our theory does not impose any statistical independence assumptions on two variables as long as these variables do not describe any of the two decision variables that are being ranked in relation to one another. However, the independence assumptions on the latter are that none of the variables or products of variables in $\{ I_1, I_2, F_1^I, F_2^I \}$ is dependent on any other variable or product of other variables in $\{ I_i, F_i^I : \ i = 1, 2, .., N\}$..

Assumption: None of the variables or products of variables in $\{ I_1, I_2, F_1^I, F_2^I \}$ is statistically dependent on any other variable or product of other variables in $\{ I_i, F_i^I : \ i = 1, 2, .., N\}$.

Admittedly, these independence requirements are frequently violated in practical settings. For example, situations in which the probability of a false positive is different from the probability of a false negative are known to be common. Technically, I_1 and F_1^I (or I_2 and F_2^I) are not statistically independent. Likewise, we often encounter situations in which two decision variables are not independent, such that the matching dichotomous variables are not independent either, e.g., the number of bedrooms in a property may be statistically dependent on its location such that the respective dichotomous variables may not be statistically independent either. However, in fact, the ranking that our model provides is valid under a range of statistical dependencies. In addition, our theoretical work indicates

that we can maintain the simplicity of our model while relaxing these assumptions if the data are assumed to be used by multiple decision rules, and, accordingly, if what is of interest is the average damage rather than the damage to a single decision (Askira Gelman 2008).

Ranking (damage type II): A ranking of the damage which errors in the recorded values of I_1 and I_2 inflict on the recorded values of O_2 is determined using equations (1)-(2):

$$\partial p_2^{FO} / \partial p_1^{FI} = p_2^I + p_2^{FI} \cdot (1 - 2(p_1^I + p_2^I - p_1^I p_2^I)) \quad (1)$$

$$\partial p_2^{FO} / \partial p_2^{FI} = p_1^I + p_1^{FI} \cdot (1 - 2(p_1^I + p_2^I - p_1^I p_2^I)) \quad (2)$$

Equation (1) quantifies the damage that errors in the recorded values of I_1 inflict on the recorded values of O_2. Similarly, (2) quantifies the damage of errors in the recorded values of I_2 to the recorded values of O_2. Interestingly, (1) implies that the damage that errors in the recorded values of I_1 produce does not depend on the error rate in that variable, p_1^{FI}. A similar observation applies to I_2. In other words, everything else being equal, when we lower the error rate in an input (i.e., I_1 or I_2), the damage to the decision does not change (linearity).

A critical component of this theory is the understanding that, given a conjunctive decision rule that employs additional decision variables apart from V_1 and V_2, *a ranking of the damage type II that is determined through computation of (1) and (2) is guaranteed to be preserved in the entire decision rule under broad conditions*. In particular:

Proposition 1 (damage type II): Suppose that $\partial p_2^{FO} / \partial p_1^{FI} \geq \partial p_2^{FO} / \partial p_2^{FI}$. Then, $\partial p_N^{FO} / \partial p_1^{FI} \geq \partial p_N^{FO} / \partial p_2^{FI}$ for any $N \geq 2$ if either of (3) or (4) is satisfied:

$$p_2^{FI} - p_1^{FI} \geq 0 \quad (3)$$

$$p_2^I - p_1^I \geq (p_1^{FI} - p_2^{FI})(1 - 2(p_1^I + p_2^I - 2p_1^I p_2^I)) \quad (4)$$

In fact, Proposition 1 and equations (1)-(2) imply a simple ranking rule:

$$| p_1^{FI} - p_2^{FI} | \leq p_2^I - p_1^I \Rightarrow \partial p_N^{FO} / \partial p_1^{FI} \geq \partial p_N^{FO} / \partial p_2^{FI} \quad (5)$$

In words, *regardless of the number of decision variables that a decision accounts for, the damage type II of errors in the recorded values of I_1 is higher than the damage type II of errors in the recorded values of I_2 if the difference $p_2^I - p_1^I$ is higher than the absolute value of the difference $p_1^{FI} - p_2^{FI}$.*

As a later example demonstrates, rule (5) alleviates the need to obtain precise estimates of the parameters (p_1^I, p_2^I,..., p_N^I) for the purpose of ranking.

