NISS

Data Quality: A Statistical Perspective

Alan F. Karr, Ashish P. Sanil and David L. Banks

Technical Report Number 151
March 2005

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Data Quality: A Statistical Perspective

Alan F. Karr and Ashish P. Sanil
National Institute of Statistical Sciences
{karr, ashish}@niss.org

David L. Banks
Duke University
banks@stat.duke.edu

Abstract

We present the old-but–new problem of data quality from a statistical perspective, in part with the goal of attracting more statisticians, especially academics, to become engaged in research on a rich set of exciting challenges. The data quality landscape is described, and its research foundations in computer science, total quality management and statistics are reviewed. Two case studies based on an EDA approach to data quality are used to motivate a set of research challenges for statistics that span theory, methodology and software tools.

1 Introduction

Data quality is an old problem that has acquired urgent new dimensions. Once it was largely a scientific issue, with roots in measurement error and survey uncertainty. But for today's world of massive electronic data sets and difficult policy decisions, data quality (DQ) problems can create significant economic and political inefficiencies. Modern research on DQ improvement draws upon multiple disciplines and presents a rich set of scientific, technological, and process control challenges to statisticians.

DQ merits more attention from the statistical community, especially among academics. Faculty engagement in the problem has been virtually nil. At a National Institute of Statistical Sciences (NISS)-sponsored session on DQ at the 2002 Joint Statistics Meetings, only two of more than 70 attendees were from universities, and both of them had prior connection with NISS.

This paper is meant to help stimulate academic attention, in part by framing modern DQ from a statistical perspective, which becomes increasing quantitative as the paper progresses. In §2, we present a conceptual overview of DQ, as well a view of DQ as a multi-disciplinary problem that merges ideas from computer science, quality control, human factors research, and the statistical sciences. In any application, of course, these all linked by domain knowledge to the specific context of the problem. §3 addresses measurement and assessment of DQ, but still largely from a conceptual viewpoint.

In §4 we focus on three aspects of the DQ process: preliminary screening for DQ, the role of exploratory data analysis (EDA), and characterization and improvement of DQ. All of these, but especially EDA, are illustrated in the case studies presented in §5. In particular, we show how statistical visualization provides powerful tools for DQ improvement.

Finally, we outline in §6 DQ research challenges for statisticians that range from theory and methodology to new software tools.

2 What is Data Quality?

The organizing principle underlying our views about DQ is that DQ problems and actions are *driven by decisions based on the data*. In some contexts, for example, whether to undertake military action on the basis of intelligence reports or whether to cease marketing a drug because of reported side effects, both the nature and the import of the decisions are clear. In others, such as much “basic” scientific research, the decisions and consequences are more nebulous.

This organizing principle does, however, have tangible implications. In particular, “good” or “improved” DQ is often not an end in itself, because it is the quality of the decisions, not the data, that ultimately matters. For instance, as discussed further in §6, data clean-up efforts—especially if they are costly—must be justified on the basis of better decisions.

Like any principle, this one breaks when taken to extremes. Cleaning up a customer database to remove duplicates, for example, has economic benefits even in the absence of links to explicit decisions. And good scientific data are inherently valuable.

2.1 Overview

We begin with a definition that supports our position that DQ should always be embedded in a decision-theoretic context:

> *Data quality is the capability of data to be used effectively, economically and rapidly to inform and evaluate decisions*. Necessarily, DQ is multi–dimensional, going beyond record-level accuracy to include such factors as accessibility, relevance, timeliness, metadata, documentation, user capabilities and expectations, cost and context-specific domain knowledge.

DQ concerns are problems of large-scale machine and human generation of data, the assembly of large data sets, data anomalies, and organizational influences on data characteristics such as accuracy, timeliness and cost. The impact of poor DQ and the potential benefit of good DQ have implications beyond the ambit of standard statistical analyses.

DQ has dramatic implications. Some people blame the U.S. government’s failure to avert the terrorist attacks of September 11, 2001 on DQ problems that prevented the easy availability of prompt, accurate, and relevant information from key federal databases. In a different context, the Fatality Analysis and Reporting System discussed in §5 may not be of sufficient quality—vehicle make-model data were rife with errors—to support rapid identification of such problems as those associated with Firestone tires on Ford Explorers. At a more mundane level, nearly every major company loses significant income because of data errors—they send multiple mailings to a single person, mishandle claims, disaffect customers, suffer inventory shortfall, or simply spend too much on corrective data processing.

The impetus for improved DQ is especially strong in federal agencies. Managers there are struggling with declining survey response rates and consequent diminished quality. Simultaneously, Congress and the executive branch are using the Government Performance Results Act (GPRA) to require clear linkage between regulation and measurable outcomes. This confluence of a decreasing ability to obtain accurate measurements and increasing management accountability for achieving data-determined goals has compelled federal managers to address DQ more directly than ever before.

Some federal organizations are trying to formalize aspects of the DQ process. For example, the Bureau of Transportation Statistics (BTS) has tried to evaluate databases using a data quality report card (DQRC), on which a database receives a rating of 2 (good performance), 1, or 0 (substantial failure) on eight criteria that span many of the DQ dimensions discussed in §3.

Although the scientific community has developed improved measuring devices, such as automatic sensors that create computer records directly, DQ remains an issue in this setting. Indeed, technological advances may have created a false sense of security. Nonetheless, DQ in scientific research is a different problem from DQ in government and industry.

Industries stand between the Federal government and academic research in terms of DQ. While the quality of their scientific measurements may be generally good, significant problems arise in inventory management, customer service, billing, and marketing databases. Medical data are a particularly thorny problem: they are copious, complex, hard to verify, and entered by many uncoordinated hands. Nonetheless, many problems in industry data have bounded impact (Example: billing errors affect one customer at a time), and there are feedback mechanisms that can support correction (Example: customers report overcharges and incorrect addresses).

Even though the goal is remote, we believe that ultimately DQ must be given a decision-theoretic formulation (see §6.1), allowing use of tools based on statistical decision theory. Decisions drive both the generic need and context-specific requirements for DQ. Decisions are of two kinds: those *based on the data* (Example: government policies) and those *about the data*. Decisions of both types depend on cost, but despite its importance this has not yet stimulated significant research into the economics of DQ. For example, it would be good to know:

- What cost is needed to achieve a specified level of DQ;
- What are the financial benefits of improved DQ; and
- What are the costs of poor DQ.

To complicate the problem, note that costs may be shifted among multiple stakeholders (Example: more burden on survey respondents can reduce clean-up cost of survey data).

Human factors are a challenging part of DQ. People are the key links in many data generation processes, and the ultimate customers,even if decisions are made on their behalf by another entity. The case studies in §5 illustrate this.

Domain knowledge is central to DQ. Data can be useless for one purpose but adequate for others, and domain knowledge is necessary to distinguish these situations. Anomalies in data, a central theme of §5.2, make no sense in the absence of domain expertise. Possibly the most difficult challenge in creating the data quality toolkit (DQTK) outlined in §6.3 is to automate generic aspects for characterizing and measuring DQ but to build in the right "hooks" for domain knowledge.

2.2 DQ as a Multi-Disciplinary Problem

Besides problem-specific domain knowledge, DQ rests on three disciplines, each of which brings its own perspective. We discuss them in order of increasing familiarity to statisticians.

2.2.1 Computer Science

Ever since organizations started collecting and storing their data electronically, information technology (IT) departments have served as custodians of the data. Consequently, computer scientists and information technologists have had considerable exposure to some DQ issues, and they have developed the most mature set of technologies to deal with them—from an IT/database perspective. We present a brief sketch of the computer science perspective on DQ.

Database management is concerned with ensuring data correctness at the collection or entry stage. The general approach has been to devise sound database design guidelines and to establish correctness-enforcing application development practices. This thinking is often incorporated into development environments of relational database management systems (RDBMSs).

Data Entry. Structured Query Language (SQL) (Date, 1999; Groff and Weinberg, 1999), the *lingua franca* of the database community, provides a powerful and flexible syntax to create data tables in which data type and other constraints on data attributes (from the metadata specification) are enforced. "Good practice" guidelines exist for designing user-entry forms for data-entry (Kendall and Kendall, 2001). Example: when a person's contact information is being input, the user should be made to select an enumerated-type data element such as State only via a menu of allowable choices (eliminating mis-spellings, spelling variants and other errors).

Database Design. The central question that drives design considerations for relational databases is how the logical database should be divided physically into separate tables or flat files, as in spreadsheets, in order to minimize duplication and updating anomalies without information loss (i.e., one must be able to link data in individual tables in order to answer queries that span the entire logical database). The theory of "database normalization" provides methodology to address the design problem (Date, 1999). This theory defines a set of increasingly complex "normal forms" (first normal form, second normal form, . . .) such that databases in higher normal forms have fewer sources of anomalies. Database designs that satisfy the third normal form requirements are generally considered adequate.

Transactions and Business Rules. Modern database management systems include powerful features for ensuring correct and consistent data. For instance, sophisticated On-Line Transaction Processing (OLTP) systems handle extremely complex transactions and are able to enforce elaborate business rules and operational constraints. Recent versions of SQL implement constructs for incorporating non-trivial rules and constraints (Groff and Weinberg, 1999).

The *data warehousing* process involves assembling data from a variety of sources and consolidating them into a central data store for future analysis or decision support. Even if the data were all of high quality, integration from disparate sources (Example: payroll, manufacturing, and laboratory data), each with their own idiosyncracies and standards, is challenging. The need to incorporate quality checks in the traditional ETL (Extract, Transform, and Load) process is increasingly recognized, and a number of commercial software solutions in the form of add-on modules are emerging.

The DQ issues that the warehousing process attempts to address revolve around identifying data elements that are *invalid* because they violate physical, logical, or metadata-based constraints that can be specified independently of the data actually observed. The EDA-based approach described in §4.2 and used in §5 complements the warehousing process since it is geared toward finding anomalies in possibly valid data.

Standardization and Duplicate Removal. Free format data (Example: addresses) may require complex parsing in order to be put into standardized form. Commercial DQ software tools contain procedures to do this parsing, and also metadata on commonly occurring data types that allow automatic correction or flagging of some errors. Following standardization, suspected duplicate records which may be processed manually or in a semi-automated manner, but the current tools are not very powerful.

Record Matching. Integrating data from multiple sources requires correctly linking each record in one database with the corresponding record in another. In many cases, the shared attributes in the two sources are not exactly the same (Example: one database may contain full names and the other only initials and the last name). Record linkage has received attention from computer scientists as well as statisticians, and a number of methods that span a range of complexity and applicability have been devised (Winkler, 2000a,b). Record matching techniques that lie within the IT domain typically rely on key-generation and string matching. String matching methods of varying levels of complexity determine if string-valued attributes from the two sources are "close enough" to be declared a match. The key-generation approach is based on generating a surrogate key for string-valued data that is less ambiguous than the original, and then using the key for record matching.

2.2.2 Total Quality Management

Statistical quality control has had profound effects on industrial production. Some researchers and practitioners believe that the same ideas and techniques can be applied to DQ. However, inability to model, or even quantify, costs is one reason why the total quality management (TQM) paradigm may be difficult to implement for DQ.

The chain of reasoning is straightforward: Data are a product, with producers and customers; therefore data have both cost and value. In conceptually the same way as physical products, data have quality characteristics resulting from the processes by which they are generated. In principle, DQ can be measured and improved. Finally, the same financial incentives that lead to high quality products also apply to DQ.

One approach (English, 1999) relates DQ to information quality in a TQM setting: "quality in all characteristics of information, such as completeness, accuracy, timeliness, clarity of presentation" should "consistently meet knowledge worker and end-customer expectations." A key tenet, which is admirable if not always attainable, is that DQ should be present from the start, rather than created by re-work. The need to measure the costs and benefits associated with different levels of DQ is critical in this setting

An alternative approach is called Total Data Quality Management (TDQM). It involves concepts such as multi-dimensional data quality, data quality metrics, evaluation of the user's assessment of DQ, and data production maps (Huang et al., 1999; Wang, 1998; Wang et al., 2000).

2.2.3 Statistics

Statisticians have always worked for better DQ. That is why we worry about outliers and long-range dependence and exploratory methods. It is why we study robust statistics, and stress the need to understand the science and to talk to domain experts before undertaking an analysis. An indeed, some statistical contributions to DQ are substantial.

Data Editing. Statistical data editing is the automated process of stepping through the data records and correcting them if they violate pre-specified constraints. See §4.4 for details.

Probabilistic Record Linkage. Probabilistic approaches to the record linkage problem can be traced to Fellegi and Sunter (1969) and Newcombe and Kennedy (1962). To link records in file **A** with those in file **B**, the Fellegi–Sunter framework provides a method for evaluating the likelihood that pairs of records in **A** × **B** are matches, as well as a corresponding optimal linkage rule for specified Type I and Type II errors. Extensions of the Fellegi–Sunter methods are still being developed (Winkler, 1994, 2000a).

Measurement Error and Survey Methodology. Statisticians involved with survey data have devoted enormous attention to DQ (Groves, 1987). In this setting, DQ effects arise from modes of collection (de Leeuw and van der Zouwen, 1988), interviewers (Groves et al., 1981) and survey design (Tucker, 1992). DQ concerns for surveys have focused primarily on non-response problems (Hidiroglou et al., 1993), coverage biases, (Duncan and Stasny, 2001) and measurement error (Biemer et al., 1991).

2.3 DQ and Other Problems

The relationship of DQ to other problems provides a way to leverage research on DQ, and raises intriguing research challenges.

