Dual Assessment of Data Quality in Customer Databases

ADIR EVEN
Ben-Gurion University of the Negev
and
G. SHANKARANARAYANAN
Babson College

Quantitative assessment of data quality is critical for identifying the presence of data defects and the extent of the damage due to these defects. Quantitative assessment can help define realistic quality improvement targets, track progress, evaluate the impacts of different solutions, and prioritize improvement efforts accordingly. This study describes a methodology for quantitatively assessing both impartial *and* contextual data quality in large datasets. Impartial assessment measures the extent to which a dataset is defective, independent of the context in which that dataset is used. Contextual assessment, as defined in this study, measures the extent to which the presence of defects reduces a dataset's utility, the benefits gained by using that dataset in a specific context. The dual assessment methodology is demonstrated in the context of Customer Relationship Management (CRM), using large data samples from real-world datasets. The results from comparing the two assessments offer important insights for directing quality maintenance efforts and prioritizing quality improvement solutions for this dataset. The study describes the steps and the computation involved in the dual-assessment methodology and discusses the implications for applying the methodology in other business contexts and data environments.

Categories and Subject Descriptors: E.m [**Data**]: Miscellaneous

General Terms: Economics, Management, Measurement

Additional Key Words and Phrases: Data quality, databases, total data quality management, information value, customer relationship management, CRM

Authors' addresses: A. Even, Department of Industrial Engineering and Management (IEM), Ben-Gurion University of the Negev, Beer-Sheva, 84105, Israel; email: adireven@bgu.ac.il; G. Shankaranarayanan (corresponding author), Technology, Operations, and Information Management (TOIM), Babson College, Babson Park, MA 02457-0310; email: gshankar@babson.edu.

1. INTRODUCTION

High-quality data makes organizational data resources more usable and, consequently, increases the business benefits gained from using them. It contributes to efficient and effective business operations, improved decision making, and increased trust in information systems [DeLone and McLean 1992; Redman 1996]. Advances in information systems and technology permit organizations to collect large amounts of data and to build and manage complex data resources. Organizations gain competitive advantage by using these resources to enhance business processes, develop analytics, and acquire business intelligence [Davenport 2006]. The size and complexity make data resources vulnerable to data defects that reduce their data quality. Detecting defects and improving quality is expensive, and when the targeted quality level is high, the costs often negate the benefits. Given the economic trade-offs in achieving and sustaining high data quality, this study suggests a novel economic perspective for data quality management. The methodology for dual assessment of quality in datasets described here accounts for the presence of data defects in that dataset, assuming that costs for improving quality increase with the number of defects. It also accounts for the impact of defects on benefits gained from using that dataset.

Quantitative assessment of quality is critical in large data environments, as it can help set up realistic quality improvement targets, track progress, assess impacts of different solutions, and prioritize improvement efforts accordingly. Data quality is typically assessed along multiple quality dimensions (e.g., accuracy, completeness, and currency), each reflecting a different type of quality defect [Wang and Strong 1996]. Literature has described several methods for assessing data quality and the resulting quality measurements often adhere to a scale between 0 (poor) and 1 (perfect) [Wang et al. 1995; Redman 1996; Pipino et al. 2002]. Some methods, referred to by Ballou and Pazer [2003] as structure-based or structural, are driven by physical characteristics of the data (e.g., item counts, time tags, or defect rates). Such methods are *impartial* as they assume an objective quality standard and disregard the context in which the data is used. We interpret these measurement methods as reflecting the *presence* of quality defects (e.g., missing values, invalid data items, and incorrect calculations). The extent of the presence of quality defects in a dataset, the impartial quality, is typically measured as the ratio of the number of nondefective records and the total number of records. For example, in the sample dataset shown in Table I, let us assume that no contact information is available for customer A. Only 1 out of 4 records in this dataset has missing values; hence, an impartial measurement of its *completeness* would be $(4-1)/4 = 0.75$.

Other measurement methods, referred to as content-based [Ballou and Pazer 2003], derive the measurement from data content. Such measurements typically reflect the impact of quality defects within a specific usage context and are also called contextual assessments [Pipino et al. 2002]. Data-quality literature has stressed the importance of contextual assessments as the impact of defects can vary depending on the context [Jarke et al. 2002; Fisher et al. 2003]. However, literature does not minimize the importance of impartial

Table I. Sample Dataset

Customer	Income	Children	Contact
A	200,000	0	*<Missing>*
B	80,000	5	*<Available>*
C	160,000	1	*<Available>*
D	60,000	4	*<Available>*

assessments. In certain cases, the same dimension can be measured both impartially and contextually, depending on the purpose [Pipino et al. 2002]. Given the example in Table I, let us first consider a usage context that examines the promotion of educational loans for dependent children. In this context, the records that matter the most are the ones corresponding to customers *B* and *D*: families with many children and relatively low income. These records have no missing values and hence, for this context, the dataset may be considered complete (i.e., a completeness score of 1). For another usage context that promotes luxury vacation packages, the records that matter the most are those corresponding to customers with relatively higher income, *A* and *C*. Since 1 out of these 2 records is defective (record *A* is missing *contact*), the completeness of this dataset for this usage context is only 0.5.

In this study we describe a methodology for the dual assessment of quality; dual, as it assesses quality both impartially and contextually and draws conclusions and insights from comparing the two assessments. Our objective is to show that the dual perspective can enhance quality assessments and help direct and prioritize quality improvement efforts. This is particularly true in large and complex data environments in which such efforts are associated with significant cost-benefit trade-offs. From an economic viewpoint, we suggest that *impartial assessments can be linked to costs*. The higher the number of defects in a dataset, the more is the effort and time needed to fix it and the higher the cost for improving the quality of this dataset. On the other hand, depending on the context of use, improving quality differentially affects the usability of the dataset. Hence, we suggest that *contextual assessment can be associated with the benefits gained* by improving data quality. To underscore this differentiation, in our example (Table I), the impartial assessment indicates that 25% of the dataset is defective. Correcting each defect would cost the same, regardless of the context of use. However, the benefits gained by correcting these defects may vary, depending on the context of use. In the context of promoting luxury vacation, 50% of the relevant records are defective and correcting them will increase the likelihood of gaining benefits. In the context of promoting educational loans all the relevant records appear complete. The likelihood of increasing benefits gained from the dataset by correcting defects is low.

Using the framework for assessing data quality proposed in Even and Shankaranarayanan [2007] as a basis, this study extends the framework into a methodology for dual assessment of data quality. To demonstrate the

Table II. Attributing Utility to Records in a Dataset

Customer	Income	Children	Contact	Context - Lux. Vac.	Context - Edu. Loans	Overall
A	200,000	0	<Missing>	80 (40%)	0 (0%)	80 (16.0%)
B	80,000	5	<Available>	32 (16%)	150 (50%)	182 (36.4%)
C	160,000	1	<Available>	64 (32%)	30 (10%)	94 (18.8%)
D	60,000	4	<Available>	24 (12%)	120 (40%)	144 (28.8%)
Total				200 (100%)	300 (100%)	500 (100%)

methodology, this study instantiates it for the specific context of managing alumni data. The method for contextual assessment of quality (described later in more detail), is based on *utility*, a measure of the benefits gained by using data. Information economics literature suggests that the utility of data resources is derived from their usage and integration within business processes and depends on specific usage contexts [Ahituv 1980; Shapiro and Varian 1999]. The framework defines data utility as a nonnegative measurement of value contribution attributed to the records in the dataset based on the relative importance of each record for a specific usage context. A dataset may be used in multiple contexts and contribute to utility differently in each; hence, each record may be associated with multiple utility measures, one for each usage context.

We demonstrate this by extending the previous example (see Table II). In the context of promoting luxury vacations, we may attribute utility, reflecting the likelihood of purchasing a vacation, in a manner that is proportional to the annual income; that is, higher utility is attributed to records A and C than to records B and D. In the context of promoting educational loans, utility, reflecting the likelihood of accepting a loan, may be attributed in a manner that is proportional to the number of children. In the latter case, the utilities of records B and D is much higher than that of A and C. We hasten to add that the numbers stated in Table II are for illustration purposes only. Several other factors which may affect the estimation of utility are discussed further in the concluding section.

The presence of defects reduces the usability of data resources [Redman 1996] and hence, their utility. The magnitude of reduction depends on the type of defects and their impact within a specific context of use. Our method for contextual assessment defines quality as a weighted average of defect count, where the weights are context-dependent utility measures. In the preceding example (Table II), the impartial completeness is 0.75. In the context of promoting luxury vacations, 40% of the dataset's utility (contributed by record A) is affected by defects (missing *contact*). The estimated contextual completeness is hence 0.6. In the context of promoting educational loans, utility is unaffected (as record A contributes 0 to utility in this context) and the estimated contextual completeness is 1. Summing up both usages, 16% of the utility is affected by defects; hence, the estimated contextual completeness is 0.84. This illustration highlights a core principle of our methodology: high variability in utility-driven scores, and large differences between impartial and contextual

scores may have important implications for assessing the current state of a data resource and prioritizing its quality improvement efforts.

