Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2008 Proceedings

European Conference on Information Systems (ECIS)

2008

Comparative Analysis of Data Quality and Utility Inequality Assessments

Adir Even
Ben-Gurion University of the Negev, adireven@bgu.ac.il

G. Shankaranarayanan
Boston University, gshankar@bu.edu

COMPARATIVE ANALYSIS OF DATA QUALITY AND UTILITY INEQUALITY ASSESSMENTS

Even, Adir, Ben Gurion University of the Negev, Department of Industrial Engineering and Management, P.O. Box 653, Beer-Sheva, 84105, Israel, adireven@bgu.ac.il

Shankaranarayanan, G., Boston University School of Management, Information Systems Department, 595 Commonwealth Ave., Boston, MA, 02215, USA, gshankar@bu.edu

Abstract

Given that the volumes of organizational data resources are rapidly increasing, achieving and sustaining high data quality are becoming much more challenging tasks. In the face of this growing challenge, this study posits the need to introduce a robust economic thinking into the process of improving and maintaining data quality. Economic thinking requires investigating and assessing the business-value contribution of data resources, conceptualized as utility. We show that quantitative assessments of inequality in the utility of data resources, together with assessments that reflect the presence and the impact of defects, can provide key insights into the current state of data quality. A comparative analysis of such assessments can also direct the development of data quality maintenance policies and help prioritize quality improvement efforts. In this study, we demonstrate the application of such a comparative analysis in a real-life CRM context, using samples from a large data repository used for managing alumni relations. We show that the results of such a comparative analysis have important managerial implications for data quality management within the evaluated environment. We also discuss its applicability in other data management contexts.

Keywords: Data Quality Management, Database, Information Value, Inequality, Gini Index

1 INTRODUCTION

Achieving and sustaining high data quality has long been recognized as a key challenge in managing organizational data resources. Low data quality damages the efficiency and the effectiveness of business operations, hinders decision making, breeds mistrust and frustration and, consequently, damages revenues and profitability. Managing data quality has become increasingly difficult in recent years, as information systems and technologies such as data warehouses, ERP/CRM systems, RFID, and Click-streams have mandated advanced data processing and analysis, driving the need to manage large data repositories. The increasing costs associated with managing large and complex data resources necessitate introducing a more robust economic thinking into the continuous process of improving and maintaining data quality. This requires a better understanding of the business benefits gained by using data resources, conceptualized as utility in this study, as well as the costs of maintaining the quality of these resources at a high level. Improving quality may increase utility but the associated costs might offset the added benefits. We suggest that understanding this link between quality, utility, and cost - and the possible tradeoffs between them - is critical for managing the quality of large data resources and can direct quality management policies and prioritize improvement efforts.

To that end, our objective is to better understand the utility of data resources and its implications for data quality management (DQM). We specifically examine the magnitude of utility inequality – the extent to which records in a dataset differ in their utility contribution. We introduce analytical tools for modeling inequality and measuring it in large datasets. We show that quantitative assessments of inequality, together with assessments that reflect the presence and the impact of defects, can provide key insights into the current state of data quality, direct the development of data quality maintenance policies, and help prioritize improvement efforts. We demonstrate this comparative assessment in the context of Customer Relationship Management (CRM) and show that, in this context, such analysis offer superior benefits compared with those offered by traditional data quality assessment methods.

In the remainder of this paper, we first discuss economic aspects of managing the quality of large data resources and review research on quality assessment that influences our study. We then propose a methodology for assessing quality and utility inequality. We demonstrate its application using large samples of alumni data and use the results to formulate recommendations for improving the quality of this data resource. We finally discuss managerial implications and offer our concluding thoughts.

2 BACKGROUND

Economic aspects, such as the benefits and the costs associated with improving data quality, are critical for successful DQM and yet, have not been sufficiently researched (Even and Shankaranarayanan, 2007; Heinrich et al., 2007). This gap needs addressing, as data resources are critical to organizations and the costs of managing them are rapidly increasing. Data quality has been defined at a high level as “fitness for use”, reflecting the ability to satisfy customer needs (Wang, 1998). Poor data quality lowers satisfaction, increases costs, and breeds mistrust towards IS. Conversely, high data quality improves decision making, empowers organizational strategy and, hence, can sustain competitive advantage (Redman, 1996). DQM literature has proposed a plethora of algorithmic methods and techniques for data inspection and correction (e.g., Redman, 1996). Taking a broader perspective, the Total Data Quality Management (TDQM) paradigm (Wang, 1998) views data environments as manufacturing processes and their outputs as information products. It perceives DQM as an ongoing cycle of defining, measuring, analyzing and improving data manufacturing processes for continuously improving the quality of information products. This study adopts the TDQM view and suggests that managing the quality of data processes, resources and products can benefit from the adoption of an economic perspective. This perspective has to acknowledge possible utility/cost effects and tradeoffs in evaluating DQM procedures, policies, and technologies.

