An Empirical Study of the GIGO Axiom in Satisficing Decisions

Irit Askira Gelman
Research and Solutions, Tucson, AZ, United States., askirai@dqiqsolutions.com

An Empirical Study of the GIGO Axiom in Satisficing Decisions

Irit Askira Gelman
DQIQ Research and Solutions
askirai@dqiqsolutions.com

ABSTRACT

An effort to improve data accuracy that yields poorer information accuracy when the data are processed would normally be labeled a major failure. While popular belief discounts the likelihood of such an event, research of conjunctive and disjunctive decision rules suggests that a negative association between input accuracy and decision accuracy is a deeply rooted phenomenon. In this paper we extend the understanding of this phenomenon through an empirical investigation of conjunctive decision rules using Monte Carlo simulations. The implications of this research are not limited to data accuracy; other data deficiencies can generate a comparable effect.

Keywords

Satisficing decisions, Multi-criteria decisions, Information accuracy, Data quality management, Resource allocation, Garbage in garbage out, GIGO, Simulation

INTRODUCTION

Common sense advises us that an effort to improve data quality should take into account the expected utility of the improvement when the data are used. The recent emphasis in data quality (DQ) research on the uses of data is in line with this understanding. In particular, contrary to solutions that do not differentiate among errors (e.g., Janson, 1988; Parsaye and Chignell, 1993) recent years have witnessed the growing popularity of approaches that take into account the intended use of the data (e.g., Lee, Strong, Kahn, and Wang, 2002; Pipino, Lee, and Wang, 2002; Wang, Reddy, and Kon, 1995) and differentiate accordingly among errors (e.g., Askira Gelman, 2010; Ballou and Tayi, 1989; Even, Shankaranarayan, and Berger, 2007). Obviously, an organizational effort to improve input data accuracy that yields poorer output information accuracy when the data are processed would be labeled a major failure by common standards, if detected. However, the popular belief discounts the likelihood of such an event, as expressed by the acronym GIGO (Garbage In, Garbage Out). Originally coined in the computer industry, this acronym, which indicates a strong positive link between input accuracy and output accuracy, is now widely accepted. For the most part, scientists have embraced the popular belief in GIGO and have treated GIGO as an axiom. Nonetheless, the scientific understanding of the relationship between input accuracy and output accuracy is still partial, and the sign of that relationship (positive or negative), in particular, is not well understood. Contrary to the widespread belief in GIGO there is, in fact, a growing literature that hints at a more complex association between input accuracy and output accuracy.

One theory in this category, established in several research domains, says that statistical dependencies between data sources or data errors can dramatically affect output accuracy (e.g., Askira Gelman, 2004; Ali and Pazzani, 1996; Barabash, 1965; Berg, 1993; Clemen and Winkler, 1985; Cover, 1974; Elashoff, Elashoff and Goldman, 1967; Fang, 1979; Frantsuz, 1967; Kuncheva, Whitaker, Shipp and Duin, 2003; Ladha, 1995; Nitzan and Paroush, 1984) This theory implies that higher data accuracy can produce lower information accuracy, subject to variations in statistical dependencies. A second theory originates in studies of prediction model-building paradigms, which indicate that adding noise to a data sample that serves in the construction of a model can improve the accuracy of the model (e.g., Bishop, 1995; Raviv and Intrator, 1996). Apparently, controlled levels of noise can compensate for limitations of the model-building algorithms.

A recent DQ study (Askira Gelman, 2011) examines the association between input accuracy and output accuracy, especially the sign of that association, from a new angle. That research explored the sign of the association in a highly used class of applications, namely, applications consisting of dichotomous decisions that are implemented through logical conjunction or disjunction of selected criteria. A mathematical-statistical analysis showed a surprising result: the sign of the relationship between input accuracy and decision accuracy varies. Notably, this discrepancy with GIGO is not explained by variations in statistical dependencies or information processing limitations, but rather by inherent properties of the logical conjunction and disjunction operations. The study proposed a model for determining the sign of the relationship and concluded that a negative sign is not reversed when new decision variables are added.

In this paper we extend and complement the work in (Askira Gelman, 2011) through an empirical investigation using Monte Carlo simulations. We verify several of the findings of the former study regarding conjunctive decision rules, and, mainly, we extend the qualitative understanding of its results through (1) a study of the source of the change in decision accuracy, (2) a study of the change in the likelihood of a discrepancy with GIGO when the number of decision criteria increases, and (3) a discussion of the validity of the theory under conditions that are outside the scope of the formal study.

The new theory can be used for improving the effectiveness and efficiency of resource allocation in data quality management settings. Consider a catalogue or an inventory database, for example. Users of these databases often employ conjunctive decision rules for the purpose of item selection or item screening. Our theory can assist both users and data owners to prioritize resource allocation among different database attributes when the objective is high information quality. Essentially, the findings of this research imply a small set of rules that can guide resource allocation decisions from an information quality perspective. Furthermore, the increasing use of noise injection for improving the accuracy of learning algorithms (Bishop, 1995; Raviv and Intrator, 1996) points to a potentially profitable, counter-intuitive use of our findings, which may be investigated through future work. Namely, future research should examine the value of error injection for improving decision accuracy.

The rest of this paper is structured as follows. A description of the conceptual foundation is given, followed by an overview of the intuitive insights that underlie this research. We then present the model and data that are used by the simulations, introduce our hypotheses, and describe the simulation results. Finally, we offer a discussion of the managerial implications of this work, as well as an analysis of the validity of this theory under conditions that are outside the scope of the formal study. A detailed review of related literature is provided by (Askira Gelman, 2011).

