Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2008 Proceedings

Americas Conference on Information Systems (AMCIS)

2008

A Model of Error Propagation in Satisficing Decisions and its Application to Database Quality Management

Irit Askira Gelman
University of Arizona, askirai@email.arizona.edu

A Model of Error Propagation in Satisficing Decisions and its Application to Database Quality Management

Irit Askira Gelman
University of Arizona
askirai@email.arizona.edu

ABSTRACT

This study centers on the accuracy dimension of information quality and models the relationship between input accuracy and output accuracy in a popular class of applications. Such applications consist of dichotomous decisions or judgments that are implemented through conjunction of selected criteria. Initially, this paper introduces a model that designates a single decision rule which employs a single binary conjunction operation. This model is extended to handle multiple, related decision rules that consist of any number of binary conjunction operations. Finally, application of the extended model is illustrated through the example of an online hotel reservation database. This example demonstrates how the new model can be utilized for ranking and quantifying the damage that errors in different database attributes inflict. Numerical estimates of the model can be integrated into cost-benefit analyses that assess alternative data accuracy enhancements or process or system designs.

KEYWORDS

Data quality management, information accuracy, databases, Statistical-Mathematical model, satisficing decisions.

INTRODUCTION

Nearly every organization is plagued by bad data, resulting in higher costs, angry customers, compromised decisions, and greater difficulty for the organization to align departments. The overall cost of poor data quality to businesses in the US has been estimated to be over 600 billion dollars a year (Eckerson, 2002), and the cost to individual organizations is believed to be 10%-20% of their revenues (Redman, 2004). Apparently, such estimates are not impressive enough to drive organizations to action—most of them do not have corporate data quality programs in place. Moreover, many have no plans for managing or improving data quality in the future (Eckerson, 2002).

The disregard that organizations show for the quality of their data is explained by the general difficulty of assessing the economic consequences of the data quality factor and the substantial cost that can be involved in achieving high data quality (Eckerson, 2002, Redman, 2004). The economic aspect of data quality has been drawing a growing research interest in recent years. An understanding of this aspect can be crucial for convincing organizations to address the data quality issue. It can guide decisions on how much to invest in data quality and how to allocate the limited organizational resources (Wang and Strong, 2006).

The economics of data quality, however, is partly determined by the relationship between the quality of the data and the quality of the information that the information system outputs. This is because data often undergo various processing before any actual use. An increasing number of Management Information Systems (MIS) studies have explored this relationship, mainly from a methodological perspective, but also from empirical and, occasionally, analytical perspectives. Parallel questions have been studied in numerous research areas. However, our grasp of the relationship between an information system's input quality and its output quality is still often limited.

This study centers on the accuracy dimension of information quality and uncovers the relationship between input accuracy and output accuracy in a popular class of applications. Such applications consist of dichotomous decisions or judgments that are implemented through logical conjunction of selected criteria. Decision-making instances that are implemented through conjunction are often labeled "satisficing." This term was 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. Evidence in this direction has been found in consumer choice settings, medical diagnosis, job preference decisions, university admission decisions, residential rental searches, political leaders' decision-making, and in many other domains (e.g., Einhorn, 1971,1972, 1973; Lussier and Olshavsky,1979; Mintz, 2004; Park, 1976; Payne, 1976; Phipps, 1983).

Consider, for example, a decision regarding the reservation of a hotel room. Nowadays, this decision is often carried out using an online hotel reservation system such as www.hotels.com and akin websites. Suppose that, due to the high number of alternatives, a satisficing decision strategy is employed for the initial screening of alternatives (e.g., Lussier and Olshavsky, 1979; Payne, 1976) or throughout the entire selection process. Presumably, for instance, three widespread decision variables are location, price, and the star ranking of the hotel, and the decision rule that combines such variables is conjunctive. A particular decision maker may, for example, look for a hotel room in downtown Tucson, Arizona, that has a star ranking of three stars or more, in the price range of $90-$150 per night. This decision could be implemented as follows: The values of each variable would be tested against the matching criterion (the registered location of each hotel in the given database would be verified against downtown Tucson, Arizona; its star ranking would be verified against three stars or more, and so on). These tests would determine the values of three matching dichotomous variables (e.g., a test of a hotel in New York would determine that the value of the corresponding variable is "false" while a test of a hotel in downtown Tucson, Arizona would produce the value "true"). Subsequently, the values of the dichotomous variables which are determined in this way would be combined using conjunction operation(s) to produce the outcome of the decision. A hotel that satisfies the entire rule would result in a positive decision and would be included in the resultant hotel set. Obviously, in agreement with the common experience, one may assume that the hotel reservation database is not free of errors and such errors can lead to incorrect hotel selection and rejection decisions.