For example, software quality bears a strong resemblance to DQ. Both "products" are electronic rather than physical, so that quality is a characteristic of a class rather than instances. For both, quality is highly situational: in the same way that a database may be adequate for one purpose but not another, the same software may serve adequately in one setting, but not in another. Also for both, humans (data generators and software developers) are central to the production process as well as an essential source of variability. The protracted (and still incomplete) effort to produce metrics and standards for software quality (Kan, 1994) may provide insight into DQ metrics.

Data confidentiality, a long-standing problem for government agencies, is burgeoning in the world of E-commerce and has dramatic implications for homeland security. The relationship between data confidentiality (DC) and DQ is complementary: the same tools, such as those for record linkage, that increase DQ threaten DC. Conversely, tools that alter data but allow informative inference (Example: swapping some attributes between records (Sanil et al., 2003)), or that maximize data utility subject to constraints on disclosure risk (National Institute of Statistical Sciences, 2004), suggest ways to characterize how much information can be extracted from low quality data.

3 Measurement and Assessment of DQ

As noted in §1, to become a science DQ must have a foundation built on measurement. This section focuses mainly on a conceptual scheme for what *should* be measured, with less attention to *how*.

For the purposes of this discussion, we assume that the data consist of, or are derived from, *records* with several *attributes*. We assume that attributes are type-structured (Examples: numerical values, categorical responses, dates, times). Existing abstractions and tools for DQ seem to be handle to handle text data only in highly structured cases such as addresses.

A *table* is a collection of records with the same attributes, and a *database* consists of one or more related tables (Example: the Fatality Analysis Reporting System (FARS), which consists of four tables containing data keyed by accidents, drivers, persons and vehicles). Whether a data are stored in a RBDMS is not material at this point.

3.1 Dimensions of DQ

Measurement and assessment of DQ are based on a set of conceptual dimensions of DQ. We distill the DQ literature into a number of dimensions grouped into three hyperdimensions:

Process: Dimensions of DQ related to the generation, assembly, description and maintenance of data—Reliability (with several subdimensions), Metadata, Security and Confidentiality. With very few exceptions, these dimensions can be assessed only qualitatively, and in most cases highly subjectively. As discussed in §3.2, such assessments also require contextual knowledge, and they frequently are made in the absence of detailed—or possibly any—examination of the data themselves.

Data: Dimensions of DQ specifically associated with the data themselves. At the record/table level, these comprise Accuracy, Completeness, Consistency and Validity. The database level dimensions are Identifiability and Joinability. Only the Data hyperdimension seems meaningfully amenable to quantitative measurement.

User: Dimensions of DQ related to use and users—Accessibility, Integrability, Interpretability, Rectifiability, Relevance and Timeliness. Like the Process hyperdimension, User seems susceptible only to qualitative assessment.

None of these is new here. Many appear in sources such as Brackstone (1999) and Wang et al. (2000) and Federal agencies DQ and information quality (IQ) guidelines (Bureau of Economic Analysis, 2002; Bureau of Justice Statistics, 2002; Bureau of Labor Statistics, 2002b,a; Census Bureau, 2002; Department of Education, 2002; Energy Information Administration, 2002; National Agricultural Statistics Service, 2002; National Center for Health Statistics, 2002; Office of Management and Budget, 2002b).

The Office of Management and Budget (OMB) Guidelines for IQ (Office of Management and Budget, 2002a), which address conclusions drawn from data as well as the data themselves, employ related hyperdimensions of:

Objectivity: "whether disseminated information is accurate, reliable, and unbiased" in terms of both substance and presentation.

Utility: "usefulness of the information for the intended audience's anticipated purposes."

Integrity: "protection of information from unauthorized, unanticipated or unintentional [...] falsification or corruption."

We mention the OMB guidelines since many agencies' guidelines seem to be derived from them.

Specifying DQ dimensions: (1) Facilitates assessment and measurement; (2) Provides a framework for stipulating guidelines and DQ improvement plans, and (3) Provides a path for actions to improve DQ. Our dimensions, since they closely follow the "information chain" (Loshin, 2001), are better suited for (1) and (3). Using our dimensions, DQ may be either assessed and measured either qualitatively or quantitatively. Qualitative assessments are nearly always categorical (Examples: ratings from 1 to 10), often coarse (Example: categories are "Good," "Fair" and "Poor") and commonly subjective. Qualitative assessments serve as an alternative to (currently, weak) quantitative assessments, fulfill administrative requirements, and reflect simple common sense.

Quantitative measurements are made by means of *DQ metrics*. Ideally, measurements would enable comparison of DQ across databases and quantify improvements (or decreases) over time in DQ, as well as support prediction of the effects of DQ improvement strategies. Different DQ metrics apply at different scales—individual records, databases and integrated databases. Some other useful characteristics of DQ metrics are articulated in Nousak and Phelps (2002): they should not depend on the size of the database; metrics should reflect users' needs for the data; different metrics should measure different DQ dimensions; and the number of metrics employed should not be too large. How any of these can be assured is not clear.

To date, approaches have been substantially *ad hoc*. Moreover, specifics about calculation of any implied metrics seem to be documented only rarely. Important issues are often downplayed or ignored: for example, accuracy can be computed only by means of exogenous data, which may have their own DQ problems.

One clear gap is the absence of metrics related to *inferences* based on the data. Although to some degree the question is unanswerable, at least absent comparable data without DQ problems, one ought to ask: To what extent are conclusions drawn from the data affected by their quality? For example, when a model is fit to the data, how does DQ change the model, which can be measured in various ways. Additional discussion appears in §6.

DQ metrics remain speculative. It is not clear that they can be used even to compare one database at two different times (to ascertain, for example, whether efforts to improve DQ succeeded), let alone one database to another. Nor is it clear how to determine what value of a metric is acceptable.

3.2 The Process Hyperdimension

Process dimensions of DQ are, nearly uniformly, conceptualized as application of "best practice" or "state of the art" methods and reasoning from some field (Examples: statistics, computer security). As a result, they are assessed only qualitatively, usually categorically, and subjectively, often by means of DQ or other checklists. Rarely is the assessment finer-grained than ternary (Example: good, poor, unacceptable) and often it is binary (Example: was X done or not?).

Quantitative assessment of Process dimensions of DQ is beyond the current state of the art. This is yet another reason why the potential for TQM to affect DQ at an operational level seems limited. TQM works for physical production because *process control* works: (1) Processes can be measured and monitored; and (2) There is scientific and empirical evidence that if the process is "in control," then product quality necessarily will be acceptable. Neither of these is currently true with respect to DQ.

Reliability. Both within and outside the federal statistical agencies, reliability is construed as application of appropriate, justified, state-of-the-art statistical methods. There are three principal subdimensions of reliability, corresponding to generation, editing and analysis, although the latter is more properly associated with IQ than DQ.

To a significant extent, for federal statistical agencies, reliability at the data generation stage means conducting surveys in a manner designed to maximize DQ. Historically, statisticians involved with survey data have devoted an enormous amount of attention to DQ (Groves, 1987). A significant portion of this effort has been devoted to understanding the DQ effects that arise from different modes of data collection (de Leeuw and van der Zouwen, 1988), interviewers (Groves et al., 1981) and survey design (Fowler, 2002). Survey instrument design—question wording, question sequence, response format, cognitive guidelines—is a well-researched area (Converse and Presser, 1986). Selection of the data gathering mode is another established area. This includes guidelines for selection between collection mechanisms such as face-to-face, telephone, mail or more sophisticated ones such as Computer-Aided Telephone Interviews. Strategies for minimizing non-response and the processing of data (editing, coding and entry) have also been researched extensively.

Virtually every federal statistical agency has strong expertise in survey design. The Bureau of Labor Statistics (BLS), Census Bureau (Census) and National Center for Education Statistics (NCES) design very complex surveys. Both BLS and Census have units dedicated to cognitive issues. Private sector organizations such as NORC, RTI International and Westat design and conduct large-scale, complex surveys. Statistical software packages such as R, SAS, S-Plus and SPSS contain tools for survey design. The American Statistical Association Section on Survey Research Methods (American Statistical Association, 2002) provides access to a wealth of information. Both specialized and generic software for survey authoring are available; see SPSS, Inc. (2002) for example.

Characterization of the effects of data collection methods (especially surveys) on DQ also has a lengthy history, sound basis and plethora of techniques. Such effects arise from modes of collection, interviewers and survey design. DQ concerns for surveys have focused primarily on non-response problems (Hidiroglou et al., 1993), coverage bias (Duncan and Stasny, 2001) and measurement error (Biemer et al., 1991).

Statistical and other approaches to editing of data are discussed in §4.4.2. Imputation provides a wide variety of methods to deal with missing data attributes, especially for survey data. At the simplest level, imputation "fills in" missing values, conditional on the values of non-missing attributes, using distributions derived from the same data("hot deck imputation") or other data ("cold deck imputation") (Little, 1992; Little and Rubin, 1987; Rubin, 1987).

In a more general setting of IQ, in which disseminated information includes statistical analyses of data, the analyses themselves are also part of reliability. Examples of strategies include adjustment of uncertainties in statistical estimates to reflect non-response and dissemination of detailed response information with data, in order to allow users to conduct meaningful analyses.

Metadata address the problem of ensuring that the content, collection process, ownership and reliability of data are documented clearly, unambiguously and in a form that is conveniently accessed by users.

The quality and accuracy of data documentation critically impact DQ via data usability. Organizationally, having a group assigned to review metadata can be an effective and relatively inexpensive strategy. This group could also formulate guidelines and checklists to be followed at the data recording time. Information systems analysis and design principles (Kendall and Kendall, 2001) advocate the use of "data dictionaries" as a formal means for recording metadata. At a more mechanical level, recording metadata is facilitated by software tools associated with RDBMSs.

A more sophisticated approach is to define precise standards for data documentation, and then create software for verification, parsing, processing and generation of metadata. The US Geological Survey (USGS) metadata initiative for spatial data (US Geological Survey, 2003) is a striking example of this approach. USGS, recognizing the need for uniform documentation of spatial data required for each of the several uses, such as spatial and non-spatial analyses, mapping, . . ., has devised comprehensive and widely accepted standards. USGS also distributes software tools. The USGS metadata standards are perhaps the only standards that have been widely adopted by a user community, and have taken root to the point that universities devote courses (or portions of courses) to them. By contrast, the extremely detailed Census Metadata Repository (G. J. Lestina, Jr. et al., 1997) seems to be used essentially only internally.

The recent development and widespread adoption of the extensible markup language (XML) standard is likely to provide a powerful tool for developing metadata standards. XML (World Wide Web Consortium, 2004) uses a schema or a Document Type Definition (DTD) that enables one to describe a document or data set in a precise, structured manner. Representing data by means of XML allows one to bundle the metadata with the data such that the data values can be readily accessed and processed (for data integration, for example) by XML-aware software. The "Dublin Core Metadata Initiative" (DCI) (Dublin Core Metadata Initiative, 2003) is an open forum of various parties (such as government agencies, library scientists, information networkers) engaged in the development of "interoperable online metadata standards that support a broad range of purposes and business models." This involves setting up XML-based frameworks, the necessary ontologies and software tools. The DCI is currently in its infancy, but holds great promise for the future.

An indirect but key reason for structured and precise metadata is that they are absolutely essential for successful consolidation and integration of data from various sources (§3.4).

Security. Physical and electronic security lie outside the scope of this paper. We believe that existing hardware (Example: firewalls) and software tools (Examples: RDBMSs that limit database access and intrusion detection systems), used properly, are able to cope with the onslaught of attacks.

Confidentiality. Protection of confidentiality by statistical agencies is mandated by regulation, so there is a clear trade-off between privacy protection and confidentiality on the one hand and user access to high-quality statistical data on the other (Duncan et al., 1993). At the simplest level, poor DQ may actually improve confidentiality: for example, adding noise to data (either bias or increased variability) reduces the probability of re-identification (Fienberg et al., 1997). To some, the error rate of approximately 10% in the 1990 US Decennial Census gave better protection to Census Bureau releases than data swapping alone would have provided (Anderson and Fienberg, 2001b).

Conversely, improved DQ threatens DC: decreased error rates of approximately 6% in the 2000 Decennial Census add concerns to the Bureau's re-examination of its disclosure limitation strategy (Anderson and Fienberg, 2001a; Census Bureau Executive Steering Committee for Accuracy and Coverage Evaluation Policy, 2001). Similarly, tools to improve DQ, such as record linkage (§2.2.3), are powerful means of breaking confidentiality.

There are a voluminous literature and active statistical disclosure limitation (SDL) research community addressing the problem of ensuring that disseminated data and information do not violate confidentiality, in which the NISS plays a leading role (National Institute of Statistical Sciences, 2004). For example, theory and methodology have been developed, and software systems built, that maximize the utility (read "quality") of disseminated data subject to risk constraints (Gomatam et al., 2003; Dobra et al., 2002, 2003; Karr et al., 2001a, 2003; Sanil et al., 2003) or characterizing the effects of SDL on released data (Gomatam et al., 2003, 2004; Lee et al., 2001). Although these methods and techniques have not yet been applied to assess or measure DQ, there is significant opportunity to do so.

Much less understood is the "front end" interaction between confidentiality and DQ, in which more protection of confidentiality improves DQ: data subjects who fear that their confidentiality is not protected adequately may provide false information. Beyond awareness of the problem, little has been done regarding this important question.

3.3 The Data Hyperdimension

3.3.1 Record/Table Level

Data dimensions are largely binary at the record level (Example: an attribute value is either correct or not) and summarized at the table level by DQ metrics such as the fraction of data records that are correct.

Accuracy. The literal notion is that a data record is "correct" or not. In some cases this is unambiguous: a person's age or income either is or is not reported correctly. In others, such as opinion polls or surveys about behavior in hypothetical circumstances ("What would you do if . . .") correctness seems problematic even conceptually.