In this study, we demonstrate dual assessment in a real-world data environment and discuss its implications for data quality management. We show that dual assessment offers key insights into the relationships between impartial and contextual quality measurements that can guide quality improvement efforts. The key contributions of this study are: (1) it extends the assessment framework proposed in Even and Shankaranarayanan [2007] and illustrates its usefulness by applying it in a real-world Customer Relationship Management (CRM) setting. (2) It provides a comparative analysis of both impartial and contextual assessments of data quality in the context of managing the alumni data. Importantly, it highlights the synergistic benefits of the dual assessments for managing data quality, beyond the contribution offered by each assessment alone. (3) Using utility-driven analysis, this study sheds light on the high variability in the utility contribution of individual records and attributes in a real-world data environment. Further, the study also shows that different types of quality defects may affect utility contribution differently (specifically, missing values and outdated data). The proposed methodology accounts for this differential contribution. (4) The study emphasizes the managerial implications of assessing the variability in utility contribution for managing data quality, especially, for prioritizing quality improvement efforts. Further, it illustrates how dual assessment can guide the implementation and management of quality improvement methods and policies.

In the remainder of this article, we first review the literature on quality assessment and improvement that influenced our work. We then describe the methodology for dual assessment and illustrate its application using large samples of alumni data. We use the results to formulate recommendations for quality improvements that can benefit administration and use of this data resource. We finally discuss managerial implications and propose directions for further research.

2. RELEVANT BACKGROUND

We first describe the relevant literature on managing quality in large datasets and assessing data quality. We then discuss, specifically, the importance of managing quality in a Customer Relationship Management (CRM) environment, the context for this study.

2.1 Data Quality Improvement

High-quality data is critical for successful integration of information systems within organizations [DeLone and McLean 1992]. Datasets often suffer defects such as missing, invalid, inaccurate, and outdated values [Wang and Strong 1996]. Low data quality lowers customer satisfaction, hinders decision making, increases costs, breeds mistrust towards IS, and deteriorates business performance [Redman 1996]. Conversely, high data quality can be a

unique source for sustained competitive advantage. It can be used to improve customer relationships [Roberts and Berger 1999], find new sources of savings [Redman 1996], and empower organizational strategy [Wixom and Watson 2001]. Empirical studies [Chengalur-Smith et al. 1999; Fisher et al. 2003; Shankaranarayanan et al. 2006] show that communicating data quality assessments to decision makers may positively impact decision outcomes.

Data Quality Management (DQM) techniques for assessing, preventing, and reducing the occurrence of defects can be classified into three high-level categories [Redman 1996].

(1) *Error Detection and Correction.* Errors may be detected by comparing data to a correct baseline (e.g., real-world entities, predefined rules/calculations, a value domain, or a validated dataset). Errors may also be detected by checking for missing values and by examining time-stamps associated with data. Correction policies must consider the complex nature of data environments, which often include multiple inputs, outputs, and processing stages [Ballou and Pazer 1985; Shankaranarayanan et al. 2003]. Firms may consider correcting defects manually [Klein et al. 1997] or hiring agencies that specialize in data enhancement and cleansing. Error detection and correction can also be automated; literature proposes, for example, the adoption of methods that optimize inspection in production lines [Tayi and Ballou 1988; Chengalur et al. 1992], integrity rule-based systems [Lee et al. 2004], and software agents that detect quality violations [Madnick et al. 2003]. Some ETL (Extraction, Transformation, and Loading) tools and other commercial software also support the automation of error detection and correction [Shankaranarayanan and Even 2004].

(2) *Process Control and Improvement.* The literature points out a drawback with implementing error detection and correction policies. Such policies improve data quality, but do not fix root causes and prevent recurrence of data defects [Redman 1996]. To overcome this issue, the Total Data Quality Management (TDQM) methodology suggests a continuous cycle of data quality improvement: define quality requirements, measure along these definitions, analyze results, and improve data processes accordingly [Wang 1998]. Different methods and tools for supporting TDQM have been proposed, for example, systematically representing data processes [Shankaranarayanan et al. 2003], optimizing quality improvement trade-offs [Ballou et al. 1998], and visualizing quality measurements [Pipino et al. 2002; Shankaranarayanan and Cai 2006].

(3) *Process Design.* Data processes can be built from scratch or, existing processes redesigned, to better manage quality and reduce errors. Process design techniques for quality improvement are discussed in a number of studies (e.g., Ballou et al. [1998], Redman [1996], Wang [1998], and Jarke et al. [2002]). These include embedding controls in processes, supporting quality monitoring with metadata, and improving operational efficiency. Such process redesign techniques can help eliminate root causes of defects, or greatly reduce their impact.
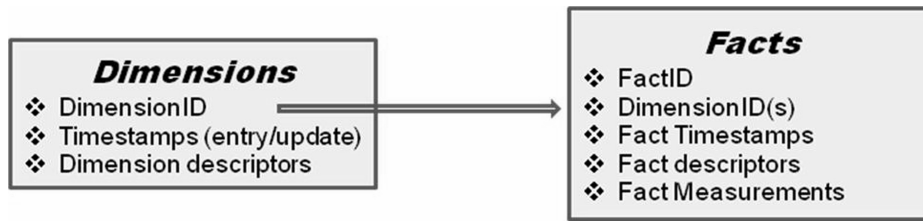
Fig. 1.   Dimension and fact tables.

Organizations may adopt one or more quality improvement techniques, based on the categories stated previously, and the choice is often influenced by economic cost-benefit trade-offs. Studies have shown that substantial benefits were gained by improving data quality [Redman 1996; Heinrich et al. 2007], although the benefits from implementing a certain technique are often difficult to quantify. On the other hand, quality improvement solutions often involve high costs as they require investments in labor for monitoring, software development, managerial overheads, and/or the acquisition of new technologies [Redman 1996]. To illustrate one such cost, if the rate of manual detection and correction is 10 records per minute, a dataset with 10,000,000 records will require ~16,667 work hours, or ~2,083 work days. Automating error detection and correction may substantially reduce the work hours required, but requires investments in software solutions. We suggest that the dual, assessment methodology described can help understand the economic trade-offs involved in quality management decisions and identify economically superior solutions.

2.2 Improving the Quality of Datasets

This study examines quality improvement in a tabular dataset (a table), a data storage structure with an identical set of attributes for all records within. It focuses on tabular datasets in a Data Warehouse (DW). However, the methods and concepts described can be applied to tabular datasets in other environments as well. Common DW designs include two types of tables: *fact* and *dimension* (Figure 1) [Kimball et al. 2000]. Fact tables capture data on business transactions. Depending on the design, a fact record may represent a single transaction or an aggregation. It includes numeric measurements (e.g., quantity and amount), transaction descriptors (e.g., time-stamps, payment and shipping instructions), and foreign-key attributes that link transactions to associated business dimensions (e.g., customers, products, locations). Dimension tables store dimension instances and associated descriptors (e.g., timestamps, customer names, demographics, geographical locations, products, and categories). Dimension instances are typically the subject of the decision (e.g., target a *specific subset* of customers), and the targeted subset is commonly defined along dimensional attributes (e.g., send coupons to customers *between 25–40 years of age and with children*). Fact data provide numeric measurements that categorize dimension instances (e.g., the frequency and the total

amount of past purchases). This study focuses on improving the quality of *dimensional data*. However, in real-world environments, the quality of fact data must be addressed as well, as defective fact data will negatively impact decision outcomes.

Improving the quality of datasets (dimension or fact) has to consider the *targeted quality level* and the *scope of quality improvement*. Considering *quality target*, at one extreme, we can opt for perfect quality and at the other, opt to accept quality as is without making any efforts to improve it. In between, we may consider improving quality to some extent, permitting some imperfections. Quality improvement may target multiple quality dimensions, each reflecting a particular type of quality defect (e.g., completeness, reflecting missing values, accuracy, reflecting incorrect content, and currency, reflecting how up-to-date the data is). Studies have shown that setting multiple targets along different quality dimensions has to consider possible conflicts and trade-offs between the efforts targeting each dimension [Ballou and Pazer 1995; 2003]. Considering the *scope of quality improvement*, we may choose to improve the quality of all records and attributes identically. Alternately, we may choose to differentiate: improve only certain records and/or attributes, and make no effort to improve others.

From these considerations of target and scope, different types of quality improvement policies can be evaluated.

(a) *Prevention.* Certain methods can prevent data defects or reduce their occurrences during data acquisition, for example, improving data acquisition user interfaces, disallowing missing values, validating values against a value domain, enforcing integrity constraints, or choosing a different (possibly, more expensive) data source with inherently cleaner data.

(b) *Auditing.* Quality defects also occur during data processing (e.g., due to miscalculation, or mismatches during integration across multiple sources), or after data is stored (e.g., due to changes in the real-world entity that the data describes). Addressing these defects requires auditing records, monitoring processes, and detecting the existence of defects.

(c) *Correction.* It is often questionable whether the detected defects are worth correcting. Correction might be time consuming and costly (e.g., when a customer has to be contacted, or when missing content has to be purchased). One might hence choose to avoid correction if the added value cannot justify the cost.

(d) *Usage.* In certain cases, users should be advised against using defective data, especially when the quality is very poor and cannot be improved.