We particularly focus on quantitative assessments of data quality, which is critical for understanding the current quality state of a data resource, setting improvement targets, and tracking the progress toward these targets. Data quality can be measured along different dimensions (e.g., completeness, accuracy, and currency), each reflecting a specific type of quality defect (missing content, incorrect values, and outdated data items, respectively). Targeting high quality along multiple dimensions may introduce economic tradeoffs (Ballou and Pazer, 1995), and recent studies have suggested that methods for assessing data quality should consider economic impacts and tradeoffs (Even and Shankaranarayanan, 2007; Heinrich et al., 2007). DQM literature differentiates between *impartial* and *contextual* assessments (Pipino et al., 2002) - the former reflects the *presence* of quality defects (e.g., missing or incorrect values), while the latter their *impact*, which may vary with usage contexts. In this study, we apply a quantitative framework that permits both contextual and impartial assessments (Even and Shankaranarayanan; 2007). In addition, we evaluate the magnitude of utility inequality – the extent to which records in a dataset differ in utility contribution (Even et al., 2007). Previous studies have evaluated these two forms of assessments independently. Here, we consolidate them into a single framework showing that a comparative evaluation can offer key insights for DQM decisions.

We demonstrate a comparative analysis of quality and inequality assessments in the context of Customer Relationship Management (CRM). Data quality is a critical issue in CRM environments, and may damage the associated economic benefits (Heinrich et al., 2007). In this study, we address a specific type of quality defect, *missing content*, and the associated *completeness* dimension. Missing content is common defect in CRM environments - certain attributes (e.g., income and credit score) may not be available when initiating a customer record. The firm may choose to leave these unfilled and update them later. Existing CRM data is also enhanced with new attributes (e.g., new contact and demographic data), and the corresponding values may initially be null. They may remain null for certain customers, as the firm may choose to not update them due to high data acquisition costs. Our comparative analysis can help assess the extent and the impact of missing values in data attributes and, accordingly, prioritize improvement initiatives associated with addressing these missing values.

3 UTILITY-DRIVEN ASSESSMENT

The methodology that we present is based on assessing *utility* – a numeric measure that reflects the importance and value contribution of information resources. Utility may reflect enhancements to business performance, improvements to the decisions made, or the data consumer's willingness to pay (Ahituv, 1980). In many data-usage contexts, utility allocation can reflect monetary assessment (e.g., revenue potential). However, the tools that we describe do not depend on the utility units. A plethora of attribution methods, which reflect relative importance and value, have been discussed in the literature and may be adapted for the purpose of assessing and allocating utility - e.g., Customer Lifetime Value (Berger and Nasr, 1998) and Recency/Frequency/Monetary (R.F.M.) (Petrison et al., 1997) analysis in database marketing, and ABC classification in inventory management.

In this section, we briefly review a method for utility-driven assessment of data quality (Even and Shankaranarayanan, 2007) and tools for quantifying utility inequality (Even et al., 2007). We then illustrate the use of these tools, and discuss insights that can be gained from a comparative evaluation for prioritizing data improvement efforts. We assess utility in a tabular dataset in which records have an identical attribute structure. CRM and other data environments manage large tabular datasets (e.g., customer lists, sale transactions, stock price quotes, RFID readings) and these are vulnerable to quality defects. While records in tabular dataset are similar in structure, their content variability differentiates their relative importance to data consumers and, hence, their associated utility. For brevity, we describe a single utility attribution that reflects one usage or an aggregation of usages. As shown in our previous studies, these formulations can be extended to address multiple usages.

3.1 Utility-Driven Assessment of Data Quality

The framework proposed in (Even and Shankaranarayanan, 2007) permits both impartial and contextual assessment of quality along different dimensions. We consider a dataset with N records (indexed by *[n]*) and M attributes (indexed by *[n]*), and denote the content of attribute *[m]* in record *[n]* as $f_{n,m}$. We assign each record a non-negative utility measure, $u_n \geq 0$, assuming additivity: $u^D = \Sigma_n u_n$, where u^D is the dataset utility. Utility is at its maximum when the entire dataset is available and may reduce to some extent if some dataset content is defective. The attribute quality measure $q_{n,m}$ reflects the extent to which attribute *[m]* in record *[n]* is defective (between *0,* if very defective, and *1* if perfect). Similarly the record quality measure Q_n reflects the extent to which the entire record is defective (between *0* and *1*). The record quality measure Q_n is derived from aggregation of the associated attribute quality measures: $Q_n = f(\{q_{n,m}\}_{m=1..M})$. Even and Shankaranarayanan (2007) describe different possible forms of aggregation, and in this study we consider *(a) absolute* – indicating *any* missing content in a record (*0* if at least one attribute is missing, *1* otherwise), and *(b) proportion* - the *ratio* of non-missing content (between *0 – all missing,* and *1- none missing*).

Using the set of record-utility attributions $\{u_n\}$ as weights, we can formulate a dataset quality measure with respect to attribute *[m]:*

$$(1) \qquad Q_m^D = \left(\sum_{n=1..N} u_n q_{n,m}\right) \Big/ \left(\sum_{n=1..N} u_n\right) = \left(1/u^D\right)\left(\sum_{n=1..N} u_n q_{n,m}\right)$$

As shown in (Even and Shankaranarayanan, 2007), this formulation reflects loss in utility due to the presence of quality defects; hence, since utility is context-dependent, this measure can be interpreted as contextual quality assessment for attribute *[m]* at the dataset level. Similarly, we can formulate a dataset quality measure that reflects quality assessment of the entire record:

$$(2) \qquad Q^D = \left(\sum_{n=1..N} u_n Q_n\right) \Big/ \left(\sum_{n=1..N} u_n\right) = \left(1/u^D\right)\left(\sum_{n=1..N} u_n Q_n\right)$$

When utility is allocated independent of attribute content (i.e., constant $u_n = u^D/N$), it can be shown the result is an impartial measure that reflects a ratio between counts of perfect items and total items, which is consistent with the impartial definition of quality (e.g., (Redman, 1996; Pipino et al., 2002)):

$$(3) \qquad Q_m^D = \left(1/N\right)\sum_{n=1..N} q_{n,m} \quad \text{, and}$$

$$(4) \qquad Q^D = \left(1/N\right)\sum_{n=1..N} Q_n$$

These definitions permit assessment along different quality dimensions, depending on the type of defect reflected by the quality assessments $\{q_{n,m}\}$, in this study - missing attribute values.