Much more important, however, is *determination* of correctness, which ranges from difficult to impossible. Resort to exogenous data (possibly with their own DQ problems) or repeating the data collection may be feasible in principle, but are often ruled out by cost considerations. Whether it is cost-effective even to measure correctness has never been addressed. The structure outlined in §6.2 provides one path to treat such questions.

If correctness could be determined at the record level, it would be measured at the table level, for example as the fraction of records in which *all* attributes are accurate. An alternative metric, which we feel is less insightful, is the fraction, relative to records × attributes, of attribute values that are accurate (Wang et al., 2000). Comparison across either time (improvement or degeneration) or databases is possible. What constitutes acceptable accuracy is not clear.

A common concept of accuracy in agency guidelines is statistical correctness of the data—how well do the data measure what they purport to measure. It is usually characterized by statistical errors associated with the data such as coverage biases, sampling defects, and nonresponse.

Completeness. This is one dimension that seems essentially unambiguous: each record is either complete—with no missing values—or not. However, problems can arise, for example with confounding between missing and legitimate data values. Two metrics are possible: the fraction of records with *no* missing attribute values, which we favor, and the fraction of attributes fields containing values (Wang et al., 2000).

Consistency is an analogue of validity (see below) pertaining to *intra-record relationships* among attribute values rather than individual attribute values. Examples range from "hard constraints" (in the FARS, the accident time must precede the EMS arrival time) to "plausible relationships" such as that between driver height and weight (see Figure 6). Given consistency criteria, each record is consistent or not, and the resultant DQ metric is the fraction of records that are consistent. There is also a notion of consistency across tables, discussed below. Strategies for dealing with some consistency problems are described in §4.4.2.

Validity is a weakened but more readily measured form of accuracy. Attribute values may be valid without being correct, but not *vice versa*. Importantly, validity can be determined from the data table itself, without need for external data. An attribute value is defined to be *valid* if it falls in some exogenously defined and domain-knowledge dependent set of values. Validity can range from mechanical (Example: 18/19/2002 is not a well-formed and hence not a valid date) to logical (Example: -5 is not a valid age) to the domain-derived (Example: 1234 pounds is not a valid weight for a person). As always, there are complications: 16:12 may be a valid time in one database but not in another. Two metrics make sense for validity: the fraction of records for which *all* entries are valid and the fraction of attribute values that are valid (relative to records × attributes) (Wang et al., 2000).

Multidimensional DQ Metrics. Several attempts have been made to construct metrics that combine DQ dimensions. Here are representative examples from the disciplines underlying DQ.

From computer science, a conceptual approach to measurement of DQ at the warehousing stage has been proposed using dimensions of timeliness, verity, semiotic, medium and novelty (Helfert, 2001). TQM ideas from Huang et al. (1999) are also present in this approach, including strong reliance on subjective measurement.

TQM ideas may be applied at the data collection phase by monitoring such characteristics as the number of invalid or nonconforming records generated per day (Loshin, 2001). These can be tracked not only over time, but also against geography or the amount of money expended to acquire the data.

Two examples from statistics appear in §6.1.

3.3.2 Database Level

With a database Identifiability and Joinability both address the issue of whether multiple data tables that should be combined actually can be combined. (Example: in the FARS, the accident time in the one table and the driver's age in another, so investigating a possible relationship between the two requires combining the two tables.) Proper database design using modern RDBMSs effectively precludes problems with either Identifiability or Joinability.

Identifiability. Each record in a data table must have a unique identifier, which may be either a single attribute (Example: Social Security number) or a combination of several attributes (Example: in the FARS accident file, the state, case number and vehicle number within case), known as the *primary key*. Potential problems include lack of a primary key and duplicate keys.

Simplistic metrics have been performed (Example: "the percent of records having a unique identifier" (Department of Defense, 2002)), but they miss the point. Identifiability assessment is binary at the table level: either there is or there is not a proper primary key.

Joinability of two data tables requires first that the primary key of one be an attribute in the other (where it is termed a *foreign key*) (Ullman and Widom, 1997). Like Identifiability, this is a binary assessment. In addition, tables that ought to be in one-to-one correspondence should be checked to be so, and the extent to which this is not realized forms a natural DQ metric. To illustrate, attempts to join the driver and vehicle files available on the now defunct Intermodal Transportation Database (ITDB) version of the FARS (Bureau of Transportation Statistics, 2002) for 1999 revealed that despite proper keys, there was *no overlap* between the two files.

3.4 The User Hyperdimension

All of these operate at multiple levels ranging from IT to human-centric, with such items as human-computer interfaces in between. At best, they can be assessed qualitatively and subjectively, with strong input from multiple user communities. Commendably, the federal statistical agencies are acutely aware of the need for strong data customer relations.

Accessibility. At the IT level, accessibility is physical and structural. Are dissemination systems accessible? Can servers support the traffic? Are file formats modern? Legacy database management systems that use outmoded software (Example: databases with fixed width attributes still exist) and run on old-fashioned platforms may be useless. Even systems with modern software may suffer from poorly designed interfaces.

The human-centric aspect of accessibility raises concerns such as diverse user communities (technical, non-technical, handicapped, . . .), but these are issues of IQ rather than DQ.

Integrability. Integration of multiple databases is increasingly vital and almost invariably difficult. Issues range from differing attribute definitions (Example: death from an automobile accident must occur within 30 days, but from a train wreck death can occur up to a year later) to inability to join tables in different databases because they use different keys. (Example: it is not possible to join data tables from the FARS and the National Transit Database (NTD) (Federal Transit Administration, 2002), which would be needed to relate accidents involving transit vehicles to other transit system characteristics.) Tools for integration, especially RDBMSs and record linkage, are rather well-developed. The impediments to integrability are often in the databases themselves, at the conceptual and design levels.

Interpretability has multiple components. One, which bridges DQ and IQ, is consistency: if data are reported in multiple ways, or if interacting databases use different definitions or if attribute definitions change over time, then the value of the data to users is reduced.

Another aspect is the ability to extract information from data, which is essential for decisions. Statisticians construe this as “inference,” but a broader perspective is necessary. One strategy is to develop inferential tools that are resistant to poor quality (§6.1). Adequate metadata and documentation are also necessary.

Rectifiability. The establishment of procedures for users as well as affected parties to request correction of information made public by statistical agencies is critical to usability, correctness and overall quality of the data dissemination process. The OMB guidelines emphasize this, and almost all agencies’ IQ guidelines provide details on how they implement correction processes. Some agencies guarantee a correction/appeals process and provide unambiguous information on who should be contacted.

In many commercial settings (Examples: customer service, billing, and marketing databases), there is near-total reliance on customer feedback to correct data errors (Example: customers report overcharges and wrong addresses). Other than by application of the powerful consistency checks associated with double-entry bookkeeping, most accounting errors are corrected as the result of feedback from those affected by them.

Federal agencies are at a clear disadvantage in regard to Rectifiability. Few people or organizations wish to or even can access federal agency data about them; indeed, confidentiality protections may prevent data providers from recognizing even themselves.

Relevance. This dimension seems as elusive as it is important. At the highest level, relevance addresses whether the data in the database are the data that users want, as well, perhaps, only the data users want. The “omission” component of relevance can be assessed only through feedback from users, while the “commission” aspect (the database contains irrelevant data) might be assessed by checking that each item captured is needed by some clearly-defined class of customers.

Inertia in data collection processes works against the need for relevant data. As databases age, they often serve different (or more) purposes from those originally intended, or the problem that drove the initial use may change.

Timeliness. Outdated data can be worse than no data. For example, they may create false certitude, or may mask problems and delay necessary action (Example: slow availability of data hampered rapid identification of such problems as those associated with Firestone tires on Ford Explorers, although in this case there were also substantial DQ problems). Proposals to quantify timeliness exist (Example: “percent of data available within specified time” (Department of Defense, 2002; Wang et al., 2000)), but have never been operationalized usefully. Nor might quantification even be possible: to a significant degree, timeliness is domain-specific and user-centric, and is therefore categorical and subjective.

3.5 Organizational Issues

The overall commitment of an organization to DQ cuts across all of the dimensions in §3.2–3.4. TQM advocates of DQ (English, 1999; Huang et al., 1999; Wang et al., 2000) are relentless about this.

Specific manifestations of such commitment range from allocation of adequate human and financial resources to DQ to review processes to a customer-centric viewpoint of DQ, especially the User hyperdimension (Examples: data dissemination in multiple formats, help desks).

As with checklists, some aspects of organizational commitment seem more effective at calling attention to issues than at resolving those issues. For example, Accuracy could be assessed (to some degree) by taking a random sample of the records and independently verifying their correctness. Doing so however, requires that such independent checks be available and affordable, which we believe is often simply not the case.

Other than isolated examples, we are aware of no attempts to document these issues, let alone assess or measure their importance.

4 Components of the DQ Process

In this section we highlight four components of the DQ process: preliminary screening for DQ (§4.1), the use of EDA for assessing DQ (§4.2), identification of anomalous data elements (§4.3) and selected methods for improvement of DQ (§4.4).

4.1 Preliminary Screening for DQ

One central application of DQ measurements and assessments is preliminary screening for DQ—an initial determination of the quality of a database

Consistent with §2 and 6.2, we formulate preliminary screening as a decision problem leading to one of four outcomes ordered by increasing desirability:

Discard the Data (or do not disseminate or use them), on the grounds that DQ is so poor that no sensible use of them is possible.

Characterize but Do Not (Seek to) Improve DQ, which applies when DQ is not high but the means or resources to improve it are not available or not cost effective. For example, the Toxic Release Inventory (TRI) data discussed in §5.2 have evident DQ problems, but there is no clear path to resolving them.

Improve DQ, which is feasible when techniques such as those in §4.4 are applicable and commitment of necessary resources (personnel, money, time) can be justified. Of course such improved DQ must then also be characterized.

Improvement Not Required, the best but often the least likely outcome. In this case, detailed characterization of DQ may not be necessary.

Of course, not all outcomes may be feasible in some contexts.

Preliminary screening is especially relevant to organizations with limited control of data generation and assembly processes, especially the Reliability and Metadata dimensions in §3.2. This is because sufficient attention to such issues usually ensures that data will not need to be discarded.

The decision process is sequential. First, the determination is made whether the data are to be discarded. In some cases, this decision can defensibly be made on the basis of metadata and domain knowledge alone, i.e., in effect on the Process hyperdimension in §3.1. (Example: the response rate in a survey may be so low that no reliable inference is possible about the population of interest.) In other cases, however, some examination of the data themselves (Example: exploratory analyses described in §4.2) may be needed.

Second, if data are not discarded, a decision must be made whether only to characterize DQ or to consider expenditure of resources to improve DQ. In some circumstances, basing this decision on the Process hyperdimension alone may be justifiable, but we feel strongly that at least some examination of the data is necessary.[1]

[1] Of course, if improvement is impossible or infeasible, then there is no decision to be made.

The final decision, if improvement is to be considered, of what exactly what improvements are required, cannot be made absent detailed examination of the data.

Many federal statistical agencies employ some form of DQ checklist, perhaps derived in part from the OMB Guidelines, for either preliminary screening or elsewhere in the DQ process. For example, BTS has proposed using a DQRC, on which a database receives a rating of 2 (good performance), 1, or 0 (substantial failure) on eight dimensions that either coincide with ours (Example: whether the database is compatible with other databases in the US Department of Transportation—our Integrability dimension) or address organizational issues discussed in §3.5 (Example: whether a database is subject to periodic review by data managers and suppliers). Reference books on DQ (English, 1999; Loshin, 2001; Wang et al., 2000) often include prototype checklists.

While clearly useful, checklists are inadequate to cope with all DQ issues. In particular, they usually lack guidance for assessment, criteria for evaluation and actionable responses. A harsh but not altogether incorrect summary is that existing DQ checklists are statements of good intentions and admonishments to pay attention to DQ issues. The statistical standards of the NCES (National Center for Education Statistics, 2004) are a notable exception.

4.2 Exploratory Analyses

Tools from EDA (Tukey, 1977) provide effective means for developing a high-level view of DQ from a statistical perspective. In particular, as illustrated in §5, EDA strategies with a strong component of visualization, augmented by consideration of issues involving relational databases, are especially effective for addressing:

Metadata Characteristics, including information about the database, its structure, tables, and attributes, as well as missing and incomplete values.

Characteristics of Individual Attributes, such as the legitimacy and distribution of their values.

Relationships among Attributes, whose existence may be suggestive of either good or poor DQ. This also includes consistency checks between attributes, possibly in different relational tables.

Relational Characteristics, within the database, such as primary and foreign keys and Joinability of tables.

EDA produces useful insights into the data that can improve quality and increase value to the users. It has a history of success and strong proponents (Tukey, 1977) within the statistics community. Critics raise concerns over multiplicity (How many uninteresting features are implicitly ignored in the process of finding one interesting feature?), as well as the need to separate exploration from inference. Neither criticism seems entirely on-target in the context of DQ.

4.3 Anomaly Detection

When a database is analyzed without access to other, related databases, then arguably the only approach to DQ is to detect and characterize of anomalous aspects of the data, at multiple levels—attribute, data record and database.

The rules-based correction strategies and statistical DQ edits described in §4.4.2 implicitly detect anomalies in the process of correction. In many other cases, however, anomalies cannot be fixed, but must be detected and characterized in order to assess their effects on inferences based on the data.

Other anomalies are statistical outliers that can be detected and, to some degree, coped with by means of robust statistical methods; see §6.1. However, real-world data anomalies correspond to error structures that even robust models seem unable to accommodate, so there is need for alternatives. Visualizations (a form of EDA) can be effective in some circumstances, as illustrated in Figures 6 and 9. The former, taken from the FARS driver file, is a scatterplot of driver height and weight, and shows readily a number of anomalous values. Those equal to 999 in fact represent missing data. Others seem simply to be wrong. Figure 9, derived from the TRI Environmental Protection Agency (2002) of the US Environmental Protection Agency (EPA), shows facility-level time series of releases of lead. The anomalies there would not be handled by many time series analyses.