Our study examines the effect of errors in the classification of the values of a specific decision variable as fulfilling or not fulfilling the decision criterion on the accuracy of the outcome of a conjunctive decision. Our objective is to enable comparisons and quantification of the damage that errors in different inputs inflict on the outcomes of conjunctive decisions. For example, given a hotel reservation database as above, this research provides the tools for identifying in which of the decision variables in use (location, price, etc.) errors would be most detrimental to users' decision validity. Specifically, this work focuses on quantifying the *change in decision error probability* due to a *change in the classification error probability* of a given input. This focus reflects the perception that the magnitude of the change in decision error probability corresponds to the damage that the respective classification errors, and, consequently, data errors, inflict. Hence, a higher effect on the decision error rate implies that the respective input data errors are more detrimental. The association between errors in the values of the decision variables and associated classification errors is not considered in this paper.

Numerical estimates of the relationships that are uncovered through this study can be useful for data and information quality assessments as well as for policy-making purposes. They can be integrated into cost-benefit analyses that assess alternative data quality enhancements or process or system designs.

The structure of this paper is as follows: The next section offers a short summary of relevant literature. Then, we introduce the model that designates a single conjunctive decision rule which is implemented through one or more binary conjunction operations. This model is later extended to the case of multiple, related decision rules. The implementation of the extended model is illustrated through the example of an online hotel reservation database as described above. The final section concludes this paper with a discussion of, its limitations and future research directions.

LITERATURE SURVEY

The literature on the relationship between input accuracy and output accuracy is vast. This relationship has been investigated in countless problem domains, assuming numerous information-processing models and data and error characteristics, and an assortment of accuracy measures. Some of these problem domains are Condorcet's jury theorem (e.g., Condorcet, 1785; Grofman, Owen, and Feld, 1983), feature selection (e.g., Cover, 1974), expert resolution (e.g., Clemen and Winkler 1985), and ensemble learning (e.g., Kuncheva and Whitaker, 2003). This brief survey does not attempt to discuss that rich literature. Importantly, however, our extensive literature scans have indicated the uniqueness of our research.

In MIS, the problem of the relationship between an information system's input quality and its output quality has received much attention through the Data Quality (DQ) literature. For the most part, research has maintained a methodological nature (e.g., Motro and Rakov, 1997; Ballou, Wang, Pazer, and Tayi, 1998; Naumann, Leser, and Freytag, 1999; Avenali, Bertolazzi, Batini, and Missier, 2004; Ballou and Pazer, 2005; Ballou, Chengalur-Smith, and Wang, 2006), but some studies

have examined this problem empirically (e.g., Klein and Rossin, 1999a; Klein and Rossin, 1999b) and, occasionally, using an analytical perspective (e.g., Askira Gelman, 2007a).

This study is a direct extension of (Askira Gelman 2007a, 2007b, forthcoming). Unlike this paper which is largely focused on multiple, related, conjunctive decisions, earlier work examined single satisficing decisions that consist of a sequence of conjunction and disjunction operations. Mainly, earlier work suggests a theory which helps identify conditions in which the Garbage In Garbage Out (GIGO) assumption is violated (Askira Gelman, 2007a) and provides simple principles for ranking the damage that errors in different inputs inflict on the outcome of a satisficing decision (Askira Gelman, forthcoming). Askira Gelman offers initial empirical validation of these theories through simulations (Askira Gelman, 2007b).

A model that is partly compatible with this research is described by Ballou and Pazer (Ballou and Pazer, 1990). Similar to our work, Ballou and Pazer propose a framework for assessing the effect of input errors on the accuracy of conjunctive decisions. However, their assumptions are substantially different from this research. For example, they consider only numerical decision variables and account for errors in the data as well as errors in the decision criteria.

Various frameworks are related to our online hotel reservation database example. In particular, such frameworks address the relationship between the quality of the raw data and the quality of the output of a relational database query. Parssian et al. (Parssian, Sarkar, and Varghese, 2004) assess the relationship between the quality of the data and the quality of the output of a query. The quality dimensions of interest there are completeness, accuracy, and membership. Our model, in contrast, is limited to the dimension that they label "membership." In addition, while Parssian et al.'s unit of analysis is the relation as a whole, the unit of analysis here is the individual attribute. Motro and Rakov (Motro and Rakov, 1997) describe a data analysis method that identifies data subsets which are homogeneous in their soundness or completeness. They employ aggregates of the data quality estimates that their method generates to assess the quality of query answers. Additional instances of work that accounts for the relationship between the quality of raw data and the quality of the output of queries include (Wang, Reddy, and Kon, 1995; Naumann et al., 1999; Avenali et al., 2004). In contrast to that research stream's concentration on single queries, the present study enables an overall evaluation of the effect of errors in specific attributes and its sensitivity to a change in the error rate.