Since the concept of damage is more consistent with common perceptions than the concept of damage type II, a ranking of damage may be preferred over a ranking of damage type II. Proposition 2 handles the ranking of damage. Proposition 2 suggests that a ranking of the damage is derived from a ranking of the respective damage type II and an additional factor. Clearly, a ranking of damage involves additional effort. Let α_1 denote the change in the probability of error in the recorded value of I_1 due to an increase in the probability of error in the recorded value of V_1. Let α_2 denote the change in the probability of error in the recorded value of I_2 due to an increase in the probability of error in the recorded value of V_2. Proposition 2 stipulates that the damage of errors in the recorded values of V_1 has a higher ranking than the damage of errors in the recorded values of V_2 if the damage type II of

errors in the recorded values of I_1 has a higher ranking than the damage type II of errors in the recorded values of I_2 and, in addition, $\alpha_1 \geq \alpha_2$.

Proposition 2 (damage): Suppose that $p_1^{FI} = f(p_1^{FV})$ and $p_2^{FI} = g(p_2^{FV})$, and let $\alpha_1 = \partial p_1^{FI} / \partial p_1^{FV}$, $\alpha_2 = \partial p_2^{FI} / \partial p_2^{FV}$. If $\partial p_N^{FO} / \partial p_1^{FI} \geq \partial p_N^{FO} / \partial p_2^{FI}$ and $\alpha_1 \geq \alpha_2$ then $\partial p_N^{FO} / \partial p_1^{FV} \geq \partial p_N^{FO} / \partial p_2^{FV}$.

Unfortunately, the additional requirement that $\alpha_1 \geq \alpha_2$ means that when $\partial p_2^{FO} / \partial p_1^{FI} \geq \partial p_2^{FO} / \partial p_2^{FI}$ and $\alpha_1 < \alpha_2$, the ranking of the damage is undefined. This limitation of the ranking of damage may increase the attractiveness of a ranking of damage type II.

Example: Suppose that a ranking of the damage of errors in the recorded zip code (denoted next V_1) versus the damage of errors in the recorded number of bedrooms (V_2) is of interest. Assume, in particular, a database that covers a broad geographic area. In this case, the percentage of properties in a given zip code is probably low, while the percentage of properties that have 2-3 bedrooms may be much higher since such properties are common. For instance, one may find out by consulting a data analyst that $p_1^I \leq 0.05$ and $p_2^I \geq 0.5$. Suppose that the data analyst also estimates that the error rates in this database are not extremely high, e.g., p_1^{FV}, $p_2^{FV} \leq 0.1$. Given that the independence assumptions of this theory hold true, the premise of (5) is easily satisfied, since, by definition, $p_1^{FI} \leq p_1^{FV}$ and $p_2^{FI} \leq p_2^{FV}$, and, therefore, $| p_1^{FI} - p_2^{FI} | \leq 0.1$. Clearly, there is no need to obtain precise measurements of the specified parameters in order to determine that the damage type II associated with the zip code data is higher. In other words, errors in the classification of the recorded zip codes as equal or not equal to 85719 are more damaging than errors in the corresponding classification of the recorded bedroom values. In general, (5) suggests that, if the data are combined through conjunction, then classification *errors in the input that is less likely to satisfy the decision criterion are more detrimental to decision accuracy*.

An estimate of α_1 (the increase in the error rate of the classification of zip codes as equal or not equal to 85716 that results from an increase in the error rate of the zip code data) may be produced based on an estimate of the ratio $p_1^{FI} : p_1^{FV}$. If the estimates of α_1 and α_2 indicate that α_1 is not lower than α_2, then errors in the zip code data are more damaging to the real estate decision than errors in the bedrooms data.

In the following sections we test the validity of our ranking theory. We validate that a ranking of damage based on (1)-(2) is correct when the condition of Proposition 1 is satisfied. In addition, since this condition is a sufficient conditions but it is not a necessary condition, a ranking based on (1) and (2) may be valid even when Proposition 1 does not guarantee that the ranking is valid. Therefore, our simulations explore the conditions in which a ranking of damage type II based on equations (1) and (2) is, indeed, incorrect.

Monte Carlo Simulation

The method employed by this study is Monte Carlo simulation. Monte Carlo simulation is a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. The inputs are generated pseudo-randomly from selected probability distributions to simulate the process of sampling from an actual population. The model is evaluated for each simulated input set, and the result is taken as an average over the number of data points in the sample (Fishman 1995). The elements that comprise our simulation method are described below.

Instantiation of the Variables

The simulations examine conjunctive decision rules with up to ten decision variables, $V_1, .., V_N$, where $2 \leq N \leq 10$. We have implemented the decision variables as dichotomous variables that accept the values zero and one, i.e., $V_i = 0$ or $V_i = 1$. Notably, when the decision variables are dichotomous, $\alpha_1 = \alpha_2 = 1$, i.e., a data error always translates into a false negative or false positive when the chosen decision criterion is verified against the recorded data value. By utilizing this simple relationship we focus our tests on a critical element of this theory, namely, Proposition 1.