In other cases, the adage LOOK AT THE DATA, especially after the data have been sorted, can be effective. Figure 10, also based on TRI data, shows anomalies similar to those in Figure 9.

Some data mining techniques could be potentially useful for anomaly detection. For example, a low depth CART tree (Breiman et al., 1983) fitted to the data and has leaf nodes with sample size of one is a indication that the leaf node is an outlier. Similarly, singleton clusters produced by clustering algorithms could be indicators of anomalies. Also, anomalies may be defined relative to "patterns" in the data identified by application of data mining tools for discovery.

Applicable visualization tools include Advizor (Visual Insights, Inc., 2002), JMP, S-Plus and R. These tools, used manually and guided by domain knowledge can work on a small scale for low-dimensional data. XGobi (AT&T Labs Research, 2003) has facilities for visualizing higher-dimensional real data, and also some facilities to automate finding interesting features in the data. Some EDA functionality is provided by data warehousing add-ons. There are not, however, scalable implementations—targeted to DQ or not—that work automatically, even to some degree, on large problems.

4.4 DQ Improvement

Here we lay out the state of the art and practice of improving DQ. Appendix A in Karr and Sanil (2003) contains a tabular version of the same information.

4.4.1 Standardization and Duplicate Removal

Placing free format but structured data (Example: addresses) into standardized form is a ubiquitous need. Standardization enhances the usability of the data by making the data less ambiguous. It is a crucial step for data integration. It is also the key to identifying duplicate records—this is of economic significance in Customer Relationship Management (CRM) systems.

Commercial DQ software tools such as Dataflux from SAS (SAS Institute, Inc., 2002), Evoke (Evoke Software, 2002), and Trillium (Trillium Software, 2002) contain procedures to parse data such as dates and addresses, with some capability for metadata-based correction; see §4.4.2. Relatively free-format data are broken down into smaller units to which precise tags can be attached. For example, free-format names that are parsed into prefix, first, middle, last and suffix,in order to execute much more sophisticated searches, queries, and matching algorithms based on the individual components.

Following standardization, refined record matching and data integration procedures may be applied more effectively. Suspected duplicate records can be flagged and then processed manually or in a semi-automated manner.

4.4.2 DQ Edits

The purpose of DQ edits is to identify and, if possible, correct inconsistencies among data attributes or between data values and data definitions.

Different but overlapping approaches arise from computer science and statistics. The computer science approaches are typically embedded in warehousing tools, and are used to some degree because "they are there." Rationales for statistics-based approaches include domain knowledge (of circumstances when they are effective), logic and experience with other, similar databases.

Commercial can automatically correct or flag some errors. Sophisticated string matching algorithms can detect very subtle problems (Example: that the street name `Fxx Craft` in Durham, NC should be `Foxcroft`). The constraint-checking component of the data warehousing process discussed in §2.2.1 verifies that the data domain, type, and numeric range constraints specified in the metadata are satisfied. Some software tools also allow the user to specify domain-knowledge-based constraints.

Statistical data editing is the automated process of stepping through the data records and correcting them if they violate pre-specified constraints known as edit rules. There are many, sometimes *ad hoc*, methods based essentially on sets of if-then-else rules, as in the warehousing approach. A formal framework for data editing (Fellegi and Holt, 1976) allows one to specify edit rules or constraints to identify invalid data records (Examples: lists of impossible attribute combinations for discrete data (Winkler, 1995) and sets of linear inequality constraints on ratios of continuous variables (Greenberg and Petkunas, 1990)). Making the edits involves first generating all the implied edit rules from the explicitly specified rules and then solving optimization problems to determine the minimum-change edit for every record that violates some edit rule. There are formidable computational challenges in implementing an automatic edit system, and the development of efficient algorithms and heuristics remains an active area of research (Winkler and Chen, 2001).

There also exist less formal, strongly domain-knowledge based approaches to DQ editing, whose principal practitioners are federal statistical agencies. For example, a medical data record with `Sex = Male` and `Procedure = Hysterectomy` might routinely be changed to `Sex = Female`. Manual intervention is typically required.

Specialized, non-production software is available (Greenberg and Petkunas, 1990; Winkler, 1995; Winkler and Chen, 2001); transferability to other contexts is uncertain.

5 Case Studies

Two case studies of DQ undertaken by NISS are described here. The first treats a version of the FARS once maintained in the ITDB of the BTS, and the second addresses the TRI maintained by the EPA. The methods used in handling the case studies illustrate the power that modern visualization tools and a statistical perspective offer for finding, measuring, and correcting DQ problems.

Two case studies cannot span modern DQ in its entirety. Examples of DQ problems encountered by NISS in other contexts include inconsistencies between documentation and data files; cumbersome data representations (Example: tables converted to vectors in the NTD Federal Transit Administration (2002)); records in which the attributes in which the cannot be compared (Example: multiple paired attributes of the form "(Mode, Expenditure)," also in the NTD, such that bus expenditures by different organizations may in different attributes); and databases with attributes having identical names and meaning (Example: Number of Students Enrolled in Second Grade in 2000–01) but different values, because they came from different data sources at different times.

Building on §4.2, the case studies employ EDA strategies having a strong component of visualization.

5.1 Case Study 1: The ITDB Version of FARS

Background. The FARS (National Highway Traffic Safety Administration, 2002) contains a census of fatal traffic accidents within the 50 States, the District of Columbia and Puerto Rico. Data for each year, which are derived from police and emergency medical system (EMS) reports, are in four files: **FARS-A**, the accident file, **FARS-V**, the vehicle file, **FARS-D**, the driver file, and **FARS-P**, the person file. The FARS is assembled and maintained by the National Highway Traffic Safety Administration (NHTSA), and is available the NHTSA web site. The version analyzed was downloaded from the no-longer-extant ITDB (Bureau of Transportation Statistics, 2002). As will be seen below, the two versions differ significantly, possible because "DQ problems are acute when [data] are collected in one organization for use by another" (Galway and Hanks, 1996).

Metadata Issues. Inconvenient coding and poor data representations are not simply annoyances, but can inhibit analysis. For example, people who use FARS data in conjunction with a geographical information system (GIS) would be served better if the geographical attributes State and County were coded using Federal Information Processing System (FIPS) codes rather than General Services Administration (GSA) codes, which are effectively unintelligible (see Figure 2) without a data dictionary. Similarly, in the ITDB, time values are coded as 0147 for 1:47 AM, whereas the NHTSA version of FARS treats hours and minutes as separate attributes. Without correct and cumbersome parsing, subtraction of these values to compute the time difference between two events produces erroneous results.

A pervasive DQ problem is the failure of documentation to distinguish clearly between "original data" attributes and *derived attributes* calculated from original data. At the very least, making this distinction informs users of what might be redundant (or informative!) consistency checks, and also flags some kinds of analyses as inappropriate (Example: those where both derived and original attributes are used as predictors in a regression).

Poor coding of missing and incomplete values, a venerable problem, seriously reduces the usability of the data. In the ITDB version of FARS, missing values are padded with "0" (but this usually means "not applicable"), "9" (which usually means "missing"),"*" and blanks. Aggravating this inconsistency, there are partially missing attributes in the ITDB version of FARS, such as dates of the form *10991999*; here the missing month is replaced by "99." (The NHTSA version of FARS has month, day and year as separate attributes, and so has no partially missing values.) This again unnecessarily burdens the user to parse data values correctly. Figure 1 shows the prevalence of these partially missing values in the **FARS-P** file for 1999. Moreover, legitimate data values become confounded with the missing value code (Example: in FARS, Age = 99 years is indistinguishable from Age = missing). To some degree, these are legacy problems associated with fixed-width attributes, and would vanish if data were maintained in a modern RBDMS.

Metadata should address possible systematic influences on missing data. Figure 2 plots the percentage of missing EMS arrival times by state. A state-to-state effect is apparent; some analyses would be flawed without accounting for it.

Even such seemingly simple metadata as file sizes can help identify DQ problems. Figure 3 shows the

(DeathDate, DeathTime)	Number of Records
(0,0)	29,701
(MM**1999,*)	29,091
(99991999,*)	188
(99999999,*)	3
(,*)	643

Figure 1: Summary of various forms of DeathDate in the 1999 FARS.

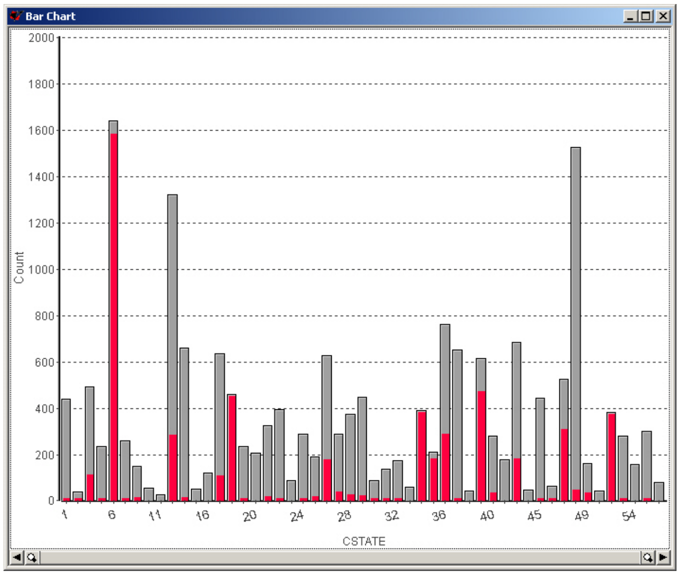

Figure 2: Bar chart showing missing EMS arrival times by state. Accidents having missing times are shown in dark gray. There are clear systematic state-to-state effects. Numerical state labels rather than USPS codes (It is not easy to know that 6 = CA!) make the data harder to use.

sizes of the four FARS files available for download on the ITDB. The small sizes of the files for 1989, 1991, 1992, 1993 1998 and 1999 clearly merit investigation and, as discussed in **Relational Database Issues** below, signal a dramatic DQ problem. Figure 4 reveals that the problem affects all states.

Single Attribute Analyses. A natural first check is to ensure that data values are legitimate (Examples: date values should represent *bona fide* dates, and purportedly positive-valued data should indeed be positive). Consistency of attribute representation must also be checked, especially for databases such as FARSthat are collected from disparate sources with differing data-collection procedures. Problems do exist: for some records in **FARS-A**, the accident time and EMS arrival time have differing representations, usually with one on a 24-hour clock (1602 Hours) and other on a 12-hour clock (0402 PM).

A simple, effective strategy to find anomalous values of individual attributes consists of three steps. First, one should sort and list unique values, which reveals suspicious values and can also indicate subtle DQ problems. To illustrate, looking at extreme values of the derived attribute LTIME in **FARS-P**, the lag

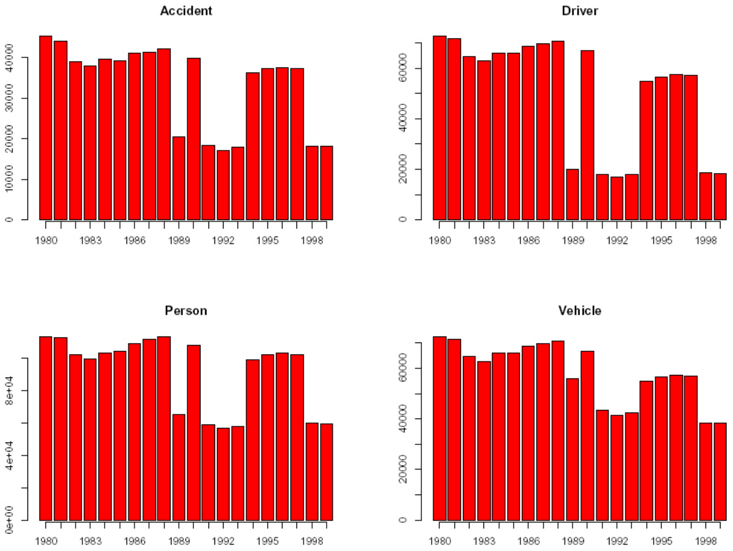

Figure 3: Sizes of FARS files downloaded from the former ITDB Web site (Bureau of Transportation Statistics, 2002) for the years 1980–1999. The inexplicable deficiencies for some years clearly require explanation.

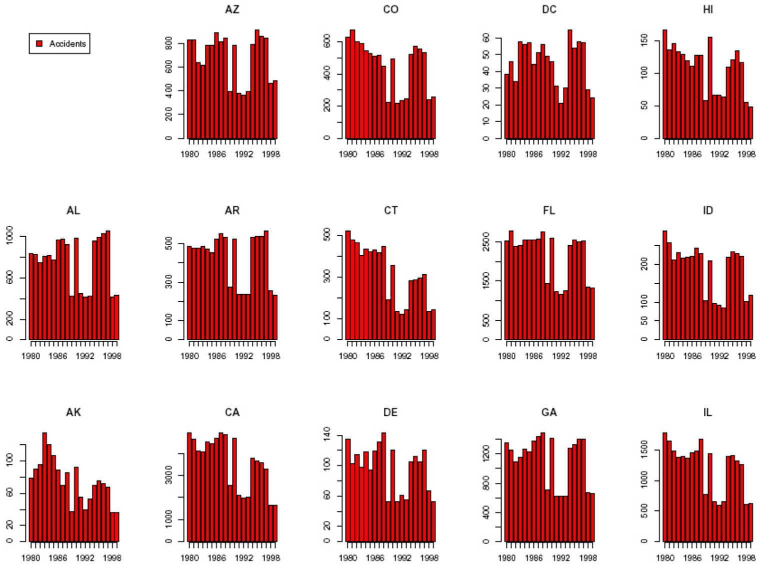

Figure 4: Accident counts by year in **FARS-A** files downloaded from the former ITDB Web site (Bureau of Transportation Statistics, 2002) for selected states. It is clear that the deficiency problem in Figure 3 affects all states.