3.2 Inequality in the Utility of Data

Given the variability in the utility of dataset records, will the overall utility depend on the entire dataset, or only on a small subset of records? The magnitude of inequality in the utility of data records may have implications for managing data resources (Even et al., 2007). *Lorentz curve* (1905) and the associated *Gini index* (1912), statistical tools for modeling and measuring inequality in value distributions, can be adapted to assess utility inequality in datasets.

For a large dataset (large N), we represent the utility of records as a random variable u with a known probability density function (PDF) $f(u)$. From the PDF we obtain the mean $\mu = E[u]$, the cumulative distribution function (CDF) $F(u)$, and the percent point function (PPF, the inverse of CDF) $G(p)$. Here we demonstrate computations for a discrete distribution (Figure 1), later used for our evaluation. Similar computations can be applied to continuous distributions (e.g., Pareto, Normal, or Weibull). Inequality can also be assessed for a random data sample, without assuming a specific distribution.

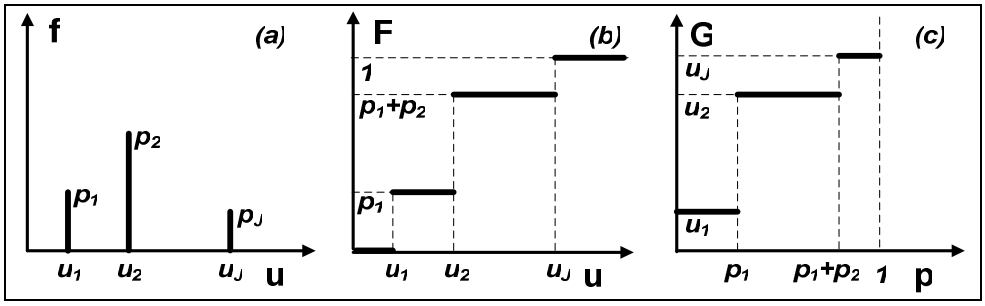

Figure 1. A Discrete Distribution: (a) PDF, (b) CDF and (c) PPF.

A variable with a discrete distribution (Figure 1) has a finite set of J possible values $u_1 \ldots u_J$ (the index $[j]$ reflects sorting in an *increasing* order), with probabilities of $p_1 \ldots p_J$, respectively $(\Sigma_j p_j = 1)$:

$$f(u) = \begin{cases} p_j & u = u_{j=1..J} \\ 0 & otherwise \end{cases}, \qquad F(u) = \begin{cases} 0 & u < u_1 \\ \sum_{k=1}^{j} p_k & u \in [u_j, u_{j+1})_{j=1..J-1} \\ 1 & u \geq u_j \end{cases},$$

$$G(p) = u_j \quad p \in \left(\sum_{k=1}^{j-1} p_k, \sum_{k=1}^{j} p_k\right]_{j=1..J}, \qquad \mu = \sum_{j=1}^{J} p_j u_j \tag{5}$$

To assess the extent to which records vary in their utility, we define R, the proportion of highest-utility records, as a $[0,1]$ ratio between the N^* records of highest utility (i.e., the top N^* when rank-ordered in descending order) and N, the total number of records (e.g., $R=0.2$ for a dataset with $N=1{,}000{,}000$ records and $N^*=200{,}000$ records that offer the highest utility).

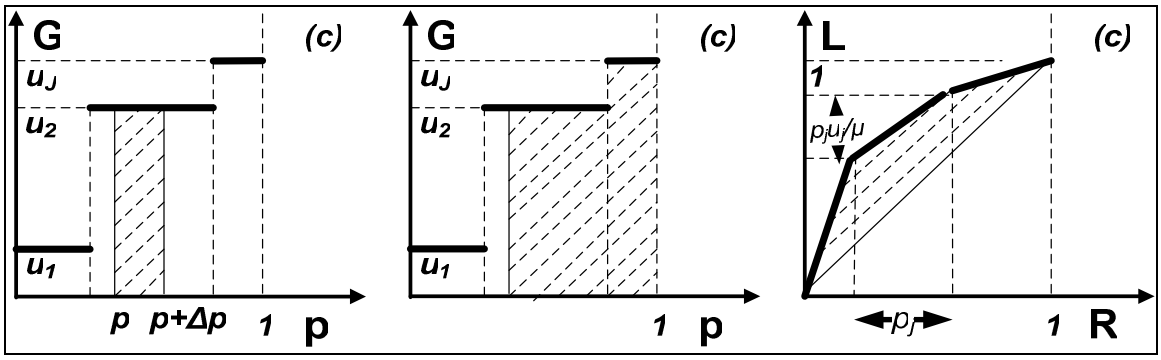

Figure 2. Calculating the Cumulative Utility Curve and the Inequality Index.