The values of $V_1,..,V_N$ are generated pseudo-randomly according to pre-determined distributions. In particular, these values are generated from distributions that are determined separately for each simulation. p_i^V, the expected value of V_i, is chosen randomly in each simulation such that $0 < p_i^V < 1$ (note that $p_i^V = E(V_i) = \Pr(V_i = 1)$). The values of F_i^V, which inform us of the occurrence of errors in the recorded, possibly incorrect version of V_i, are generated from distributions that, again, are determined individually for each simulation. p_i^{FV}, the expected value of F_i^V, is chosen randomly such that two value ranges are explored. In one simulation set that includes 9,500 simulations, $0< p_i^{FV} <0.10$, and in a second simulation set, which includes, again, 9,500 simulations, $0< p_i^{FV} <0.20$. Table 1 summarizes the simulation parameters.

Table 1. Implemented parameter values and number of simulations

N (# of decision variables)	p_i^V (i=1,2,..,N)	p_i^{FV} (i=1,2,..,N)	M (Sample size)	Total # of simulations
9 simulation sets: N=2,3,4,5,6,7,8,9,10	random value in the interval (0,1)	2 simulation sets: $0< p_i^{FV} <0.10$ $0< p_i^{FV} <0.20$	$5\cdot10^9$	$9\cdot2\cdot1,000 = 18,000$
10	random value in the interval (0,1)	2 simulation sets: $0< p_i^{FV} <0.10$ $0< p_i^{FV} <0.20$	$5\cdot10^{12}$	$2\cdot500 = 1,000$

Altogether, 19,000 simulations of conjunctive decisions were carried out. The simulations were conducted using MATLAB, a programming language and interactive environment that enables us to perform computationally intensive tasks.

Sample size: Each of 18,000 simulations produces $M=5\cdot10^9$ input instances of each variable, while in the remaining 1,000 simulations each simulation produces $M=5\cdot10^{12}$ input instances of each variable.

Simulation Outputs

For calculating the actual damage ranking, each simulation computes a base decision error rate, f_b^o, which is calculated from input samples that exhibit the randomly selected error rates p_i^{FV} (i =1,2,..,N). In addition, a simulation computes a set of decision error rates, f_i^o (i =1,2,..,N), one for each decision variable. In these computations, all the input error rates are the same as in the base except for the error rate of the chosen decision variable, which is 0.01 higher than the randomly selected rate. The damage that errors in the recorded values of V_i inflict on the recorded values of O_N, denoted by Δ_i, is estimated as:

$$\Delta_i = f_i^o - f_b^o \qquad (6)$$

The value of f_b^o and the value of f_i^o are each estimated through a suitable implementation of (7)

$$\overline{\Pr(F_N^O = 1)} = \frac{1}{M}\sum_{j=1}^{M} F_{N,j}^O \tag{7}$$

For each pair of decision variables, V_i and V_j, the actual damage ranking based on Δ_i and Δ_j is compared to the damage ranking that is predicted by (1) and (2).

Simulation Model

The simulation model implements the conjunction of $V_1,..,V_N$ and, analogously, the conjunction of $V_1^R,..,V_N^R$ (the observed, possibly incorrect versions of V_i). In the first step, V_i is mapped to I_i. Precisely, the value of I_i is set equal to the value of V_i:

$$I_i = V_i \tag{8}$$

Equation (8) forms an inconsequential simplification of a "correct" simulation model in which, in some simulations that are selected randomly, I_i is determined to be equal to V_i, while in the remaining simulations, I_i is set equal to 1-V_i. Next, since the possible values of V_i are limited to zero and one, every error in the recorded value of V_i produces an error in the classification of the value as satisfying or not satisfying a decision criterion. Therefore:

$$F_i^I = F_i^V \tag{9}$$

The value of I_i^R, the recorded, possibly incorrect portrayal of I_i, is derived from I_i and F_i^I using (10):

$$I_i^R = I_i \cdot (1 - F_i^I) + F_i^I \cdot (1 - I_i) \tag{10}$$

If the value of F_i^I is zero, that is, if this variable indicates that no error has occurred, then (10) is reduced to $I_i^R = I_i$. However, if the value of F_i^I indicates the occurrence of an error, then (10) assigns a value of one to I_i^R if I_i is zero and a value of zero if I_i is one.