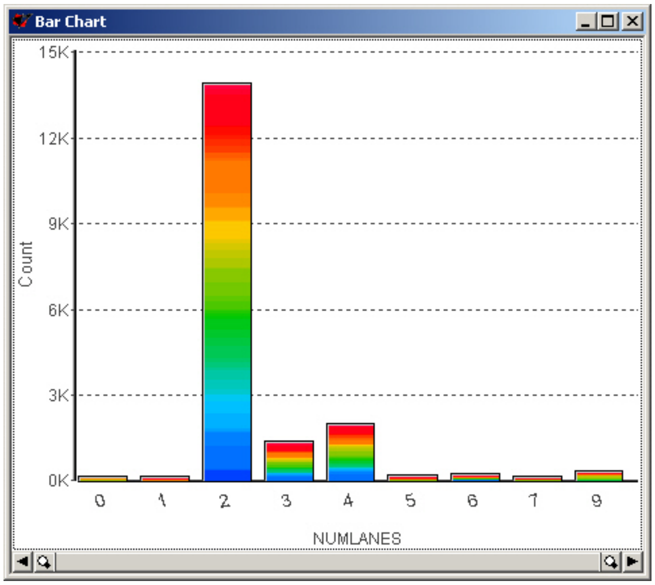

Figure 5: Bar chart showing distribution accident counts as a function of the number of lanes, from **FARS-A**. Colors (here translated to gray levels) correspond to states. As discussed in the text, at least two (7 = "≥ 7" and 9 = "Missing") and possibly three (0) of the seemingly numerical values are actually categorical.

time between accident time and death time, reveals several anomalies. Two of the extreme values of LTIME are both 72000 (which means, although this is not obvious, 720 hours and 00 minutes). Closer examination reveals that both of these are associated with the same accident—in State = 54 having Case No. = 321—which took place at 19:27 hrs on 11/7/97. Both fatalities are recorded in the **FARS-P** file with death date *12*/7/97 and death time 19:27. This clear case of a mis-recorded month could not readily have been detected otherwise.

Second, one can visualize distributions of data values for each attribute. Figure 5 shows the distribution of the number of lanes on the road on which accidents occurred, taken from **FARS-A**. The relatively large number of zeroes ("not applicable") might lead to further investigation whether, for instance, "0" and "9" were both used to denote missing values. The category "7" in Figure 5 actually represents "seven or more," which cannot be determined without close reading of the documentation. An analysis that ignored this would be incorrect. The problem could readily have been avoided simply by using "7+" as the attribute value, which also prevents the attribute from being treated as a purely numerical quantity.

Third, for multi-year data, one can examine the longitudinal behavior of individual or aggregated values, which we do repeatedly in §5.2.

Multiple Attribute Analyses. Extending the notion of examining distributions of individual attributes, one can examine bivariate relationships. Looking at every possible pair of attributes may not be feasible or meaningful, but two special cases are relevant. Nor need the relationships be "systematic;" for example, (see Figure 7) variation in one attribute (Example: numbers of lanes associated with accidents by state) may reflect an "underlying" variation (Example: states' differing road inventories).

The first case is pairs of attributes for which there is a "physical" basis to expect a positive or negative relationship. An example is driver height and weight in **FARS-D**, which are shown in Figure 6, where we

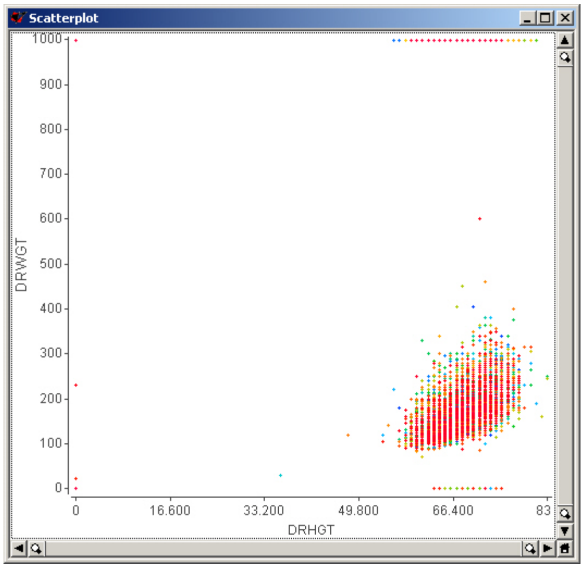

Figure 6: Scatterplot of driver height and weight, from **FARS-D**. As expected, these attributes are positively correlated. Missing weights, denoted by 999, are visually apparent, but would not be rejected or ignored by many statistical packages.

see the expected positive relationship. Moreover, outlying and boundary cases are readily visible and can be investigated further.

The second case concerns relationships based on domain knowledge. For example, data can be checked for “plausible” and “implausible” correlations. As examples, we present in Figures 7 and 8 visualizations for pairs of attributes in **FARS-A**. Both plausible (for example, between state and number of lanes) and implausible (for example, between state and day of week) pairwise relationships are easily assessed from the visualizations.

Relational Database Issues. Checking RBDMS properties of the database is essential to appraising its quality. Principles of RBDMS design (cf. §2.2.1) provide a framework to ensure consistency and to identify or even resolve many data anomalies. The principal steps are:

Primary Keys: Check if the attributes that are stated to be or logically ought to be primary keys for a given file actually are. For example, in **FARS-A**, check that State and Case No. uniquely determine an accident record.

Normalization: Check if the database is properly normalized. If not, check that the non-normalized aspects do not harbor inconsistencies.

Foreign Keys: Verify that there are foreign keys or other attributes that allow files to be joined. For example, in FARS, verify that an accident record can be linked with the driver, vehicle, and person records involved in that accident.

Joins: Perform joins of tables to verify that records can be *correctly* linked.

To illustrate, attempts to join the Accident, Driver, Vehicle and Person files in the ITDB version of the

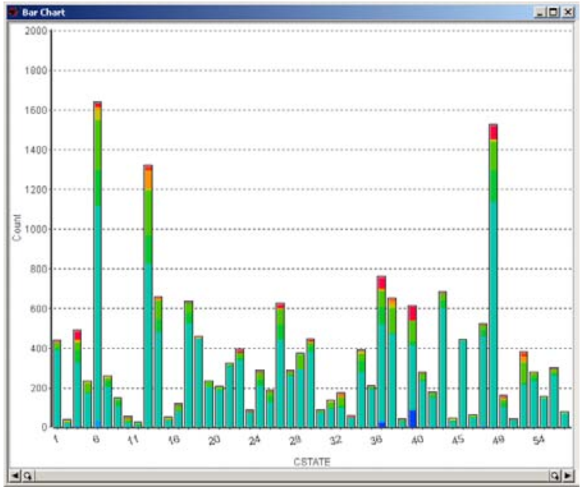

Figure 7: Bar chart of accident counts showing existence of a relationship between state and number of lanes. Colors (gray levels) correspond to the number of lanes. (This bar chart is the "transpose" of that in Figure 5.) This relationship is plausible since the distribution of the number of lanes varies by state.

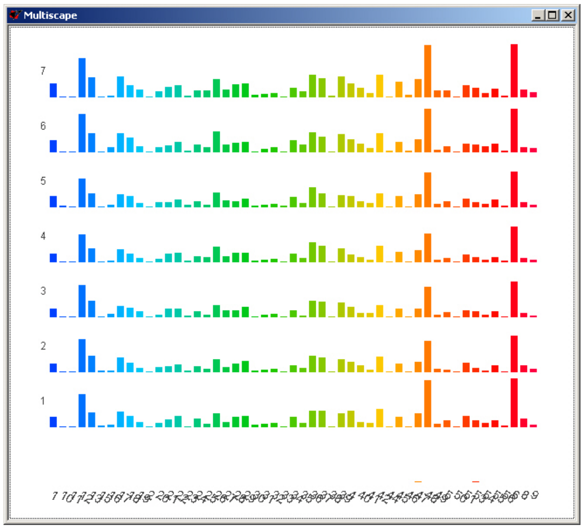

Figure 8: Plot of accident counts by state (horizontal axis) and day of week (vertical axis), showing no relationship.

State Code	Case Number	Vehicle Number in **FARS-D**	Vehicle Number in **FARS-V**	Vehicle Number in **FARS-P**
1	85	2,4,6,7	1,3,5	1,3,5,6,7
1	91	1	2	0,2
1	92	2	1,3	1,3
6	1345	2,3	1	1,2
6	1349	2,3	1	1,1,2,3,3

Table 1: Excerpt from Joins in FARS 1999 downloaded from the ITDB Web site. The vehicle numbers are those present in each of the files. The driver and vehicle files have *no records in common*.

FARS for those years with significantly less data (Figure 3) reveal very serious problems, which explain the anomalies discussed under **Metadata Characteristics.** Specifically, for a given accident we would expect a 1-to-1 correspondence between vehicle records and driver records. However, for 1999 FARS data downloaded from the ITDB in 2002, there is *no overlap* between the driver and vehicle files! In fact,

$$\begin{aligned}
\#\{\textbf{FARS-V}\} &= 38{,}268 \qquad \text{Key = (CSTATE, CNUM, VNUM)} \\
\#\{\textbf{FARS-D}\} &= 18{,}403 \qquad \text{Key = (CSTATE, CNUM, VNUM)} \\
\#\{\textbf{FARS-D} \bowtie \textbf{FARS-V}\} &= 0
\end{aligned}$$

where $\#\{A\}$ denotes the number of records in database A and $A \bowtie B$ represents the linkage of records in databases A and B. Table 1 shows details of a few of the records. This diagnoses but does not explain the problem in Figures 3 and 4: no accident record appears in both the driver and vehicle files.

5.2 Case Study 2: The TRI

Background. The TRI database (Environmental Protection Agency, 2002) was established by Federal law in 1986 to compel industrial and government facilities to disclose releases of any of more than 300 registered toxic chemicals into the air and water. Reports must be made for any facility that manufactures or processes more than 25,000 pounds (annually) of a registered chemical. (Starting 2001, the threshold for lead was reduced to 100 pounds, leading to 9800 additional facilities being required to report.) The TRI is used by the public and government agencies for purposes ranging from clean-up planning to lawsuits involving environmental justice.

Two characteristics of the TRI dominate assessment and amelioration of DQ problems. First, the TRI was created without identifying any specific uses of the data, which militates against user-centric approaches. Second, TRI data are self-reported, using a complex, five-page form that must be completed for *every facility and chemical* for which releases must be reported. There is no strong incentive even to complete the report, let alone do so accurately, which leads to a variety of data anomalies, most of which are not readily handled by standard statistical methods. In many cases, the form appears to have been completed by an engineer, often with substantial substitution of engineering judgement for “hard numbers.”

Rather than repeat all steps of the strategy outlined in §5.1, we focus here on detection and characterization of anomalies specific to one plausible use of the TRI data—to estimate regional-level trends. Both detection and characterization are addressed in the context of data viewed as facility-level time series. The data employed were downloaded using the TRI Explorer (Environmental Protection Agency, 2002) and cover air, water, on-site land, and off-site land releases of lead in "Original Industries" (SIC 20xx–39xx) for the years 1988–1998. Since the TRI Explorer contains only separate files for each year, time series analyses require addition of Year attributes to each file, followed by merging of the annual files, a non-trivial impediment for some users.

Data Anomalies. The most striking characteristic of TRI data is extreme variation and gaps in the facility-level time series. Figure 9 shows several examples, all of which are both incomplete and exhibit dramatic year-to-year variability.

There is clear (but sometimes misleading) evidence of systematic constancy in the TRI data. Other than one aberrant year (a different completer of the form?), the facility at the top of Figure 10 seems to have constant, albeit suspiciously rounded releases. In fact, the values of "500" are categorical responses (corresponding to the range 100–999) to Question 5.1 on Form R (Figure **??**) that appear in the TRI database as numerical values! Beyond this, multiple measurement methods may be employed by the form-completer, and while the method used must be entered on Form R, it is not contained in the downloadable data.

There is also evidence of "systematic change:" the facility at the bottom of that figure (in addition to the isolated 444,000 pound land release in 1990) has examples for which (with R_t the release in year t), $R_t = (2/3) * R_{t-1}$ and $R_t = (1/2) * R_{t-1}$. In these cases, the completer of Form R seems to have adjusted the previous year's values to approximate current releases.

Another human-engendered anomaly is disparities between on-site and off-site releases. Figure 11 illustrates this for several facilities; the same behavior is also present in Figure 10.

These anomalies, whose detection again requires use of visualization methods, have hard-to-model consequences. Most methods for time series analysis, for instance, are not designed to handle the kinds of error structures encountered above.

Regional-Level Trend. EPA administrators and others are concerned with pollution trends, at local and more aggregated geographical scales. The capability of TRI to provide useful information about regional trends, therefore, is one measure of its quality. Even though facility-level latitude and longitude are available from the TRI Explorer, without a GIS, true geographical aggregation is difficult. A simple approach aggregates facilities on the basis of the first digit of their Zip codes (the first 5 characters of the TRIFID in Figure 10), in effect dividing the country into ten regions and pooling releases within region.

This EDA-like strategy again illuminates the central issue. In many cases, the regional sum is effectively the maximum over facilities in the region. That is, regions are dominated by massive facilities. Therefore, facility-level anomalies are transferred to the regional level, and there is no error-cancellation from aggregation. Figure 12 illustrates this point for the regions defined by Zip codes 6**** (Midwest) and 7**** (south Central).

