



Applying Control Chart Methods to Enhance Data Quality

L. Allison Jones-Farmer, Jeremy D. Ezell & Benjamin T. Hazen


To cite this article: L. Allison Jones-Farmer, Jeremy D. Ezell & Benjamin T. Hazen (2014) Applying Control Chart Methods to Enhance Data Quality, *Technometrics*, 56:1, 29-41, DOI: [10.1080/00401706.2013.804437](https://doi.org/10.1080/00401706.2013.804437)



To link to this article: <http://dx.doi.org/10.1080/00401706.2013.804437>

 View supplementary material 



 Accepted author version posted online: 28 May 2013.

 Submit your article to this journal 

 Article views: 644

 View related articles 

 View Crossmark data 

 Citing articles: 2 View citing articles 

Applying Control Chart Methods to Enhance Data Quality

L. Allison JONES-FARMER and Jeremy D. EZELL

Department of Aviation and Supply Chain Management
Auburn University, Auburn, AL 36849
(jde009@auburn.edu)

Benjamin T. HAZEN

916th Maintenance Squadron, United States Air Force
Seymour Johnson Air Force Base, NC 27531
(benjamin.hazen@us.af.mil)

As the volume and variety of available data continue to proliferate, organizations increasingly turn to analytics in order to enhance business decision-making and ultimately, performance. However, the decisions made as a result of the analytics process are only as good as the data on which they are based. In this article, we examine the data quality problem and propose the use of control charting methods as viable tools for data quality monitoring and improvement. We motivate our discussion using an integrated case study example of a real aircraft maintenance database. We include discussions of the measures of multiple data quality dimensions in this online process. We highlight the lack of appropriate statistical methods for the analysis of this type of problem and suggest opportunities for research in control chart methods within the data quality environment. This article has supplementary material online.

KEY WORDS: Attributes control chart; Data analytics; Data production process; Process improvement; Quality management.

1. INTRODUCTION

It has been noted that in today's information age, everything can be monitored and measured, but it is increasingly difficult to use, analyze, and make sense of the data (Lohr 2009). The proliferation of data has led to increased global spending on data analytics software by 16.4% in 2011 to a total of \$12.2 billion (Kalakota 2012). For instance, IBM has invested more than \$14 billion in analytics in the past five years, mostly in analytics-related company acquisitions (IBM 2011). Additionally, organizations, such as the United States Air Force, have adopted several initiatives to expand their analytical capabilities in order to enhance logistics and maintenance effectiveness (Nunnally and Thoele 2007). However, as organizations seek to improve their analytics capabilities, it must be emphasized that gathering, storing, and making sense of a large amount of data is a complex process and there are many opportunities throughout for data quality (and thus analytical usefulness) to be compromised (Tayi and Ballou 1998).

Both academic and practitioner literature have long stated the need for improved data quality for effective decision making; however, the literature stops short of recommending generally applicable methods for measuring, monitoring, and improving data quality (Bose 2009; Warth, Kaiser, and Kugler 2011). Whereas the use of statistical process control (SPC) methods to monitor and improve the quality of manufacturing and service processes is well researched and implemented in practice, there has been only limited and rudimentary usage of SPC methods to monitor and improve the quality of the data itself. The purpose of this article is to introduce the data quality problem to researchers in statistical process control and to examine how control charts may be used to monitor and ultimately improve data quality.

The remainder of this article begins with a short overview of the data production and warehousing process. Then, we define the accepted dimensions of data quality and discuss ways in which data quality can be measured. We follow with a review of

existing applications of control chart methods to the data quality problem. Throughout, we integrate a discussion of a data production process based on an actual repair and refurbishment facility that maintains jet engine components for United States Air Force cargo aircraft. In doing so, we give an example illustrating how to measure data quality within a data production process. Next, we discuss how the unique nature of the "data about the data" provides additional challenges to data quality process improvement and illustrate the application of a control chart to establish a monitoring scheme within the aircraft maintenance data warehouse. Finally, we discuss future opportunities for statistical research in control charting as it applies to the data quality problem.

2. THE DATA PRODUCTION PROCESS

We assume the reader is familiar with SPC and basic control charting methods, but possibly unfamiliar with data quality, information systems, and data warehousing concepts. Thus, we begin with a brief introduction and overview of the methods for gathering and storing data in large organizations, such as in our aircraft maintenance example. In this article, we consider the data as both the raw material input and, in its transformed state, the output of the process. We refer to the transformed output of the data production process as a data product (Pitt, Watson, and Kavan 1995; Wang and Strong 1996; Kahn, Strong, and Wang 2002). Although the terms *information* and *data* are often used interchangeably in the literature and in practice, we restrict our terminology to *data* as it implies a more raw form.

In many organizations, data regarding transactions, customer relations, and company performance are ultimately stored within a computer-based repository known as a data warehouse (Wixom and Watson 2001; Breslin 2004; March and Hevner 2007). In the past, this data warehouse would often take the form of a single, specialized organizational data storage system. Today's data warehouses used by larger organizations regularly encompass data integrated from across physically distributed systems, for example, cloud-based storage.

Inmon (1992) noted that a key characteristic of the data warehouse is that it can be used to integrate data from across multiple organizational information systems through the extraction, transformation, and loading (ETL) process. Extraction refers to the process of obtaining data from either internal or external organizational sources such as Enterprise Resource Planning systems, Customer Relationship Management systems, third party vendors, click-stream data from customer visits to websites, or many other sources. Once extracted from these sources, the data are placed in an operational data store where the data will be processed during the transformation stage using a series of rules that will attempt to eliminate duplicates, missing records, and standardize the language. In the final stage, the data are loaded into the data warehouse. Every stage of the ETL process has the possibility of introducing error that can affect the quality of the final data product and the quality of the eventual analysis of this data.

Ideally, modern data processing systems include automated data quality checks. Even moderately complex business rules can be programed, helping to facilitate high quality data entry and extraction. Automated logical checks occurring early in the data production process can assist with continuous data collection improvement and the prevention of later data quality issues. Figure 1 gives a general overview of data warehousing, from data input to the output of a transformed data product.

The aircraft maintenance data management system that we examine is used for making repair decisions regarding both aircraft engines and engine subcomponents. Variations of these engines have been in service since the late 1960s, and this particular data management system has been in use since the 1990s.

This system contains data regarding unique, individual jet engine subcomponents that are serviced at the maintenance facility. Examples of such subcomponents include compressor disks, turbine disks, spools, shafts, seals, and similar items that combine to create one of over 600 jet engines in the inventory. Engines are removed from aircraft when they have reached a given number of flying hours and are due for scheduled maintenance (approximately every five years), or when a malfunction is suspected. After removal from the aircraft, engines require complete disassembly, inspection, and overhaul maintenance, and are subsequently shipped to the maintenance facility for service. As an engine is disassembled and its constituent subcomponents inspected and overhauled, facility personnel hand-enter data for each individual subcomponent.

Consistent with Figure 1, the raw maintenance data are input into both internal and external databases. Externally, data are recorded by field-level technicians and engine managers at air force installations throughout the world. Externally recorded data include number of cycles and hours of use at both the engine and component level and data are generally uploaded after each flight. Internally, technicians at the repair facility record data by hand upon the arrival and disassembly of aircraft engines. Internally hand-recorded data include the condition and preventive maintenance tasks at the component level. Through this process, technicians add internally recorded data properties to those externally recorded, forming complete data records. For this example, we define each record as a row of data containing 14 separate internally and externally recorded data properties that describe the individual engine subcomponent, to include serial number, date of manufacture, time since last overhaul, status, etc. See Table 1 for a definition of the terms used throughout our example.