The variables in each of $\{ I_i \}$ and $\{ I_i^R \}$ are joined iteratively through a sequence of logical conjunction operations. The algorithm treats the output of one binary operation as an input of a subsequent binary operation. For instance, the output of combining the values of I_1 and I_2, which we have denoted by O_2, is treated as one of the inputs of a binary operation whose second input is I_3. The ideal conjunction output—where inputs are error-free—is computed using (11):

$$O_i = O_{i-1} \cdot I_i \tag{11}$$

The consistency of (11) with the definition of logical conjunction can be quickly verified through a systematic evaluation of O_i for each possible combination of the values of O_{i-1} and I_i. Analogously, the observed decision is derived through:

$$O_i^R = O_{i-1}^R \cdot I_i^R \tag{12}$$

O_i^R designates the output of a decision that joins the first i observed, possibly incorrect inputs ($O_1^R \equiv I_1^R$). Finally, for calculating the occurrence of a decision error F_i^O the simulations use equation (13):

$$O_i^R = (1 - F_i^O) \cdot O_i + (1 - O_i) \cdot F_i^O \tag{13}$$

The logic of (13) is comparable to the logic of (10).

Results

Table 2 and Figure 1-Figure 4 summarize the results of the simulations. Table 2 portrays the inconsistency rates that have been registered for the ranking based on equations (1) and (2) versus the simulation results. It also shows the percentage of the former inconsistencies which were inconsistent with Proposition 1 as well, i.e., when Proposition 1 implied that the ranking based on (1)-(2) was correct. In the simulations that generated smaller input samples ($M=5\cdot10^9$), the average rate of inconsistency with (1)-(2) is 1.2% when the input error rates are lower (p_i^{FV} <0.1), and 2.3% when the input error rates are higher (p_i^{FV} <0.2). Rates increase with the number of decision variables. The maximal inconsistency rates, 3.2% (p_i^{FV} <0.1) and 3.7% (p_i^{FV} <0.2), were demonstrated in simulations of decisions that join 10 decision variables.

Many of the instances that exhibited inconsistency with a ranking based on (1)-(2) also disagreed with Proposition 1. However, the proportion of this "dual disagreement" varied dramatically depending on the sample size that the simulation generated. One possible explanation of the inconsistency with Proposition 1 is that the ranking theory is invalid. However, a second explanation that attributes this inconsistency to the limitations of simulation seems to fit the results better. Primarily, since the size of the input set that a simulation generates is not infinite, random variations can affect the result of the simulation. However, results are swayed more when the sample size is smaller, or when the numbers that the simulation calculates are smaller. Therefore, the finding that the proportion of dual disagreement falls dramatically in the simulations that generate a larger sample size is very much in line with the explanation that the inconsistencies are due to the limitations of simulation. Most importantly, in the simulations with the larger sample size ($M=5\cdot10^{12}$) the disagreement with Proposition 1 is negligible.

Table 2. Inconsistency rates (model vs. simulations)

p_i^{FV}	M (Sample size)	N (# of decision variables)	Inconsistency with equations (1)-(2) (average over all the values of N)	Inconsistency with equations (1)-(2) N=10	Inconsistency with Proposition 1 (average over all the values of N)	Inconsistency with Proposition 1 N=10
p_i^{FV} <0.10	$5\cdot10^9$	2-10	1.2%	3.2%	28% of the inconsistencies with (1)-(2)	65% of the inconsistencies with (1)-(2)
p_i^{FV} <0.10	$5\cdot10^{12}$	10	----	1.1%	----	3% of the inconsistencies with (1)-(2)
p_i^{FV} <0.20	$5\cdot10^9$	2-10	2.3%	3.7%	10% of the inconsistencies with (1)-(2)	25% of the inconsistencies with (1)-(2)
p_i^{FV} <0.20	$5\cdot10^{12}$	10	----	2.8%	----	4% of the inconsistencies with (1)-(2)

On the other hand, numbers are smaller when the parameter values are lower, for instance. Therefore, if the explanation that links the dual disagreements with limitations of simulation is correct, then, everything else being equal, simulations with p_i^{FV} <0.1 should demonstrate a higher discrepancy with Proposition 1 than those with p_i^{FV} <0.2. The simulations indeed demonstrate that pattern. When p_i^{FV} <0.1 the rate of dual disagreement is 28%,

and when p_i^{FV} <0.2 it is only 10%. In addition, numbers are smaller when the number of decision variables grows. Therefore, everything else being equal, the discrepancy when taking the average over all the decision rules should be lower than the discrepancy when *N*=10. The results of the simulations agree with this perception as well.