The TRI exercise shows that simple strategies—visualization and manual examination guided by domain knowledge—can work on a small scale. It also highlights the need to detect and characterize anomalies in data, raising in turn the need to characterize the impact of anomalies on inference from the data. Many of the same questions raised by the analysis of the ITDB arise again: What strategies scale to large problems? What strategies can be automated? These kinds of issues inform the research challenges described in the following section.

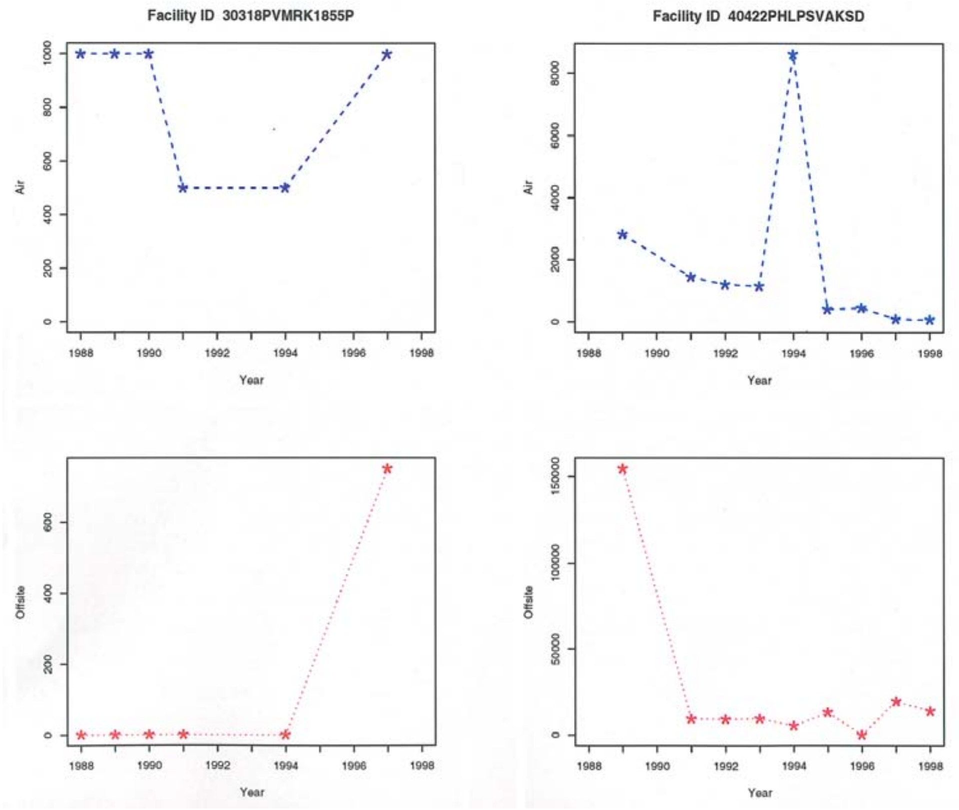

Figure 9: TRI Data showing annual air (top) and off-site releases of lead for two representative facilities.

6 Research Challenges

Here we present statistical research challenges associated with DQ, grouped into theory and methodology (§6.1), a decision-theoretic formulation for DQ (§6.2) and the software tools (§6.3) that are necessary to address real DQ problems.

Framing these challenges is the need to *integrate generic and problem-specific* components into the DQ system. The case studies in §5 confirm that domain knowledge is an essential component of DQ. This is necessary at both the theory and methodology and software tools levels. How to do it, however, is not clear.

6.1 Statistical Theory and Methodology

DQ Metrics are necessary for both summary and decision purposes. Here are two initial examples, which were developed by NISS for BTS (Karr and Sanil, 2001).

First, the *Intactness* of a table is the fraction of records that pass completeness, consistency and plausibility checks of the sort described in §5. In general, low intactness is symptomatic of low DQ. Accounting only for missing values other than latitude/longitude and inconsistent times (Example: EMS arrives before the time of the accident), the intactness of the 1999 **FARS-A** file is 0.3. Were the latitude/longitude included,

TRIFID	Year	Air	SurfH$_2$O	Land	OnSite	OffSite	Total
07029KRYNS936HA	1989	500	0	0	500	0	500
07029KRYNS936HA	1990	500	0	0	500	0	500
07029KRYNS936HA	1991	156	0	0	156	0	156
07029KRYNS936HA	1992	500	0	0	500	0	500
07029KRYNS936HA	1993	500	0	0	500	0	500
07029KRYNS936HA	1994	500	0	0	500	0	500
07029KRYNS936HA	1995	500	0	0	500	0	500
07029KRYNS936HA	1996	500	0	0	500	0	500
07029KRYNS936HA	1997	500	0	0	500	0	500
07029KRYNS936HA	1998	500	0	0	500	0	500

TRIFID	Year	Air	SurfH$_2$O	Land	OnSite	OffSite	Total
38109RFNDM257WE	1988	15300	0	0	15300	22163	237643
38109RFNDM257WE	1989	15300	0	0	15300	148151	164451
38109RFNDM257WE	1990	10200	0	440000	450200	19444	469644
38109RFNDM257WE	1991	3800	0	0	3800	76850	80650
38109RFNDM257WE	1992	1900	0	0	1900	313750	315650

Figure 10: Constancy and systematic changes in TRI data. *Top:* A facility with one seemingly anomalous year, which may be misleading because values of 500 actually correspond to the range 10–999. *Bottom:* A facility on which year t air releases appeared to be derived by judgement (Examples: two-thirds and one-half as much) from year $t-1$ releases.

intactness would be less than one-half of one percent.

Second, given a table with D attributes, the *Dimensional Efficiency* D^* of D may be defined in both a simple way, as the number of attributes that are neither constant across all records (and hence contain no discriminatory or predictive information) nor missing to such an extent that they are useless nor derived from other attributes, and in a complex way, as the number of dimensions needed to explain a fixed proportion of the variability in principal components analysis. The latter provides sharper information about the amount of independent information among the attributes.

Other, yet-to-be-developed metrics must accommodate such DQ characteristics as those laid out in §3.

We also urge, and are engaged in, development of DQ metrics that account explicitly for inferences drawn from the data. An example is described in §6.2.

Quality Resistant Inference. Despite attractions of the TQM approach to DQ, it will always be necessary to use poor quality data. In light of this, inference procedures are needed that mitigate poor DQ.

One strategy is to use robust estimators. Some procedures (Example: S-estimators (Rousseeuw and Yohai, 1984; Sakata and White, 1998)), can perform well for single parameters even when just more than half of the data are correct, giving robust estimates of central tendency, dispersion, and association. (Of course, one would want to know first that the data are this bad.) These techniques perform less well, of course, for multivariate parameters such as covariance matrices: breakdown points depend on the dimension,

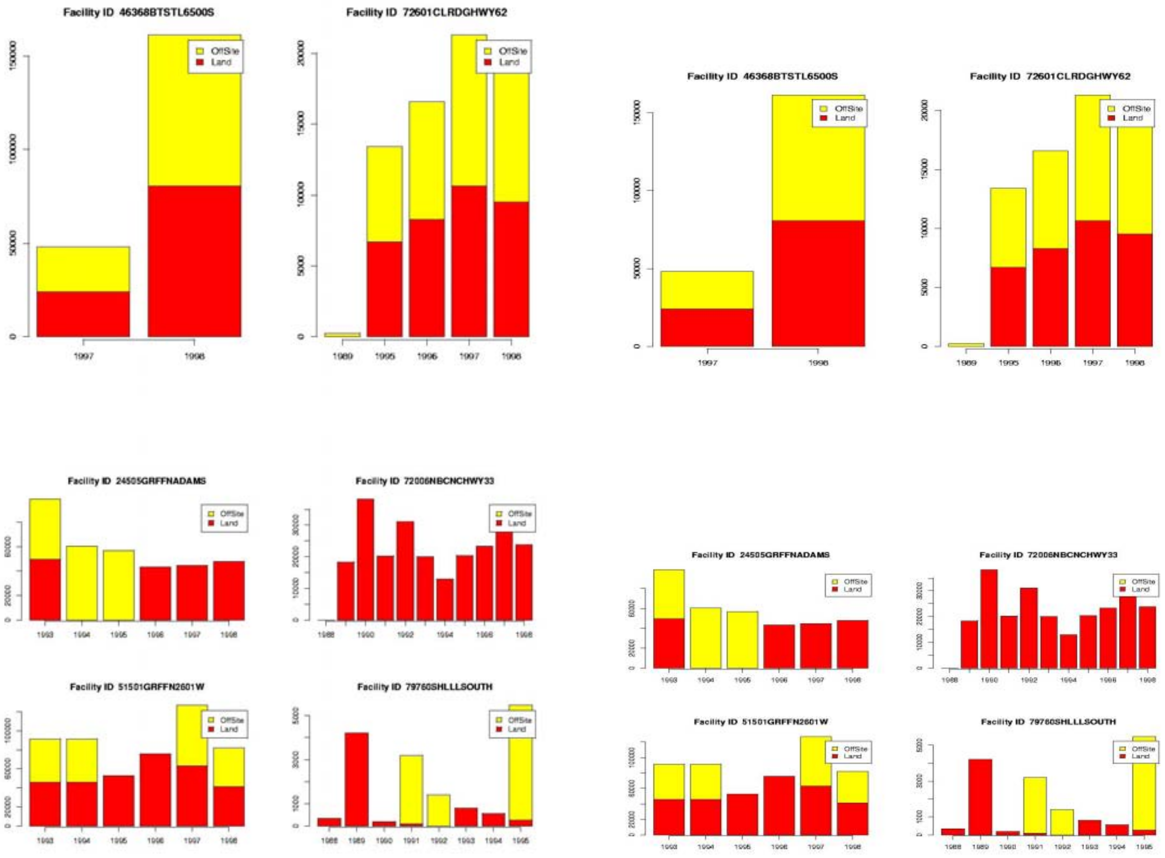

Figure 11: Stacked bar chart showing on-site and off-site disparities for selected facilities. Questionable aspects include Off-Site ≡ On–Site, Off-Site oscillating between 0% and 100%.

and most methods cannot accommodate a very large percentage of poor-quality data.

Robust methods can also assist in detecting “outliers” and other data oddities, but as the TRI case study shows, real-world data anomalies correspond to error structures that even robust models seem unable to accommodate, so there is much room for innovation.

One specific research issue is that there is no robust procedure for estimating a proportion from survey data, a ubiquitous issue in federal statistics. The impact would be enormous: estimates of proportions arguably play a larger role in framing national policy than estimates of location or dispersion.

Crosswalks. Data collection processes are not static, and this introduces a special kind of DQ problem. For example, the Census Bureau recently changed the way in which race is reported: for many decades, people could report membership in only a single race, but in 2000 they were permitted to report any combination of six racial categories. Such changes are made for good reasons and are likely to improve DQ in the long run, but the immediate impact is to make it difficult to characterize trends. Crosswalks are meant to address such problems. A crosswalk is a plan to collect data by both the new and old methods for a period

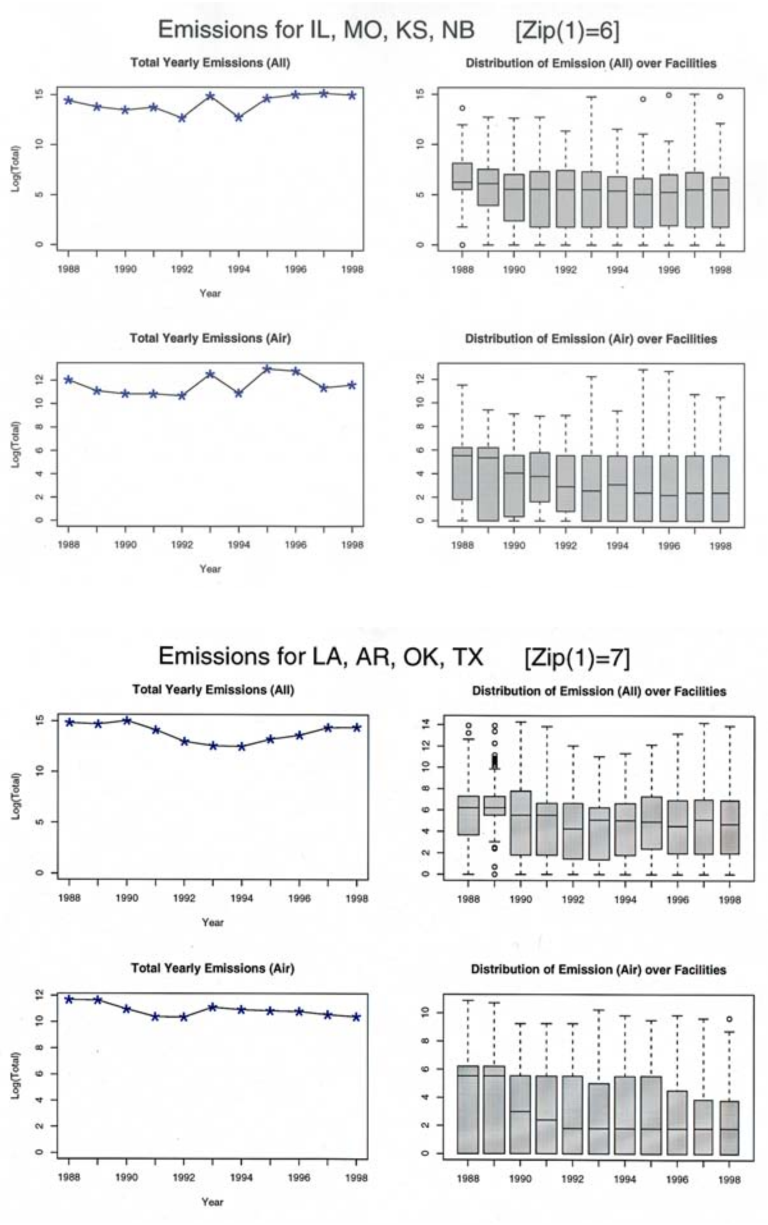

Figure 12: Facility-level distribution (Box plots) and regional totals (lines for Zip codes 6xxxxx (Top) and 7xxxx (Bottom).

of time, so that the old time series and the new time series can be calibrated (Bradley and Earle, 2001).