Data from both sources are extracted, transformed, and loaded into a data warehouse. From this warehouse, users at many levels and locations can query the database to obtain desired information. There are several data consumers, from Air Force leaders who require a real-time picture of the maintenance status of engines and components, to mid-level managers who use the data

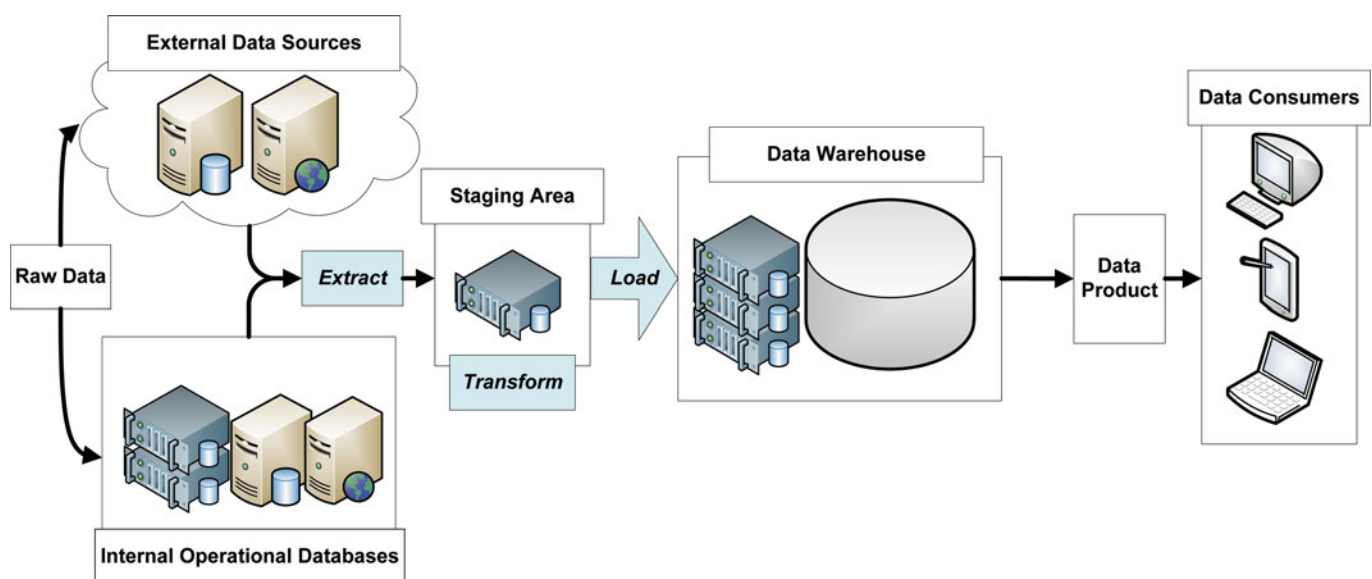


Figure 1. An example data warehouse/data production process.

Table 1. Clarification of terminology used in the aircraft maintenance example

Term	Definition
Record	Row in the aircraft maintenance database that stores personnel hand-entered servicing information on one individual subcomponent.
Data property	Column in the aircraft maintenance database that represents one specific piece of information about an individual subcomponent part, e.g., item serial number.
Field	A single “cell” of information contained at the intersection of a Record and Data Property, e.g., the date of manufacture for a single, specific subcomponent part.
Part number	Manufacturer identification code assigned to a family of subcomponent types. One or more part numbers are present for each type of subcomponent; multiple part numbers are driven by the use of more than one subcomponent manufacturer or use of newer models of the same subcomponent.
Item serial number	Identification code assigned to each individual subcomponent. Our example contains several hundred unique individual item serial numbers.
Date of manufacture	Reported manufacture date for each individual subcomponent.
Time since last overhaul	The amount of time that has elapsed since the most recent previous service event for an individual subcomponent part. Measured in cycles. Each cycle denotes one use of the engine (start/shut-down) in an aircraft.
Status	Code representing the service personnel’s overall judgment of the quality and further potential use of the individual engine subcomponent, based on technical guidance. There are four status levels, ranging from condemned to like-new.

for decision making (i.e., deciding when to remove an engine from an aircraft and send it to the repair facility), to field-level technicians who use historical information for troubleshooting. In sum, once aggregated into the data warehouse, data are used by a variety of consumers for analysis and decision making. Unfortunately, as with any data management program, the process described above presents many opportunities for the data quality to be compromised.

Several researchers have suggested that, like a physical product, data are the end-result of a manufacturing process, with raw data as the input, and a polished, transformed data product as the output (Emery 1969; Huh et al. 1990; Ronen and Spiegler 1991; Arnold 1992; Wang and Kon 1993; Wang, Storey, and Firth 1995; Ballou et al. 1998; March and Hevner 2007). Wang (1998) adapted the define, measure, analyze, improve, control (DMAIC) cycle, as prescribed by Six Sigma and made popular by Motorola, for the data production process, referring to it as the total data quality management (TDQM) methodology. For a discussion of the DMAIC cycle see, for example, Montgomery and Woodall (2008). However, unlike the DMAIC cycle from Six Sigma, there is no control stage recommended in the TDQM cycle. This important omission leaves practitioners with few methods for monitoring, improving, and controlling data quality over time.

Later, we discuss how control charts can be used for monitoring and controlling data quality outcomes; however, many of the current control chart methods are not directly applicable to the unique aspects of the “data about the data” that we observe when monitoring data quality. To gain a complete perspective on data quality, multiple, correlated dimensions are measured and the measures are often categorical variables of mixed types including dichotomous variables, count variables, and sometimes quantitative metrics. Batini et al. (2009) compared many methodologies for improving data quality, and listed several open research issues. These include the identification of more precise statistical methods for assessing data quality, empirical validation of data quality methods, and the extension of existing

data quality methods to handle the unique aspects of unstructured data.

Perhaps the most pertinent, yet challenging obstacle in monitoring data quality relates to the difficulty of measuring the “data about the data” for data quality. A common phrase of quality control practitioners is “you cannot improve that which you cannot measure.” Thus, some attempt must be made to operationally define and measure data quality. As with measuring the quality of a physical product, data quality is a multidimensional problem (Garvin 1984, 1987). In the next section, we review the mainstream literature on the dimensions of data quality to gain insight into the accepted data quality characteristics.

3. DEFINING AND MEASURING DATA QUALITY

A review of the literature regarding measures of data quality reveals more than 60 articles published since 1985 that investigate various categories, dimensions, and impact of data quality on business decision making. This stream of research has shown data quality to be multidimensional and sometimes difficult to measure (Ballou and Pazer 1985; Redman 1996; Wang and Strong 1996; Wand and Wang 1996; Ballou et al. 1998; Huang et al. 1999; Pipino, Lee, and Wang 2002). Despite the volume of research on measures of data quality, the suggested dimensions have only been discussed in very general terms, and we emphasize these dimensions should be operationally defined in order to be applied to specific monitoring situations.

Both Wang and Strong (1996) and Lee et al. (2002) organized data quality dimensions into two categories: *intrinsic* refers to data qualities that are objective and native to the data; *contextual* refers to data qualities that are dependent on the context in which the data are observed or used. Contextual dimensions of data quality lend themselves more toward *information*, formed by placing data within a situation or problem specific context (Davenport and Prusak 2000; Batini et al. 2009; Haug, Arlbjørn, and Pedersen 2009; Watts, Shankaranarayanan, and Even 2009). Contextual dimensions include relevancy, value-added quantity

(Wang and Strong 1996), believability, accessibility, and reputation of the data (Lee et al. 2002, 2004). Measurement of these dimensions has relied heavily on self-report surveys and user questionnaires, as they rely on subjective and situational judgments of decision makers for quantification (Batini et al. 2009). The scope of our article is to consider the quality of data, not information; thus, we limit our study to consider the more general intrinsic measures of data quality discussed next.

Four intrinsic data quality dimensions have received the greatest attention in the data quality research: accuracy, timeliness, consistency, and completeness (Scannapieco and Catarci 2002; Parsian 2006; Batini et al. 2009; Haug and Arlbjørn 2011). Ballou and Pazer (1985) were among the first to investigate methods of measuring the quality of each of these four dimensions, and the view that these four dimensions are the core, intrinsic aspects of data has remained stable since their early work (Wang and Strong 1996; Kahn, Strong, and Wang 2002; Lee et al. 2002). For recent and notable research from which we source our following core-dimension definitions and their proposed measures, see, for example, Batini et al. (2009), Haug and Arlbjørn (2011), and Warth, Kaiser, and Kugler (2011).

Accuracy is one of the oldest aspects of data quality investigated in the literature, with research regarding its measurement and assessment occurring from the 1960s onward (Morey 1982; Laudon 1986). Accuracy refers to data that are as equivalent to their corresponding “real” values as possible (Ballou and Pazer 1985), or simply data that match up to other external values considered to be correct (Redman 1996, pp. 255–256). A simple example would be the degree to which a current data record in our aircraft maintenance database regarding a specific aircraft engine component (including serial number, part number, date of manufacture, etc.) correctly matches an externally comparable database, i.e., the original component manufacturer’s database.

Timeliness implies that data are as up-to-date as possible. Haug, Arlbjørn, and Pedersen (2009) make the argument for the intrinsic nature of the timeliness dimension. Depending on the nature of the data, infrequently updated data records may hamper effective managerial decision making (e.g., outdated data or errors that occur in the data may be missed, potentially preventing the early correction of any operational issues). It is noteworthy that the quality inferred from a measure, such as timeliness, depends heavily on the nature of the data. For example, the original date of manufacture of an aircraft engine component will not change, and need not be updated frequently; however, the date of the most recent cycle time must be kept current. Ballou et al. (1998, p. 468) also demonstrated a timeliness measure calculated by finding the difference in time between record delivery and entry into the storage system. This difference is added to the time elapsed from the occurrence of the real-world event represented by the record.