In conclusion, the simulations validate our ranking theory. They also reveal that, under the conditions of this study, the proposed ranking model predicts the ranking correctly in an overwhelming majority of the cases.

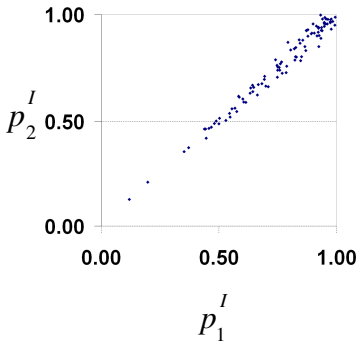

Figure 1. Distribution of Ranking Failures ($p_i^{FV} < 0.1$)

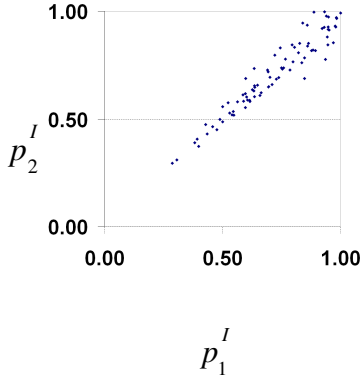

Figure 2. Distribution of Ranking Failures ($p_i^{FV} < 0.2$)

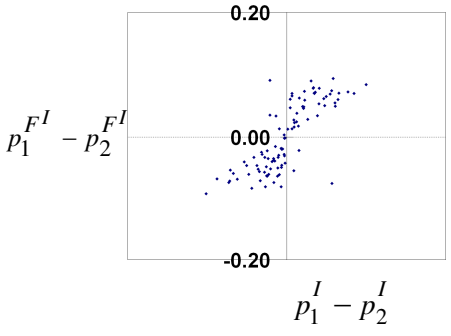

Figure 3. Distribution of Ranking Failures ($p_i^{FV} < 0.1$)

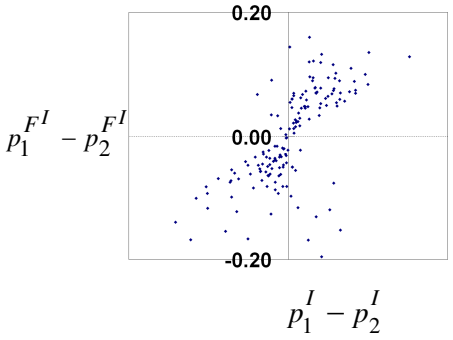

Figure 4. Distribution of Ranking Failures ($p_i^{FV} < 0.2$)

We turn next to a study of the conditions in which the ranking model fails to predict the correct ranking. These conditions are captured by Figure 1-Figure 4. Figure 1-Figure 4 are exclusively based on the simulations with the larger samples. In these simulations, the inconsistencies that have been recorded between the ranking model and the

simulations have systematically been in line with Proposition 1. Subsequently, we assume that these inconsistencies actually reflect failures of the ranking model, i.e., instances in which equations (1)-(2) do not provide the correct ranking.

One noticeable pattern in Figure 1 and Figure 2 is the accumulation of ranking failures in decision variable pairs where the probability of satisfying the criterion in one variable is similar to the probability of satisfying the criterion in the second variable. Namely, ranking failures are characterized by decision variable pairs whose values have a similar probability of satisfying the respective decision criteria ($p_1^I \approx p_2^I$). This pattern is coherent with (5), which implies that the ranking model is valid when the difference $p_2^I - p_1^I$ is higher than the absolute value of the difference $p_1^{FI} - p_2^{FI}$. If the difference $p_2^I - p_1^I$ is high, as was the case in the earlier described ranking example involving the zip code attribute and the number of bedrooms attribute, the validity of the ranking model is typically ensured.

A careful observation of Figure 1 and Figure 2 reveals a second pattern. The ranking failures are largely concentrated among higher values of p_1^I and p_2^I. When the probabilities of satisfying the decision criteria are low, our ranking model is reliable. In the property decision ranking example, $p_1^I \leq 0.05$. Therefore, a ranking failure is unlikely.

Figure 3 and Figure 4 suggest a third pattern of the ranking failures. Ranking failures are more common when $p_1^I - p_2^I$ and $p_1^{FI} - p_2^{FI}$ have the same sign. That is, one decision variable must show consistently low parameter values relative to the other decision variable. In the ranking example, again, since $p_1^I \leq 0.05$ and $p_1^I \geq 0.5$, such that $p_2^I > p_1^I$, if there is good reason to believe that the rate of error in the classification of the zip code of properties as equal or not equal to 85719 is higher than the corresponding error rate in the number of bedrooms, then a ranking failure is unlikely.