An implicit assumption of this inquiry is that errors can be differentiated based on the intended use of the data. Counter to an approach that does not differentiate between errors (e.g., Janson, 1988; Parsaye and Chignel, 1993), an approach that differentiates between errors based on the intended use of the data is consistent 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). A recent work that resembles the viewpoint that underlies our research introduces a data quality assessment method for database settings that accounts for variations in the potential utility of the data (Even and Shankaranarayan, 2007). In general, nowadays there are various tools and methods that guide the design and resource allocation from a data utilization perspective. For instance, Ballou and Pazer (Ballou and Pazer, 1985) belong in this class; they propose a framework for tracking numeric data errors through an information system to assist with estimating the impact of errors on the output. Notably, (Ballou et al. 1998) and various other studies (e.g., Shankaranarayan, Zaid, and Wang, 2003) have extended the model of Ballou and Pazer in several directions. Prioritization of data quality issues according to users' perceptions and needs is assisted today by several methods and tools (e.g., Wang and Strong, 1996; Lee, Strong, Kahn, and Wang, 2002).

A MODEL OF A SINGLE DECISION RULE

This section demonstrates the formulation of the change in decision error probability due to a change in the probability of a classification error of a specific input of a given decision rule. The interest in this relationship is motivated by the perception that the magnitude of the change in decision error probability expresses the damage that the respective input classification errors, and, consequently, data errors, inflict. As mentioned in the introduction, the term classification error refers to a condition where a value of a decision variable is incorrectly judged as fulfilling the matching decision criterion or not fulfilling it. A decision error can be a Type 1 error; e.g., when including a hotel that does not satisfy the criteria, or it can be a Type 2 error; e.g., when rejecting a hotel that has the desired attributes.

Initially we explore a decision rule which consists of one binary conjunction operation. Later, we generalize our study to a decision rule which consists of a sequence of binary conjunction operations. The ensuing analysis exploits statistical properties of random variables, mainly their expected values. The variables in use by this analysis are listed and defined below:

- U, V: The correct input classifications. These are dichotomous random variables that accept the values 1 and 0. These values correspond to *true*, i.e., the value of the decision variable fulfills the relevant selection criterion, and *false*, i.e., the

value of the decision variable does not fulfill the criterion, respectively. The terms "input" and "input classification" will be used interchangeably.

- W: The correct decision outcome; W is a dichotomous random variable that accepts the values 1 (*true*, i.e., the correct decision is positive) and 0 (*false*, i.e., the correct decision is negative).

- D_U, D_V: These are dichotomous random variables that accept the values 1 and 0, which correspond to *error* and *no error*, respectively. Namely, these variables inform us of the occurrence of an error in the observed value of U and V, respectively.

- U_a, V_a: The observed, possibly incorrect representation of U and V; U_a and V_a are dichotomous random variables that accept the values 1 (*true*) and 0 (*false*).

- W_a: The decision output that is generated based on the observed data; W_a is a dichotomous random variable that accepts the values 1 (*true*) and 0 (*false*).

- D_W: A dichotomous random variable that accepts the values 1 (*error*) and 0 (*no error*). This variable tells us if the observed decision, W_a, is correct or not.

Statistical parameters:

- p_U , p_V , p_W , p_{D_U} , p_{D_V} , p_{D_W} : Expected values; subscripts identify the relevant random variables. For example, the expected value of U is denoted by p_U , i.e., $p_U = E(U) = \Pr(U = 1)$. Note that p_U (as well as p_V) is equal to the probability that a given value satisfies the criterion on the corresponding decision variable; this is the same as the fraction of true values that satisfy the criterion. Also, the expected value of a random variable that represents the occurrence of an error is the same as the probability of occurrence of that error. The terms "error probability" and "error rate" will be used interchangeably.

To simplify the analysis, we impose certain statistical independence requirements. We will discuss these assumptions shortly in the last section of this paper.

Statistical Independence Assumption (SIA) 1: None of the variables in $\{U, V, D_U, D_V\}$ or products of these variables is statistically dependent on any other variable in $\{U, V, D_U, D_V\}$ or any product of the variables.[1]

The ideal conjunction operation—where inputs are error-free—is captured by:

$$W = UV \tag{1}$$

The consistency of (1) with the definition of logical binary conjunction can be verified through a systematic evaluation of W for each possible combination of the values of U and V. Analogously, the observed operation is given by:

$$W_a = U_a V_a \tag{2}$$

The relationship among U_a, D_U, and U is given by:

[1] SIA 1 does not contradict the understanding that an error in a binary variable is inherently negatively correlated with the true variables (Aigner, 1973). While such understanding relates to the magnitude of the error, this work accounts for the incidence of an error, not its magnitude.

$$U_a = (1 - D_U)U + D_U(1 - U) = U + D_U - 2UD_U \quad (3)$$

If the value of D_U is zero, that is, if this variable indicates that no error has occurred, then (3) is reduced to U_a=U, i.e., the observed input is the same as the correct input. However, if the value of D_U indicates the occurrence of an error, then (3) assigns a value of one to U_a if U is zero and a value of zero if U is one. An equivalent relationship exists among V_a, D_V, and V, and among W_a, D_W, and W.