CONCEPTUAL BACKGROUND

Decision-making instances that are implemented through logical conjunction and/or disjunction are often classified as “satisficing,” a term 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, 1955). Research indicates that satisficing rules agree with human choices in diverse situations, often involving complex problems, such as when the number of alternatives or decision criteria is high (Payne, Bettman and Johnson, 1993). Evidence in this direction has been found in consumer choice settings, medical diagnoses, job preference decisions, university admission decisions, residential rental searches, political leaders’ decision-making, and in many other domains (e.g., Einhorn, 1970, 1971, 1972; Lussier and Olshavsky, 1979; Mintz, 2004; Park, 1976; Payne, 1976; Phipps, 1983).

In many practical situations the source of the decision input data is a database. In these circumstances a decision variable corresponds to a suitable database attribute or a collection of database attributes. (Note that 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.) Consider, for example, a website that presents data about residential properties for sale in a chosen geographic area. Websites in this category feature varying amounts of data and property selection criteria. The selection criteria are compatible with the understanding that, due to the high number of available properties, the decision maker employs a satisficing decision strategy for the initial screening of alternatives (e.g., Lussier and Olshavsky, 1979; Payne, 1976) or throughout the entire selection process. In our example, a relational real estate database provides a rich set of selection criteria. However, consistent with modern understanding of the cognitive limits on active human memory (see the “magical number seven, plus or minus two” principle, Miller, 1956), users, for the most part, ignore many of the available selection criteria. For instance, an investor may limit his or her initial decision variable set to three variables such as location, number of bedrooms, and price. In the remainder of this paper we will assume a decision maker who is looking for a property in zip code 85716 with two bedrooms or more, in the price range of $0-$50,000. This preference is expressed by a conjunctive decision rule that combines three criteria, i.e., zip code = 85716 *and* number of bedrooms ≥ 2 *and* price ≤ $50,000.

Obviously, in agreement with common experience, one may assume that the data are not free of errors. These errors can lead to incorrect classifications of input values as fulfilling or not fulfilling a decision criterion. Classification errors can, in turn, generate decision errors. A false positive real estate investment decision includes an unsuitable property in the short list of suitable properties, while a false negative decision excludes a property that may actually embody the decision maker’s dream house. *Accuracy* is defined by this work 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. While the measure of input data accuracy can vary, the accuracy of the output information is taken to be measured by decision error probability. 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. While the important question of how to derive the respective measures is outside the scope of this paper, there is a growing literature that offers practical solutions (e.g, Ballou, Chengalur-Smith and Wang, 2006; Hipp, Guntzer and Grimmer, 2001; Motro and Rakov, 1997 Parssian, 2006).

Our approach to studying the validity of the GIGO assumption utilizes a concept labeled *damage* that has been offered for prioritization and resource allocation purposes in data quality management settings (e.g., Askira Gelman, 2010). The damage that errors in an input inflict on output accuracy is defined as the *change in output accuracy due to a change in the accuracy of that input*. A negative damage is equivalent to a negative association between input accuracy and output accuracy, while a positive damage is equivalent to a positive association between input accuracy and output accuracy. Hence, the belief in GIGO is interpreted by this work as an assumption that the damage is positive—a finding of negative damage is inconsistent with this belief.

THE DISCREPANCY WITH GIGO: INTUITIVE INSIGHTS

We argue that a negative association between input accuracy and output accuracy stems from fundamental properties of the conjunction and disjunction operations. Therefore, it is not limited to a specific type of decision variable (e.g., numeric, categorical), nor is it exceedingly sensitive to statistical independence assumptions. This section lays out our understanding of the negative damage phenomenon in conjunctive decision rules. An equivalent explanation applies to disjunctive decision rules.

Consider the truth table 1(a) below. This table refers to the logical conjunction (AND) operation where the inputs of the operation are denoted by p and q. Let $\hat{p}$ denote an incorrect representation of p. Let $\hat{q}$ denote an incorrect representation of q. Table 1(b) describes a scenario in which an AND operation combines $\hat{p}$ and q. Similarly, 1(c) portrays an AND operation that combines p and $\hat{q}$, and 1(d) captures an AND operation in which both inputs are incorrect representations of the original inputs. As a whole, tables 1(b), 1(c), and 1(d) cover all the possible input error combinations.

p	q	p and q
FALSE	FALSE	FALSE
FALSE	TRUE	FALSE
TRUE	FALSE	FALSE
TRUE	TRUE	TRUE

(a) AND operation, correct inputs

$\hat{p}$	q	$\hat{p}$ and q
TRUE	FALSE	FALSE
TRUE	TRUE	TRUE
FALSE	FALSE	FALSE
FALSE	TRUE	FALSE

(b) AND operation with $\hat{p}$ and q

p	$\hat{q}$	p and $\hat{q}$
FALSE	TRUE	FALSE
FALSE	FALSE	FALSE
TRUE	TRUE	TRUE
TRUE	FALSE	FALSE

(c) AND operation with p and $\hat{q}$

$\hat{p}$	$\hat{q}$	$\hat{p}$ and $\hat{q}$
TRUE	TRUE	TRUE
TRUE	FALSE	FALSE
FALSE	TRUE	FALSE
FALSE	FALSE	FALSE

(d) AND operation with $\hat{p}$ and $\hat{q}$

Table 1. Truth tables (AND)