The cumulative utility curve $L(R)$ (equivalent of Lorentz curve) is defined as a $[0,1]$ proportion of the overall utility as a function of R. $L(R)$ can be calculated from the percent point function $G(p)$. For a large N, the added utility for a small probability interval $[p, p+\Delta p]$ can be approximated by $NG(p)\Delta p$ (Figure 2a). By taking $\Delta p \rightarrow 0$, integrating the PPF over $[1-R, 1]$ (Figure 2b), and dividing the result by the total utility (approximated by $N\mu$), the cumulative utility curve $L(R)$ (Figure 2c) is:

$$L(R) = N\int_{1-R}^{1} G(p)dp \Big/ N\mu = \int_{1-R}^{1} G(p)dp \Big/ \mu \text{, where,} \tag{6}$$

R – The $[0,1]$ proportion of highest-utility records
$L(R)$ - The cumulative utility curve of the utility variable u, within $[0,1]$
N - The number of dataset records
u, μ - The utility variable and its mean
$G(p)$ - The proportion point function of the utility variable u

Gini index *(φ)*, which is derived from Lorentz curve, is a commonly used measure of inequality. It measures the relative area between *L(R)* and *f(R)=R* (highlighted in Figure 2c):

$$(7)\quad \varphi = \left(\int_0^1 L(R)dR - \int_0^1 RdR \right) \Big/ \int_0^1 RdR = (2/\mu)\int_0^1 pG(p)dp - 1$$

The value of φ is within *[0,1]*, where a higher value indicates a greater inequality. The lower bound, $\varphi \rightarrow 0$, indicates perfect equality – dataset records with identical and deterministic record utilities and a curve that approaches *L(R)=R*. The upper bound, $\varphi \rightarrow 1$, indicates a high degree of inequality - a small portion of records with a relatively high utility, while the utility of most other records is substantially lower. The corresponding curve approaches *L(R)=1*.

The cumulative utility for a discrete distribution is a piecewise-linear curve (Figure 2c). Each segment is associated with a single value in the set of *J* possible values. The curve is obtained by backwards integration of the PPF; hence, the segments are sorted in a *decreasing order of utility values* (i.e., in a reverse order of the index *[j]*). The length of the horizontal axis of each segment is the relative proportion of the dataset, or the probability p_j associated with the utility value u_j. The length of the vertical axis of each segment is the relative utility contribution of value *[j]*: $(p_j*u_j)/(\Sigma_j p_j*u_j)=(p_j*u_j)/\mu$. It can be shown that the Gini index for a discrete distribution can be calculated as:

$$(8)\qquad \varphi = 1 + (1/\mu)\sum_{j=1..J} p_j^2 u_j - (2/\mu)\sum_{j=1..J} p_{J-j+1} u_{J-j+1} \sum_{w=1..J-j} p_w$$

In the case of two categorical values ("High" vs. "Low"), this can be simplified to $\varphi=p_2(u_2/\mu-1)$, where u_2 is the higher utility among the two, and p_2 is the associated probability.

3.3 Comparing Quality Assessments and Utility-Inequality

Comparative analysis of *impartial quality*, *utility-driven quality*, and *magnitude of inequality* measures can help better understand the quality of the data and prioritize improvements. To demonstrate it, we use an illustrative customer dataset (Table 1). Each record (identified by a unique *Customer ID*) has four attributes - *Gender*, *Marital Status*, *Occupation*, and *Income* – each associated with a finite set of possible values. Here, we assume two possible values per attribute (e.g., I1 - "high" vs. I2 - "low" *Income); however,* similar calculations can address a larger set of values. Some attribute values are missing (marked as *<null> in Table 1*). The *absolute* measure indicates missing content in a record (*0* if at least one attribute is missing, *1* otherwise), and the *proportion* measure reflects the ratio of non-missing content (between *0 – all missing,* and *1- none missing*). The utility measure that we use in this example reflects relative purchase power per customer (e.g., based on past sale transactions).

ID	Gender	Marital	Occupation	Income	Absolute	Proportion	Utility
1	G1	M1	O1	I1	1	1.00	10
2	G2	M1	O2	<null>	0	0.75	280
3	G2	M1	<null>	<null>	0	0.50	0
4	G2	M2	O1	I2	1	1.00	110
5	G1	M2	O2	<null>	0	0.75	150
6	G1	M2	O2	<null>	0	0.75	180
7	G1	M1	<null>	I1	0	0.75	10
8	G1	<null>	O2	<null>	0	0.50	230
9	G1	M2	<null>	I1	0	0.75	30
10	<null>	<null>	<null>	<null>	0	0.00	0
Dataset:							***1000***
Imp. Comp.	***0.900***	***0.800***	***0.600***	***0.400***	***0.20***	***0.675***	
U.D. Comp.	***1.000***	***0.770***	***0.960***	***0.160***	***0.12***	***0.723***	
Inequality	***0.057***	***0.110***	***0.208***	***0.438***			***0.548***

Table 1. An Illustrative Sample of the Customer Dataset

Impartial data quality assessment [Eq. 3, 4] reflects the *presence* of defects and can help assess the *efforts* and the *costs* for improving quality. In our dataset, *8* out of *10* records have missing values; hence, the impartial completeness corresponding to the *absolute* measure is *(10-8)/10=0.2*. Considering *proportion*, *12* out of *40* values are missing; hence, the corresponding impartial score is *(40-12)/40=0.675*. Further insights can be gained by assessing impartial quality per attribute - *Gender* has 1 missing value, *Marital Status* - 2, *Occupation* - 4, and *Income* – 6 (impartial scores - *0.9*, *0.8*, *0.6*, and *0.4*, respectively). This implies that correcting missing content in certain attributes (e.g., *Gender*) is likely to be significantly easier and cheaper than in others (e.g., *Income*).