Consistency accounts for how closely successive and related data records match in terms of format and structure. Ballou and Pazer (1985) defined consistency as when the “representation of the data value is the same in all cases” (p. 153). Batini et al. (2009) developed the notion of both intra-relation and inter-relation constraints on the consistency of data. The former assesses adherence of the data property to a range of possible values, that is, a value domain (Coronel, Morris, and Rob 2011), and the latter requires that tuples/rows from two or more

datasets must represent the data with the same structure. An example of this would be that an aircraft engine component, currently in use, would have for “date of manufacture” a possible value range of 1980–2012 (intra-relation constraint), while that component’s record in two different datasets would, in both cases, have a field for date of manufacture, and both fields would intentionally represent the component’s date of manufacture in the same format (inter-relation constraint).

Completeness refers to data that are full and complete in content, with no missing data. This dimension describes a data record that captures the minimally required amount of information needed in order to technologically reflect some “real-world system” (Wang and Wang 1996), or data that have had all values captured (Gomes, Farinha, and Trigueiros 2007). Every field in the data record is needed in order to capture the complete picture of what the record is attempting to represent in the “real world.” In the aircraft maintenance example, if a particular component’s record includes a date of manufacture and serial number, but no part number, then that record is considered incomplete. The minimum amount of data needed for a complete record is not present. A simple ratio of incomplete versus total records can then form a potential measure of the dataset’s completeness.

A discussion on measures of data quality would not be complete without a brief mention of the importance of metadata to measuring and monitoring data quality. Quite simply, metadata are data about the data (McNurlin and Sprague 2006). Kimball et al. (1998) described metadata as all data in a database that are not the specific data themselves. Metadata are stored in a metadata repository within the warehouse and have been described as consisting of two different types: technical metadata and business metadata (Vetterli, Vaduva, and Staudt 2000). Technical metadata provide information such as field names, field lengths, number of records, etc. Business metadata can be more flexible and are defined to reflect specific aspects of the data that are important to a particular business context. For example, business metadata may include the measures of data quality such as timeliness, completeness, accuracy, and consistency of the data. The metadata repository can be stored as a separate database, or distributed throughout the various toolsets available in a modern data warehousing environment, with implementations varying in practice (Shankaranarayanan and Even 2004). It is a valuable resource for gathering, tracking, and monitoring data quality, as the stored metadata can be used to determine the levels of all four aforementioned quality dimensions. It is within this repository that data about data quality (metadata) can be gathered for quality control applications.

4. STATISTICAL MONITORING OF DATA QUALITY USING CONTROL CHARTS

Of the few published applications of control charts to data quality that we could locate, most used univariate Shewhart charts, and only one recent paper referenced univariate exponentially weighted moving average (EWMA) and cumulative sum (CUSUM) methods. None used multivariate methods to control the multidimensional data quality process, but instead relied on multiple univariate control charts, one (Pierchala et al. 2009) with over 7500 charts applied simultaneously. Because of the dichotomous and/or categorical nature of the data quality

measures, recent applications relied heavily on Shewhart attribute charts. A brief review of the applications of control charting to data quality outcomes suggests many opportunities for researchers to improve upon the methods used in these scenarios.

The earliest use of control charts for monitoring data quality that we could locate dates back to the 1940 census when Deming and Geoffrey (1941) recommended using p -charts for monitoring clerical accuracy in data entry with punch cards. Neter (1952) reviewed the literature on the use of statistical methods applied to monitoring and controlling clerical accuracy in auditing and data entry applications. Examples in Neter (1952) included Bell System companies' use of continuous inspection techniques for posting entries of workers' time reports, Standard Register Company's control of accuracy of sales invoices, and United Airline's use of control charts to monitor accuracy of plane reservations.

More recent applications of Shewhart control charts to data quality are included in the work of Redman (1992, 1996, 2001, 2008) that recommended using p -charts to monitor the proportion of inaccurate (or accurate) records. Pautke and Redman (1990) recommended using Shewhart attribute charts along with a method developed at Bell Labs referred to as tracking. This method requires sampling a set of records entering a data production process and tracking the observations through the entire process. The accuracy of the data is measured during and/or at the end of the process, and control charts can be employed to monitor the accuracy of the records over time. Redman (1992) notes that one problem with their tracking methodology is that, unlike a production process, the observations do not travel through the information system at the same rate; thus, the sample size at the beginning is likely to be reduced substantially when observations are made at the end of the data production process.

Pierchala et al. (2009) published a thorough report on the use of p -charts and c -charts at the National Highway Traffic Safety Administration (NHTSA) to monitor the data quality in the fatality analysis reporting system (FARS). They reported using as many as 7569 control charts annually to monitor data quality for the FARS system, and they adjust the limits to $\pm 5\sigma$ to better control for false alarms. Additionally, they do not directly measure data quality metrics, but measure, for example, the proportion of vehicle occupants who were not wearing seatbelts. Signals on the control charts may be due to changes in the process (e.g., seatbelt law enforcement tactics) or may be due to data quality problems, and must be investigated by the data owners. It would be preferred to use more direct measures of data quality such as the metrics suggested in the next section rather than process measures that confound actual process changes with data quality changes.

Shewhart \bar{X} control charts with time varying limits have been reportedly used to monitor the quality of temperature readings at the California Irrigation Management Information System (CIMS) weather stations (Eching and Snyder 2004). Like the chart applications in the FARS database, direct measures of data quality are not used, rather measures of hourly temperature are taken, and signals outside of the control limits ($\pm 2\sigma$) are considered indicators of data quality problems.

Sparks and OkuGami (2010) considered the use of CUSUM and EWMA charts for flagging biased measures in monitor-

ing the consistency of large volumes of data. They made the distinction between spatial consistency (e.g., similar thickness measures at several locations on a sheet of metal), temporal consistency (e.g., measures consistent with their one-step-ahead forecast values), and multivariate consistency (e.g., measures consistent with related measures, such as rainfall consistent with temperature and humidity). Sparks and OkuGami (2010) recommended using EWMA charts for monitoring departures from temporal consistency, and using CUSUM charts for detecting persistent bias in a measurement device. They also recommended the use of the exponentially weighted moving variance (EWMV) chart to detect persistent increases in variance of a measurement device. Lenz and Borowski (2010) developed an automatic error tracking and correction application based on Oracle's Warehouse Builder (Oracle Corporation 2012). Their method developed a system of rules for defining errors in various data quality dimensions and then used $\pm 3\sigma$ Shewhart-type rules to signal potential data quality problems.

In some of the examples presented above, direct measures of data quality, such as accuracy and consistency, are used in conjunction with control charts to monitor data quality. In others, indirect process measures are monitored with control charts, with signals to process changes possibly confounded with data quality problems. Whenever possible, we recommend the use of direct measures of data quality, with the focus on the intrinsic and contextual measures that are supported through the information technology (IT) literature and discussed in the previous section.

One practical barrier to the use of control charts with direct measures of data quality is that the measures are often correlated dichotomous or count variables (see, e.g., Pipino, Lee, and Wang 2002; Even and Shankaranarayanan 2009; Blake and Mangiameli 2011). Further, for a given data production process, there are often many dimensions and subdimensions of data quality that must be simultaneously monitored (see, e.g., Haug, Arlbjørn, and Pedersen 2009). As noted in the Pierchala et al. (2009) example above, more than 7500 control charts were constructed and monitored for just *one* database. Despite the use of $\pm 5\sigma$ limits, the false alarm rate would be notably high for this application and there would be little power to detect actual process changes. Multivariate control chart methods would help in this situation, but control charts for mixed attribute data types to monitor high-dimensional processes are limited or nonexistent in the literature.

In our own data quality example from the aircraft maintenance database, we have multiple measures of the intrinsic dimensions of data quality that can be observed over time. In the next section, we discuss how data quality is measured in the aircraft maintenance example, and discuss an application of a control chart for monitoring the data quality measures.