An earlier study (Askira Gelman, 2011) uses Table 1 to explain why (a) a negative damage is possible when the probability of satisfying one of the decision criteria is high, approaching one, while the probability of satisfying the other criterion is low, approaching zero, and (b) given two inputs that exhibit such a disparity, a negative damage is evidenced only when we change the accuracy of the input where the probability of satisfying the decision criterion is high; when we change the accuracy of the second input the damage is positive, as one would normally expect. As an example, consider the real estate investment search (two bedrooms or more in the price range of $0-$50,000 in zip code 85716). Ignore the zip code variable for the moment. Suppose that the decision maker is assisted by a property listing database where the large majority of the properties have two bedrooms or more, i.e., the criterion on the number of bedrooms is typically satisfied, while only a small number of properties are priced at $50,000 or less, such that the price criterion is rarely met. Given this scenario, the intuitive explanation suggested by (Askira Gelman, 2011) implies that when the accuracy of number of bedrooms data deteriorates, the output of a decision based on number of bedrooms and price would typically improve. That explanation also clarifies that when the accuracy of price data deteriorates, the accuracy of such a decision would deteriorate.

Negative Damage: The Source of the Change in Decision Accuracy

If the findings of (Askira Gelman, 2011) hold true then an empirical study of a decision that is implemented through conjunction of two inputs should show the following: When lower accuracy of input increases decision accuracy, the decrease in decision error rate is exhibited as a lower rate of false positive decisions; the rate of false negative decisions does not decline. The logic behind this statement is simple. According to (Askira Gelman, 2011), a negative damage is evidenced when we change the accuracy of the input where the probability of satisfying the decision criterion is high. The author explains that decision accuracy increases because errors in such an input offset errors in the second input. However, due to the nature of the second input (i.e., the probability that it satisfies the decision criterion is low), a decision error that is caused by an error in that input is typically a false positive decision. Therefore, when such an input error is offset by an error in another input, the outcome is that the false positive decision is reversed.

Negative Damage in Decision Rules that Combine Many Criteria

Suppose that a negative association between input accuracy and output accuracy is shown in a conjunctive decision that employs two decision variables. What happens if that decision rule is extended through conjunction with additional variables? Is the sign of the damage preserved through the entire multi-criteria decision rule? The mathematical statistical analysis by (Askira Gelman, 2011) shows that the answer to this question is commonly positive. A discrepancy with GIGO is a lasting phenomenon in this sense. The following paragraphs explain and justify these results.

Assuming that a discrepancy with GIGO is observed in the output of a conjunction operation involving two inputs, these inputs are often characterized, as explained above, by a substantial inequality in the probability of satisfying the decision criterion. Therefore, the proportion of input instances where both variables satisfy the criteria, such that the inputs of the logical conjunction operation are both “true,” is small. However, a conjunction operation has a well known property, which is illustrated by Table 1(a)—the output is “true” only if both inputs are “true.” Subsequently, the output of the conjunction operation exhibits a small proportion of “true” values when a discrepancy with GIGO comes up, and that proportion keeps decreasing when new variables are added to the decision rule. Now, suppose that two decision variables that have a conjunction showing a negative damage are combined with a third variable. Obviously, a negative sign of the damage will be preserved only if the association between the accuracy of the output of the first conjunction operation and the accuracy of the output of the second conjunction operation is positive—otherwise, the sign of the damage is reversed. A reversal is unlikely to happen, however. Since, as clarified above, the probability of “true” values for the output of the first conjunction operation is low, research (Askira Gelman, 2011) shows that errors in the output of the first conjunction are generally detrimental to the output of the second conjunction. That is, the association between the accuracy of the output of the first conjunction operation and the accuracy of the output of the second conjunction operation is positive. Therefore, the negative sign of the damage is preserved through a second conjunction operation for combining a third variable.

The Likelihood of a Negative Damage When Adding Criteria

A different question regarding a multi-criteria, conjunctive decision rule is whether the likelihood that errors in a given input will produce a negative damage increases, decreases, or does not change when we add criteria to the rule. Evidently, the answer to this question varies in general. Since an important prerequisite of a negative damage is a highly unequal probability of satisfying the decision criterion, and since the output of a conjunction operation exhibits decreasing probability of satisfying the decision criteria, approaching zero, that answer depends to a large extent on the likelihood that the decision variable will show a high probability, approaching one, of satisfying the matching criterion. Suppose that we adopt a “neutral” assumption that says that decision variables satisfy the criteria imposed on them with probabilities that vary randomly in the range between zero and one. One can see that, under these conditions, the likelihood of a negative damage increases when decision variables are added to the rule. In essence, as we continue to add criteria to a conjunctive decision rule, the probability that all the criteria are met decreases steadily, approaching zero. This downtrend, however, is independent of the probability that the criterion on a newly added variable is satisfied. Given that this probability varies randomly in the range between zero and one, the likelihood that the decision output satisfies the decision criteria with a sufficiently low probability and the chosen variable satisfies its matching criterion with a sufficiently high probability to generate a negative damage increases as we continue to add new decision criteria.

MODEL

The empirical tests of the theory center on conjunctive decision rules that join *N* decision variables ($2 \leq N \leq 10$) which are ordered one way or another. We denote the ideal, error-free variables V_i, $i = 1, 2, .., N$. The implementation of a conjunctive decision is as follows. Initially, for every *i*, the value of V_i is tested against the corresponding decision criterion (or criteria). The outcome of this test—zero for "false" or one for "true"—is captured by a matching, dichotomous variable I_i. The values of the dichotomous variables 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 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 O_1 is combined with the value of I_2, and the output is given by O_2. In the second iteration, the value of O_2 is combined with the value of I_3, and the output is given by O_3, and so on. It is easy to see that O_N registers the outcome of a conjunctive decision that accounts for all *N* decision variables.