Determining the target and scope of quality improvement efforts has to consider the level of improvement that can be achieved, its impact on data usability, and the utility/cost trade-offs associated with their implementation [Even et al. 2007]. Our dual-assessment methodology can provide important inputs for such evaluations.

2.3 Managing Data Quality in CRM Environments

We apply the dual-assessment methodology in a CRM setting. The efficiency of CRM and the benefits gained from it depend on the data resources: customer profiles, transaction history (e.g., purchases, donations), past contact efforts, and promotion activities. CRM data supports critical marketing tasks, such as segmenting customers, predicting consumption, managing promotions, and delivering marketing materials [Roberts and Berger 1999]. It underlies popular marketing techniques such as the RFM (Recency, Frequency, and Monetary) analysis for categorizing customers [Petrison et al. 1997], estimating Customer Lifetime Value (CLV), and assessing customer equity [Berger and Nasr 1998; Berger et al. 2006]. Blattburg and Deighton [1996] define customer equity as the total asset value of the relationships which an organization has with its customers. Customer equity is based on customer lifetime value and understanding customer equity can help optimize the balance of investment in the acquisition and retention of customers. A key concern in CRM is that customer data is vulnerable to defects that reduce data quality [Khalil and Harcar 1999; Coutheoux 2003]. Datasets that capture customer profiles and transactions tend to be very large (e.g., the Amazon Web site (www.amazon.com), as of 2007, is reported to manage about 60 million active customers). Maintaining such datasets at high quality is challenging and expensive.

We examine two quality defects that are common in CRM environments: *(a) Missing Attribute Values:* some attribute values may not be available when initiating a customer profile record (e.g., income level and credit score). The firm may choose to leave these unfilled and update them later, if required. Existing profiles can also be enhanced with new attributes (e.g., email address and a mobile number), and the corresponding values are initially null. They may remain null for certain customers if the firm chooses not to update them due to high data acquisition costs. *(b) Failure to Keep Attribute Values Up to Date:* some attribute values are likely to change over time (e.g., address, phone number, and occupation). If not maintained current, the data on customers becomes obsolete and the firm looses the ability to reach or target them. A related issue in data warehouses is referred to as "slowly changing dimensions" [Kimball et al. 2000]. Certain dimension attributes change over time, causing the transactional data to be inconsistent with the associated dimensional data (e.g., a customer is now married, but the transaction occurred when s/he was single). As a result, analyses may be skewed. In this study, we focus on assessing data quality along two quality dimensions that reflect the quality defects discussed before: completeness, which reflects the presence of missing attribute values, and currency, which reflects the extent to which attribute values or records are outdated.

With large numbers of missing or outdated values, the usability of certain attributes, records, and even entire datasets is considerably reduced. Firms may consider different quality improvement treatments to address such defects, for example, contact customers and verify data or hire agencies to find and validate the data. Some treatments can be expensive and/or fail to achieve the desired results. A key purpose of the dual-assessment method proposed

is to help evaluate the different quality improvement alternatives and assess their costs and anticipated impact.

3. DUAL ASSESSMENT OF DATA QUALITY

The dual-assessment method described includes a comparative analysis of impartial and contextual measurements. To facilitate description, we use an illustrative CRM-like context with two tables: *(a) Customers*, a dimensional dataset with demographic and contact data. Each record has a unique customer identifier (ID) and, for simplicity, we will assume that only three customer attributes are captured: *Gender*, *Marital Status*, and *Income Level*. The dataset includes an *Audit Date* attribute that captures the date on which a customer profile was most recently audited. We use this attribute to assess currency. *(b) Sales*, a fact dataset containing sale transactions. Besides a unique identifier (*Sale ID*), this dataset includes a *Customer ID* (a foreign key that links each transaction to a specific customer record), *Date*, and *Amount*. This fact dataset is not a target for quality improvement, but used for assessing the relative contribution of each customer record and for formulating quality improvement policies accordingly.

3.1 Evaluation Methodology and Operationalization of Utility

The methodology includes assessing impartial and contextual data quality. *Impartial quality assessment* reflects the presence of quality defects in a dataset. We consider a dataset with N records (indexed by $[n]$), and M attributes (indexed by $[m]$). The quality measure $q_{n,m}$ (0, severe defects and 1, no defects) reflects the extent to which attribute $[m]$ of record $[n]$ is defective. An impartial measurement reflects the proportion of defective items in a dataset. Accordingly, the quality Q_n^R of record $[n]$, the quality Q_m^D of attribute $[m]$ in the dataset, and the quality of the entire dataset Q^D are defined as

$$(a)\, Q_n^R \ = \ (1/M) \sum\nolimits_{m=1..M} q_{n,m},$$
$$(b)\, Q_m^D \ = \ (1/N) \sum\nolimits_{n=1..N} q_{n,m},\, \text{and}$$
$$(c)\, Q^D \ = \ (1/MN) \sum\nolimits_{n=1..N} \sum\nolimits_{m=1..M} q_{n,m}$$
$$= \ (1/N) \sum\nolimits_{n=1..N} Q_n^R = (1/M) \sum\nolimits_{m=1..M} Q_m^D. \tag{1}$$

With a binary quality indicator (i.e., $q_{n,m} = 0$ or $q_{n,m} = 1$), this formulation is equivalent to a ratio between the count of perfect items and the total number of items. This ratio is consistent with common structural definitions of quality measures (e.g., Redman [1996]; Pipino et al. [2002]).

We now illustrate the impartial assessments for missing values and the extent to which data values are up-to-date along with the corresponding dimensions of completeness and currency. To differentiate between the two dimensions, we replace the (Q, q) annotations in Eq. (1) with (C, c) for completeness, and with (T, t) for currency. For completeness, we assign binary

indicators: $c_{n,m} = 1$ if the value of attribute $[m]$ in record $[n]$ exists, and $c_{n,m} = 0$ if missing. Using Eq. (1), we compute the completeness of records $\{C_n^R\}$, attributes $\{C_m^D\}$ and the entire dataset (C^D) as the $[0, 1]$ proportions of nonmissing values. We term these proportion-based measures *ranked completeness*. For comparison, we also use an alternate measure termed *absolute completeness*: $C_n^{R/a} = 1$ if no attribute values are missing in record $[n]$ and $C_n^{R/a} = 0$ otherwise.

To measure currency, the extent to which data values are up-to-date, we use the *Audit Date* to calculate the record's age (in years). The *Audit Date* time-stamp applies to the entire record, not to specific attributes, hence, our currency calculations are only at the record level. We use both absolute and ranked measures for currency. For *absolute currency*, we assign record currency as $T_n^{R/a} = 1$, if the record $n$has been audited within the last 5 years, and $T_n^{R/a} = 0$, if not. For *ranked currency*, we use the exponential transformation suggested in Even and Shankaranarayanan [2007] to convert the record age to a $[0, 1]$ measure.

$$t_n = \exp\left\{-\alpha \left(Y^C - Y_n^U\right)\right\} \text{ where,} \qquad (2)$$

$Y^C =$ the current year (in this example, we assume $Y^C = 2006$),

$Y_n^U =$ the last year in which record $[n]$ was audited,

$\quad \alpha =$ the sensitivity factor that reflects the rate at which profiles get outdated. Here, assuming that *between* 20% and 25% of the profiles become outdated every year, we chose $\alpha = 0.25$ $(e^{-0.25} =\sim 0.77)$,

$\quad t_n =$ the up-to-date rank of record $[n]$, ~ 0 for a record that have not been audited for a while (i.e., $Y^C >> Y_n^U$) and 1 for an up-to-date record (i.e., $Y^C = Y_n^U = 2006$).

We use t_n as a measure for the ranked currency T_n^R of record $[n]$. We compute absolute and ranked dataset currency ($T^{D/a}$ and T^D, respectively) as an average over all records. To demonstrate the calculations for completeness and currency (up-to-date), we use the sample data in Table III. For illustration, we assume that some attribute values are missing (highlighted) and some records have not been audited recently. We observe that 2 of 4 records are missing values for *Gender* and hence, the impartial completeness of *gender* is 0.5. Similarly, the impartial completeness of *Marital Status* and *Income Level* are 0.75 and 0.25, respectively. The absolute record-level completeness is 0 if at least one attribute is missing and 1 otherwise. The ranked record-level completeness is a $[0, 1]$ proportion of nonmissing values. Accordingly, the absolute dataset completeness (averaged over all records) is 0.25, and the ranked completeness is 0.5. A record's absolute currency score is 1, if it is audited within the last 5 years, and 0 otherwise. The ranked currency is computed using the currency transformation (Eq. (2)). The impartial currency is computed by averaging the corresponding currency score over all records. Accordingly, the absolute and ranked impartial currency scores are 0.75 and 0.58.