$$V_a = (1 - D_V)V + D_V(1 - V) = V + D_V - 2VD_V \quad (4)$$

$$W_a = (1 - D_W)W + D_W(1 - W) = W + D_W - 2WD_W \quad (5)$$

Assuming (1), the definition of Expected Value, and SIA 1, we easily derive the correct probability of a positive decision:

$$p_W = p_V p_U \quad (6)$$

Furthermore, using (1)-(5), the link among the probability of a decision error, the correct probabilities of satisfying the criteria, and classification error probabilities, is described by (7). Equation (7) asserts that the probability of a decision error is equal to an aggregate of products of one or more of the probabilities of satisfying the decision criteria and/or one or more of the classification error probabilities. In particular, these products include $p_V p_{D_U}$, $p_U p_{D_V}$, $p_{D_U} p_{D_V}$, $p_U p_{D_U} p_{D_V}$, $p_V p_{D_U} p_{D_V}$, and $p_U p_V p_{D_U} p_{D_V}$. The sign of the terms $p_V p_{D_U}$, $p_U p_{D_V}$, $p_{D_U} p_{D_V}$, and $p_U p_V p_{D_U} p_{D_V}$ is positive, and the remaining terms are negative.

$$p_{D_W} = p_V p_{D_U} + p_U p_{D_V} + p_{D_U} p_{D_V} - 2p_U p_{D_U} p_{D_V} - 2p_V p_{D_U} p_{D_V} + 2p_U p_V p_{D_U} p_{D_V} \quad (7)$$

The proof is shown in (Askira Gelman 2007a). The change in decision error probability due to a change in the classification error probability of a chosen input can now be easily calculated through partial derivation of (7). The derivative of p_{D_W} with respect to p_{D_U} is given by:

$$\partial p_{D_W} / \partial p_{D_U} = p_V + p_{D_V} - 2p_{D_V}(p_U + p_V - p_U p_V) \quad (8)$$

The derivative of p_{D_W} with respect to p_{D_V} is equivalent to its derivative with respect to p_{D_U} (an appropriate change of notation is required).

The calculation of the change in decision error probability can be easily extended to decision rules which consist of successive applications of the binary conjunction operation. In scenarios involving successive applications of an operation, such calculation is implemented by applying equations (6)-(8) repeatedly. The output of one binary operation is treated as an input of a subsequent binary operation.

MULTIPLE, RELATED DECISION RULES

The model that we have described in Section 3 may be useful if data utilization is limited to a single decision rule. Evidently, however, the use of data is rarely limited to one decision rule. In the majority of practical settings, data are used over and over again in different ways. In this section we extend our model to account for conditions where data serve N satisficing decision rules that bear partial resemblance. Specifically, we focus on such decisions that are all based on the same set of decision variables and the same binary operations, but the criteria that are tested against the decision variables vary across the decisions. As an example, think about a catalog database where products are chosen based on a subset of a given set of product attributes, but each buyer can impose different requirements on any such attribute. Evidently, an error in the classification of a value as matching, or not matching, a criterion that is imposed in one decision instance may or may not occur with respect to a criterion that is utilized in another instance.

The mathematical notation that has been introduced earlier is modified to accommodate the broader scenario. Instead of U, V, W, U_a, V_a, and W_a, we refer to U_j, V_j, W_j, U_j^a, V_j^a and W_j^a, respectively (j=1,..,N); each set of these random variables designates a distinct decision. In addition, a significant technical distinction from the earlier analysis is that p_U, p_V, and

p_W are treated as random variables rather than parameters. Each of these variables has *N* observed values, all in the range [0,1]. In a similar fashion, instead of D_U , D_V, and D_W , we refer to D_j^U , D_j^V , and D_j^W , respectively (*j*=1,..,*N*), and p_{D_U} , p_{D_V} , and p_{D_W} are treated as random variables that accept values in the range [0,1].

Assumption SIA 1 is relaxed. Instead of the requirement that U_j, V_j, D_j^U , and D_j^V maintain statistical independence, a sufficient requirement is that dependencies offset each other such that, *on the average*, the effect of errors on the output error probability is the same as when SIA 1 holds true.

Statistical Independence Assumption (SIA) 1*: Under a conjunction operation: $E(p_{D_W}) = E(p_V p_{D_U} + p_U p_{D_V} + p_{D_U} p_{D_V} - 2p_U p_{D_U} p_{D_V} - 2p_V p_{D_U} p_{D_V} + 2p_U p_V p_{D_U} p_{D_V})$.

In addition, we assume that p_U , p_V , p_{D_U} , and p_{D_V} are not involved in any statistical dependencies. This convenient assumption does not seem to be particularly restrictive when decisions are independent.

Statistical Independence Assumption (SIA) 2: None of the variables in { p_U , p_V , p_{D_U} , p_{D_V} } or products of these variables is statistically dependent on any other variable in { p_U , p_V , p_{D_U} , p_{D_V} } or any product of these variables.