Symbol	Meaning
V_i , V_i^R	Decision variable (random variable); V_i describes the correct data; V_i^R describes the observed, possibly incorrect data
I_i , I_i^R	Informs us whether V_i (/ V_i^R) passes the decision criterion or not (dichotomous random variable)
O_i , O_i^R	The output of a decision based on V_1 ,.., V_i , (/ V_1^R ,.., V_i^R) (dichotomous random variable)
F_i^V	Offers a measure of the error in the value of V_i^R
F_i^I	Informs us whether I_i^R is correct or not (dichotomous random variable)
F_i^O	Informs us whether O_i^R is correct or not (dichotomous random variable)
p_i^I , p_i^{FI} , p_i^O , p_i^{FO}	Expected (mean) values

Table 2. Notation

The observed, possibly incorrect representation of V_i is denoted by V_i^R ; the observed representation of I_i is denoted by I_i^R ; and the observed representation of O_i is denoted by O_i^R . The symbol F_i^V identifies a variable that offers a measure of the fault, or error, in the value of V_i^R . The value domain of F_i^V varies. The symbol F_i^I refers to a variable that informs us about the occurrence of error in the value of I_i^R ; F_i^I=0 if $I_i^R = I_i$, and F_i^I=1 otherwise. Likewise, F_i^O=0 informs us that the decision expressed by O_i^R is correct (i.e., $O_i^R = O_i$) and F_i^O=1 points to an error in that decision (a false negative decision or a false positive decision).

In order to simplify the model and increase the generality of the results, we limit the scope of the empirical study in terms of the variables that it explores. The tests are based on the understanding that the relationship between F_i^V and F_i^I is positive. That is, all other things being equal, a higher value of F_i^V implies a higher probability of error in the value of I_i^R, in agreement with GIGO.[1] This understanding enables us to exclude V_i and F_i^V from this study; recall that the explanation in an earlier section (see intuitive insights) does not account for these variables. Thus, the simulations center on the relationship between the accuracy of I_i^R, derived from V_i^R, and the accuracy of the decision, O_N^R. This choice minimizes the model and frees us of the need to make specific assumptions about the nature of the decision variables and errors in their representation. Ultimately, the model that is used in the simulations consists of four fundamental equations. The value of I_i^R is derived from I_i and F_i^I using (1):

$$I_i^R = I_i \cdot (1 - F_i^I) + F_i^I \cdot (1 - I_i) \qquad (1)$$

The ideal conjunction output—where inputs are error-free—is computed using (2):

$$O_{i+1} = O_i \cdot I_{i+1} \qquad (2)$$

Analogously, the observed decision is derived through (3):

$$O_{i+1}^R = O_i^R \cdot I_{i+1}^R \qquad (3)$$

Finally, for calculating the occurrence of a decision error F_i^O the simulations use equation (4):

$$O_i^R = (1 - F_i^O) \cdot O_i + (1 - O_i) \cdot F_i^O \qquad (4)$$

A detailed explanation of equations 1-4 is provided by [3].

The variables in $\{ I_i, I_i^R, F_i^I,\ O_i,\ O_i^R,\ F_i^O:\ i = 1, 2, .., N\}$ are viewed as random variables that accept the values zero and one. For each of the variables in $\{ I_i, F_i^I, O_i, F_i^O:\ i = 1, 2, .., N\}$, 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 2).

Monte Carlo Simulation

The validity of the insights that we have proposed in a previous section are examined empirically using the Monte Carlo simulation method. Monte Carlo simulation is a method for iteratively evaluating a deterministic model using sets of random numbers as inputs. The inputs are generated 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)

Instantiation of the Input Variables

The simulations center on conjunctive decision rules with up to ten decision variables. The values of $I_1, .., I_N$ $(2 \le N \le 10)$ are generated randomly according to distributions that are determined separately for each simulation. Mainly, p_i^I, the expected value of I_i, is chosen randomly in each simulation such that $0 < p_i^I < 1$ (note that $p_i^I = E(I_i) = \Pr(I_i = 1)$, i.e., p_i^I is equal to the probability that the value of the decision variable satisfies the criterion on that variable). Likewise, the values of F_i^I are determined individually in agreement with p_i^{FI}, the expected value of F_i^I $(p_i^{FI} = E(F_i^I) = \Pr(F_i^I = 1)$. In particular, p_i^{FI} is chosen randomly in each simulation such that two value ranges are explored. In one simulation set, which consists of 5,400 simulations, 0< p_i^{FI} <0.1, and in a second simulation set, which also consists of 5,400 simulations, 0< p_i^{FI} <0.2. Table 3 summarizes the simulation parameters.

[1] An error in the derived value of the dichotomous variable is caused by an error in the recorded value of the matching decision variable, although not every error in the recorded value of the decision variable results in an error in the respective dichotomous variable.

Sample Size

Each simulation produced $5 \cdot 10^8$ instances of each input variable. In addition, simulations with results that were inconsistent with our theory were repeated using a larger sample size of $5 \cdot 10^{11}$ instances.