Contextual quality assessments reflect not only the presence of defects in a dataset, but also their impact on the usage of this dataset in a specific context. The framework in Even and Shankaranarayanan [2007] suggests measuring

Table III. Impartial versus Utility-Driven Data Quality Assessments

ID	Gender	Marital Status	Income Level	Completeness (Absolute)	Completeness (Ranked)	Last Audit	Up-to-date (Absolute)	Up-to-date (Ranked)	Monet. Utility
A	Male	Married	Med.	1	1	2006	1	1.00	20
B	Female	Married	NULL	0	0.667	2003	1	0.47	80
C	NULL	Single	NULL	0	0.333	2005	1	0.78	0
D	NULL	NULL	NULL	0	0	1996	0	0.08	0
Impartial Attribute Completeness	0.50	0.75	0.25					Total	100
Utility-Driven Attribute Completeness	1.00	1.00	0.20						
Dataset	**Impartial Dataset Completeness**			0.25	0.50	**Impartial Currency**	0.75	0.58	
	Utility-Driven Dataset Completeness			0.20	0.73	**Utility-Driven Currency**	1.00	0.58	

this impact in terms of utility degradation, that is, to what extent is utility reduced as a result of defects. The framework assumes that the overall dataset utility U^D can be attributed among the N records $\{U_n^R\}$, based on relative importance such that $U^D = \Sigma_{n=1..N} U_n^R$. The presence of defects in a record lowers the record's utility by some magnitude. The framework assumes that this magnitude is proportional to the record's quality level Q_n^R (or to the data item quality $q_{n,m}$, for a specific attribute value). It can be shown that, under this assumption, the attribute quality Q_m^D, and the dataset quality Q^D calculations in Eq. (1) can be revised to a weighted-average formulation, using the utilities allocated to each record as weights.

$$
\begin{aligned}
(a)\ Q_m^D &= \left(\sum\nolimits_{n=1..N} U_n^R q_{n,m}\right) \Big/ \left(\sum\nolimits_{n=1..N} U_n^R\right), \text{and} \\
(b)\ Q^D &= \left(\sum\nolimits_{n=1..N} U_n^R Q_n^R\right) \Big/ \left(\sum\nolimits_{n=1..N} U_n^R\right) \\
&= (1/M) \left(\sum\nolimits_{n=1..N} U_n^R \sum\nolimits_{m=1..M} q_{n,m}\right) \Big/ \left(\sum\nolimits_{n=1..N} U_n^R\right) \qquad (3) \\
&= (1/M) \sum\nolimits_{m=1..M} Q_m^D
\end{aligned}
$$

These utility-driven formulations assess the impact of defects in terms of utility degradation. Since utility and its allocation depend on the context of

usage, these formulations are treated as *contextual assessments* of quality (in the remainder of this article, the terms contextual assessment and utility-driven assessment are used interchangeably). The utility-driven assessments use the same quality indicators as the impartial assessments. The scores at the dataset level are weighted averages that use the utility allocations per record as weights (Eq. (3)). To extend our example (Table III) with utility-driven measurements, we use the last year's total sales amount per customer as a proxy for utility. In our example, only two of the four records are associated with utility. Considering attributes *Gender* and *Marital Status*, these two records have no missing values. Hence, for these attributes, utility-driven completeness is 1. Conversely, one of the two utility-contributing records is missing the Income value. Using the utility scores as weights, the income-level completeness is (1*20+0*80)/100=0.2. At the record level, the absolute completeness is (1*20+0*80)/100=0.2 and ranked completeness is (1*20+0.667*80)/100=0.73. Similarly, the absolute currency is (1*20+1*80)/100=1.0 and the ranked currency is (1*20+0.47*80)/100=0.58.

In this example, some utility-driven assessments are relatively close to the corresponding impartial assessments, but others are substantially different. This relationship depends on the distribution of utility between records, and on the association between utility and quality. When utility is distributed equally between records, impartial and utility-driven assessments are expected to be nearly identical. The same is also true when the association between utility and the presence of defects is weak. However, in large real-world datasets, it is more likely that utility is unequally distributed among records. Further, the relationship between utility and the presence of defects may be nontrivial. Recognizing a record as one that offers a higher utility may encourage focused efforts to reduce defects in it. In such cases, utility-driven assessments are likely to be substantially different from corresponding impartial assessments.

Acknowledging these factors, the comparison of utility-driven (contextual) assessments to impartial assessments can provide key insights for managing quality in large datasets. At a high level, this comparison can yield three scenarios.

(a) *Utility-Driven Assessments are Substantially Higher than Impartial Assessments.* This indicates that records with high utility are less defective. Two complementary explanations are possible:
  (i) Defective records are less usable to begin with, hence, have low utility.
  (ii) Efforts have been made to eliminate defects and maintain records with high utility at a high quality level.

(b) *Utility-Driven Assessments are not Significantly Different from Impartial Assessments.* Indicating that utility is evenly distributed across all records, and/or that the association between defect rates and utility is weak.

(c) *Utility-Driven Assessments are Substantially Lower than Impartial Assessments.* This abnormality may indicate a very unequal distribution of utility in the dataset (i.e., a large proportion of utility is associated with a small number of records) and some significant damage to high-utility records (e.g., due to systemic causes). Understanding the relationship between

Table IV. Profile Attributes Evaluated

Category	Attribute	Description
Graduation	Graduation Year	The year of graduated
	School	The primary school of graduation
Demographics	Gender	Male or Female
	Marital Status	Marital status
	Ethnicity	Ethnic group
	Religion	Religion
	Occupation	The person's occupation
	Income	Income-level category
Contact	Home Address	Street address, city, state and country
	Business Address	Street address, city, state, and country
	Home Phone	Regular and/or cellular phone
	Business Phone	Regular or cellular phone

impartial and utility-driven assessments, discussed later in more detail, can guide the development of data quality management policies.

4. DUAL ASSESSMENT OF ALUMNI DATA

We apply our dual-assessment methodology and examine its implications for data quality management using large samples from real-world datasets. The datasets are part of a system used to manage alumni relations. This form of CRM is owned by an academic institution and helps generate a large proportion of its revenue. The data is used by different departments for managing donors, tracking gifts, assessing contribution potential, and managing pledge campaigns. For the purpose of this study, we interacted with and received exceptional support from 12 key users, including the administrators of the alumni data, and alumni-relations managers who use this data often.

4.1 Alumni Data

This study evaluates sizably large samples from two key datasets in the alumni system:

(a) *Profiles* (358,372 records) is a dimensional dataset that captures profile data on potential donors. Besides a unique identifier (Profile ID), this dataset contains a large set of descriptive donor attributes. We evaluate 12 of these attributes (listed in Table IV) for quality. Key alumni data users indicated that the selected attributes were among the ones most commonly used for managing alumni relations and/or classifying profiles. These attributes can be classified by:
  (i) *Graduation:* Values for graduation year and school are included when a record is added, and are unlikely to change later.
  (ii) *Demographics:* Some demographic attribute (e.g., *Gender*, *Religion*, and *Ethnicity*) values are available when a record is added and rarely change later. Others (e.g., *Income, Occupation*) are updated only later and may change over time and,

(iii) *Contact:* Values for *Home Address* and *Home Phone* number are typically included when a record is added, but may change over time. In most cases, values for *Business Address* and *Business Phone* are added only later. These values are typically unavailable when the record is created for full-time students (both graduate and undergraduate degrees), as these students, even if employed during their studies, typically change jobs when they graduate. The *Business Address and Phone* values are available more often for part-time students. However, the vast majority of profile records belong to full-time students.

Two other profile attributes play special roles in our evaluation. *Audit Date*, used to assess currency, reflects the date on which a profile was most recently audited. During the audit of a profile, some attribute values may change (e.g., if the person has moved to a new address, and/or changed marital status). The other is *Prospect*, an attribute which classifies donors and reflects two fundamentally different data usages. Some donors (11,445, ~3% of the dataset) are classified as prospects, based on large contributions made or on the assessed potential for a large gift. Prospects are typically not approached during regular pledge campaigns, but have assigned staff members responsible for maintaining routine contact (e.g., invitations to special events and tickets to shows/sporting events). Nonprospects, (~97% of the dataset), are approached (via phone, mail, or email) during pledge campaigns that target a large donor base.

(b) *Gifts* (1,415,432 records) is a fact dataset that captures the history of gift transactions. Besides a unique identifier (*Gift ID*), this dataset includes a *Profile ID* (foreign key linking each gift transaction to a specific profile record), *Gift Date*, and *Gift Amount*. In addition, this dataset includes administrative attributes that describe payment procedures, not used in our evaluation.

In this study, we focus on improving the quality of *Profiles* dataset. The *Gifts* dataset, though not targeted for improvement, is used for assessing the quality *of Profiles* and formulating quality improvement policies for it. Both datasets include data from 1983 to 2006. In 1983 and 1984, soon after system implementation, a bulk of records that reflect prior (pre-1984) activities were added (203,359 profiles, 405,969 gifts), and since then both datasets have grown gradually. The average annual growth of the *Profiles* dataset is 7,044 records (STDEV: 475). The *Gifts* dataset grows by 45,884 records annually (STDEV: 6,147). Due to confidentiality, the samples shown include only ~40% of the actual data. Some attribute values have been masked (e.g., actual addresses and phone numbers) and all gift amounts have been multiplied by a constant factor.

4.2 Analytical and Statistical Methods Used

The purpose of the data quality measurement methodology described is to measure and understand the current quality of the evaluated dataset. It can help identify key quality issues and prioritize quality improvement efforts. The

methodology neither identifies explicitly the root causes for data quality issues nor sets an optimization objective (other that stating that the purpose of quality improvement is to achieve high data quality). However, as described later, the measurements can promote discussions with and among key stakeholders that manage and use the data resources. Such discussions can direct the investigation into identifying root causes.