Table 3. Parameter values that produced ranking failures

p_1^I	p_2^I	p_1^{FI}	p_2^{FI}
0.977	0.931	0.070	0.021
0.931	0.997	0.021	0.090
0.602	0.603	0.018	0.062
0.749	0.738	0.052	0.035
0.847	0.784	0.077	0.007
0.547	0.520	0.097	0.025
0.983	0.967	0.098	0.052
0.805	0.833	0.025	0.072
0.648	0.659	0.063	0.078
0.635	0.655	0.004	0.039
0.664	0.623	0.091	0.022

Table 3 shows a small, arbitrary subset of the parameter value combinations that produced ranking failures. This subset has been taken from the simulations that created a large sample size and $p_i^{FV} < 0.1$. It is easy to see that, in

all of these parameter value combinations, the values of p_1^I and p_2^I are not far apart from each other; they are relatively high, and whenever p_1^I is greater than p_2^I, p_1^{FI} is greater than p_2^{FI} and vice versa.

Concluding Remarks

The simulations validate the ranking theory. They also uncover the fact that, under the conditions of this study, the proposed ranking model predicts the ranking correctly in an overwhelming majority of the cases. We have studied the rare instances in which the ranking model has offered incorrect predictions of the ranking and pointed to failure patterns that can guide assessments of specific parameter combinations.

Despite the fact that the ranking that our model produces is valid under a range of statistical dependencies, the statistical independence assumptions that underlie this model are an important limitation. Another limitation of this model is its focus on a single decision rule. Our forthcoming theoretical work addresses ranking when the data are used by multiple, diverse decision rules, and, accordingly, the average damage is of interest instead of the damage to a single decision. That work shows that many of the statistical independence assumptions can be relaxed under this more realistic scenario.

Evidently, a ranking which refers to errors in the recorded data that match the decision variables is more in line with common perceptions than the alternative ranking that this paper implies ("damage type II"). However, such a preference complicates the ranking, as it requires an estimate of α_1 and α_2, and may also produce inconclusive results while the alternative ranking would be clear. There are many scenarios in which the additional effort that is required for evaluating α_1 and α_2 is unjustifiable. A ranking based on damage type II, which does not require estimates of α_1 and α_2, can be just as useful. Suppose, for example, that, in addition to damage, the cost of cleaning the data is an important factor. Take, for instance, the property data ranking problem that we have examined earlier. From a damage perspective, that example indicated that the damage (type II) of errors in the zip code data is higher than the damage (type II) of respective errors that originate in records of the number of bedrooms. Furthermore, the discussion of the ranking model (equations (1)-(2)) implies that, if nothing else changes, then the damage (type II) of the errors in the zip code will have same magnitude as we keep improving the accuracy of this data set. The same is true for errors in the data on the number of bedrooms: if nothing else changes, then the damage (type II) will have the same magnitude as we keep improving the accuracy of this data set. As for costs, suppose that a preliminary study of alternative methods of cleaning the data has shown that, while the cost of cleaning the bedroom data would be quite high, the zip code errors can be treated effectively using an inexpensive method (an automated program that verifies the zip code based on a combination of street address, subdivision, and related data, i.e., such data are used for extracting the correct zip code from a reliable information source). In conclusion, a study of the damage type II and the relevant costs leads, in this scenario, to an unequivocal recommendation to treat the zip code data first. Mainly, *since the goal of the treatment is to effectively clean the data source of all errors*, the values of α_1 and α_2 are irrelevant.

Obviously, a ranking of the damage based on the proposed model may not always provide an answer—this model offers only partial ranking. Our model can also quantify the damage rather than rank it. While such a model is higher in its input requirements and is more complex, it can be useful in circumstances in which a more light-weight ranking model fails. It also has the advantage that a quantitative estimate of the damage can be more compatible with a broad quantitative assessment that analyzes various factors apart from damage.

Another direction that is currently under study is a ranking that distinguishes between decision error type 1 (false positive), e.g., when a property that does not satisfy the criteria is included in the short list of suitable properties, and decision error type 2 (false negative), e.g., when a property that has the desired attributes is excluded from that list. This distinction is motivated by the understanding that, in real world settings, the implications of a false negative can differ greatly from the implications of a false positive.