N (Number of decision variables)	p_i^I (i=1,2,..,N)	p_i^{FI} (i=1,2,..,N)	Sample size (M)	Total # of simulations
9 simulation sets: N=2,3,4,5,6,7,8,9,10	random value in (0,1)	2 simulation sets: 0< p_i^{FI} <0.1 0< p_i^{FI} <0.2	$5 \cdot 10^8$	$9 \cdot 2 \cdot 600 = 10,800$
Simulations with findings that disagreed with the theory were repeated using a larger sample size.			$5 \cdot 10^{11}$	

Table 3. Implemented parameter values and number of simulations

Damage Estimates

The simulation model implements the conjunction of $I_1,..,I_N$ and, analogously, the conjunction of $I_1^R,..,I_N^R$, based on equations (1)-(4). For estimating the damage values, each simulation first computes a base decision error probability f_b^o using (5) below. In addition, each simulation computes a series of decision error probabilities, f_i^o (i =1,2,..,N), one for each input variable, in the following manner. For each input variable, in turn, the decision error probability is estimated again using (5), such that all the input samples are the same as the base samples, except for the sample that matches the chosen input. That sample is generated anew, such that the error probability there is 0.01 higher than the original error probability p_i^{FI} .

$$\overline{\Pr(F_N^O = 1)} = \frac{1}{M} \sum_{j=1}^{M} F_{N,j}^O \qquad (5)$$

The values of $F_{N,j}^O$ in (5) are derived by applying (1)-(4) on the artificially generated input values; the letter M denotes the input sample size.

The *damage* that errors in I_i^R inflict on the accuracy of O_N^R, denoted by Δ_i , is computed as:

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

We have also computed the change in the rate of false positive errors and the change in the rate of false negative errors due to a change in input accuracy by classifying each decision error and applying (5) and (6) on each of these error classes.

HYPOTHESES

Given the simulation settings described here, we pose the following three hypotheses.

HYPOTHESIS 1: If the damage that errors in I_1^R inflict on the accuracy of O_2^R is negative then the change in the probability of a false positive error in O_2^R is negative; however, the change in the probability of a false negative error in O_2^R is not negative.

Hypothesis 2 expresses an insight that was suggested in our section on "Negative Damage in Decision Rules that Combine Many Criteria." Namely, if a negative damage is exhibited in the output of a conjunctive decision rule and that rule is extended through conjunction to combine additional variables, then the sign of the damage is preserved through the entire multi-criteria rule.

HYPOTHESIS 2: Suppose that, when O_i^R is combined with I_{i+1}^R through conjunction, the damage that errors in the values of O_i^R inflict on the accuracy of O_{i+1}^R is negative. Assume that O_{i+1}^R is combined through conjunction with I_{i+2}^R to produce O_{i+2}^R. Then, the damage that errors in O_i^R inflict on the accuracy of O_{i+2}^R is negative.

Finally, we will test the understanding that was introduced in the section on "The Likelihood of a Negative Damage When Adding Criteria," according to which, the likelihood of encountering a negative association between input accuracy and output accuracy increases as we add criteria to the decision rule.

HYPOTHESIS 3: As the number of decision variables N grows higher**,** the probability that errors in the observed values of a given decision variable will produce a negative damage increases.

RESULTS

Hypotheses 1-3 are supported by the results of the simulations. The findings are detailed below.

Hypothesis 1

The findings regarding the conditions that produce a discrepancy with GIGO are consistent with Hypothesis 1. Hypothesis 1 has been directly validated through the computations of the change in the rate of false positive errors and the change in the rate of false negative errors. The rate of false positive errors decreased in all the instances in which the damage was negative, while the rate of false negative errors increased at the same time. The decrease in the first number was always greater than the increse in the second number.

Hypothesis 2

In a large majority of the cases, the simulations that utilized the smaller input sample size were consistent with Hypothesis 3. However, a small fraction of the input pairs that exhibited a discrepancy with GIGO did not reveal a similar discrepancy when they were combined with additional inputs. A similar inconsistency with Hypothesis 3 was registered in 8% of the input pairs, which had conjunctions that challenged the common GIGO assumption. These simulations were repeated using the larger input sample size. All the inconsistencies with Hypothesis 3 were resolved at that stage.

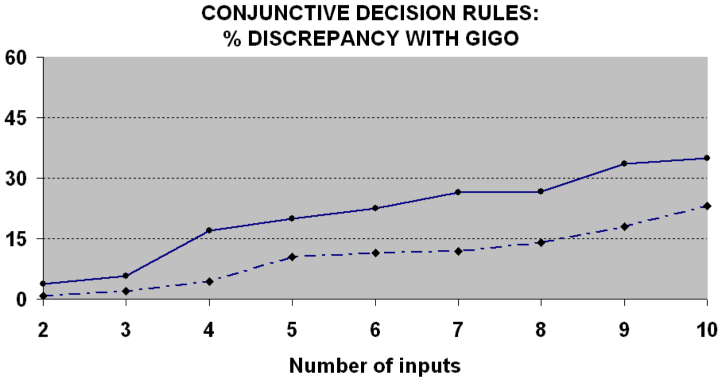

Figure 1. The average percentage of inputs that produce negative damage (dashed line represents input error rates up to 10%; continuous line represents input error rates up to 20%)

Hypothesis 3

Figure 1 portrays the average percentage of inputs that exhibit negative damage as a function of the number of inputs. The dashed line refers to the simulations that employed the lower input error rates (p_i^{F1} <0.1), while the continuous line corresponds to the findings of simulations that used the higher input error rates (p_i^{F1} <0.2). Clearly, the rate of inputs that exhibit negative damage rises as the number of inputs increases. At the lower input error rates, the average rate is lower than 1% in rules that combine two inputs (i.e., a probability of less than 0.01). However, this average grows dramatically to nearly 24% in rules that combine ten inputs. At the higher input error rates, the average percentage of inputs that exhibit negative damage is higher, varying from just below 4% to 35%. In conclusion, even at the lower input error rates that we studied, negative damage turned out to be a widespread phenomenon when the number of decision variables was high enough.