Utility-driven data quality assessments [Eq. 1, 2] reflect the *impact* of defects and can help assess the possible *benefits* from improving quality. Some insights can be gained by comparing utility-driven to impartial scores. *(a) A significantly higher utility-driven score* indicates that high-utility records are less defective. Two complementary explanations are possible. First, defective records are less usable to begin with and hence, their utility is inherently lower. Second, efforts may have been made to maintain high-utility records at a high quality level by eliminating defects. *(b)Insignificant difference between scores* indicates that utility is evenly distributed across all records and/or a weak association between defect rates and utility. *(c) A significantly lower utility-driven score:* This is an abnormality indicating a high utility inequality and some substantial damage to high-utility records, possibly due to a systematic cause of defects. In the example above, considering the *absolute* measure, the utility-driven completeness is *0.12* (*880* out of *1000* utility units are damaged), slightly lower than its corresponding impartial score *(0.2)*. Considering *proportion*, the utility-driven completeness is *0.723*, slightly higher than the corresponding impartial score. In both cases the margin between utility-driven and impartial scores is minor and, as shown later, this cannot be explained by equal distribution of utility among records. We may however have a weak association between the presence of defects and utility, when measured at the *record level*. Measuring at the *attribute level* clarifies this further. For *Gender* and *Marital Status*, the utility-driven scores are not substantially different than the impartial (*1.0* vs. *0.9*, and *0.77* vs. *0.8*, respectively). Conversely, the utility-driven score for *Occupation (0.96)*, is significantly higher than the impartial *(0.6)*, suggesting strong association between missing *Occupation* data and lower utility. The utility-driven score for *Income (0.16)* is significantly higher than the impartial score *(0.4)*, suggesting some possible abnormality that needs further investigation.

More insights can be gained by also measuring utility inequality [Eq. 8]. High inequality implies that benefits can be gained by focusing on a relatively small number of records. Further, high utility inequality in a specific attribute may suggest that the attribute is a strong *predictor of utility* - certain attribute values are significantly associated with higher utility. Considering the utility among records, the inequality score *(0.548)* is relatively high. Refining inequality assessments to specific attributes - *Gender*, *Marital Status*, *Occupation*, and *Income* – the scores are *0.057*, *0.110*, *0.208*, and *0.438*, respectively. These scores can further help prioritize quality improvement efforts:

- *Income* must get the highest priority. The rate of defects in this attribute is high and defects are associated with utility loss. Further, the magnitude of utility inequality for income is the highest, suggesting that it has the strongest predictive capability among the four attributes evaluated.
- *Occupation* has to get a high priority, having relatively high defect rate and inequality score. Unlike *Income*, the utility-driven score for *Occupation* is higher than the impartial score, showing strong positive association between quality and utility.
- *Marital Status* and *Gender* should get lower priority. Their defect rates are relatively low, and the utility-driven completeness is similar to the impartial. The magnitude of inequality is low, suggesting that the values stored in these attributes possibly cannot predict utility very well.

As illustrated, attributing utility to records and using this for quality and inequality assessments offers benefits, beyond those offered by impartial assessment alone. Table 2 summarizes some high-level guidelines for using the comparative analysis to gain insights, guide prioritization, and develop quality improvements policies. In the following section, we describe the comparative analysis using real-life alumni data and discuss its implications for managing quality in the context of alumni data.

Impartial Score	*Utility-Driven Score*	*Inequality Score*
High **–** low rate of defects; a high data quality can be reached with relatively low efforts and costs	***High*** **–** defects are not significantly associated with loss in utility	***High*** **–** strong utility differentiation; current quality maintenance efforts are achieving good results, and it is important to keep maintaining this attribute at a high quality level
		Low **–**weak utility differentiation; Improving the quality further is possibly not going to add much contribution.
	Low **–** defects are associated with significant loss in utility, high-utility records are more defective	***High*** **–**strong utility differentiation; The association between higher utility and higher defect rate is possibly due to a systematic cause. Requires focused attention on high-utility records
		Low **–**weak utility differentiation; The association between higher utility and higher defect rate is possibly incidental
Low **–** high rate of defects; reaching a high level of quality will likely require significant efforts and costs	***High*** **–** high utility is associated with lower defect rates	***High*** **–** strong utility differentiation and strong association between the higher defects and lower utility. Needs to get a high priority
		Low **–**weak utility differentiation; may indicate that some quality improvements are made *after* utility contribution, but the cost of additional improvement might not justify the added benefit
	Low **–** defects are not significantly associated with loss in utility	***High*** **-** strong utility differentiation; a high potential for improving utility by correcting missing values.
		Low **–**weak utility differentiation; improvement efforts are not likely to contribute further.