A major concern for the application of the ranking model is the need to obtain parameter estimates. While data quality measurement methods are outside the scope of this work, a growing number of studies explore this issue (Naumann and Rolker, 2000). For example, (Ballou et al 2006; Motro and Rakov 1997; Parssian 2006) refer to the use of high quality data samples, while (Hipp et al. 2001) study a less costly data quality assessment through data mining. This approach detects errors through associations between different data. Specifically, deviations from such

patterns are perceived to be errors. Other common methods which are usually not too costly include user and expert evaluations (Naumann and Rolker, 2000). As hinted in the discussion of the property data ranking example and elsewhere, we believe that such methods can also be useful for the purpose of implementing this ranking theory. However, future research should include an investigation of real world cases that can instruct us about practical parameter assessment methods, and, in general, shed light on implementation issues as well as the overall usefulness of this model.

Acknowledgement

I am deeply grateful to Professor David E. Pingry for enabling this research by providing access to the High Performance Computing (HPC) system at the University of Arizona.

References

Askira Gelman, I., "A Model of Error Propagation in Satisficing Decisions and its Application to Database Quality Management," *14th Americas Conference on Information Systems* (AMCIS), 2008.

Askira Gelman, I., "Setting Priorities for Data Accuracy Improvements in Satisficing Decision-making Scenarios: A Guiding Theory," *Decision Support Systems*. Forthcoming.

Ali, K. M., and Pazzani M. J., "Error Reduction through Learning Multiple Descriptions." *Machine Learning* (24:3), 1996, pp. 173-202.

Avenali, A., Batini, C., Bertolazzi, P., and Missier, P. "Brokering Infrastructure for Minimum Cost Data Procurement Based on Quality–Quantity Models," *Decision Support Systems* (45:1), 2008.

Ballou D. P., and Pazer, H. L. "Modeling Data and Process Quality in Multi-input, Multi-output Information Systems," *Management Science* (31:2), 1985.

Ballou D. P., and Pazer, H. L. "A framework for the analysis of error in conjunctive, multi-criteria, satisficing decision processes," *Decision Sciences* (21:4), 1990.

Ballou D. P., and Tayi, G.K. "Methodology for Allocating Resources for Data Quality Enhancement," *Communications of the ACM* (32:3), 1989.

Ballou D. P., Wang, R. Y., Pazer, H. L., and Tayi, G.K. "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science* (44:4), 1998.

Ballou D. P. , Chengalur-Smith, I. N., and Wang, R. Y. "Sample-Based Quality Estimation of Query Results in Relational Database Environments." *IEEE Transactions on Knowledge and Data Engineering* (18:5), 2006.

Bevington, P. R., *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York, 1969.

Clemen, R.T., and Winkler, R.L. "Limits for the precision and value of information from dependent sources." *Operations Research*, (33:2), 1985, pp. 427-442.

Condorcet , Nicolas Caritat de (1785). *Essai sur l'application de l'analyse a la probabilité des décision rendues à la pluralité des voix*. Paris.

Cover, T. "The best two independent measurements are not the two best." *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-4, No. 1, 1974, pp. 116-117.

Cushing, B. E. "A Mathematical Approach to the Analysis and Design of Internal Control Systems," *Accounting Review* (49:1), 1974.

Eckerson, W. "Achieving Business Success through a Commitment to High Quality Data," *TDWI Report Series, The Data Warehousing Institute*, 2002.

Einhorn, H.J. "The use of nonlinear, Noncompensatory Models in Decision Making," *Psychological Bulletin* (73:3), 1970.

Einhorn, H.J. "The Use of Nonlinear, Noncompensatory Models as a Function of. Task and Amount of Information," *Organizational Behavior and Human Performance* (6:1), 1971.

Einhorn, H.J. Expert "Measurement and Mechanical Combination," Organizational *Behavior and Human Performance* (7), 1972.

Elashoff, J.D., Elashoff, R.M., and Goldman, G.E. "On the choice of variables in classification problems with dichotomous variables," *Biometrika* (54), 1967, pp. 668-670.

Fang, G.S. "A Note on Optimal Selection of Independent Observables," *IEEE Transactions on Systems, Man and Cybernetics* Vol. SMC-9, No. 5, 1979, pp. 309-311.

Fishman, G.S. *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Verlag, New York, 1995.

Grofman, B., Owen, G., and Feld, S.L. "Thirteen Theorems in Search of the Truth." *Theory and Decision* 15, 1983.

Hamlen, S. S. "A Chance-Constrained Mixed Integer Programming Model for Internal Control Design," *Accounting Review* (55:4), 1980.

Hevner, S. March, J. Park, and S. Ram, "Design Science Research in Information Systems," *Management Information Systems Quarterly* (28:1), March 2004, pp. 75-105.