DISCUSSION

An effort to improve input data accuracy that yields poorer information accuracy when the data are processed would normally be labeled a major failure. While the popular belief discounts the likelihood of this scenario, this research suggests, in contrast, that a negative association between input accuracy and decision accuracy is a deeply rooted phenomenon in such rules. It is not limited to a specific type of decision variable, it is not exceedingly sensitive to variations in statistical dependencies, and it does not tend to disappear as the number of decision variables grows higher. Askira Gelman (2011) offers the following guideline for data quality resource allocation and design decisions: When the percentage of the values that satisfy the decision criterion varies significantly across decision variables, efforts to improve the accuracy of the output of a conjunctive decision should assign lower priority to errors in variables where a high percentage of the values meet the criterion. The findings of (Askira Gelman, 2011) imply an equivalent rule for disjunctive decision rules: When the percentage of the values that satisfy the decision criterion varies dramatically across decision variables, efforts to improve the accuracy of the output of a disjunctive decision should assign low priority to errors in variables where a *low* percentage of the correct values meet the criterion. The results of the empirical investigation presented here indicate, furthermore, that the likelihood of a negative damage may increase considerably as the number of decision variables goes up. In an environment where the probability of satisfying the decision criterion varies extensively, the likelihood that errors in a chosen input will produce a negative damage is expected to grow higher as the number of variables increases, such that a negative damage may actually be common. Therefore, resource allocation for data quality purposes should be made with great care when decision rules employ many variables and the probability of satisfying the decision criterion is known to vary significantly. Alternatively, a potentially profitable, counter-intuitive use of our theory will be investigated through future work, namely, future work will examine the value of error injection for improving decision accuracy.

The observed improvement in decision accuracy when input error rate goes up has been shown to be driven by a lower rate of false positive decisions. In real-world settings the implications of a false positive error can differ greatly from the implications of a false negative error. In some domains, such as in medical diagnosis, the risk of a false negative is typically far greater than the risk of a false positive. A doctor may feel that missing an opportunity for early diagnosis could mean the difference between life and death, while a false positive, on the other hand, might result only in a routine biopsy operation. In other domains, such as in criminal law, the risk of a false positive is considered more consequential. In our real estate scenario, a wasted visit to inspect an unsuitable property may be judged to be significantly less costly than missing one’s ideal home when it is offered for sale at a bargain price. Accordingly, decision makers that adopt this standpoint will consider a false positive decision much less costly than a false negative decision. The magnitude of the economic outcome of accuracy improvement will vary depending on the relative weights assigned to the former two error types. However, the guidelines that (Askira Gelman, 2011) offers should normally apply, even if the importance assigned to a false negative decision is substantially higher than the importance assigned to a false positive decision. This is true because an input where errors generate a negative damage shows only small improvement in the rate of false negative decisions relative to other inputs.

A practical estimate of the damage must be based on an understanding of the input set whose accuracy is projected to change. While our running example conveniently assumes that input accuracy improvements cover an entire database, this assumption may be false in practice when the change in accuracy targets a well-known, pre-defined subset of the database. Clearly, the choice of input set determines the values of relevant parameters. In the remainder of this section we go over the conditions that have been associated with a negative damage and the assumptions of this research in general. We examine their similarity to real-world scenarios, as well as the validity of this theory under conditions that are outside the scope of the formal study.

Wide Disparity in the Probabilities of Satisfying the Decision Criteria

We begin with the condition of a wide disparity in the probabilities of satisfying the decision criteria. We believe that decision scenarios that demonstrate great inequality in these probabilities are common. Take, for instance, the real estate property investment decision. Real estate investment decisions are often aided by online databases, which can cover sizeable geographic areas. Therefore, a decision criterion that targets a specific zip code may actually match a minute fraction of the properties in the database. Likewise, depending on the geographic area, a price range of $0-$50,000 may match only a small percentage of the items. In contrast, the requirement of two bedrooms or more might be satisfied by a large majority of the properties. Therefore, when using this decision rule one may see a gap that is wide enough to produce a negative damage. Any one of the following two input pairs can demonstrate such a gap: the pair consisting of number of bedrooms and property price, and the pair consisting of number of bedrooms and zip code. Any one of the respective conjunctions can be characterized by a negative damage to decision accuracy due to errors in the data on the number of bedrooms. Furthermore, Hypothesis 3 suggests that a negative damage can surface at a later stage, when combining the conjunction of price and zip code (i.e., the two criteria that are satisfied by low percentages of the population) with the criterion for the number of bedrooms (which is satisfied by a high percentage of the population). Ultimately, this understanding implies that an investor who employs the former decision rule may be better off if s/he assigns a lower priority to the accuracy of the data on the number of bedrooms than the accuracy of the data about the price or zip code. Notably, however, if the accuracy of these two sources is brought to a high enough level, then a negative damage by errors in the data on the number of bedrooms should turn positive.

For the sake of completeness we need to qualify the requirement of a wide disparity in the probabilities of satisfying the decision criteria. We found that, if error rates are extremely high then a negative damage is possible even when this condition is not satisfied. This is largely due to the fact that, under these circumstances, new errors tend to have no effect at all.

Input Errors

We now turn to a discussion of a second factor that has been identified by this theory, namely, the rates at which tests of individual decision variables show incorrect conclusions. Previous results (Askira Gelman, 2011) indicate that these rates form an important determinant of the sign of the damage. A given gap between the probability of satisfying the criterion on one variable and the probability of satisfying the criterion on another variable can be wide enough, or not wide enough, to produce negative damage, depending on the error rate in the input with the lower probability.