When considering multiple decision rules rather than a single decision rule, we modify the objective of the analysis accordingly. Instead of quantifying the change in decision error probability due to a change in classification error probability of a single input, we formulate the change in the *expected value* of the decision error probability due to a change in the *expected value* of classification error probability of a chosen input. In the case of decisions that are implemented through one binary conjunction operation, we can easily infer from (7), SIA 1*, and SIA 2 that the mean decision error probability is given by:

$$\begin{aligned} E(p_{D_W}) = {} & E(p_V)E(p_{D_U}) + E(p_U)E(p_{D_V}) + E(p_{D_U})E(p_{D_V}) - 2E(p_U)E(p_{D_U})E(p_{D_V}) \\ & - 2E(p_V)E(p_{D_U})E(p_{D_V}) + 2E(p_U)E(p_V)E(p_{D_U})E(p_{D_V}) \end{aligned} \tag{9}$$

Consequently, by applying a partial derivation operation, we derive the relationship that is of interest to us:

$$\partial E(p_{D_W}) / \partial E(p_{D_U}) = E(p_V) + E(p_{D_V}) - 2E(p_{D_V})[E(p_U) + E(p_V) - E(p_U)E(p_V)] \tag{10}$$

Finally, analogous to equation (6), equation (11) provides an additional tool that is needed in order to address *N* similar decision rules, each consisting of multiple binary operations:

$$E(p_W) = E(p_V)E(p_U) \tag{11}$$

By using equations (9)-(11) repeatedly, comparable to equations (6)-(8), we acquire the capacity to handle *N* similar decision rules, each consisting of a sequence of binary operations. An illustration of the application of our extended model follows next.

ILLUSTRATION: AN ONLINE HOTEL RESERVATION SYSTEM

The following example refers to an online hotel reservation system. That is to say, a website that presents information about the hotels in a chosen geographic area and provides the tools for reserving hotel rooms online. Two popular instances of such systems are WWW.HOTELS.COM and WWW.ORBITZ.COM . Apparently, different websites in this category feature different hotel selection criteria. Some are limited to a minimal number of attributes (e.g., location, dates, and number of people), while others allow the user to express various preferences, such as price range, desired amenities, hotel type (e.g., resort hotel or bed and breakfast), average guest ranking, star ranking, and other preferences.

In our scenario, the hotel reservation system offers a rich set of selection criteria. However, consistent with research findings (Markey, 2007), users largely ignore many of the available selection criteria. Users' choices are mostly determined based on a combination of the following three criteria: location (city), price range, and average guest ranking. A preferred ranking is

specified in terms of a pre-determined scale, e.g., 1 to 5, and the user chooses a minimum value, e.g., 2.6 or 4. A typical search reflects some flexibility regarding the desired price and guest ranking. To simplify this illustration we will ignore all other decision rules that users employ, i.e., we will assume that all users base their choices on a conjunction of the above three criteria. This assumption can be relaxed without much difficulty—weaker assumptions will be briefly discussed at the end of this section.

HOTELS.COM and ORBITZ are supported by large databases, enabling potential customers to reserve hotel rooms in many cities worldwide. Our database is large enough that a particular location usually corresponds to a small percentage of the registered hotels. In contrast, each of the price range and guest ranking preferences ordinarily matches a significantly higher percentage of the entries. Suppose that, by monitoring users' interactions with the system through query logs or in some other manner, database managers have collected data about users' interactions with the system and have generated unbiased estimates of relevant parameters. The average fraction of database records that satisfy the location requirement, denoted next by $\overline{p}_U^a$, is equal to 0.007 ($\overline{p}_U^a$=0.007), i.e., a user's location criterion matches, on the average, 0.7% of the database entries. The average fraction of records that satisfy the price range requirement, $\overline{p}_V^a$, is equal to 0.11 ($\overline{p}_V^a$=0.11), and the average fraction of records that satisfy the ranking requirement, $\overline{p}_X^a$, is equal to 0.21 ($\overline{p}_X^a$=0.21).

As is often the case, our database is not free of errors. Inaccurate guest ranking is mainly due to false rankings (either favorable or unfavorable), motivated by interests which are unrelated to serving the community. Errors in price and location are rarer. Errors in price are often related to deficiencies in the update process. Errors in location originate in weaknesses of data collection and entry processes. Obviously, errors in the data produce errors in the classifications of items as either passing, or not passing, the criteria of users. Such errors can lower the perceived reliability of the data and lead to missed business opportunities. From a data quality manager's perspective, identifying which errors are most detrimental to users' choice validity and quantifying the damage that errors in different attributes generate can be very useful. In particular, in this instance management plans to integrate these quantitative estimates in a cost-benefit analysis that would determine how resources that are available for data accuracy enhancements should be allocated.

In agreement with our research model, we assume that the statistical independence assumptions SIA 1* and SIA 2 are met. A study of the potential validity of these assumptions is beyond the scope of this short paper. Clearly, however, while in some cases our independence assumptions may be close enough to real world conditions, there is a need to generalize our model to conditions in which these assumptions do not hold true. The concluding section addresses this issue.