There are limits as to how successful a crosswalk can be. Suppose that at time c the process changes, and that the parallel collection for the crosswalk lasts for m time periods after the change. Thus the original data collection system captures $X_1, \ldots, X_c, X_{c+1}, \ldots, X_{c+m}$, while the new collection system captures $Y_c, Y_{c+1}, \ldots$. Suppose that the original data collections were used to produce summary estimates (Example: the proportion of people whose primary racial identification falls into each of six categories). Denote that series of estimates by $\hat{\theta}_1, \ldots, \hat{\theta}_c$. Then the goal of the crosswalk is to find a function $f(\cdot)$ that operates upon the Y_t to produce estimates $\hat{\theta}_t$, thereby allowing the analyst to provide comparable estimates for time period $c + 1$ and beyond.

The ideal statistical strategy is to ensure that the m is large, so that there is extensive overlap in the crosswalk. That would enable analysts to use nonparametric models such as MARS (Friedman, 1991) to find the function that best extends the series. However, this ideal is almost never achieved in practice: typically, because of financial considerations, $m \leq 2$. Moreover, only rarely are both systems run at a full level. As a result, in practice many organizations apply a simple proportional correction, assuming that correction factor is stable across future survey conditions. For cases such as race in the Census, this assumption is suspect: people's attitudes about race are changing at the same time that the racial composition of the US is changing.

New research ideas are needed to attempt to resolve such multiple effects. Almost inevitably, the models will be Bayesian, in order to incorporate both prior information and hierarchical structure.

6.2 Decision–Theoretic Formulation

A full-blown decision–theoretic formulation of DQ is a daunting research challenge. For example, when decisions depend sensitively on factors other than data issues, it may not be worthwhile to improve DQ, but it might nevertheless be essential to characterize DQ. In other instances, where databases drive decisions that have enormous consequences, one could, with the right models, make a strong case for significant investment of resources to improve DQ.

The principal components of such a structure are, of course, models, loss functions and characterizations of uncertainties. We discuss these in turn, and illustrate with problem of whether application of strategies to improve DQ, such as those in §4.4, is justified. Let $\mathcal{D}^{\mathrm{true}}$ be the true database, which of course may exist only conceptually. Let $\mathcal{D}^{\mathrm{pre}}$ be the database prior to clean-up. Let S be a clean-up strategy, and let $\mathcal{D}^{\mathrm{post}}(S)$, which depends on both $\mathcal{D}^{\mathrm{pre}}$ and S, be the database resulting from applying S to $\mathcal{D}^{\mathrm{pre}}$. Many such strategies S are in some sense parameterized by a level of intensity, which we denote by θ, and which itself is of course a decision variable.

Predictive models for DQ that reflect the nature of the processes by which data are generated represent an immense opportunity for statisticians. The needs encompass modeling the role of people in data production processes (Example: Form R for TRI), new kinds of errors (Example: transposition of digits by a data entry clerk, or the incorrect linkage of records) with unusual dependences (Example: TRI time series), and novel concepts of "noise."

To illustrate, to the extent that DQ is a function of the quality of individual elements of a database, not all elements are equally likely to have poor quality. (Example: the accuracy of a particular item in a survey depends upon such factors as the complexity of the question, the sensitivity or delicacy of the information, the burden of the survey, and the demographics of the respondent.) One path would be to build a logistic regression model that predicts the probability that a data item is correct as a function of suitably quantified explanatory variables. These variables are context-specific, but such abstractions as complexity, sensitivity, and burden are significantly generic. Ultimately, this kind of approach will help in targeting quality improvement efforts.

Models for changes in data quality under transformations (Examples: aggregation, joins of relational tables within the same database, and integration of multiple databases) of data are also essential.

In our example, effectiveness of the clean-up strategy S is measured by

$$\mathrm{Eff}(S) = d(\mathcal{D}^{\mathrm{post}}(S), \mathcal{D}^{\mathrm{true}}), \tag{1}$$

where d is a DQ metric. Following on comments in §6.1, d should reflect inferences drawn from the data. To be more specific, let P be an inference procedure that can be applied to the data—either $\mathcal{D}^{\mathrm{pre}}$ or $\mathcal{D}^{\mathrm{post}}(S)$, and let d_P be a function measuring the difference in the results of P applied to two different databases. For example, if P represents a linear regression, d_P could measure the difference, for example, the L^2-distance, between the resulting estimates of the coefficients, perhaps with some accounting for differing uncertainties in the estimates. Of course, d_P is not be a true metric; in particular, low values of $d_P(\mathcal{D}^{\mathrm{true}}, \mathcal{D}^{\mathrm{pre}})$ indicate that $\mathcal{D}^{\mathrm{pre}}$ is closer to $\mathcal{D}^{\mathrm{true}}$.

If the true database $\mathcal{D}^{\mathrm{true}}$ were known, then a plausible estimate of the effectiveness of the strategy S that converts $\mathcal{D}^{\mathrm{pre}}$ to $\mathcal{D}^{\mathrm{post}}(S)$ is

$$\mathrm{Eff}(S, P, \mathcal{D}^{\mathrm{pre}}) = d_P(\mathcal{D}^{\mathrm{true}}, \mathcal{D}^{\mathrm{pre}}) - d_P(\mathcal{D}^{\mathrm{true}}, \mathcal{D}^{\mathrm{post}}(S)). \tag{2}$$

Positive values of $\mathrm{Eff}(S, P, \mathcal{D}^{\mathrm{pre}})$ indicate that S has improved the data: the conclusions drawn using P on $\mathcal{D}^{\mathrm{post}}(S)$ are closer to the truth than those drawn using P on $\mathcal{D}^{\mathrm{pre}}$. Of course, without knowledge of $\mathcal{D}^{\mathrm{true}}$, one cannot apply (2) directly, and alternatives would need to be constructed.

The most immediate need, however, is for models of the form that support principled selection of the parameter θ in a parameterized family $\{S(\theta) : \theta \in \Theta\}$ of clean-up strategies. The complication is that $\mathrm{Eff}(S(\theta))$ is known, in the sense that there is data (i.e., a corresponding $\mathcal{D}^{\mathrm{post}}(S(\theta))$ for at most a few values of θ. Thus the need becomes predictive statistical models for effectiveness, which would have form

$$\widehat{\mathrm{Eff}}(\theta) = f(\theta) + \text{uncertainty}. \tag{3}$$

Such models represent an entirely new set of problems for statistical modeling and estimation. Among obvious issues are the "form" of the model, the nature of the uncertainties and what data are necessary to fit the model. It seems inevitable, and desirable, to use a Bayesian approach. Validation of such models will also be a major challenge.

Quantification of costs and benefits of DQ is also necessary. The issues are complex: multiple costs (Examples: costs incurred by data generators, data maintainers, data users and data subjects), elusive costs (Examples: how much does it cost one to receive duplicate mailings? How much would it cost EPA to remove anomalies from the TRI? How much does it cost anyone if a person is wrongly suspected of terrorism on the basis of low quality data?) and elusive benefits (Examples: how much would a telephone operating company gain if it could link its cellular and landline customers accurately? How much would the public benefit if TRI were of higher quality?). Given the complexity of the issues, a meaningful starting point is to conduct cost-benefit case studies similar to those in §5.

Similarly, and linking to cost considerations, regression models could be used to predict the cost of raising DQ to a given level, at either the item or the database level. Such models, tailored to specific applications, can also help prioritize resource allocation for quality improvement.

Returning to the illustration,a cost model would have the form

$$\mathrm{Cost}(\theta) = g(\theta) \qquad \text{[+ uncertainty]}. \tag{4}$$

Given both this and the estimated effectiveness function $\widehat{\mathrm{Eff}}(\theta)$ of (3), there are three ways, at least, to select an optimal value θ^*. The first of these is to minimize cost subject to a lower bound on effectiveness. This might be termed a "regulatory" approach, especially if the effectiveness threshold were set externally. The second approach is to maximize effectiveness subject to an upper bound on cost. This is a resource-constrained approach, which might be appropriate in many scientific settings. The final approach is to select a value θ^* on the cost-effectiveness frontier, which consists of (Cost, Effectiveness) pairs that are undominated in the sense that no other value of θ can produce both lower cost higher effectiveness.

In general, construction or estimation of $\mathrm{Cost}(\theta)$ in (4) is an extremely difficult problem. At least two kinds of costs must be accounted for. "Process" costs of implementing S at intensity θ may be accessible because in some sense θ itself might be a cost. On other hand, "opportunity" costs of poor data quality are, and seem likely to remain, inaccessible in most organizations.

6.3 Software Tools

A NISS DQ Workshop in 2000 (Karr et al., 2001b) identified as a central research need "Software tools that solve real data quality problems. Such tools (even in prototype form) can also be used to evaluate and improve new theory and methodology. Issues include algorithms that cope with complexity of the techniques as well as the scale of the data and problems of human–computer interaction such as presentation and visualization."

Such tools exist already, as discussed in §2 and 4.4, but are oriented more to the computer science view of DQ than the statistical perspective. To make the challenges concrete, we frame them in terms of a *data quality toolkit*—an extensible set of software tools that automates the generic components of the strategy described in §5, performing both automatic and on-demand computations, including visualizations, but contains "hooks" for relevant domain knowledge. The DQTK has not yet been built, although NISS has worked to develop many of its components.

Functionality. At a high level, a DQ evaluation using the DQTK would consist three phases. Importantly, the DQTK creates a complete log of the process that contains user-entered metadata and results.

The first phase is user metadata entry and characterization. For the database, this includes names of tables, file names, and formats. For tables, it includes dimensions (records × attributes), and for each attribute, such information as name, unit of measurement, data type (Examples: numerical, text, date, time, latitude/longitude, Boolean, "other" with parsing rule), constraints (Examples: "must be positive," "must be an integer"), whether the attribute is original or derived, and representation of missing and N/A values.

An essential functionality is to help users understand metadata such as survey response rates and respondent burden (§2). One very intriguing way to do this is to use data collection instruments such as the TRI's Form R (Figure **??**) to visualize such metadata, for example, by coloring items according to the response rate. This makes not only the collection instrument but also key metadata about the data generation step immediately accessible to DQTK users.

In the second phase, DQ evaluation uses automatic computations at two stages: following metadata entry and at the end of the process to calculate DQ metrics. At the first stage, the files are read into the RBDMS. Then, for each attribute, the DQTK computes the percentages of missing and N/A values, and creates one key visualization, based on attribute type (Examples: histogram, bar chart, map). Consistency checks are then performed, first for user-selected individual attributes. For instance, the DQTK checks for "illegal values" relative to the attribute type (Example: time = 2585) or relative to bounds specified by user on the basis of domain knowledge (Example: weight $\geq$ 500 pounds).

In the second stage of the second phase, the DQTK checks inter-attribute consistency (Examples: temporal consistency, mathematical relationships such as $A_1 \geq A_2 + A_{15}$, and logical relationships such as A_1 = (Latitude, Longitude) belongs to A_{25} = State). Derived attributes are checked as well, with the user specifying the "formula" to compute a derived attribute from others. The DQTK also checks missingness, to detect systematic features such as those in Figure 2.

In the third phase, the user follows through EDA DQ strategy, seeing the results of selected automatic computations and directing the system to perform additional computations. As this process occurs, the user is provided text and visual summaries of problems, with lists containing details available as one drills down. All results, including visualizations, are written to the log file.

RDB characteristics are treated similarly. For instance, for each table the user specifies the attribute meant to be its primary key, and the DQTK provides either a verification or a summary of problems (with details via drill-down). Joins are performed, with the user specifying which tables are to be joined, the joining attributes, and attribute filters, and the DQTK returns the cardinality of the join and other summary information with lists of problematic records as drill-down.

Finally, the DQTK computes automatically DQ metrics (§6.1) for each table in the database.

Structure. The main components of the DQTK are:

User interface (UI), with the look and feel a graphical SQL client, allowing access by means of a standard Web browser.

Logging engine to record the entire session.

Computational engine to perform necessary calculations, calling as necessary on (1) An RBDMS to access and manipulate the data, which are converted to relational form by a **preprocessor**; (2) A GIS for manipulating and displaying geographical data; and (3) Visualization components to provide graphical output to the user.

Postprocessor to generate reports from the logs.

Building even a prototype DQTK raises research issues that span the three contributing disciplines of DQ. Beyond those implicit in the discussion of its core functionality, these include selection of software and hardware platform, design of the user interface, and incorporation of additional statistical and data manipulation capabilities. In the long run, an expert system wrapper for the system, which would suggest analyses to the user on the basis of the results of other analyses, seems very desirable. Making the DQTK scalable to handle large data sets is a challenge in its own right.

7 Conclusions

Too many organizations run on poor quality data. The appeal of TQM approaches notwithstanding, the error rates in most databases (with rare exceptions) are so great that it would be prohibitively expensive to attempt correction. Root cause analyses, although intriguing, seem problematic.

Instead, organizations have developed informal coping mechanisms that limit the sensitivity of their key decisions to the influence of bad data. Business managers look to their databases for suggestive trends, but are skeptical of detailed conclusions. In government, policy is almost never framed solely on the basis of statistics, or even largely upon statistics, but rather arises from compromises among stakeholders, legal experts and pragmatists, all of whom are skeptical (perhaps for different reasons) about even their own experts' analyses.