5. MEASURING DATA QUALITY: AIRCRAFT MAINTENANCE

For our example, we retrieved the maintenance history for one type of jet engine subcomponent (compressor stage-2 disks) from the aircraft maintenance data management system described earlier. Historical records were collected between 2000 and 2009. During this time, the maintenance and record-keeping procedures had remained unchanged, limiting the likelihood of

ITEM SER	PART NUM	TSNC	NHA SER	SERIAL	ESN	TSOH	BASE	ITEM CEI	DATE	DLOH	TSNH	STAT	CHNG	STATUS
00CAV51873	9661M82G40	314	00EM010077	70000457	00GE441744	12586	ZHTV	GE0020A	8/25/2004		12586	4342		M
00GWNK7780	9661M82G33	1350	00LPE00381	00GE441233	00GE441233	27369	FJXT	GE0020A	4/23/1989	12/21/1999	33615	4008		M
00GWNG2332	9658M84P20	3097	00LPE00686	87000033	00GE441670	34222	YTPM	GE0020A	4/21/1988	9/18/1997	29347	7157		M
00CAV5089	9661M82G40	688	00LPE01067	68000215	00GE441291	25228	ZHTV	GE0020A	12/5/2001		25228	7291		M
00LPE00394	9661M82G33	4430	00LPE01164	00GE441378	00GE441378	5797	UHHZ	GE0020A		9/30/2006	113496	8126		M

Figure 2. Sample rows of maintenance facility data.

changes in data quality over time occurring due to changes in maintenance facility procedures. The example dataset contains data on several hundred individual compressor stage-2 disks. Each record has 14 separate data properties that describe that single disk when the record was created. Within each record, there are multiple data property fields that can be measured with quality metrics of accuracy, completeness, and consistency. Although the aircraft maintenance database did include data properties, such as time since engine overhaul, date of manufacture, and time since subcomponent was new, the database did not contain the date when the actual record was entered into the system. Thus, we are not able to examine timeliness in this example. Measures for each of the three extrinsic data quality dimensions available for this example were guided by the literature and business rules were developed by maintenance personnel. All data properties for each dimension were measured at the field level.

Seven of the 14 data properties for each record were used to measure accuracy. For the remaining seven data properties, inaccuracy measures were either redundant or no standard was available for comparison. Definitions of the data property terms and details of the measures for the entire dataset can be found in the supplementary materials. Air Force technicians heuristically judge the accuracy (or inaccuracy) of a record's data property by comparing its value to that of another variable. For example, for each individual record, the subcomponent's serial number field should not contain the same value as the overall engine serial number field. If the values are equal, a data entry error on the part of maintenance personnel has occurred, and the field is considered to be inaccurate. Following established guidelines, inaccuracy was measured for each data property and defined as

$$IA_{ij} = \begin{cases} 0 & \text{if field is accurate} \\ 1 & \text{if field is inaccurate,} \end{cases}$$

for $i = 1, \dots, 7$ fields, within $j = 1, \dots, N_R$ part records.

Completeness (or incompleteness) was measured according to the incompleteness of each of the 14 possible data property fields in a record. These 14 data properties are detailed in the supplementary materials. The incompleteness is recorded as

$$IC_{ij} = \begin{cases} 0 & \text{if record is complete} \\ 1 & \text{if record is incomplete,} \end{cases}$$

for $i = 1, \dots, 14$ fields and $j = 1, \dots, N_R$ part records.

Next, 13 data properties were used to compute measures of consistency (or inconsistency) for each record. The 13 data properties and definitions of the inconsistency measures are given in the supplementary materials. Several maintenance business rules guided the development of consistency metrics. For example, the values for some data properties were deemed consistent if they contained a value of a certain character length, such as 8

or 10 characters for serial and item numbers, respectively. Other data record properties, such as the date of the last subcomponent overhaul (DLOH), were deemed inconsistent if they were not in a numeric format. In all cases, inconsistency was defined as

$$ICN_{ij} = \begin{cases} 0 & \text{if field is consistent} \\ 1 & \text{if field is inconsistent,} \end{cases}$$

for $i = 1, \dots, 13$ fields, within $j = 1, \dots, N_R$ part records. Figure 2 displays a few sample rows of the data collected from the maintenance facility. Inaccuracies in the SERIAL (SERIAL should not equal ESN) and TSOH (TSOH should not be greater than TSNH) data properties are highlighted in bold. Incomplete fields are observed for the DATE and DLOH properties as indicated by blank fields. The three fields surrounded by a bold border indicate inconsistent fields (ITEM SER should contain a 10-character value, and SERIAL should contain an 8-character value).

The data quality measures for the aircraft maintenance database can be summarized to include 34 dichotomous variables (7 measures of inaccuracy, 14 measures of incompleteness, 13 measures of inconsistency). In the current dataset, there are $N_R = 603$ records. The complete data are provided as supplementary materials to this article. To establish a baseline, we will use the first 547 observations, which were recorded prior to January 1, 2009.

Interestingly, in 20 of the 34 variables, no nonconforming items in data quality were observed in the baseline sample. In a high-quality process, where the probability of observing a nonconforming item is very small, very large baseline samples are often required in order to observe even a single nonconforming. Thus, nonconforming items may be possible in the process, but the baseline sample size is not large enough to observe them. Some authors have approached the problem of no observed nonconforming in the baseline sample by suggesting that any subsequent observation of a nonconforming item would constitute a signal to an out-of-control event (see, e.g., Yang et al. 2002; Chakraborti and Human 2006). Recently, Zhang et al. (2012) suggested a Bayes estimator for the proportion nonconforming when all items in the baseline sample are conforming. Using a beta prior, the Bayes estimator is given by

$$\hat{p}_{0B} = \frac{N + a}{m + a + b},$$

where there are N nonconforming items observed in a sample of m items, and a and b represent prior numbers of observations that are nonconforming and conforming, respectively. Zhang et al. (2012) studied different values for a and b with respect to univariate geometric control charts using estimated parameters. They suggested using $a = 1$ and $b = 999$ (1 prior nonconforming observation out of 1000) because these parameter values resulted

Table 2. Estimated proportion nonconforming from the baseline sample of size $m = 547$ from the aircraft maintenance database. Values indicated with asterisks are maximum likelihood estimates. All other values are Bayes estimates using the method suggested by Zhang et al. (2012)

	Variable	\hat{p}_0
Inaccuracy	IA_1	0.0006
	IA_2	0.3272*
	IA_3	0.1818*
	IA_4	0.4790*
	IA_5	0.1993*
	IA_6	0.2834*
	IA_7	0.4662*
Incompleteness	IC_1	0.0006
	IC_2	0.0006
	IC_3	0.0006
	IC_4	0.0006
	IC_5	0.0006
	IC_6	0.0006
	IC_7	0.0006
	IC_8	0.0006
	IC_9	0.0006
	IC_10	0.2687*
	IC_11	0.1042*
	IC_12	0.0018*
	IC_13	0.0006
Inconsistency	ICN_1	0.0018*
	ICN_2	0.0006
	ICN_3	0.0006
	ICN_4	0.0006
	ICN_5	0.3272*
	ICN_6	0.0006
	ICN_7	0.0006
	ICN_8	0.0006
	ICN_9	0.0006
	ICN_10	0.2687*
ICN_11	0.1042*	
ICN_12	0.0018*	
ICN_13	0.0006	

ables (Cohen et al. 2003, pp. 30–31), and can be computed from a 2×2 table as follows

	X2 = 0	X2 = 1	
X1 = 0	A	B	A+B
X1 = 1	C	D	C+D
	A+C	B+D	

$$\phi = \frac{AD - BC}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$$

The coefficient, ϕ , has a similar interpretation to Pearson’s correlation coefficient, with high positive values indicating agreement (many incidents of $X1 = X2 = 0$, or $X1 = X2 = 1$), and low negative values indicating disagreement (many incidents of $X1 = 0$ and $X2 = 1$, or $X1 = 1$ and $X2 = 0$). The ϕ coefficient does not exist in cases in which no nonconforming items are observed in a sample; thus, Table 3 gives the mean square contingency coefficients for 14 of the 34 variables. It is noteworthy that several pairs of variables are perfectly correlated. For example, for cases in which an inaccurate (or accurate) value is observed in field two (IA_2), an inconsistent (or consistent) value is also observed in field 5 (ICN_5). The variable IA_2 corresponds to an inaccurate item serial number, and ICN_5 corresponds to an inconsistent end item serial number. One possibility is that this collinearity is artificially created by the rules-based measurement system. However, further investigation into the rules showed that the variables are, indeed, measuring different aspects of the data quality, yet inaccurate item serial numbers and inconsistent serial numbers tend to occur simultaneously in this process. For both IA_2 and ICN_5, 33% of the observed fields were nonconforming in the baseline sample, indicating a high potential for improvement in the data quality regarding item serial numbers. Similar collinearity was observed between IC_10 and ICN_11, IC_11 and ICN_12, as well as IC_12 and ICN_13. For other variables, the mean square contingency coefficients ranged from near zero to as high as 0.96 in absolute value.

6. MONITORING DATA QUALITY: AIRCRAFT MAINTENANCE

The aircraft maintenance data example contains 34 dichotomous variables, 20 of which presented no nonconforming items in the baseline sample of size $m = 547$. Further inspection of Table 3 reveals perfect associations between four pairs of the variables in which nonconforming items were observed, with other pairs of variables ranging from nominally to significantly related. The first step to establishing a monitoring scheme for this process is to search the literature for methods that may be directly applicable to this process. Thus, we begin with an overview of existing methods that might apply to this specific aircraft maintenance data quality example. First we discuss methods for establishing a baseline, in-control reference sample. This is followed with a discussion of prospective monitoring tools that might apply to this process. Although our review of the literature reveals that there are few directly applicable control chart methods that can be readily applied to our

in more consistent control chart performance when the observed proportion defective was less than the true proportion defective.