Suppose now that, by analyzing a sample of user query logs, or by using some other method, database managers produce error averages that are unbiased estimates of the mean error probabilities.[2] These estimates are: $\overline{p}_{D_U}$=0.001 (for location), $\overline{p}_{D_V}$=0.0009 (price range), and $\overline{p}_{D_X}$=0.01 (average guest ranking). Since these are unbiased estimates, the following equalities hold true: $E(\overline{p}_{D_U}) = E(p_{D_U})$, $E(\overline{p}_{D_V}) = E(p_{D_V})$, and $E(\overline{p}_{D_X}) = E(p_{D_X})$.

It can be easily shown that, under our assumptions, $E(\overline{p}_V^a) = E(p_V)$, and, similarly, $E(\overline{p}_U^a) = E(p_U)$ and $E(\overline{p}_X^a) = E(p_X)$. Namely, the expected values of the average fractions based on the recorded, inaccurate data are equal to the expected values of the true fractions.

Given all of the above, we will now demonstrate the implementation of our model for producing quantitative estimates of the detrimental effects of the data errors. (Note that the algorithm assumes that the parameter estimates are good enough approximations of the actual parameters, such that the parameter notation is used in place of the notation of the estimates.) The process is carried out in three steps:

Inputs of the algorithm: $E(p_U)$, $E(p_V)$, $E(p_X)$, $E(p_{D_U})$, $E(p_{D_V})$, and $E(p_{D_X})$.

Outputs of the algorithm: $\partial E(p_{D_Z}) / \partial E(p_{D_U})$, $\partial E(p_{D_Z}) / \partial E(p_{D_V})$, $\partial E(p_{D_Z}) / \partial E(p_{D_X})$, where Z refers to the output of applying the conjunction operation on U, V, and X.

[2] Note that a growing number of DQ studies examine sampling methods for data quality assessment (e.g., Ballou et al. 2006; Parssian 2006).

Step 1.

Input: $E(p_U)=0.007$, $E(p_V)=0.11$, $E(p_{D_U})=0.001$, and $E(p_{D_V})=0.0009$.

Process:

1.1 Equation (11) enables us to derive $E(p_W)$ from the values of $E(p_U)$ and $E(p_V)$.

1.2 Equation (9) supports the calculation of $E(p_{D_W})$ from the values of $E(p_U)$, $E(p_V)$, $E(p_{D_U})$, and $E(p_{D_V})$.

1.3 Equation (10) enables us to derive $\partial E(p_{D_W})/\partial E(p_{D_U})$ and $\partial E(p_{D_W})/\partial E(p_{D_V})$, respectively, from the values of $E(p_U)$, $E(p_V)$, $E(p_{D_U})$, and $E(p_{D_V})$.

Output: $E(p_W)=0.00077$, $E(p_{D_W})=0.00012$, $\partial E(p_{D_W})/\partial E(p_{D_U})=0.11069$ and $\partial E(p_{D_W})/\partial E(p_{D_V})=0.00777$.

Step 2.

Input: $E(p_W)=0.00077$, $E(p_{D_W})=0.00012$, $E(p_X)=0.21$, and $E(p_{D_X})=0.01$.

Process: Analog to Step 1:

2.1 Equation (11) serves for calculating $E(p_Z)$.

2.2 Equation (9) serves for calculating $E(p_{D_Z})$.

2.3 Equation (10) enables us to derive $\partial E(p_{D_Z})/\partial E(p_{D_W})$ and $\partial E(p_{D_Z})/\partial E(p_{D_X})$.

All three calculations require a suitable change of notation in the equations.

Output: $E(p_Z)=0.00016$, $E(p_{D_Z})=0.000033$, $\partial E(p_{D_Z})/\partial E(p_{D_W})=0.21579$ and $\partial E(p_{D_Z})/\partial E(p_{D_X})=0.00084$.

Step 3.

Input: $\partial E(p_{D_Z})/\partial E(p_{D_W})=0.21579$, $\partial E(p_{D_W})/\partial E(p_{D_U})=0.11069$, and $\partial E(p_{D_W})/\partial E(p_{D_V})=0.00777$.

Process:

3.1 The value of $\partial E(p_{D_Z})/\partial E(p_{D_U})$ is derived using the composite function differentiation chain rule. $\partial E(p_{D_Z})/\partial E(p_{D_U}) = \partial E(p_{D_Z})/\partial E(p_{D_W}) \cdot \partial E(p_{D_W})/\partial E(p_{D_U})$.

3.2 Equivalently, $\partial E(p_{D_Z})/\partial E(p_{D_V}) = \partial E(p_{D_Z})/\partial E(p_{D_W}) \cdot \partial E(p_{D_W})/\partial E(p_{D_V})$.

Output: $\partial E(p_{D_Z})/\partial E(p_{D_U})=0.02389$, $\partial E(p_{D_Z})/\partial E(p_{D_V})=0.00168$.