Table 2. Comparative Analysis of Attribute Scores

4 ANALYSES OF ALUMNI DATA

To demonstrate comparative analysis, we evaluate data samples from a real-life system, used for managing alumni relations. This system, a form of CRM, helps generate a major portion of the organization's revenues. The data managed by this system is used by different departments for contacting donors, tracking gifts, assessing contribution potential, and initiating pledge campaigns.

4.1 Assigning Measures to Data

We evaluate large samples from two datasets: *(a) Donor Profiles (358,372 records):* Besides a unique identifier *(Profile ID)*, this dataset contains a large set of attributes. We evaluate *12* attributes, which were identified by users as being commonly used for managing alumni relations and/or classifying profiles. These can be classified as: *(1) Graduation – Year* and *School* are included when a record is added and are unlikely to change. *(2) Demographics* – some (e.g., *Gender, Marital Status, Religion*, and *Ethnicity*) are available when a record is added, others (e.g., *Income*, *Occupation*) are updated later, and *(3) Contact –Home Address* and *Phone* are typically included when a record is added, but may change later. *Business Address* and *Phone* are added only later. Some donors (*11,445* out of *358,372*, *~3%* of the dataset) are classified as *prospects*, based on large gift-giving potential. Donors who are not classified as prospects (*~97%* of the dataset, referred to as *non-prospects*) are typically approached during pledge campaigns (e.g., via phone, mail, or email). Since data collection and usage for prospects versus non-prospects are fundamentally different, we analyze these two categories separately. *(b) Gift Transactions (1,415,432 records):* besides a unique identifier (*Gift ID*), this dataset includes a *Profile ID* (linking each transaction to a specific profile), *Gift Date*, and *Amount* (besides other administrative attributes, not used here). Here we focus on improving the quality of *Profiles*. *Gifts* data, while not being a target for improvement, is used for assessing the utility of the profiles.

To permit calculations, the following measures were added to each profile record:

- *Attribute Completeness:* Each attribute (a total of *12*) is assigned *1* if a value exists, *0* otherwise.

- *Record Completeness: Absolute* measure is *1* if the completeness indicators for *all 12 attributes* are *1* and is *0* otherwise. *Proportion* measure is the average of the *12* attribute-completeness indicators.
- *Utility:* Donations are relatively consistent over time, as reflected by the correlations (Table 3) between annual *amounts* and *inclinations* (*1* – if a donation was made, *0* if not). The correlations between inclinations are positive, high and significant. The amounts are also positively and significantly correlated, but with lower scores. Assuming that gift-giving will be consistent, we calculated (per profile) the average amount in the recent *5* years, as a proxy for utility. The average utility for prospects is *1,303* (STDEV – *15,506*), and for non-prospects it is *6.9* (STDEV – *38.1*).

		Prospects (11,445 Records)				***Non-Prospects (346,927 Records)***			
	Year	**2002**	**2003**	**2004**	**2005**	**2002**	**2003**	**2004**	**2005**
Inclination	**2003**	0.566				0.529			
	2004	0.545	0.563			0.510	0.521		
	2005	0.519	0.555	0.550		0.473	0.504	0.503	
	2006	0.493	0.508	0.533	0.516	0.442	0.466	0.498	0.503
Amount	**2003**	0.240				0.399			
	2004	0.199	0.139			0.359	0.389		
	2005	0.157	0.061	0.146		0.301	0.351	0.412	
	2006	0.016	0.010	0.062	0.020	0.271	0.315	0.341	0.386

Table 3. Correlations between Annual Inclinations and Amounts (all obtained with P-value≈0)

4.2 Results

The evaluation results are summarized in Table 4. Actual addresses and phone numbers were not provided (only a *0/1* indicator), hence, inequality could not be evaluated for these attributes.

		Prospects (11,445 Records)				***Non-Prospects (346,927 Records)***			
		Missing Values	**Impartial Complexness**	**Utility-Driven Completeness**	**Inequality Index**	**Missing Values**	**Impartial Complexness**	**Utility-Driven Completeness**	**Inequality Index**
Attribute	**Graduation Year**	0	1.000	1.000	0.555	24	0.999	0.999	0.282
	Graduation School	0	1.000	1.000	0.291	24	0.999	0.999	0.240
	Gender	30	0.997	0.999	0.084	3,252	0.991	0.996	0.072
	Marital Status	316	0.972	0.981	0.172	37,768	0.891	0.964	0.214
	Ethnicity	3,837	0.665	0.514	0.079	141,039	0.594	0.627	0.062
	Religion	2,776	0.757	0.774	0.328	138,598	0.601	0.709	0.113
	Occupation	7,512	0.344	0.326	0.613	297,036	0.144	0.275	0.226
	Income	1,251	0.891	0.837	0.029	130,687	0.623	0.909	0.088
	H. Address	95	0.992	0.997	N/A	27,074	0.920	0.995	N/A
	B. Address	1,469	0.872	0.925		180,341	0.480	0.811	
	H. Phone	2,035	0.822	0.873		150,840	0.565	0.837	
	B. Phone	2,059	0.820	0.816		219,946	0.366	0.735	
Rec.	**Absolute**	9,624	0.159	0.179	0.954	236,950	0.058	0.125	0.957
	Proportion	115,960	0.844	0.853		2,836,495	0.681	0.821	