Because we observed no nonconforming items in 20 of the variables, we use the recommended values of $a = 1$ and $b = 999$ with the Bayes estimator of Zhang et al. (2012) to estimate the proportion nonconforming in these 20 variables. We use the maximum likelihood estimator ($\frac{N}{m}$) to estimate the proportion nonconforming in the remaining 14 variables. The estimates of the proportion nonconforming are given in Table 2, with asterisks indicating those computed using the maximum likelihood estimator. Close inspection shows that in many of the data quality variables, nonconforming items are estimated to occur very rarely. However, in some cases, the proportion of nonconforming fields is alarmingly high.

In order to assess the degree of correlation among the variables in our example, we computed the mean square contingency coefficients between the variables. The mean square contingency coefficient (also known as the phi coefficient) is a special case of Pearson’s correlation coefficient for dichotomous vari-

Table 3. Mean square contingency coefficients from the baseline sample of $n = 547$ observations from the aircraft maintenance database. The mean square contingency coefficients are measures of association for dichotomous variables (Cohen et al. 2003, pp. 30–31). Bold font indicates absolute coefficients greater than 0.80, italics indicate absolute coefficients in the range of [0.40, 0.80)

	IA_2	IA_3	IA_4	IA_5	IA_6	IA_7	IC_10	IC_11	IC_12	ICN_1	ICN_5	ICN_11	ICN_12	ICN_13
IA_2	1.00													
IA_3	-0.06	1.00												
IA_4	-0.13	-0.12	1.00											
IA_5	0.18	-0.08	<i>-0.40</i>	1.00										
IA_6	0.05	-0.26	0.08	0.01	1.00									
IA_7	0.10	0.06	-0.21	0.09	<i>0.67</i>	1.00								
IC_10	0.04	-0.28	0.09	0.00	0.96	<i>0.65</i>	1.00							
IC_11	0.07	<i>0.46</i>	-0.33	-0.04	-0.13	0.36	-0.13	1.00						
IC_12	0.06	0.09	-0.04	0.09	0.07	0.05	-0.03	-0.01	1.00					
ICN_1	-0.03	-0.02	0.04	-0.02	-0.03	-0.04	-0.03	-0.01	0.00	1.00				
ICN_5	1.00	-0.06	-0.13	0.18	0.05	0.10	0.04	0.07	0.06	-0.03	1.00			
ICN_11	0.04	-0.28	0.09	0.00	0.96	<i>0.65</i>	1.00	-0.13	-0.03	-0.03	0.04	1.00		
ICN_12	0.07	<i>0.46</i>	-0.33	-0.04	-0.13	0.36	-0.13	1.00	-0.01	-0.01	0.07	-0.13	1.00	
ICN_13	0.06	0.09	-0.04	0.09	0.07	0.05	-0.03	-0.01	1.00	0.00	0.06	-0.03	-0.01	1.00

aircraft maintenance data quality example, we close the section with an example analysis of the aircraft maintenance data. This discussion highlights many opportunities for research pertaining to control chart use in monitoring and improving data quality.

6.1 Phase I Analysis

In Phase I, a reference sample is retrospectively studied to identify the in-control state of the process and to estimate the process parameters. In Phase II, the process is prospectively monitored for departures from the in-control state. Ideally, we should consider a Phase I analysis to establish the in-control state of the process and to estimate the process parameters. Then we use the results of this Phase I analysis to establish an appropriate monitoring scheme for our process variables. Numerous researchers have studied the retrospective use of control charts and shown that control charts designed for Phase II monitoring perform poorly when used retrospectively (see, e.g., Yang and Hillier (1970), Borror and Champ (2001), Nedumaran and Pignatiello (2005), and Jones-Farmer, Jordan, and Champ (2009)). Thus, it is necessary to consider methods developed specifically for use in Phase I.

We found no Phase I method for analyzing *multivariate* attribute data. In fact, we found only very sparse information for conducting a Phase I analysis with *univariate* attribute data. Borror and Champ (2001) studied the Phase I use of univariate p -charts, and showed that they did not work well, with false alarms occurring with near certainty for reference samples of size $m = 500$. In their study of univariate geometric control charts with estimated parameters, Zhang et al. (2012) noted that very large in-control sample sizes are needed to accurately estimate the process parameters. There is a clear need in the literature for more research on Phase I methods for both univariate and multivariate attribute processes.

In another stream of literature, several authors have investigated the use of both model-based and model-free clustering methods for either conducting a retrospective Phase I analysis or establishing a prospective (Phase II) monitoring scheme, or both. Although none of the clustering methods fit the aircraft

maintenance data quality scenario we describe here, the philosophy behind the approaches may be useful for motivating research in multivariate Phase I methods.

Thissen et al. (2005) used Gaussian mixture models to establish an in-control reference sample. Mixture modeling is a model-based clustering method that can be used alone or in conjunction with regression models where the distribution of the independent variables is considered a mixture of two or more distributions that may differ in location, scale, correlation structure, or all three. For more details on mixture modeling, the interested reader is referred to McLachlan and Peel (2000), or Fraley and Raftery (2002). Although the model-based control chart method used by Thissen et al. (2005) is developed for continuous multivariate distributions, it is possible that a similar framework could be developed for multivariate discrete or attribute distributions. For example, Muthén, du Toit, and Spisic (1997) showed that mixture modeling could be generalized using weighted least squares to include binary, ordered, and continuous outcomes. Another possible research opportunity is to investigate the use of mixture modeling for Phase I analysis of processes measured by continuous variables, discrete variables, or a mixture of variables of different types. Noteworthy limitations to this approach should be considered, including the possible requirements of impractically large sample sizes, and/or the possible identification of a set of discontinuous in-control baseline samples.

Some model-free clustering methods have been applied to quality control scenarios, adaptations of which may be useful in the data quality framework. Sullivan (2002) introduced a clustering method to detect multiple outliers in a univariate continuous process. More recently, Zhang et al. (2010) introduced a univariate clustering-based method for finding an in-control reference sample from a long historical stream of univariate continuous data. Jobe and Pokojov (2009) introduced a computer intensive multi-step clustering method for retrospective outlier detection in multivariate processes. Although model-free clustering methods have been used for some univariate and multivariate Phase I applications, there remain many opportunities for research in this area. In particular, there is a need to investigate the practicality of model-free clustering methods for Phase

I analysis of univariate and multivariate attribute processes, as well as multivariate processes with mixed data types. There are many opportunities to investigate the strengths and limitations of these methods, including sample size and the appropriateness of the methods in terms of correctly identifying a meaningful in-control baseline sample.

6.2 Phase II Analysis

Although the Phase II multivariate control chart literature is significantly more developed than that of the Phase I multivariate control chart literature, we found few methods that were directly applicable to our example data. Here we consider the most relevant of the existing methods for monitoring multivariate attribute process, and highlight more opportunities for research in this important area.

Totalidou and Psarakis (2009) conducted an extensive literature review on multinomial and multiattribute control charts, most of which are applicable for Phase II monitoring. An early contributor in this area is Patel (1973) who suggested a Shewhart-type chart for monitoring multivariate binomial or Poisson processes using an approach similar to Hotelling's T^2 . One common monitoring approach when there are multiple types of defects in a product is the demerit chart in which a weighted sum of the defect counts is monitored (see, e.g., Montgomery 2013, pp. 30–31; or Jones, Woodall, and Conerly 1999). Lu et al. (1998) considered the design of an np -chart for multiple attributes based on a weighted sum of nonconforming units from multiple binomial variables, similar to the demerit charts. Demerit-like methods, which give heavier weights to more severe defects, are not desirable to use in this example because the establishment of the weights is arbitrary, and would be difficult to apply in the data quality scenario. It would be difficult to determine the severity of defects of a certain type (e.g., incomplete vs. inaccurate fields).

Other approaches to monitoring multiple attributes include use of chi-square-based charts for contingency-type analyses when several categories are possible including Duncan (1950), Marcucci (1985), and Nelson (1987). More recently, Ryan, Wells, and Woodall (2011) noted that a limitation of the traditional chi-square charting approach is that one must accumulate the observations into subgroups, delaying the time until a signal is observed when a quality problem is present. They proposed a multinomial chart which is an extension of Reynolds and Stoumbos' (1999) CUSUM chart for continuous inspection of a Bernoulli process. Ryan, Wells, and Woodall (2011) show that their method is preferred to using multiple Bernoulli CUSUM charts when the process shift is in the direction for which the chart is designed and recommended using multiple Bernoulli CUSUM charts when the shift direction could not be specified. Most of these multinomial-type charts have been studied on relatively small-dimensional problems; for example, Ryan, Wells, and Woodall (2011) considered up to four dimensions, noting that future research should study the applicability of their method in higher dimensions.