The descriptive (and not predictive) nature of the methodology determines the analytical and statistical methods that we chose to employ in this study. As the methodology does not involve cause-effect arguments, regression or other statistical methods that seek explanatory results do not fit. Similarly, as no objective function is defined, the analysis does not require an optimization model. Instead, in addition to the new data quality measures investigated, we use descriptive statistical methods. We compute and provide measurements (referred to also as scores) of data quality, summary statistics (averages and standard deviations of measurements). We use ANOVA (analysis of variance) to compare corresponding measurement scores or summary statistics across subsets to determine if the difference is statistically significant. We use correlation to highlight possible links between the different data quality measures. Although statistical significance is typically assured by our considerably large sample size, we have mentioned all relevant parameters where needed. All statistical analyses were conducted using SPSS, a software package from SPSS, Inc. (www.spss.com).

4.3 Impartial Data Quality Assessments

We initially considered four types of defects: *(a) Missing Values:* a preliminary evaluation indicates that certain attributes in the *Profiles* dataset are missing a large proportion of their values. *(b) Invalid Values:* an attribute value that does not conform to the associated value domain is said to be invalid [Redman 1996]. A preliminary evaluation indicates that this is not a serious issue in the *Profile* dataset, as almost 100% of the values present conform to their associated value domain. *(c) Up-to-Date:* lack of currency is a serious quality issue in the *Profile* dataset, as many profile records have not been audited in a long time (in some cases, since being added to the dataset). *(d) Inaccuracies:* the administrators of the alumni system indicated that a number of profile records contain inaccurate attribute values. However, due to the lack of appropriate baselines, the accuracy of specific instances in our samples could not be validated. Based on this preliminary assessment, we focus our evaluation on two quality dimensions: completeness, reflecting missing values, and currency, reflecting outdated records.

We first evaluated impartial completeness for *specific attributes* (Table V) for prospects and for nonprospects. The measurements exhibit high variability among the attributes. For some attributes (e.g., *Graduation-Year*, *School*, and *Gender*) the number of missing values is negligible and impartial completeness is almost 1. For others (e.g., *Occupation*, *Business Address/Phone*), the number of missing values is much higher and the impartial completeness is hence lower. For all attributes except for *Graduation Year* and *School*, the

Table V. Impartial Quality Assessment

	Prospects (11,445 Records)		Non-Prospects (346,927 Records)		Comparison (ANOVA)	
	Missing Val.	Imp. Score	Missing Val.	Imp. Score	F-Val.	P-Val.
Attribute Completeness						
Grad. Year	0	1.000	24	0.999	0.792	0.374
School	0	1.000	24	0.999	0.792	0.374
Gender	30	0.997	3,252	0.991	55	0.000
Marital Status	316	0.972	37,768	0.891	771	0.000
Ethnicity	3,837	0.665	141,039	0.594	233	0.000
Religion	2,776	0.757	138,598	0.601	1,146	0.000
Occupation	7,512	0.344	297,036	0.144	3,500	0.000
Income	1,251	0.891	130,687	0.623	3,438	0.000
Home Address	95	0.992	27,074	0.920	770	0.000
Business Address	1,469	0.872	180,341	0.480	6,924	0.000
Home Phone	2,035	0.822	150,840	0.565	3,016	0.000
Business Phone	2,059	0.820	219,946	0.366	9,960	0.000
Record Completeness						
Absolute (Records missing at least one value)	9,624	0.159	326,950	0.058	2,010	0.000
Ranked (Missing values in all attributes)	21,380	0.844	1,326,629	0.681	8,530	0.000
Record Currency						
Absolute (Records not audited in the last 5 years)	2,512	0.781	172,774	0.502	3,473	0.000
Ranked (Exp.-transformed avg. record age)	3.021	0.626	7.039	0.420	3,785	0.000

scores for prospects are significantly higher than for nonprospects (confirmed by ANOVA; P-values of ∼0). For *Graduation Year* and *School,* the difference in the impartial measurements (which indicate near-perfect data) between the two groups was insignificant (confirmed by ANOVA, P-value > 0.1).

We then evaluated a few quality measures at the record level (Table V). First, we evaluated impartial completeness at the record level. The absolute completeness (0.159) and ranked completeness (0.844) for prospects are a lot higher (confirmed by ANOVA, P-values of ∼0) than the corresponding scores (0.058 and 0.681, respectively) for nonprospects. Notably, the absolute completeness in both cases is very low, indicating that a large proportion of profile records have missing values (∼84% for prospects, ∼94% for nonprospects). Figure 2 shows the distribution of ranked completeness for prospects and nonprospects. The ranked completeness is between 0 and 1, reflecting the proportion of missing values in the 12 attributes evaluated. The average ranked score is higher for prospects (0.844) than for nonprospects (0.681) and
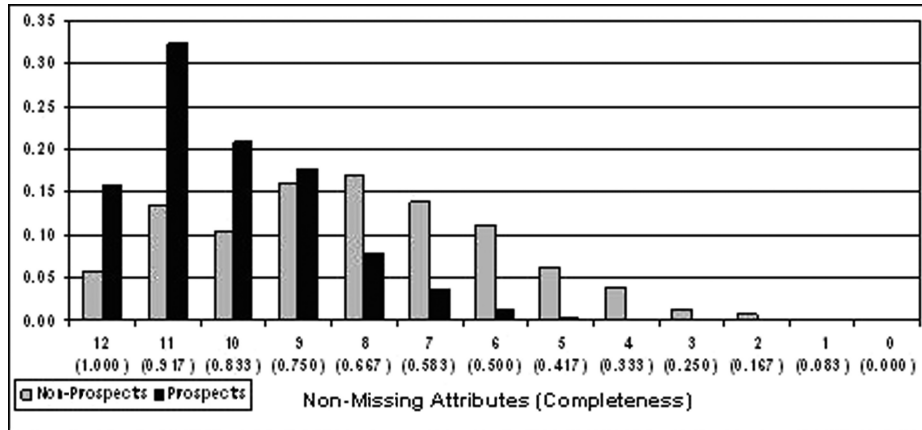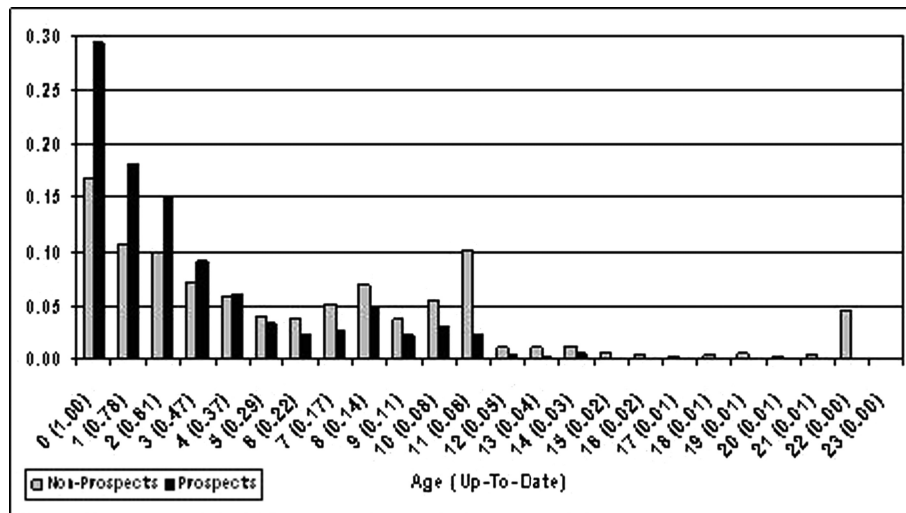
Fig. 2.   Distribution of ranked completeness.

Fig. 3.   Distribution of ranked currency.

the standard deviation is lower (0.119 versus 0.187) (confirmed by ANOVA, P-value of ∼0).