Table 4. Quality and Utility Inequality Assessments

Notably, utility-driven scores for prospects are marginally different from corresponding impartial scores, indicating no significant dependency between utility and quality defects. For non-prospects,

however, utility-driven scores are generally higher. Additional insights can be gained by closely examining the difference between impartial and utility-driven scores for non-prospects:

- For attributes with high impartial completeness (e.g., *School* and *Gender*), utility-driven scores are nearly identical to impartial scores. Some margins exist for *Marital Status* and *Home Address* but, since the impartial completeness of these attributes is relatively high, the margins are fairly small.
- For attributes with low impartial quality, we see substantial variability in margins between impartial and utility-driven scores. The margin is small for *Ethnicity*, slightly higher for *Religion*, and significantly higher for *Income*, *Occupation*, *Business Address* and *Phones*. This implies that the presence of defects in the latter attributes significantly differentiates records with relatively high utility versus records with relatively lower utility. Conversely, the presence of defects in *Ethnicity* and *Religion* do not differentiate utility contributions.
- Measuring completeness at the record level (versus measuring it for specific attributes) has an averaging effect. Some margins exist between impartial and utility-driven assessments, but they are not as high as the corresponding margins for some specific attributes.

The magnitude of inequality among records is very high. For prospects, *Occupation* and *Graduation Year* have the highest inequality, and other attributes (e.g., *Religion*, *Graduation School* and *Marital Status*) have a relatively high score as well. Other attributes – *Gender*, *Ethnicity*, and *Income* – have a relatively lower score. Attribute scores for non-prospects are, in general, not as high as for prospects. *Graduation Year* and *Occupation* have the highest score in this case as well. The *Graduation School* and *Marital Status* scores are also relatively high, and the other attributes have lower scores.

4.3 Discussion

This evaluation shows the use of the three assessment types - impartial quality, utility-driven quality, and inequality. The sample datasets allowed impartial assessment of missing values along *Profile* attributes. It also permitted allocation of utility at the record level, which was used for computing utility-driven assessments. Some important insights gained from this evaluation are:

a) High inequality in the utility of profiles can be explained by the high variability in donation amounts and by the large proportions of records associated with *0* utility (*~54%* of prospects and *~88%* of non-prospects made no contributions in the most recent *5* years). The high inequality suggests that cost-effectiveness can be improved by increasing utility contribution of records with low (or no) utility, and/or by reducing the cost of managing records with no utility improvement potential.

b) Association between quality and utility for non-prospects: results indicate that profiles with fewer missing values are associated with higher utility; hence, most utility-driven assessments are higher. Based on discussions with administrators, this association can be explained by: (1) new profiles are imported from the registration system, which only provides a subset of alumni attributes (e.g., *Income* and *Occupation* do not exist; *Ethnicity* and *Religion* are not always available). Hence, most profile records enter the system with missing attributes (2) Some profile attributes may change over time (e.g., *Address*, *Phone Numbers*, *Income*, and *Marital Status*). Failure to keep profiles up-to-date limits the ability to contact the alumni, gather data, and assess their contribution potential. (3) Data administrators and end-users tend to update profiles and fill-in missing values (e.g., by contacting the person and running a phone survey) only when a person makes a donation. So, if a person donated recently, his/her profile is likely to be up-to-date and have less missing values. Conversely, profiles of individuals who have not donated in a while are likely to deteriorate.

c) Higher impartial quality and weaker association between quality and utility for prospect profiles – Prospect profiles offer much higher utility than non-prospect profiles. Not surprisingly, the occurrence of defects in this subset is significantly lower. Each prospect is designated a manager who maintains complete and up-to-date data on the prospect. This includes a thorough investigation of the donor, involving external agencies, to estimate gift-giving potential. The weak association between quality and utility in prospect profiles appears counter-intuitive. An explanation is that the quality of prospect

profile is inherently high. So, utility degradation due to defects is less significant and harder to detect. Further, the gifting potential of prospects is not determined solely from the alumni data and uses other data resources (e.g., city assessor's database, registry of deeds) that are not part of the alumni system.

d) Significant variability in the behavior of different attributes – the evaluation shows that the presence of quality defects and their adverse effects on utility differs significantly between attributes. For some attributes with inherently-high quality, the negative impact of quality defects on utility is negligible. The utility degradation is relatively small, even for some attributes with poor quality. However, for certain attributes, the association between quality defects and utility degradation is strong (e.g., *Income*, *Business Address and Phone*). Further, some attributes could be associated with high inequality in utility for both *prospects and non-prospects*. These suggest that measuring utility at the record level alone may only provide a partial (and possibly misleading) picture of the impact of quality defects. Averaging the quality assessments of individual attributes to derive the quality assessment of the record might "soften" the effect that the quality of these attributes has on utility.

Data administrators are aware of the link between utility and quality and this drives current data management policies to some extent. However, our evaluation sheds light on issues that can guide the development of superior quality management policies for this data resource:

Differentiation: Data administrators should treat records and attributes differently when auditing, correcting defects, and implementing steps to prevent recurrence of defects. Users can be asked to avoid certain subsets of records/attributes for certain usages. Our results indicate a significant variation in utility contribution among profile records, both between and within prospects and non-prospects. The strength of association between an attribute and its utility differs significantly between attributes. With such variations, treating all records/attributes identically will not be cost-effective. Quality management (e.g., prevention, auditing, correction, and usage) must be differentially applied to subsets of records/attributes to maximize the improvement in utility for the investments made.