Recently, Li, Tsung, and Zou (2012) proposed the use of log-linear models in conjunction with an exponentially weighted moving average (EWMA) monitoring scheme for prospective control chart monitoring. Their method is strictly applicable to

prospective monitoring, and, requires a large in-control reference sample to establish the monitoring scheme. Although Li, Tsung, and Zou's (2012) log-linear model-based chart fills an important niche in multivariate control charts for categorical data, it is similar to the traditional multinomial charts in that large subgroups of observations must be accumulated prior to plotting a point on the chart, delaying the time until a signal to a problem can be observed. Additionally, Li, Tsung, and Zou (2012) developed and illustrated their method on a scenario with a relatively small dimension ($p = 3$) compared to our data quality example scenario ($p = 34$). An important area for future research is to investigate the scalability of the existing Phase II control charts for multiple attribute processes in higher dimensions (>4). In particular, it would be useful to make comparison of the statistical performance, practicality, and interpretability of the multivariate attribute methods to the use of multiple univariate charts, giving recommended approaches for practitioners.

Another stream of SPC methodology that may present opportunities for use in data quality monitoring are the techniques based on supervised learning (see, e.g., Cook and Chiu 1998; Chinnam 2002; Sun and Tsung 2003; Hwang, Runger, and Tuv 2007; Deng, Runger, and Tuv 2012). Several of these methods can be used with mixed data types and data measured using different measurement scales. Usually, an artificial dataset is generated to represent out-of-control data, which is used to train a classifier; thus converting the monitoring problem to a classification problem. Recently, Deng, Runger, and Tuv (2012) introduced a supervised learning approach based on continual updates of the classifier with real-time data. Their method is appropriate for use with multidimensional data of mixed types and may be applicable to very large data environments with rapidly occurring data. There is an opportunity to investigate the SPC methods based on supervised learning in terms of applicability for Phase I and Phase II use for high-dimensional multivariate process with attribute measures.

6.3 Example Analysis

We have highlighted a growing literature on multivariate techniques for attributes processes; however, all would require substantial adaptations and further study to be applicable to a scenario like the one presented with the aircraft maintenance data. We will make the assumption that the first $m = 547$ observations were gathered under conditions of process stability, although we admit that this assumption is untested with a Phase I study. Further, despite the multivariate nature of the process, because we have no prespecified shift direction, we will follow the recommendation of Ryan, Wells, and Woodall (2011) and construct multiple Bernoulli CUSUM charts to monitor the process. For details on constructing Bernoulli CUSUM charts, see Reynolds and Stoumbos (1999). Because the data are not gathered in subgroups, a continuous monitoring scheme for Bernoulli observations is appropriate, although our analysis is limited to the use of multiple univariate charts in place of a multivariate monitoring method.

Figure 3 shows Bernoulli CUSUM charts for the incompleteness measures for 3 of the 14 fields measured. The charts for the variables, IC_10, IC_11, and IC_12, represent the respective completeness of the date field, the field containing DLOH, and

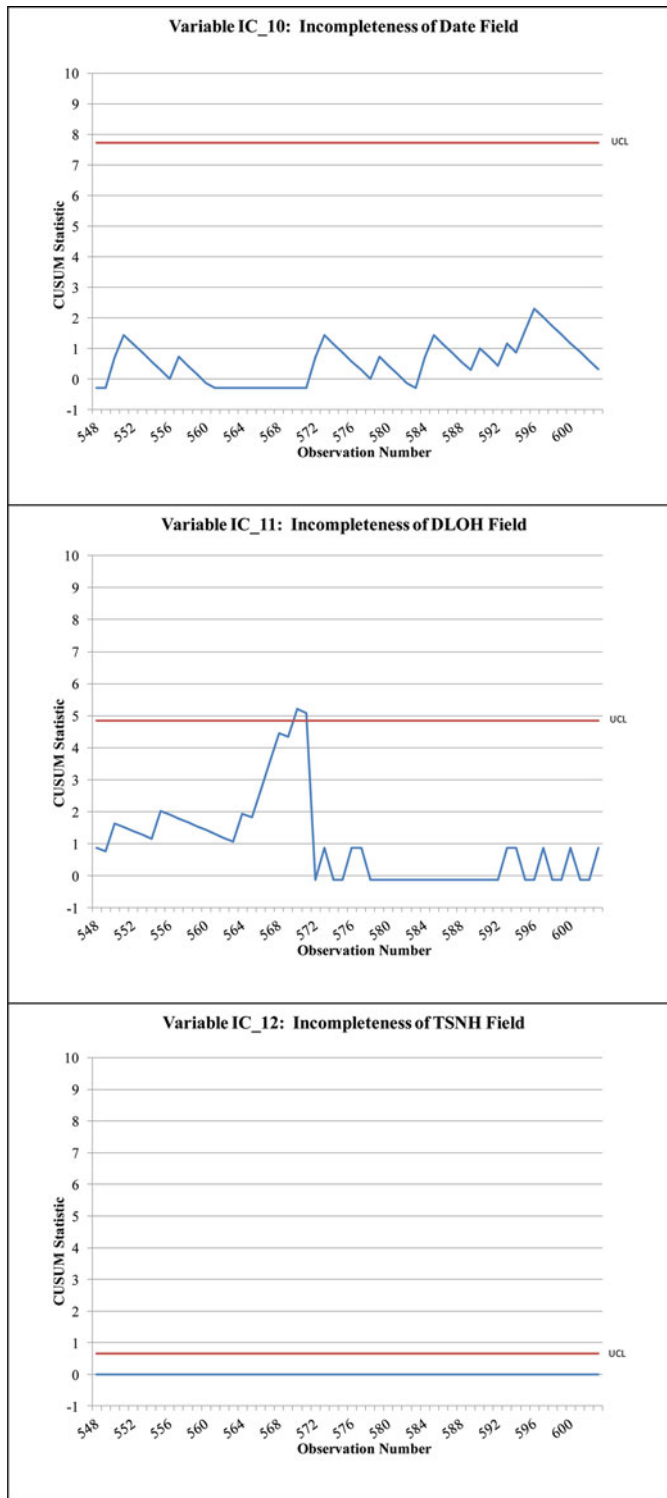


Figure 3. Bernoulli CUSUM charts for three of the measures of incompleteness in the Air Force maintenance database example. Note that, unlike more conventional CUSUM charts, the definition of the CUSUM statistic in Reynolds and Stoumbos (1999) does not immediately reset to zero, thus negative values are possible.

the time since the part was new in hours (TSNH). Each CUSUM was constructed according to the guidelines given in Reynolds and Stoumbos (1999) and was designed for an average number of observations to signal (ANOS) of 500. The initial estimates of the proportion nonconforming were obtained from the ref-

erence sample of size $m = 547$, and these values are given in Table 2. While there are 34 similar charts to fully address all variables in the example, we selected the incompleteness dimensions on these fields to illustrate the charts on variables with a very high proportion of nonconforming, a medium proportion nonconforming, and a very low proportion nonconforming.

For the date field, the Bernoulli CUSUM was tuned to detect a shift from $p_0 = 0.27$ to $p_1 = 0.30$. The process remains in-control for observations 548 through 603. For the DLOH field, the charts were tuned to detect a shift from $p_0 = 0.10$ to $p_1 = 0.14$. The chart signaled an out of control event at observation 570. Investigation into the process indicated that maintenance personnel had deviated from established guidelines regarding the entry of Julian dates representing the “Last Overhaul Date” of the engine subcomponent. Corrective action was taken and the CUSUM was reset following observation 572. The process remained in-control for the rest of the observation period. The CUSUM chart for TSNH was tuned to detect a shift from $p_0 = 0.0018$ to $p_1 = 0.0036$. The chart is uninteresting because all fields were conforming during observation period. This is not unusual, given the lack of nonconforming items in the reference sample, and the very low estimated proportion nonconforming.

7. CONCLUDING REMARKS

We have presented a unified review of the relevant literature on data quality from the perspective of process improvement. This review includes literature from the IT field that frames data production as a process and suggests measures of data quality. Although some differences exist between data production and the production of, say, an automobile, we found some similarities in the way in which the end product may be measured and monitored.

We included a review of SPC monitoring methods that have been applied to the data quality problem in the literature. We discussed the challenges of measuring data quality and gave examples of how we measured data quality in the framework of a real aircraft maintenance database example. We then explored the literature, seeking methods that might be applicable for defining an in-control reference sample and establishing a monitoring scheme for our example. This led us to many gaps in the literature that provide opportunities for future research in control chart methods.