This work examined conditions in which the upper boundary on the rate of incorrect judgments of the decision criterion is either 10% or 20%. Obviously, this boundary typically corresponds to a significantly higher data error rate, since not every error in the data causes an incorrect judgment of the respective criterion. Therefore, from a quantitative perspective, the parameter choice of this study, as well as its findings, suit conditions in which data accuracy is poor. However, regardless of data accuracy, several common data quality deficiencies, such as missing values, out-of-date data, or data that are expressed using non-standard units, can produce the same phenomenon as "ordinary" errors, i.e., a higher deficiency rate may actually improve the accuracy of a decision. Therefore, practical analyses of the sign of the damage should take into account other deficiencies as well. We discuss this issue in the final subsection.

Statistical Dependencies

Although the sign of the damage is not extremely sensitive to variations in statistical dependencies, statistical dependencies can affect the sign significantly.

Statistical dependencies are widespread in practice. For example, we often face situations in which decision variables V_1 and V_2 are not statistically independent such that I_1 and I_2 are either positively or negatively correlated. One example from the real estate domain is the dependence that is exhibited between the zip code of a property and its size, such that the probability of finding a large house in an affluent zip code is significantly higher than the overall proportion of large properties in the database (positive correlation between I_1 and I_2). Alternatively, the probability of finding a large house in a poverty stricken zip code can be significantly lower than the overall proportion of large properties in the database (negative correlation between I_1 and I_2). Likewise, we often encounter real-world settings where the probability of a false positive judgment of a value is different from the probability of a false negative judgment of the value (I_1 and F_1^I are not statistically independent). In the real estate instance, a disproportionately high percentage of the properties may be incorrectly registered under a zip code that is preferred by the local population (a trendy zip code; implies negative correlation between I_i and F_i^I), while a relatively high number of properties may be incorrectly registered *outside* other zip codes (undesirable zip codes; implies positive correlation between I_i and F_i^I). Stated another way, errors may be unevenly distributed such that some zip codes would demonstrate a relatively high rate of false positives, while other zip codes would demonstrate a relatively high rate of false negatives.

Since statistical dependencies are common in real-world settings, a deeper study of the effect of statistical dependencies on the conditions of negative damage would be useful. Nonetheless, a preliminary, qualitative understanding of this aspect can be derived based on the intuitive insight provided by this article.

	A growing correlation between I_i and F_i^I	A growing correlation between I_1 and I_2
negative damage	↑	↓

Table 4. The effect of non-zero correlation

To begin with, consider a statistical dependence between V_1 and V_2 such that the respective dichotomous criterion variables I_1 and I_2 are positively correlated. The probability of having both criteria met (or both criteria not met) is, in this case, higher than the comparable probability under statistical independence conditions. Therefore, a stream of values of these two inputs will contain a lower proportion of "true"/"false" pairs relative to the proportion of this pair under statistical independence. As highlighted earlier, this pair is, however, essential to the creation of negative damage. Subsequently, all other things being equal, a positive correlation between I_1 and I_2 *contracts* the range of the parameter values (p_1^I, p_2^I, p_2^{FI}) in which a negative damage by errors in I_1^R is observed. A similar explanation of negative correlation between I_1 and I_2 would conclude that it *expands* the range of the parameter values (p_1^I, p_2^I, p_2^{FI}) in which a negative damage by errors in I_1^R is observed.

A comparable analysis of the effect of statistical dependence between I_i and F_i^I suggests that, all other things being equal, if the correlation between I_i and F_i^I *goes up* then negative damage by errors in I_1^R is observed in a *wider range* of the parameter values (p_1^I, p_2^I, p_2^{FI}), relative to their range under conditions of independence. In this sense, a relatively high rate of false negative judgments of the individual criteria increases the likelihood of negative damage. In the case of statistical dependence between I_1 and F_1^I in particular, it is easy to see that a higher correlation increases the proportion of "good" errors.

Missing values and Other Data Quality Deficiencies

Suppose that the (correct) value of a decision variable meets the criterion on that variable. If that value is missing from the data then a decision rule that employs the criterion may subsequently show a false negative decision on that data instance (a false positive decision is possible only if a default value is used in place of the missing value). In other words, a missing value can cause a false negative error, but it cannot cause a false positive error. Hence, missing values change the balance between the rate of false positive errors and the rate of false negative errors, such that the rate of false negative errors increases. Accordingly, our analysis in the previous section indicates that missing values raise the likelihood of a negative damage.

An out-of-date value or a value that is expressed using a non-standard unit can cause either a false positive error or a false negative error. Often, however, changes in data values over time are characterized by a general trend (up or down), such as when prices trend up, or down, over time. Therefore, out-of-date values can have a similar, or an opposite, effect to that of missing values. Again, the discussion in the previous section can be used for clarifying such an effect on the sign of the damage.

Likewise, values expressed using non-standard units may have a similar, or an opposite, effect to that of missing values. A non-standard unit either inflates data values, or deflates them. Taken as a whole, the direction, if any, and magnitude of a trend depend on the choice of unit, or collection of units, that replace the standard unit.

In conclusion, several data quality deficiencies, including "simple" data errors, missing values, out-of-date data, data that are expressed using non-standard units, and possibly other deficiencies, can contribute to the phenomenon that we have examined in this work. Mainly, a higher deficiency rate can actually improve decision accuracy. Therefore, practical analyses that aim to account for that phenomenon should apply a broad assessment of the deficiencies in the data.