Attributing Utility: Our results exemplify the benefits of assessing and attributing utility. Our utility metric, reflecting the impact of quality defects on utility, permits easy assessment of utility-driven quality and inequality. Other utility assessments may provide superior insights for quality management and should be explored. For example, a refinement would consider not only past gifts, but also a prediction of future gifts (e.g. by using Customer Lifetime Value techniques) (Berger and Nasr, 1998).

Improving Completeness: Our results indicate that it is necessary to analyze the impact of missing values at the attribute level. The impartial completeness of some attributes (e.g., *School and Gender)* is inherently high, and the potential utility gain by fixing errors in these attributes is negligible. Among attributes with lower impartial completeness, some (e.g., *Occupation*, *Income*, *Business Address and Phone*) exhibit a strong association between missing values and utility. Such attributes must receive a higher priority for improvement. Other attributes (e.g., *Marital Status* and *Religion*) exhibit weaker association with utility, and yet others (e.g., *Ethnicity*) almost no association at all. In such cases, we need to examine whether to invest in any quality improvement at all. Insights for managing other attributes gained from assessing inequality are: *Occupation* deserves a high priority for quality improvement as it differentiates utility the strongest (and its quality is currently very low). *Income* has a high utility-driven score, but a low magnitude of inequality. Currently, *Income* uses only 3 values (High, Medium, Low), and it is typically added to the record only after a donation is made. Our evaluation results suggest that using this limited set of values for *Income* may be ineffective.

5 CONCLUSIONS

This study posits that the link between quality defects, improvement efforts, and economic outcomes is critical for data quality management and, so far, has been insufficiently examined. As a contribution to that end, we examine three quantitative assessments of quality in a large data resource. Impartial assessment reflects the presence of defects in a dataset and can inform the decision maker about the efforts and costs of improving quality. Contextual assessment, by attributing utility to data records,

reflects the degrading impact of defects on business value. It can help assess the benefits to be gained by improving quality. Measuring inequality in utility can highlight specific subsets of records and attributes that are associated with high utility and possibly require preferential treatment. We demonstrate the application of this comparative analysis using alumni data. We show that, in that context, such analyses can improve existing data quality management practices by offering deeper insights into the current state of quality and by identifying quality improvement priorities.

The evaluation described here only serves to demonstrate the assessment methodology and its application. It also helps get a sense of the insights to be gained from such analyses. A complete solution demands analyzing all relevant attributes, evaluating other utility assessments, examining other possible usages of this dataset, and considering all the costs involved. Quality improvement policies and prioritization decisions must also consider other constraints such as business commitments and availability of human and financial resources for data correction efforts.

Future extensions may relax certain assumptions and limitations – e.g., by examining multiple datasets, other quality dimensions, and possible interaction effects (hence, non-additive utility). This study assesses missing values – a specific data-quality defect, associated with the completeness dimension. Data resources are vulnerable to other defects (e.g., incorrect and/or outdated content). Hence, evaluating data quality along other dimensions (e.g., accuracy and currency) is equally important. As explained in (Even and Shankaranarayanan, 2007), the quantitative utility-driven framework for assessing data quality in context can address different defects. The methodology offered here can therefore be extended to include other quality dimensions as well. Importantly, the proposed comparative methodology should be evaluated in other business and data usage contexts. Evaluation in other contexts will require different methods for estimating utility and attributing it to records. Monetary utility assessments which reflect potential purchase power (similar to the one used here) are applicable in other CRM and retail contexts. However, other business environments (e.g., finance, healthcare, and insurance) will require other ways of conceptualizing utility. Efforts to address limitations and explore other business contexts are currently in progress.

References

Ahituv, N. (1980). A systematic approach towards assessing the value of IS. MISQ, 4 (4), 61-75.

Ballou, D. P., and Pazer, H. L. (1995). Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff. Information Systems Research, 6 (1), 51-72.

Berger, P.D., and Nasr, N.I. (1998). Customer lifetime value: marketing models and applications. Journal of Interactive Marketing, 12 (1), 17-30.

Even, A., and Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. The DATA BASE for Advances in Information Systems, 38 (2), 76-93.

Even, A., Shankaranarayanan, G., and Berger, P.D. (2007). Inequality in the utility of data and its implications for data management. Proceedings of the 17th Annual Workshop on Information Technologies and Systems (WITS), Montreal, Canada

Gini C. (1912). Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T), Rome: Libreria Eredi Virgilio Veschi, 1955

Heinrich, B., Kaiser, M., and Klier, M. (2007). How to measure data quality? A metric-based approach. Proceedings of the Intl. Conference on Information Systems (ICIS), Montreal, Canada

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. Publications of the American Statistical Association, 9, pp. 209-219

Petrison, L.A., Blattberg, R.C., and Wang, P. (1997). Database marketing: past present, and future. Journal of Direct Marketing 11 (4), 109-125.

Pipino L.L, Lee, Y.W. and Wang, R.Y. (2002). Data Quality Assessment. CACM, 45 (4), 211-218.

Redman, T.C. (1996). Data Quality for the Information Age. Artech House, Boston, MA.

Wang R.Y. (1998). A product perspective on total quality management. CACM, 41 (2), 58-65.