Many of these open questions are highlighted in Section 6, and include studies of both Phase I and Phase II methods for multivariate attribute processes. As control charts are used to monitor data quality, we anticipate many more interesting research questions to arise from practice. For example, when a signal is given suggesting a decrease in incompleteness, this might suggest a process improvement. It might also be “too good to be true,” suggesting that data are being falsified. Will the control charts methods be useful in distinguishing true signals from falsified data?

For the purposes of our example, we considered the quality of the data for a single subcomponent. We admit that this is an oversimplification that we used in order to make the problem reasonably straightforward for discussion, and note that even with this oversimplification we found no existing statistical methods that could be readily applied to our problem. In

addition, when considering the quality of the data in an entire database, one might find a natural hierarchical structure to the data. Here, subcomponents are nested within engines and parts are nested within subcomponents. It is possible that the data, and possibly the data quality, may correlate due to this or some other hierarchical structure. We suspect that some maintenance and transactional databases may follow a similar hierarchical structure. Thus, future research should consider appropriate ways to model the quality of data in a hierarchical structure and to monitor for continuous data quality improvement.

Finally, the scalability of the methods to large databases will certainly bring new opportunities for research concerning methods for automating the measurement of the quality dimensions, acceptable rates of false signals, and signal interpretation. Although the dataset used in our example provided us with an opportunity to clearly demonstrate the use of control charts for monitoring data quality, the smaller size of the example dataset might be seen as a limitation to the scope of these methods. As more data are acquired and stored, the scope of the data quality problem will likely increase and new monitoring procedures will need to be developed, but the principles described herein remain.

We hope that the new way of thinking about the data quality problem that we present in this paper will encourage collaboration between SPC and IT experts to further develop tangible solutions. As this article attests, examining this problem requires expertise in both areas: the IT expert provides insight into how data are collected, stored, processed, and retrieved, and the SPC expert plays an integral role in understanding the latest statistical techniques and procedures to measure, monitor, and control these processes. Working together, relevant samples can be extracted at the right place and time and meaningful process improvement efforts may emerge.

SUPPLEMENTARY MATERIALS

Technical details and dataset: An Excel spreadsheet file contains (1) a legend explaining the abbreviations used for each of the 14 data properties considered in the aircraft maintenance example, (2) details on how each of the 34 data quality measures were generated from the database, and (3) data quality measures for 603 records from the aircraft maintenance database.

ACKNOWLEDGMENTS

The authors thank the editor, associate editor, and two anonymous referees for their editorial direction that have resulted in significant improvements in the article. We also wish to thank Mr. B. Page Farmer, Jr., Senior Information Technology Architect, IBM, and Mr. Darrell Bilbrey, Vice President of Corporate Systems, HealthSouth Corporation, whose comments during the early development of this work shaped many of our ideas.

[Received January 2012. Revised February 2013.]

REFERENCES

Arnold, S. E. (1992), "Information Manufacturing: The Road to Database Quality," *Database*, 15, 32–39. [31]

- Ballou, D. P., and Pazer, H. L. (1985), "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems," *Management Science*, 31, 150–162. [31,32]
- Ballou, D. P., Wang, R., Pazer, H., and Tayi, G. K. (1998), "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science*, 44, 462–484. [31,32]
- Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009), "Methodologies for Data Quality Assessment and Improvement," *Association for Computing Machinery Computing Surveys*, 41, 1–52. [31,32]
- Blake, R., and Mangiameli, P. (2011), "The Effects and Interactions of Data Quality and Problem Complexity on Classification," *Association for Computing Machinery Journal of Data and Information Quality*, 2, 1–28. [33]
- Borrer, C. M., and Champ, C. W. (2001), "Phase I Control Charts for Independent Bernoulli Data," *Quality and Reliability Engineering International*, 17, 391–396. [36]
- Bose, R. (2009), "Advanced Analytics: Opportunities and Challenges," *Industrial Management + Data Systems*, 109, 155. [29]
- Breslin, M. (2004), "Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models," *Business Intelligence Journal*, 9, 6–20. [30]
- Chakraborti, S., and Human, S. (2006), "Parameter Estimation and Performance of the Chart for Attributes Data," *IEEE Transactions on Reliability*, 55, 559–566. [34]
- Chinnam, R. B. (2002), "Support Vector Machines for Recognizing Shifts in Correlated and Other Manufacturing Processes," *International Journal of Production Research*, 40, 4449–4466. [37]
- Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.), Mahwah, NJ: Lawrence Erlbaum Associates, Inc. [35]
- Cook, D. F., and Chiu, C. C. (1998), "Using Radial Basis Function Neural Networks to Recognize Shifts in Correlated Manufacturing Process Parameters," *IIE Transactions*, 30, 227–234. [37]
- Coronel, C., Morris, S., and Rob, P. (2011), *Database Systems: Design, Implementation, and Management*, Boston, MA: Cengage Learning. [32]
- Davenport, T. H., and Prusak, L. (2000), *Working Knowledge: How Organizations Manage What They Know*, Boston, MA: Harvard Business Press. [31]
- Deming, W. E., and Geoffrey, L. (1941), "On Sample Inspection in the Processing of Census Returns," *Journal of the American Statistical Association*, 36, 351–360. [33]
- Deng, H., Runger, G. C., and Tuv, E. (2012), "Systems Monitoring with Real Time Contrasts" *Journal of Quality Technology*, 44, 9–27. [37]
- Duncan, A. J. (1950), "A Chi-Square Chart for Controlling a Set of Percentages," *Industrial Quality Control*, 7, 11–15. [37]
- Eching, S. O., and Snyder, R. L. (2004), "Statistical Control Charts for Quality Control of Weather Data for Reference Evapotranspiration Estimation," in *ISHS Acta Horticulturae IV International Symposium on Irrigation of Horticultural Crops*, pp. 189–196. [33]
- Emery, J. C. (1969), *Organizational Planning and Control Systems: Theory and Management*, New York, NY: Macmillan. [31]
- Even, A., and Shankaranarayanan, G. (2009), "Dual Assessment of Data Quality in Customer Databases," *Journal of Data and Information Quality*, 1, 1–29. [33]
- Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [36]
- Garvin, D. A. (1984), "What Does 'Product Quality' Really Mean?," *Sloan Management Review*, 26, 25–43. [31]
- (1987), "Competing on the Eight Dimensions of Quality," *Harvard Business Review*, 65, 101–109. [31]
- Gomes, P., Farinha, J., and Trigueiros, M. J. (2007), "A Data Quality Metamodel Extension to Cwm," in *Australian Computer Society, Inc., 4th Asia-Pacific Conference on Conceptual Modeling*, pp. 17–26. [32]
- Haug, A., and Arlbjörn, J. S. (2011), "Barriers to Master Data Quality," *Journal of Enterprise Information Management*, 24, 288–303. [32]
- Haug, A., Arlbjörn, J. S., and Pedersen, A. (2009), "A Classification Model of Erp System Data Quality," *Industrial Management & Data Systems*, 109, 1053. [31,32,33]
- Huang, K., Lee, Y., and Wang, R. Y. (1999), *Quality Information and Knowledge*, Saddle River, NJ: Prentice Hall. [31]
- Huh, Y. U., Keller, F. R., Redman, T. C., and Watkins, A. R. (1990), "Data Quality," *Information and Software Technology*, 32, 559–565. [31]
- Hwang, W., Runger, G., and Tuv, E. (2007), "Multivariate Statistical Process Control With Artificial Contrasts," *IIE Transactions*, 39, 659–669. [37]
- IBM (2011), "Yale School of Management and Ibm Collaborate to Prepare Students with Analytics Skills for the Next Generation of Jobs," PR