We also evaluated impartial currency at the record level (Table V). Again, all the comparisons between prospects and nonprospects discussed in what follows are statistically significant (confirmed by ANOVA, P-values of ∼0). The absolute and ranked scores for prospects (0.781 and 0.626, respectively) are higher than for nonprospects (0.502 and 0.420, respectively). The absolute scores suggest that a large number of profile records have not been audited in the last 5years (22% of the prospect records, ∼50% of the nonprospect records). Figure 3 shows the distribution of ranked currency: a [0, 1] measure which applies the exponential transformation of the record's age (Eq. (2)). The

Table VI. Impartial Quality and Utility - Summary Statistics and Correlations

	Mean	*STDEV*	*A-CM*	*R-CM*	*A-CR*	*R-CR*	*REC*	*FRQ*	*MON*
Prospects									
A-CM	0.16	0.37		0.57	-0.06	-0.04	0.03	0.02	0.01
R-CM	0.85	0.12	0.57		-0.05	-0.05	0.09	0.08	0.01
A-CR	0.78	0.41	-0.06	-0.05		0.79	0.01	-0.01	0.03
R-CR	0.63	0.32	-0.04	-0.05	0.79		0.05	0.03	0.05
REC	1.92	2.22	0.03	0.09	0.01	0.05		0.89	0.09
FRQ	1.36	1.76	0.02	0.08	-0.01	0.03	0.90		0.10
MON	1303	15,506	0.01	0.01	0.03	0.05	0.09	0.10	
Non-Prospects									
A-CM	0.06	0.23		0.42	-0.08	-0.08	0.10	0.11	0.05
R-CM	0.68	0.19	0.42		0.11	0.11	0.23	0.23	0.13
A-CR	0.50	0.50	-0.08	0.10		0.87	0.08	0.06	0.06
R-CR	0.42	0.35	-0.08	0.11	0.87		0.10	0.07	0.06
REC	0.45	1.32	0.10	0.23	0.08	0.10		0.90	0.52
FRQ	0.28	0.88	0.11	0.23	0.06	0.07	0.90		0.60
MON	6.68	38.1	0.05	0.13	0.06	0.06	0.52	0.60	

Glossary: A/R: Absolute/Ranked, CM/CR: Completeness/Currency, REC: Recency, FRQ: Frequency, MON: Monetary. All correlations are highly significant, P-Value =~0.

average age of prospect profile records is 3.82 years, and the average currency rank is 0.626 with a standard deviation of 0.321. The proportion of up-to-date profiles (i.e., with a perfect rank of 1) is relatively high for prospects (∼29%), and sharply declines as the ranked currency decreases. The average age of nonprospects records is 7.04 years (greater than that of prospects by 85%). The average currency rank for nonprospects is 0.420 (much lower than that for prospects), with a standard deviation of 0.353 (slightly higher than that for prospects). The score distribution for nonprospects (Figure 3) is flatter than that for prospects. The proportion of up-to-date profiles (i.e., with a rank of 1) is not as high (∼17%), and the curve declines as currency rank decreases, but not as sharply or consistently as the curve for prospects.

Table VI shows the summary statistics for the four impartial measurements and the correlations between them. Overall, the impartial quality of the profiles dataset is not perfect. Some attributes are missing a large number of values and many records have not been audited recently. The quality of prospect profiles seems to be a lot higher than that of nonprospect profiles. However, even for this small subset of the dataset (∼3% of the overall), the defect rates are nontrivial. The two completeness measurements (absolute and ranked) are highly and positively correlated, as are the two currency measurements. Conversely, the correlation between completeness and currency is lower. To assess the impact of these defects on the utility that can be gained from the data, we next assess utility-driven (contextual) currency and completeness.

4.4 Utility-Driven Assessments

Utility is assessed using recent gifts associated with each alumni profile. For comparison, we consider three utility metrics per profile: Recency, Frequency, and Monetary (based on RFM analysis, a marketing technique for assessing
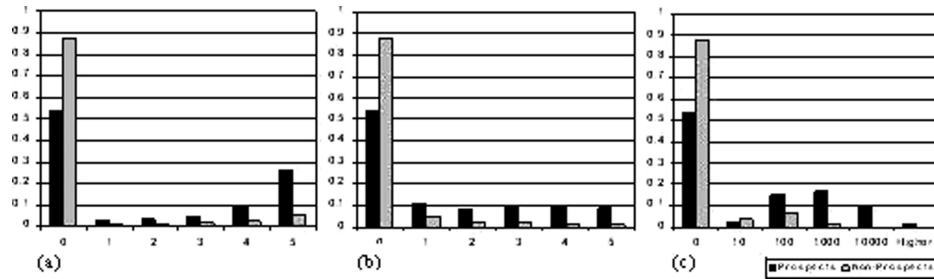
Fig. 4.   Utility Distribution: (a) Recency, (b) Frequency, and (c) Monetary.

customers' purchase power [Petrison et al. 1997]). We compute all three using the most recent 5 years of transactions in the *Gifts* dataset (2002 through 2006): (a) *Recency* determines how recent the donations associated with a profile are. It is calculated using a 0–5 scale, 5 if the last gift was in 2006, 4 if 2005, down to 0 if there were no donations in these 5 years. (b) *Frequency* counts the number of years (out of 5) a person has donated. (c) *Monetary* measures the average annual dollar donation over the last 5 years. For all methods, the utility is 0 if a person made no donation in the last 5 years, and positive otherwise.

The three utility metrics were calculated for each prospect and nonprospect profile. The distributions of these assessments are shown in Figure 4. For nonprospects, the proportion of profiles associated with 0 utility (i.e., no gifts in the last 5 years) is very high (∼88%). For prospects, it is lower (∼54%), yet certainly not negligible.

The summary of utility-driven and impartial quality measurements and the correlations between them are shown in Table VI. The three utility assessments are highly and positively correlated for nonprospects. For prospects, recency and frequency assessments are highly correlated, but their correlation with monetary assessments are lower, yet positive. For prospects, all utility-driven assessments are poorly correlated with the corresponding impartial assessments. For nonprospects, the correlations are slightly higher, with absolute completeness being most correlated.

We next computed utility-driven assessments using Recency (REC), Frequency (FRQ), and Monetary (MON) scores as weights (Table VII). For prospects, utility-driven assessments are only marginally different from impartial assessments (some higher, others lower). This may indicate a low dependency between the number of defects and utility, which confirms the low correlation between impartial quality and utility for prospects (in Table VI). A notable exception is the bigger difference in record-currency scores (absolute and ranked) for monetary (MON) utility. It appears to indicate that prospect records tied to large donations are more likely to be kept up-to-date. Utility-driven assessments for nonprospects are generally higher than their corresponding impartial assessments. This can be attributed to the positive and relatively high utility-quality correlations (in Table VI) for nonprospects.

Table VII. Impartial versus Utility-Driven Quality Assessments

| | Prospects | | | | Non-Prospects | | | |
| | Impartial | Utility-Driven | | | Impartial | Utility-Driven | | |
		REC	FRQ	MON		REC	FRQ	MON
Attribute Completeness								
Grad. Year	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999
School	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999
Gender	0.997	0.997	0.997	0.999	0.991	0.998	0.998	0.996
Marital Status	0.972	0.987	0.981	0.981	0.891	0.945	0.959	0.964
Ethnicity	0.665	0.641	0.631	0.514	0.594	0.658	0.640	0.627
Religion	0.757	0.757	0.753	0.774	0.601	0.705	0.716	0.709
Occupation	0.344	0.353	0.348	0.326	0.144	0.264	0.287	0.275
Income	0.891	0.906	0.911	0.837	0.623	0.867	0.912	0.909
Home Address	0.992	0.995	0.995	0.997	0.920	0.996	0.997	0.995
Bus. Address	0.872	0.906	0.908	0.925	0.480	0.749	0.783	0.811
Home Phone	0.822	0.885	0.890	0.873	0.565	0.829	0.840	0.837
Bus. Phone	0.820	0.858	0.869	0.816	0.366	0.674	0.708	0.735
Record Completeness								
Absolute (Records missing at least one value)	0.159	0.170	0.168	0.179	0.058	0.127	0.138	0.125
Ranked (Missing values in all attributes)	0.844	0.856	0.856	0.853	0.681	0.807	0.820	0.821
Record Currency								
Absolute (Records not audited in the last 5 years)	0.781	0.783	0.774	0.920	0.502	0.623	0.596	0.657
Ranked (Exp.-transformed avg. record age)	0.626	0.645	0.638	0.820	0.420	0.522	0.495	0.540

Some insights for managing the quality of nonprospects (∼97% of the dataset) can be gained by examining the results more closely.

—Utility-driven completeness scores, both at the attribute and at the record level, are relatively consistent along the three utility metrics. In the alumni data analyzed, there is no gain in calculating completeness using three utility metrics over measuring it along a single metric.

—For attributes with very high impartial completeness (e.g., close to 1 in *School* and *Gender*), utility-driven measurements are nearly identical to the impartial ones. Some margin exists for *Marital Status* and *Home Address*, but, since the impartial completeness is relatively high to begin with, this margin is fairly small.

—For attributes with inherently low impartial quality (e.g., *Ethnicity, Religion, Income, Occupation*), we see variability in the margins between impartial and utility-driven scores. The margins are relatively small for *Ethnicity*, and slightly larger for *Religion*. They are much larger for *Income, Occupation, Home Phone Number,* and *Business Address* and *Business Phone*. This implies that the latter attributes have a very different association with the utility gained. The completeness of *Income* and *Occupation* differentiates (along all utility measurements) profile records with relatively high utility contribution and profile records with relatively low utility contribution. Conversely, completeness of *Ethnicity* and *Religion* do not differentiate the utility contribution of profile records.

—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 specific attributes.

—Utility driven currency assessments (absolute and ranked) are not very different from corresponding impartial assessments, when using recency and frequency as weights. However, when using monetary, the utility-driven assessments are substantially higher. This implies that the extent to which a record is up-to-date is strongly associated with the amount donated. It suggests that the current practice may be to frequently audit and update the data on donors who have made large contributions or have the potential to do so (the administrators of the alumni system have confirmed this assumption). Notably, the variance of monetary utility is very large compared to the average (in Table V). This indicates a very uneven distribution of gift amounts among profile records; a small number of profiles are associated with large gifts while a large number are associated with small or no gifts.