To conclude, the procedure described above produces the values $\partial E(p_{D_Z})/\partial E(p_{D_U})=0.02389$, $\partial E(p_{D_Z})/\partial E(p_{D_V})=0.00168$, and $\partial E(p_{D_Z})/\partial E(p_{D_X})=0.00084$. These results agree with a theory that has been suggested by (Askira Gelman, forthcoming). In particular, when error rates are similar as in this scenario, then errors are more detrimental in an input in which the fraction of the values that satisfy the preference criterion is lower. That is, errors in an attribute where the matching criterion is satisfied by the smallest fraction of the data are predicted to be the most damaging to output accuracy, and so on. In this scenario, errors in location data are most detrimental, followed by errors in price range. Errors in guest ranking are the least detrimental.

Accounting for Additional Cost-Benefit Factors

The proposed algorithm can be modified to account for additional factors that affect the cost and benefit of higher data accuracy. Instead of handling one set of similar decision rules, the algorithm can be modified to account for varied user decisions. It can also be modified to figure in the distinct unit costs of data accuracy improvements, the distinct unit costs of accuracy assessment, weights that reflect variation in the importance of the decision variables, and so on. For instance, in order to handle varied user decisions, the algorithm would be modified to factor in the probabilities of the different decision rule classes and then to aggregate outcomes by decision variables. To illustrate this point, consider a situation where users of our online hotel reservation system are divided such that 80% of them adhere to a decision rule as described earlier, while the remaining users obey a somehow different decision rule. The choices of the latter are determined based on a conjunction of the desired city, price range, and the star ranking of the hotel (i.e., instead of the guest average rating). To resolve the modified scenario, the formerly specified outputs of the algorithm should each be multiplied by 0.8, i.e., $0.8 \cdot 0.02389$, $0.8 \cdot 0.00168$, and $0.8 \cdot 0.00084$; in addition, the second decision rule should be treated in a comparable manner and the subsequent algorithm outputs should each be multiplied by 0.2. Finally, related values should be aggregated, i.e., the two outputs pertaining to the location attribute should be added up, and so should the two outputs matching the price range.

CONCLUDING REMARKS

This research aims to enhance the understanding of the relationship between input accuracy and output accuracy, primarily for data quality management decision-making purposes. In particular, this work designates conditions in which data are used by one or more applications that have the form of a satisficing decision as outlined in the introduction to this study.

A potentially important limiting aspect of this study is the set of statistical independence assumptions that underlie our model. In the case of a single decision rule, the statistical independence assumptions are expressed by SIA 1. In the case of multiple decisions, the requirements on a single decision are relaxed somehow. Nonetheless, SIA 1*, which replaces SIA 1, imposes certain requirements on aggregates. We have also added independence assumptions through SIA 2. Statistical dependencies may be common in general, however. Therefore, the validity of our assumptions is an issue that deserves separate consideration. Notably, our model can be extended to account for statistical dependencies, and, furthermore, our initial investigation suggests that implementing such a model may not be too complicated.

Another possible weakness of this work is its neglect to address the relationship between classification errors and the data errors that cause them. This choice may be inconvenient due to its inconsistency with common practice and research. Mostly, the latter often refers to data errors rather than classification errors. Notably, however, the link between a classification error rate and the respective data error rate can be quite straightforward.

REFERENCES

1. Aigner, D.J. (1973) Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics*, 1, pp. 49-60.
2. Askira Gelman, I. Setting Priorities for Data Accuracy Improvements in Satisficing Decision-making Scenarios: A Guiding Theory. *Decision Support Systems*. Accepted for Publication.
3. Askira Gelman, I. (2007a) GIGO or not GIGO: Error Propagation in Basic Information Processing Operations. *Proc. American Conference on Information Systems 2007 (AMCIS 2007)*, Keystone, Colorado.
4. Askira Gelman, I. (2007b) Simulations of Error Propagation for Prioritizing Data Accuracy Improvement Efforts. *12th International Conference on Information Quality (ICIQ-07)*, MIT, Cambridge MA.
5. Avenali, A., Bertolazzi, P., Batini, C., and Missier, P. (2004) A formulation of the data quality optimization problem in cooperative information systems. *International Workshop on Data and Information Quality in conjunction with CAISE'04*, Riga, Latvia.
6. Ballou, D. P. and Pazer, H. L. (1985) Modeling Data and Process Quality in Multi-input, Multi-output Information Systems. *Management Science,* Vol. 31, No. 2, pp. 150-162
7. Ballou, D.P., and Pazer, H.L. (1990). A framework for the analysis of error in conjunctive, multi-criteria, satisficing decision processes. *Decision Sciences*, 21(4), pp. 752-770.
8. Ballou, D. P., Wang, R. Y., Pazer, H. L., and Tayi, G. K. (1998) Modeling information manufacturing systems to determine information product quality. *Management Science,* Vol. 44, No. 4, pp. 462-484.