Hipp, J. , Guntzer, U. and Grimmer, G. "Data Quality Mining: Making a Virtue of Necessity," in *Proceedings of the 6th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery* (DMKD 2001), 2001.

Janson, M. "Data quality: The Achilles Heel of End-User Computing," *Omega: International Journal of Management Science* (16:5), 1988.

Juran, J.M. *Juran on Planning for Quality* , The Free Press, New York, 1988.

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W. "Limits on the Majority Vote Accuracy in Classifier Fusion," *Pattern Analysis and Applications* (6:1), 2003, pp. 2-31.

Ladha, K. "Information Pooling through Majority-rule Voting: Condorcet's Jury Theorem with Correlated Votes." *Journal of Economic Behavior and Organization* Vol. 26, 1995, pp. 353-372.

Lee Y., Strong D., Kahn B., and Wang R., "AIMQ: A methodology for information quality assessment," *Information & Management* (40:2), 2002, pp. 133-146.

Lussier D.A. and Olshavsky, R.W. "Task Complexity and Contingent Processing in Brand Choice," *The Journal of Consumer Research* (6:2), 1979.

March, S. and Smith, G. "Design and Natural Science Research on Information Technology," *Decision Support Systems* 15, 1995, pp. 251 - 266.

Mintz, A. "How Do Leaders Make Decisions? A Poliheuristic Perspective," *Journal of Conflict Resolution* (48:1), 2004.

Mitchell H.B. *Multi-Sensor Fusion: An Introduction*. Springer-Verlag (2007)

Motro A. and Rakov, I. "Not All Answers Are Equally Good: Estimating the Quality of Database Answers," in *Flexible Query-Answering Systems* (T. Andreasen, H. Christiansen, and H.L. Larsen, Editors). Kluwer Academic Publishers, 1997, pp. 1-21.

Naumann, F. and Rolker, C. "Assessment Methods for Information Quality Criteria," in *Proceeding of the 5th International Conference on Information Quality* (ICIQ-2000) (2000).

Nunamaker, J., Chen, M. and Purdin, T. "System Development in Information Systems Research," *Journal of Management Information Systems*, (7:3), 1991, pp. 89 – 106.

Park, C. W. "The Effect of Individual and Situation-Related Factors on Consumer Selection of Judgmental Models," *Journal of Marketing Research* (13:2), 1976.

Parsaye K. and M. Chignell, "Data Quality Control with SMART Databases," *AI Expert* (8:5), 1993.

Parssian, A., Sarkar, S., and Varghese, S.J. "Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product," *Management Science* (50:7), 2004.

Parssian, A., "Managerial Decision Support with Knowledge of Accuracy and Completeness of the Relational Aggregate Functions," *Decision Support Systems* (42:3), 2006.

Payne, J.W. "Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis," *Organizational Behavior and Human Performance* 16 , 1976.

Phipps, A. G. "Utility Function Switching during Residential Search," *Geografiska Annaler. Series B, Human Geography* (65:1), 1983.

Pipino, L., Lee, Y.W. and Wang, R.Y. "Data Quality Assessment," *Communications of the ACM* (45:4), 2002.

Pokorney L.R., "Determining the Cost and Effectiveness of Enhancing the Data in the US Defense Logistics Agency Supply Chain," in *Proceedings of the 11th International Conference on Information Quality*, 2006.

Redman, T.C. "Data: an unfolding disaster," *DM Review Magazine*, 2004.

Shankaranarayan, G., Zaid, M., and Wang, R.Y. "Managing Data Quality in Dynamic Decision Environments: An Information Product Approach," *Journal of Database Management* (14:4), 2003.

Simon, H. A. *Models of Man: Social and Rational*, John Wiley and Sons, Inc, 1957.

Stratton, W.O. "Accounting Systems: The Reliability Approach to Internal Control Evaluation," *Decision Sciences* (12:1), 1981.

Toussaint, G.T. "Note on Optimal Selection of Independent Binary-Valued Features for Pattern Recognition," *IEEE Transactions on Information Theory*, Vol. IT-17, 1971, p. 618.

Wang, R.Y., "A Product Perspective on Total Data Quality Management," *Communications of the ACM* (41:1), February 1998, pp. 58-63.

Wang, R.Y., and Strong, D.M. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), 1996.

Wang, R.Y., Ziad, M., and Lee, Y.W. *Data Quality*, Springer, 2001.

Winter, R. "Guest Editorial: Design Science Research in Europe," *European Journal of Information Systems* (17:5), 2008.

Yu, S. and J. Neter, A "Stochastic Model of the Internal Control System," *Journal of Accounting Research* (11:2), 1973.