4.5 Discussion

Our evaluation demonstrated a successful application of the dual data quality assessment methodology in the context of managing alumni data. The datasets used allowed impartial assessments of the extent of missing values along different attributes and the extent to which profile records are not up-to-date. They also permitted the allocation of utility measurements at the record level and the use of these allocations as weights for assessing utility-driven quality. We highlight some important insights from our evaluation.

(a) Association between Quality and Utility for Nonprospect Profiles: the results indicate that profiles that are more up-to-date and have fewer missing attribute values are generally associated with higher utility. Accordingly, utility-driven assessments are higher than impartial assessments. These results are not surprising. Based on discussions with the data administrators,

this association between quality and utility can be explained as follows: first, new profiles are typically imported from the student registration system, which only provides a subset of the attributes required by the alumni system (e.g., *Income* and *Occupation* are not yet available when a student graduates; *Ethnicity* and *Religion* are optional attributes, which are not collected for each student). As a result, most profile records enter the system with missing attributes, which negatively affects the ability to assess the potential contribution of the donors associated with these profiles. Second, some profile attributes are likely to change over time (e.g., *Address, Phone Numbers, Income*, and *Marital Status*). Failure to keep profiles up-to-date might limit the ability to contact the alumni, gather additional data, and assess contribution potential. Finally, data administrators and end-users tend to audit profiles and fill in missing values (e.g., by contacting the person or running a phone survey) only when a person makes a donation. As a result, if a person donated recently, his/her profile data is likely to be up-to-date and have less missing values. On the other hand, if a person has not donated for a few years in a row, the quality of his/her profile data is likely to deteriorate.

(b) Higher Impartial Quality and Weaker Association between Quality and Utility for Prospect Profiles: prospect profiles represent alumni who donate larger amounts and more often; hence, they are associated with much higher utility than nonprospect profiles. Not surprisingly, the occurrence of quality defects in this subset is much lower. Typically, prospects are assigned with contact persons (alumni-office employees), who maintain their data complete and up-to-date. These efforts involve a thorough investigation of prospects' donation potential, and often require the services of external agencies.

The weak association between quality and utility in prospect profiles appears counterintuitive. A possible explanation is that the quality of prospect profile is inherently high. As a result, degradation of utility due to quality defects is less significant and hence, harder to detect. Another explanation offered by the administrators is that the gifting potential of prospects is not determined using the alumni data alone. Prospect relations are managed by dedicated staff members who use proprietary data resources (e.g., city assessor's database, registry of deeds, and data collected by investigative agencies), besides alumni data. This supplemental data is collected and maintained separately, not as part of the alumni system.

(c) High Variability in Behavior of Attributes: the results show that presence of quality defects *and* their adverse effect on utility differ between attributes. The impact of defects on utility degradation was negligible for attributes that are inherently of high quality. Even for some attributes that are low in quality, the degradation was relatively small. However, for certain attributes, quality defects degraded utility substantially. The aforesaid suggests that measuring utility solely at the record level might provide a partial and possibly misleading picture of the impact of quality defects. Measuring quality at the record level averages the assessments at the attribute level. This masks and softens the association between the quality of specific attributes and their utility.

During our study, we gathered that the key users of the alumni data understood and acknowledged the link between data quality and utility. This

is reflected to some extent in current data management policies. However, our evaluation sheds light on a few issues that can guide the development of better quality management policies for alumni data:

Differentiation.   In general, data administrators should treat records and attributes differently with respect to auditing, correcting quality defects, and implementing procedures to prevent defects from recurring. They may also consider recommending that users refrain from using certain subsets of records or attributes for certain usages (decision tasks and applications). Our results indicate high variability in utility contribution among profile records, between prospects and nonprospects, and within each of these subsets. The results also show that each attribute is associated with utility differently and that the differences, in some cases, are large. With such extensive variations, treating all records and attributes identically is unlikely to be cost effective. Data quality management efforts and policies (e.g., prevention, auditing, correction, and usage) must be differentially applied to subsets of records in a manner that is likely to provide the highest improvement in utility for the investments made.

Attributing utility.   Our results highlight the benefit of assessing and attributing utility. Our metrics, namely, Recency, Frequency, and Monetary, reflect the impact of defects on utility and hence permit convenient utility-driven assessment of data quality. Importantly, the manner in which utility was measured and attributed is specific to our evaluation context and cannot be generalized. Other real-world datasets will require different utility assessment and allocation methods. Even in our specific context, other utility assessments may provide superior insights for data quality management and should be explored. For example, utility measurement may consider not only past donations, but also a prediction of the potential for future donations. This may be done, for example, by using techniques that help assess Customer Lifetime Value (CLV) [Berger et al. 2006].

Improving completeness.   The results indicate that analyzing the impact of missing values at the record level alone is insufficient. It is necessary to analyze it at the attribute level as well. The impartial completeness is inherently high for some attributes (e.g., *School* and *Gender*, with almost no missing values) and hence, the potential to gain utility by further improving the quality of 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 contribution. Efforts to improve such attributes should receive a high priority. Other attributes (e.g., *Marital Status* and *Religion*) are weakly associated with utility and yet others (e.g., *Ethnicity*) exhibit almost no association. For the latter set of attributes, one may question whether it is worthwhile investing in any quality improvement effort at all. The data resource evaluated in this study contains many (over a hundred) other profile attributes that were not evaluated. Evaluating these attributes along the same lines will help manage the attribute configuration in the dataset and prioritize associated quality improvement efforts.

Improving currency.   Utility was strongly linked to currency, as outdated profiles are associated with lower donation amounts. This indicates a need to audit profiles more often. Currently, approximately half the profiles have not been audited in the last 5 years. The utility associated with each profile, particularly the monetary measurement, can help prioritize efforts to audit and update the profile. Another direction to explore is the ability to link donation potential to the *value* of attributes such as *Income* and/or *Occupation*. The value stored can help classify the data for setting up audit priorities, for example, frequently audit and update profile records associated with high values for *Income*. Once an attribute is selected as a classifier, its quality should be maintained at a high level. For example, if *Income* is a good predictor of utility, it should be maintained up-to-date and complete (no missing values). One may also consider refining the granularity of a classifier attribute; currently, *Income* has 3 values, "Low," "Medium," and "High," which might limit its predictive capability. Refining the classification (e.g., to 5 values instead) can increase its predictive power. One could also consider adding a dedicated time-stamp to track changes in this specific attribute (changes currently are tracked at a record level only).

It must be noted that the preceding recommendations do not define a comprehensive solution for prioritizing quality improvement efforts and defining policies for quality management in the alumni database, or even the *Profiles* dataset. These only serve to demonstrate the methodology and its application and provide a sense of the insights to be gained from such analyses. A complete solution demands analyzing all relevant attributes, evaluating other utility measurements, using statistical tools to estimate future benefits and examining all different usages of this dataset.

5. CONCLUSION

In this study, we propose a novel methodology for the dual assessment of data quality, and demonstrate its application using large data samples from a real-world system for managing alumni relations. We show that this methodology offers an in-depth analysis of the current state of data quality in this data resource and underscores possible directions for improving it. The methodology adopts existing methods for impartial quality assessment. Impartial assessment reflects the presence of defects in a dataset and we suggest that it provides an important input for estimating the *cost* of quality improvement. We also suggest that impartial quality assessments can be complemented by contextual assessments, and that some important insights can be gained from analyzing and comparing both. The methodology that we propose incorporates a novel method for contextual assessment based on attributing utility to dataset records. Such a contextual assessment reflects the impact of defects on the usability of a dataset and emphasizes the potential additional benefits to be gained by reducing the number of defects. The application of both assessments and a comparative analysis of the two point out the strengths and weaknesses of current data quality management practices in the alumni system. Further,

we show that such a dual assessment can help improve these practices and develop economically efficient policies.

Our study has some scope limitations which should be addressed in future research. It evaluates quality defects in a single tabular dataset, while data management environments often include multiple datasets and use nontabular data structures. Further, datasets used in real-world business contexts are often much larger, particularly in data warehouse environments, hence, detecting defects and quantifying their impact can be challenging. In our study, we addressed the size challenge by taking a considerably large data sample, which permitted detecting quality defects and estimating their impact in a manner that would be sufficient for our purposes. In general, for large datasets, we do suggest adopting statistical sampling methods (such as those described in Morey [1982]), for estimating the presence and the impact of defects.

The results emphasize the importance of assessing data utility. This is one of the issues that we are currently examining and we intend to focus on this issue in the next phase of this research. Our study shows that different elements (records and attributes) in a dataset may vary in their contribution to utility. In certain cases, a small subset of these elements may account for a large proportion of utility, while in other cases the utility is distributed more evenly. In a follow-up study, we will demonstrate how modeling the distribution of utility and detecting inequalities in utility contribution can improve quality management and prioritize improvement efforts.

Utility measurements used here—Recency, Frequency, Monetary—are specific to CRM. Applying our methodology in other business domains (e.g., finance or healthcare) may require fundamentally different methods for conceptualizing and measuring utility. Our research efforts are directed at identifying metrics and methods for evaluating utility in large datasets. Since utility is context sensitive, we are examining different contexts and attempting to identify assessment techniques for assessing utility in each of these different contexts. We hope that the analyses will yield insights on how to generalize the evaluation of utility in data environments.