ACKNOWLEDGMENT

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

1. Askira Gelman, I. (2004) Simulations of the relationship between an information system's input accuracy and its output accuracy, 9th *International Conference on Information Quality (ICIQ-04)*, MIT, Cambridge MA.
2. Askira Gelman, I. (2010) Setting priorities for data accuracy improvements in satisficing decision-making scenarios: a guiding theory, *Decision Support Systems (DSS)*, 48, 4, 507-520.
3. Askira Gelman, I. (2011) GIGO or not GIGO: The accuracy of multi-criteria satisficing decisions, *ACM Journal of Data and Information Quality (JDIQ)*, 2, 2.
4. Ali, K.M., and Pazzani, M. J. (1996) Error reduction through learning multiple descriptions, *Machine Learning*, 24, 3,173-202.
5. Ballou, D. P., and Tayi, G.K. (1989) Methodology for allocating resources for data quality enhancement, *Communications of the ACM* 32, 3.
6. 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* 18, 5.
7. Barabash, T. L. (1965) On properties of symbol recognition, *Engineering Cybernetics*, 71-77.
8. Berg, S. (1993) Condorcet's jury theorem revisited, *European Journal of Political Economy*, 9, 3, 437-446.
9. Bishop, C.M. (1995) Training with noise is equivalent to Tikhonov regularization, *Neural Computation*, 7, 1, 108-116.
10. Clemen, R.T., and Winkler, R.L. (1985) Limits for the precision and value of information from dependent sources, *Operations Res.*, 33, 2, 427-442.
11. Cover, T. (1974) The best two independent measurements are not the two best, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-4(1), 116-117.
12. Einhorn, H.J. (1970) The use of nonlinear, noncompensatory models in decision making, *Psychological Bulletin*, 73, 3, 221-230.
13. Einhorn, H.J. (1971) The use of nonlinear, noncompensatory models as a function of. task and amount of information, *Organizational Behavior and Human Performance*, 6, 1.
14. Einhorn, H.J. (1972) Expert measurement and mechanical combination, *Organizational Behavior and Human Performance*, 7, 86-106.
15. Elashoff, J.D., Elashoff, R.M., and Goldman, G.E. (1967) On the choice of variables in classification problems with dichotomous variables, *Biometrika*, 54, 668-670.
16. Even, A., Shankaranarayan, G. and Berger, P.D. (2007) Economics driven data management: An application to the design of tabular data sets, *IEEE Transactions on Knowledge and Data Engineering* 19, 6.
17. Fang, G.S. (1979) A note on optimal selection of independent observables, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-9(5), 309-311.
18. Fishman, G.S. (1995) *Monte Carlo: Concepts, algorithms, and applications*, Springer Verlag, New York.

19. Frantsuz, A.G. (1967) Influence of correlations between attributes on their informativeness for pattern recognition, *Engineering Cybernetics*, 4.
20. Hipp, J., Guntzer, U., and Grimmer, G. (2001) 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.
21. Janson, M. (1988) Data quality: The Achilles heel of end-user computing, *Omega: International Journal of Management Science*, 16, 5.
22. Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W. (2003) Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis and Applications*, 6,1, 2-31.
23. Ladha, K. (1995) Information pooling through majority-rule voting: Condorcet's Jury Theorem with correlated votes, *J. Econ. Behavior and Organization*, 26, 353-372.
24. Lee, Y., D. Strong, B. Kahn, and R. Wang, (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*, 6, 2, 154-165.
26. Miller, G. A. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63, 2, 81-97.
27. Mintz, A. (2004) How do leaders make decisions? A poliheuristic perspective, *Journal of Conflict Resolution*, 48, 1, 3-13.
28. Motro, A., and Rakov, I. (1997) 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, 1-21.
29. Nitzan, S., and Paroush, J. (1984) The significance of independent decisions in uncertain dichotomous choice situations, *Theory and Decision*, 17, 47-60.
30. Park, C. W. (1976) Prior familiarity and product complexity as determinants of the consumer's selection of judgmental models, *Journal of Marketing Research*, 144-151.
31. Parsaye K. and Chignell, M. (1993) Data quality control with SMART databases, *AI Expert*, 8, 5.
32. Parssian. A. (2006) Managerial Decision Support with Knowledge of Accuracy and Completeness of the Relational Aggregate Functions, *Decision Support Systems*, 42, 1494-1502.
33. Payne, J.W. (1976) Task complexity and contingent processing in decision making: An information search and protocol analysis, *Organizational Behavior and Human Performance*, 16, 366-387.
34. Payne, J.W., Bettman, J.R., and Johnson, E.J. (1993). *The Adaptive Decision Maker*, Cambridge University Press, New York.
35. Phipps, A.G. (1983) Utility function switching during residential search, *Geografiska Annaler. Series B, Human Geography*, 65(1), 23-38.
36. Pipino, L., Lee, Y.W. and Wang, R.Y. (2002) Data quality assessment, *Communications of the ACM* , 45, 4.
37. Raviv, Y., and Intrator, N. (1996). Bootstrapping with noise: An effective regularization technique, *Connection Science*, Special issue on Combining Estimators, 8, 356-372.
38. Simon, H., A. (1955). A behavioral model of rational choice, *Quarterly Journal of Economics*, 69(1), 99-118.
39. Wang, R.Y., Reddy, M.P., and Kon, H.B. (1995). Toward quality data: An attribute-based approach, *Decision Support Systems (DSS)*, 13, 3-4, 349-372.