Statisticians, in collaboration with computer scientists and domain knowledge experts, have a major role to play in the process of using data effectively—which is ultimately what DQ is all about. That role ranges from characterizing DQ to methods of inference that are resistant to poor DQ to cost-benefit analyses that determine when specific investments in quality improvement pay off. We have attempted to lay the groundwork through a concrete path of case studies, our EDA approach to DQ metrics and measurement, and the DQTK. Ultimately, a decision-theoretic formulation of DQ will lead to tools that help managers to allocate resources to the most urgent, addressable problems.

Acknowledgements

This research was supported in part by NSF grant EIA–0131884 to NISS. Support for the case studies described in §5 was provided to NISS by the BTS and EPA. Many of the ideas here were shaped by the NISS Affiliates Workshop on DQ (Karr et al., 2001b). Jon Kettenring first brought DQ to the attention of NISS. Jerome Sacks participated in the TRI case study.

References

American Statistical Association (2002). Section on Survey Research Methods. Information available on-line at www.amstat.org/sections/srms/.

Anderson, M. and Fienberg, S. E. (2001a). Counting and estimation: Methodology for improving the quality of Censuses. The US 2000 Census adjustment decision. In *Proc. International Conference on Quality in Official Statistics*. Statistics Sweden.

Anderson, M. and Fienberg, S. E. (2001b). US Census confidentiality: Perception and reality. *Bull. Internat. Statist. Inst.*, 53rd Session.

AT&T Labs Research (2003). XGobi: A system for multivariate data visualization. Information available on-line at www.research.att.com/areas/stat/xgobi/.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (1991). *Measurement Errors in Surveys*. Wiley.

Brackstone, G. (1999). Managing data quality in a statistical agency. *Survey Methodology*, 25(2):139–149.

Bradley, D. and Earle, K. (2001). Developing and explaining the crosswalk between Census 1990 and 2000 industry and occupation codes. In *Proceedings of the Joint Statistical Meetings*, Alexandria, VA. American Statistical Association.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1983). *CART: Classification and Regression Trees*. Wadsworth, Belmont, CA.

Bureau of Economic Analysis (2002). Information quality guidelines. Available on-line at www.bea.gov/bea/about/infoqual.htm.

Bureau of Justice Statistics (2002). BJS data quality guidelines. Available on-line at www.ojp.usdoj.gov/bjs/dataquality/guidelines.htm.

Bureau of Labor Statistics (2002a). BLS guidelines for informing users of information quality and methodology. Available on-line at www.bls.gov/bls/quality.htm.

Bureau of Labor Statistics (2002b). Bureau of Labor Statistics data integrity guidelines. Available on-line at www.bls.gov/bls/data_integrity.htm.

Bureau of Transportation Statistics (2002). Intermodal Transportation Database. Available on-line at www.itdb.bts.gov.

Census Bureau (2002). Census Bureau Section 515 information quality guidelines. Available on-line at www.census.gov/qdocs/www/quality_guidelines.htm.

Census Bureau Executive Steering Committee for Accuracy and Coverage Evaluation Policy (March 8, 2001). Report of tabulations of population to states and localities. Federal Register. Available on-line at www.census.gov/dmd/www/EscapRep.html.

Converse, J. M. and Presser, S. (1986). *Survey Questions: Handcrafting the Standardized Questionnaire*. Sage.

Date, C. J. (1999). *An Introduction to Database Systems, 7th Ed.* Addison-Wesley, Reading, MA.

de Leeuw, E. D. and van der Zouwen, J. (1988). Data quality in telephone and face to face surveys: A comparative meta-analysis. In *Telephone Survey Methodology*, pages 283–300, New York. Wiley.

Department of Defense (2002). DOD Guidelines on data quality management (Summary). Available on-line at http://www-datadmn.itsi.disa.mil/dqpaper.pdf.

Department of Education (2002). Information quality guidelines. Available on-line at www.ed.gov/offices/OCIO/info_quality/final_guide.html.

Dobra, A., Fienberg, S. E., Karr, A. F., and Sanil, A. P. (2002). Software systems for tabular data releases. *Int. J. Uncertainty, Fuzziness and Knowledge Based Systems*, 10(5):529–544.

Dobra, A., Karr, A. F., and Sanil, A. P. (2003). Preserving confidentiality of high-dimensional tabular data: Statistical and computational issues. *Statist. and Computing*, 13(4):363–370.

Dublin Core Metadata Initiative (2003). Dublin core metadata initiative. Available on-line at dublincore.org.

Duncan, G. T., Jabine, T. B., and de Wolf, V. A., editors (1993). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. National Academy Press, Washington. Report of a Panel on Confidentiality and Data Access, Committee on National Statistics.

Duncan, K. B. and Stasny, E. A. (2001). Using propensity scores to control coverage bias in telephone surveys. *Survey Methodology*, 27(2):121–130.

Energy Information Administration (2002). Energy Information Administration information quality guidelines. Available on-line at www.eia.doe.gov/smg/EIA-IQ-Guidelines.html.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. Wiley, New York.

Environmental Protection Agency (2002). Toxic Release Inventory Database. Available on-line at www.epa.gov/triexplorer.

Evoke Software (2002). Evoke. Information available on-line at www.evokesoft.com.

Federal Transit Administration (2002). National Transit Database. Available on-line at www.fta.dot.gov/ntl/database.html.

Fellegi, I. P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *J. Amer. Statist. Assoc.*, 71:17–35.

Fellegi, I. P. and Sunter, A. B. (1969). A theory for record linkage. *J. Amer. Statist. Assoc.*, 64:1183–1210.

Fienberg, S. E., Makov, U. E., and Sanil, A. P. (1997). A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data. *J. Official Statist.*, 13:75–89.

Fowler, Jr., F. J. (2002). *Survey Research Methods (3rd Edition)*. Sage.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, 19:1–67.

G. J. Lestina, Jr., Appel, M. V., and Gillman, D. W. (1997). Providing document retrieval through a metadata repository at the Census Bureau. Available on-line at www.census.gov/srd/www/abstract/gjlasa97.html.

Galway, L. and Hanks, C. H. (1996). *Data Quality Problems in Army Logistics: Classification, Examples & Solutions*. Rand Corporation, Santa Monica, CA.

Gomatam, S., Karr, A. F., Reiter, J. P., and Sanil, A. P. (2004). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statist. Sci.* To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

Gomatam, S., Karr, A. F., and Sanil, A. P. (2003). Data swapping as a decision problem. *J. Official Statist.* To appear. Available on-line at www.niss.org/dgii/technicalreports.html.

Greenberg, B. and Petkunas, T. (1990). SPEER (Structured Program for Economic Editing and Referrals). In *ASA Proceedings of the Section on Survey Research Methods*, pages 95–104, Alexandria, VA. American Statistical Association.

Groff, J. R. and Weinberg, P. N. (1999). *SQL: The Complete Reference*. Osborne/McGraw-Hill, Berkeley, CA.

Groves, R. M. (1987). Research on survey data quality. *Public Opinion Quarterly*, 51:156–172.

Groves, R. M., Magilavy, L. J., and Mathiowetz, N. A. (1981). The process of interviewer variability: Evidence from telephone surveys. In *ASA Proceedings of the Section on Survey Research Methods*, pages 438–443, Alexandria, VA. American Statistical Association.

Helfert, M. (2001). Managing and measuring data quality in data warehousing. In *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*, pages 55–65.

Hidiroglou, M. A., Drew, J. D., and Gray, G. B. (1993). A framework for measuring and reducing nonresponse in surveys. *Survey Methodology*, 19:81–94.

Huang, K.-T., Lee, Y. W., and Wang, R. Y. (1999). *Quality Information and Knowledge Management*. Prentice Hall, Upper Saddle River, NJ.

Kan, S. H. (1994). *Metrics and Models in Software Quality Engineering*. Addison–Wesley, Reading, MA.

Karr, A. and Sanil, A. P. (2001). Data quality for the ITDB: Issues and paths to solution. Final report to the Bureau of Transportation Statistics under contract DTTS59–01–P–00235.

Karr, A. and Sanil, A. P. (2003). Data quality: State of the art and practice. Report to the Bureau of Transportation Statistics under contract DTTS59–02–P–00339.

Karr, A. F., Dobra, A., and Sanil, A. P. (2003). Table servers protect confidentiality in tabular data releases. *Comm. ACM*, 46(1):57–58.

Karr, A. F., Lee, J., Sanil, A. P., Hernandez, J., Karimi, S., and Litwin, K. (2001a). Disseminating information but protecting confidentiality. *IEEE Computer*, 34(2):36–37.

Karr, A. F., Sanil, A. P., Sacks, J., and Elmagarmid, E. (2001b). Workshop report: Affiliates workshop on data quality. Technical Report, National Institute of Statistical Sciences. Available on-line at www.niss.org/affiliates/dqworkshop/report/dq-report.pdf.

Kendall, K. E. and Kendall, J. E. (2001). *Systems Analysis and Design, 5th Ed.* Prentice Hall, Upper Saddle River, NJ.

Lee, J., Holloman, C., Karr, A. F., and Sanil, A. P. (2001). Analysis of aggregated data in survey sampling with application to fertilizer/pesticide usage surveys. *Res. Official Statist.*, 4:101–116.

Little, R. J. A. (1992). Regression with missing X's: a review. *J. Amer. Statist. Assoc.*, 87:1227–1238.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Loshin, D. (2001). *Enterprise Knowledge Management*. Morgan Kaufmann, San Francisco.

National Agricultural Statistics Service (2002). Information quality guidelines of the US Department of Agriculture. Available on-line at www.ocio.usda.gov/irm/qi_guide/index.html.

National Center for Education Statistics (2004). Statistical Standards. Information available on-line at nces.ed.gov/statprog/stat_standards.asp.

National Center for Health Statistics (2002). Guidelines for ensuring the quality of information disseminated to the public. Information available on-line at www.hhs.gov/infoquality/nchs.html.

National Highway Traffic Safety Administration (2002). Fatility Analysis Reporting System. Available on-line at www.nhtsa.dot.gov/people/ncsa/fars.html.

National Institute of Statistical Sciences (2004). Digital Government Project II web site: Data Confidentiality, Data Quality and Data Integration for Federal Databases: Foundations to Software Prototypes. Available on-line at www.niss.org/dgii.

Newcombe, H. B. and Kennedy, J. (1962). Record linkage: Making maximum use of the discriminating power of identifying information. *Communications of the ACM*, 5.

Nousak, P. and Phelps, R. (2002). A score approach to improving data quality. Available on-line at www2.sas.com/proceedings/sugi27/p158-27.pdf.

Office of Management and Budget (2002a). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. Available on-line at www.whitehouse.gov/omb/fedreg/reproducible.html.

Office of Management and Budget (2002b). Office of Management and Budget information quality guidelines. Available on-line at www.thecre.com/pdf/20021026_omb-final.pdf.

Rousseeuw, P. J. and Yohai, V. (1984). Robust regression by means of S-estimators. In Franke, J., Hardle, W., and Martin, R. D., editors, *Robust and Nonlinear Time Series*, volume 26 of *Lecture notes in Statistics*, pages 256–272. Springer-Verlag, New York.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Sakata, S. and White, H. (1998). Breakdown point. In Kotz, S., Read, C., and Banks, D., editors, *Encyclopedia of Statistical Sciences*, volume Update Volume 2, pages 84–89. Wiley, New York.

Sanil, A. P., Gomatam, S., Karr, A. F., and Liu, C. (2003). NISSWebSwap: A Web Service for data swapping. *J. Statist. Software*, 8(7).

SAS Institute, Inc. (2002). Dataflux. Information available on-line at www.dataflux.com.

SPSS, Inc. (2002). Authoring that fits into your workflow. Information available on-line at www.spss.com/spssmr/authoring/.

Trillium Software (2002). Trillium Data Quality. Information available on-line at www.trilliumsoft.com.

Tucker, C. (1992). The estimation of instrument effects on data quality in the Consumer Expenditure Diary Survey. *J. Official Statist.*, 8:41–61.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison–Wesley, Reading, MA.

Ullman, J. and Widom, J. (1997). *A First Course in Database Systems*. Prentice-Hall, Upper Saddle River, NJ.

US Geological Survey (2003). Formal metadata: Information and tools available on this server. Available on-line at geology.usgs.gov/tools/metadata/.

Visual Insights, Inc. (2002). Visual Insights ADVIZOR. Information available on-line at www.visualinsights.com/advizor.

Wang, R. (1998). A product perspective on total data quality management. *Comm. ACM*, 41(2).

Wang, R. Y., Ziad, M., and Lee, Y. W. (2000). *Data Quality*. Kluwer, Amsterdam.

Winkler, W. E. (1994). Advanced methods for record linkage. *ASA Proceedings of the Section on Survey Research Methods*, pages 467–72.

Winkler, W. E. (1995). Editing discrete data. In *ASA Proceedings of the Section on Survey Research Methods*, pages 108–113, Alexandria, VA. American Statistical Association.

Winkler, W. E. (2000a). Machine learning, information retrieval, and record linkage. In *ASA Proceedings of the Section on Survey Research Methods*, pages 20–29, Alexandria, VA. American Statistical Association.

Winkler, W. E. (2000b). Machine learning, information retrieval and record linkage. Available on-line at www.niss.org/affiliates/dqworkshop/papers.html.

Winkler, W. E. and Chen, B.-C. (2001). Extending the Fellegi–Holt model of statistical data editing. In *ASA Proceedings of the Section on Survey Research Methods*. American Statistical Association.

World Wide Web Consortium (2004). Extensible Markup Language (XML). Information available on-line at www.w3.org/TR/REC-xml.