Further, the study evaluates utility for known usages. In many business settings, it is important to consider potential usages and associated utility predictions and develop quantitative tools to estimate them. In the alumni dataset, proxies for utility that are based on future donations were not readily available (the work for deriving such estimates is in progress). Developing estimates for potential utility (i.e., predicting future utility) is important when evaluating the quality of a new and unused data source, or enhancing an existing data source with additional records and attributes.

In this study we were able to identify and use a proxy to measure utility, the gifts, or donations received from donors. As shown in this article, this utility measure proved a powerful differentiator of the records and attributes in terms of utility contribution and allowed us to develop contextual assessments of data quality. In other real-world datasets and/or in other evaluation contexts, the identification of the right proxy for utility can be much more challenging. It is reasonable to assume that in many contexts, different proxies for utility can be considered for the same dataset. This is particularly true when

the dataset is used for multiple usages (applications) that are fundamentally different. How would one choose the right utility proxy when a few are available? Our objective for the methodology described in this study is to assess utility distribution in the dataset for better managing data quality. We would hence suggest that the best proxy is the one that best differentiates the records in that dataset, within the evaluated usage context. If different proxies exist, one would have to determine how sensitive the following are to the choice of the proxy for utility and its assessment: (a) the distribution of utility and the differentiation of records based on utility, (b) assessments of contextual quality along different quality dimensions using utility as weights, and (c) the differences between impartial and contextual assessments. If the sensitivity is low, then the proxy that is most easily assessed should be used. A lot remains to be done before we can define prescriptive guidelines for assessing utility and estimating its distribution among records and attributes in large datasets.

Finally, our evaluation highlights causality in the relationships between utility and quality. Common perceptions see quality as antecedent to utility; reducing defect rate and improving quality level increases the usability of data and hence, the utility gained. Our results suggest that in certain settings a reverse causality may exist; frequent usage and high utility encourage improvements in the quality of certain data elements, while the quality of elements that are not frequently used (e.g., profile records of donors that have not donated in a long period) is likely to degrade. This mutual dependency may have positive implications (e.g., cost-effective data quality management, as improvement efforts focus on data items that contribute higher utility) as well as negative (e.g., usage stagnation, a failure to realize the utility potential of less-frequently used items due to degradation in quality). We believe that it is important to explore and understand such causalities as they may have key implications for data quality management.

REFERENCES

Ahituv, N. 1980. A systematic approach towards assessing the value of information system. *MIS Quart. 4*, 4, 61–75.

Ballou, D. P. and Pazer, H. L. 1985. Modeling data and process quality in multi-input, multi-output information systems. *Manag. Sci. 31*, 2, 150–163.

Ballou, D. P. and Pazer, H. L. 1995. Designing information systems to optimize the accuracy-timeliness trade-off. *Inform. Syst. Res. 6*, 1, 51–72.

Ballou, D. P. and Pazer, H. L. 2003. Modeling completeness versus consistency trade-offs in information decision systems. *IEEE Trans. Knowl. Data Engin. 15*, 1, 240–243.

Ballou, D. P., Wang R., Pazer H., and Tayi G. K. 1998. Modeling information manufacturing systems to determine information product quality. *Manag. Sci. 44*, 4, 462–484.

Berger, P. D. and Nasr, N. I. 1998. Customer lifetime value: Marketing models and applications. *J. Interact. Market. 12*, 1, 17–30.

Berger, P. D., Eechambadi, M., Lehmann, G. D., Rizley, R., and Venkatesan, R. 2006. From customer lifetime value to shareholder value: Theory, empirical evidence, and issues for further research. *J. Serv. Res. 9*, 2, 87–94.

Blattburg, R. C. and Deighton, J. 1996. Manage marketing By the customer equity test. *Harvard Bus. Rev. 74*, 4, 136–144.

Chengalur, I. N., Ballou D. P., and Pazer, H. L. 1992. Dynamically determined optimal inspection strategies for serial production processes. *Int. J. Prod. Res. 30*, 1, 169–187.

Chengalur-Smith, I., Ballou, D. P., and Pazer, H. L. 1999. The impact of data quality information on decision making: An exploratory study. *IEEE Trans. Knowl. Data Engin. 11*, 6, 853–864.

Coutheoux, R. J. 2003. Marketing data analysis and data quality management. *J. Target. Measur. Anal. Market. 11*, 4, 299–313.

Davenport, T. H., 2006. Competing on analytics. *Harvard Bus. Rev. 84*, 11, 99–107.

DeLone, W. and McLean, E. 1992. Information systems success: The quest for the dependent variable. *Inform. Syst. Res. 3*, 1, 60–95.

Even, A. and Shankaranarayanan, G. 2007. Assessing data quality: A value-driven approach. *Database Adv. Inform. Syst. 38*, 2, 76–93.

Even, A., Shankaranarayanan, G., and Berger, P. D. 2007. Economics-driven data management: An application to the design of tabular datasets. *IEEE Trans. Knowl. Data Engin. 19*, 6, 818–831.

Fisher, C. W., Chengalur-Smith I., and Ballou, D. P. 2003. The impact of experience and time on the use of data quality information in decision making. *Inform. Syst. Res. 14*, 2, 170–188.

Gattiker, T. F. and Goodhue, D. L. 2004. Understanding the local-level costs and benefits of ERP through organizational information processing theory. *Inform. Manag. 41*, 431–443.

Heinrich, B., Kaiser, M., and Klier, M. 2007. How to measure data quality? - A metric based approach. In *Proceedings of the 28th International Conference on Information Systems (ICIS'07)*.

Herrmann, A., Huber, A., and Braunstein, C. 2000. Market-driven product and service design: Bridging the gap between customer needs, quality management, and customer satisfaction. *J. Prod. Econom. 66*, 77–96.

Jarke, M., Lanzerini, M., Vassiliou, Y., and Vassiliadis, P. 2002. *Fundamentals of Data Warehouses*. Springer.

Khalil, O. E. M. and Harcar, T. D. 1999. Relationship marketing and data quality management. *S.A.M. Adv. Manag. J. 64*, 2, 26–33.

Kimball, R., Reeves L., Ross M., and Thornthwaite, W. 2000. *The Data Warehouse Lifecycle Toolkit*. Wiley Computer Publishing, New York.

Klein, B. D., Goodhue, D. L., and Davis, G. B. 1997. Can humans detect errors in data? Impact of base rates incentives and goals. *MIS Quart. 21*, 2, 169–194.

Lee, Y. W., Pipino, L., Strong, D. M., and Wang, R. Y. 2004. Process-embedded data integrity. *J. Database Manag. 15*, 1, 87–103.

Madnick, S., Wang, R. Y., and Xian, X. 2003. The design and implementation of a corporate householding knowledge processor to improve data quality. *J. Manag. Inform. Syst. 20*, 3, 41–69.

Morey, R. 1982. Estimating and improving the quality of information in the MIS. *Comm. ACM 25*, 5, 337–342.

Petrison, L. A., Blattberg, R. C., and Wang, P. 1997. Database marketing: Past, present, and future. *J. Direct Market. 11*, 4, 109–125.

Pipino, L. L, Yang, W. L., and Wang, R. Y. 2002. Data quality assessment. *Comm. ACM 45*, 4, 211–218.

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

Roberts, M. L. and Berger, P. D. 1999. *Direct Marketing Management*. Prentice-Hall, Englewood, NJ.

Shankaranarayanan, G., Ziad, M., and Wang, R. Y. 2003. Managing data quality in dynamic decision making environments: An information product approach. *J. Database Manag. 14*, 4, 14–32.

Shankaranarayanan, G. and Even, A. 2004. Managing metadata in data warehouses: Pitfalls and possibilities. *Comm. AIS 14*, 247–274.

Shankaranarayanan, G. and Cai, Y. 2006. Supporting data quality management in decision making. *Decis. Support Syst. 42*, 1, 302–317.

Shankaranarayanan, G., Watts, S., and Even, A. 2006. The role of process metadata and data quality perceptions in decision making and empirical framework and investigation. *J. Inform. Technol. Manag. 17*, 1, 50–67.

Shapiro, C. and Varian, H. R. 1999. *Information Rules*. Harvard Business School Press, Cambridge, MA.

Tayi, G. K. and Ballou, D. P. 1988. An integrated production-inventory model with reprocessing and inspection. *Int. J. Prod. Res. 26*, 8, 1299–1315.

Wang, R. Y. and Strong, D. M. 1996. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst. 12*, 4, 5–34.

Wang, R. Y. 1998. A product perspective on total quality management. *Comm. ACM 41*, 2, 58–65.

Wang, R. Y., Storey, V., and Firth, C. 1995. A framework for analysis of data quality research. *IEEE Trans. Knowl. Data Engin. 7*, 4, 623–640

West, L. A. Jr. 2000. Private markets for public goods: Pricing strategies of online database vendors. *J. Manag. Inform. Syst. 17*, 1, 59–84.

Wixom, B. H. and Watson, H. J. 2001. An empirical investigation of the factors affecting data warehousing success. *MIS Quart. 25*, 1, 17–41.