9. Ballou, D. P., Chengalur-Smith, I. N., and Wang, R. Y. (2006) Sample-Based Quality Estimation of Query Results in Relational Database Environments. *IEEE Transactions on Knowledge and Data Engineering* Vol. 18, No. 5, pp. 639-650.
10. Clemen, R.T. and Winkler, R.L. (1985) Limits for the Precision and Value of Information from Dependent Sources, *Operations Research* 33(2).
11. Nicolas Caritat de Condorcet, Essai sur l'application de l'analyse a la probabilité des décision rendues à la pluralité des voix (Paris, 1785).
12. Cover, T. (1974) The Best Two Independent Measurements are Not the Two Best, *IEEE Transactions on Systems, Man and Cybernetics* SMC-4(1).
13. Eckerson, Wayne W. (2002) Achieving business success through a commitment to high quality data. *TDWI Report Series, The Data Warehousing Institute*.
14. Einhorn, H.J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, 73(3), pp. 221-230.
15. Einhorn, H.J. (1971) The use of nonlinear, noncompensatory models as a function of. task and amount of information. *Organizational Behavior and Human Performance*, Vol. 6, No. 1.
16. Einhorn, H.J. (1972) Expert measurement and mechanical combination. *Organizational Behavior and Human Performance*, Vol. 7, pp. 86-106.
17. Even, A., and Shankaranarayanan, G. (2007) Utility-Driven Assessment of Data Quality. *The DATA BASE for Advances in Information Systems*, Vol. 38, No. 2, pp. 75-93.
18. Grofman, B., Owen, G., and Feld S.L. (1983) Thirteen theorems in search of the truth, *Theory and Decision*, Vol. 15, No. 3, pp. 261-278.
19. Janson, M. (1988) Data quality: The Achilles Heel of End-User Computing, *Omega: International Journal of Management Science* 16(5).
20. Juran, J.M. (1988) *Juran on Planning for Quality* (The Free Press, New York).
21. Klein, B.D., and Rossin (1999a) D.F., Data quality in linear regression models: Effect of errors in test data and errors in training data on predictive accuracy. *Informing Science*, Vol. 2, No. 2.
22. Klein, B.D., and Rossin (1999b) D.F., Data quality in neural network models: Effect of error rate and magnitude of error on predictive accuracy. *Omega*, Vol. 27, No. 5, pp. 569-582
23. Kuncheva L.I. and Whitaker, C.J. (2003) Measures of Diversity in Classifier Ensembles and their Relationship with the Ensemble Accuracy, *Machine Learning* 51.
24. Lee, Y. Strong, D. Kahn, B. and Wang, R. (2002) AIMQ: A Methodology for Information Quality Assessment, *Information and Management* 40(2).
25. Lussier, D.A. and Olshavsky, R.W., (1979) Task complexity and contingent processing in brand choice. *The Journal of Consumer Research*, Vol. 6, No. 2, pp. 154-165.
26. Markey, K., (2007) Twenty-five years of end-user searching, Part 1: Research findings, *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 8, pp. 1071-1081.
27. Mintz, A. (2004) How do leaders make decisions? A poliheuristic perspective. *Journal of Conflict Resolution*, Vol. 48, No. 1, pp. 3-13.
28. 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.
29. Naumann, F., Leser, U., and Freytag, J. (1999) Quality-driven Integration of Heterogeneous Information Systems. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB 99.
30. Park, C. W. (1976) Prior familiarity and product complexity as determinants of the consumer's selection of judgmental models. *Journal of Marketing Research*, pp. 144-151.
31. Parsaye K. and Chignell, M. (1993) Data Quality Control with SMART Databases, *AI Expert* 8(5).

32. Parssian. A., Sarkar, S., and Varghese, S.J. (2004) Assessing data quality for information products: Impact of selection, projection, and Cartesian product. *Management Science*, Vol. 50, No. 7, pp. 967-982.
33. Parssian. A., (2006) Managerial Decision Support with Knowledge of Accuracy and Completeness of the Relational Aggregate Functions. *Decision Support Systems*, Vol. 42, pp. 1494-1502.
34. Payne, J.W. (1976) Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis. *Organizational Behavior and Human Performance*, Vol. 16, pp. 366-387.
35. Phipps, A. G. (1983) Utility function switching during residential search. *Geografiska Annaler. Series B, Human Geography*, Vol. 65, No. 1, pp. 23-38.
36. Redman, T.C. (2004) Data: an unfolding disaster. *DM Review Magazine*.
37. Shankaranarayan, G., Zaid M., and Wang, R. (2003) Managing data quality in dynamic decision environments: an information product approach. *Journal of Database Management*, Vol. 14, No. 4.
38. Simon, H., A. (1957) *Models of Man: Social and Rational*, John Wiley and Sons, Inc..
39. Wang, R., Reddy, M., and Kon, H. (1995) Toward quality data: An attribute-based approach, *Journal of Decision Support Systems*, Vol. 13, No. 3-4., pp. 349-372.
40. Wang, R.Y., and Strong, D.M. (1996) Beyond accuracy: What data quality means to data consumer. *Journal of Management Information Systems*, 12, pp. 5-34.