- Newswire Association LLC. [online], available at <http://www.prnewswire.com/news-releases/yale-school-of-management-and-ibm-collaborate-to-prepare-students-with-analytics-skills-for-the-next-generation-of-jobs-120836949.html> [29]
- Inmon, W. H. (1992), *Building the Data Warehouse*, New York, NY: Wiley. [30]
- Jobe, J. M., and Pokojov, M. (2009), "A Multistep, Cluster-Based Multivariate Chart for Retrospective Monitoring of Individuals," *Journal of Quality Technology*, 41, 323–339. [36]
- Jones-Farmer, L. A., Jordan, V., and Champ, C. W. (2009), "Distribution-Free Phase I Control Charts for Subgroup Location," *Journal of Quality Technology*, 41, 304–316. [36]
- Jones, L. A., Woodall, W. H., and Conerly, M. D. (1999), "Exact Properties of Demerit Control Charts," *Journal of Quality Technology*, 31, 207–216. [37]
- Kahn, B. K., Strong, D. M., and Wang, R. Y. (2002), "Information Quality Benchmarks: Product and Service Performance," *Communications of the Association for Computing Machinery*, 45, 184–192. [29,32]
- Kalakota, R. (2012), "Gartner Says - Bi and Analytics a \$12.2 Bln Market," *Business Analytics 3.0* [online], available at <http://practicalanalytics.wordpress.com/2011/04/24/gartner-says-bi-and-analytics-a-10-5-bln-market/> [29]
- Kimball, R., Ross, M., Thorthwaite, W., Becker, B., and Mundy, J. (1998), *The Data Warehouse Lifecycle Toolkit*, New York, NY: Wiley-India. [32]
- Laudon, K. C. (1986), "Data Quality and Due Process in Large Interorganizational Record Systems," *Communications of the Association for Computing Machinery*, 29, 4–11. [32]
- Lee, Y. W., Pipino, L., Strong, D. M., and Wang, R. Y. (2004), "Process-Embedded Data Integrity," *Journal of Database Management*, 15, 87(17). [32]
- Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002), "Aimq: A Methodology for Information Quality Assessment," *Information & Management*, 40, 133–146. [31,32]
- Lenz, H.-J., and Borowski, E. (2010), "Business Data Quality Control – A Step by Step Procedure," in *Frontiers in Statistical Quality Control* (Vol. 10), eds. H.-J. Lenz, W. Schmid and P.-T. Wilrich, Heidelberg: Springer, pp. 371–383. [33]
- Li, J., Tsung, F., and Zou, C. (2012), "Directional Control Schemes for Multivariate Categorical Processes," *Journal of Quality Technology*, 44, 136–154. [37]
- Lohr, S. (2009), "For Today's Graduate, Just One Word: Statistics," *NYTimes.com* [online], available at <http://www.nytimes.com/2009/08/06/technology/06stats.html?hpw> [29]
- Lu, X. S., Xie, M., Goh, T. N., and Lai, C. D. (1998), "Control Chart for Multivariate Attribute Processes," *International Journal of Production Research*, 36, 3477–3489. [37]
- March, S. T., and Hevner, A. R. (2007), "Integrated Decision Support Systems: A Data Warehousing Perspective," *Decision Support Systems*, 43, 1031–1043. [30,31]
- Marcucci, M. (1985), "Monitoring Multinomial Processes," *Journal of Quality Technology*, 17, 86–91. [37]
- McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York, NY: Wiley. [36]
- McNurlin, B. C., and Sprague, R. H. (2006), *Information Systems Management in Practice* (7th ed.), Upper Saddle River, NJ: Pearson Prentice Hall. [32]
- Montgomery, D. C., and Woodall, W. H. (2008), "An Overview of Six Sigma," *International Statistical Review*, 76, 329–346. [31]
- Montgomery, D. M. (2013), *Introduction to Statistical Quality Control* (7th ed.), New York: Wiley. [37]
- Morey, R. C. (1982), "Estimating and Improving the Quality of Information in a Mis," *Communications of the Association for Computing Machinery*, 25, 337–342. [32]
- Muthén, B., du Toit, S. H. C., and Spisic, D. (1997), "Robust Inference Using Weighted Least Squares and Quadratic Estimating Equations in Latent Variable Modeling With Categorical and Continuous Outcomes," accepted for publication in *Psychometrika*, preprint available at http://pages.gseis.ucla.edu/faculty/muthen/articles/Article_075.pdf [36]
- Nedumaran, G., and Pignatiello, Jr., J. J. (2005), "On Constructing Retrospective-X Control Chart Limits," *Quality and Reliability Engineering International*, 21, 81–89. [36]
- Nelson, L. S. (1987), "A Chi-Square Control Chart for Several Proportions," *Journal of Quality Technology*, 19, 229–231. [37]
- Neter, J. (1952), "Some Applications of Statistics for Auditing," *Journal of the American Statistical Association*, 47, 6–24. [33]
- Nunnally, B., and Thoele, B. (2007), "Logistics Analysis," *Air Force Journal of Logistics*, 31, 124–127. [29]
- Oracle Corporation. (2012), *Oracle Warehouse Builder*, Redwood Shores, CA: Oracle Corporation. [33]
- Parssian, A. (2006), "Managerial Decision Support With Knowledge of Accuracy and Completeness of the Relational Aggregate Functions," *Decision Support Systems*, 42, 1494–1502. [32]
- Patel, H. I. (1973), "Quality Control Methods for Multivariate Binomial and Multivariate Poisson Distributions," *Technometrics*, 15, 103–112. [37]
- Pautke, R., and Redman, T. (1990), "Techniques to Control and Improve Quality of Data Large Databases," in the *Proceedings of the Statistics Canada Symposium 90: Measurement and Improvement of Data Quality*, pp. 319–333. [33]
- Pierchala, C. E., Surti, J., Peytcheva, E., Groves, R. M., Kreuter, F., Kohler, U., Chipperfield, J. O., Steel, D. G., Graham, P., and Young, J. (2009), "Control Charts as a Tool for Data Quality Control," *Journal of Official Statistics*, 25, 167–191. [32,33]
- Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002), "Data Quality Assessment," *Communications of the Association for Computing Machinery*, 45, 211–218. [31,33]
- Pitt, L. F., Watson, R. T., and Kavan, C. B. (1995), "Service Quality: A Measure of Information System Effectiveness," *Management Information Systems Quarterly*, 19, 173–188. [29]
- Redman, T. C. (1992), *Data Quality: Management and Technology*, New York, NY: Bantam Books. [33]
- (1996), *Data Quality for the Information Age*, Norwood, MA: Artech House Publishers. [31,32,33]
- (2001), *Data Quality: The Field Guide*, Boston, MA: Digital Press. [33]
- (2008), "To Solve This Data-Driven Crises, We Need Better Data," Harvard Business School Press: Discussion Starter [online]. Available at http://blogs.hbr.org/cs/2008/09/we_need_better_data_to_solve_it.html [33]
- Reynolds, Jr., M. R., and Stoumbos, Z. G. (1999), "A Cusum Chart for Monitoring a Proportion When Inspecting Continuously," *Journal of Quality Technology*, 31, 87–108. [37]
- Ronen, B., and Spiegler, I. (1991), "Information as Inventory : A New Conceptual View," *Information & Management*, 21, 239–247. [31]
- Ryan, A. G., Wells, L. J., and Woodall, W. H. (2011), "Methods of Monitoring Multiple Proportions When Inspecting Continuously," *Journal of Quality Technology*, 43, p. 237. [37]
- Scannapieco, M., and Catarci, T. (2002), "Data Quality Under a Computer Science Perspective," *Archivi & Computer*, 2, 1–15. [32]
- Shankaranarayanan, G., and Even, A. (2004), "Managing Metadata in Data Warehouses: Pitfalls and Possibilities," *Communications of the Association for Information Systems*, 14, 13. [32]
- Sparks, R., and OkuGami, C. (2010), "Data Quality: Algorithms for Automatic Detection of Unusual Measurements," in *Frontiers in Statistical Quality Control* (Vol. 10), eds. H.-J. Lenz, W. Schmid and P.-T. Wilrich, Heidelberg: Springer, pp. 371–383. [33]
- Sullivan, J. H. (2002), "Detection of Multiple Change Points From Clustering Individual Observations," *Journal of Quality Technology*, 34, 371–383. [36]
- Sun, R., and Tsung, F. (2003), "A Kernel-Distance-Based Multivariate Control Chart Using Support Vector Methods," *International Journal of Production Research*, 41, 2975–2989. [37]
- Tayi, G.K., and Ballou, D.P. (1998), "Examining Data Quality," *Communications of the ACM*, 42, 54–57. [29]
- Thissen, U., Swierenga, H., De Weijer, A., Wehrens, R., Melssen, W., and Buydens, L. (2005), "Multivariate Statistical Process Control Using Mixture Modelling," *Journal of Chemometrics*, 19, 23–31. [36]
- Topalidou, E., and Psarakis, S. (2009), "Review of Multinomial and Multivariate Quality Control Charts," *Quality and Reliability Engineering International*, 25, 773–804. [37]
- Vetterli, T., Vaduva, A., and Staudt, M. (2000), "Metadata Standards for Data Warehousing: Open Information Model Vs. Common Warehouse Metadata," *Association for Computing Machinery Special Interest Group on Management of Data Record*, 29, 68–75. [32]
- Wand, Y., and Wang, R. Y. (1996), "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the Association for Computing Machinery*, 39, 86–95. [31,32]
- Wang, R. Y. (1998), "A Product Perspective on Total Data Quality Management," *Communications of the Association for Computing Machinery*, 41, 58–65. [31]
- Wang, R. Y., and Kon, H. B. (1993), "Towards Total Data Quality Management (Tdqm)," in *Information Technology in Action: Trends and Perspectives*, ed. R. Y. Wang, Englewood Cliffs, NJ: Prentice-Hall. [31]
- Wang, R. Y., Storey, V. C., and Firth, C. P. (1995), "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, 7, 623–640. [31]

- Wang, R. Y., and Strong, D. M. (1996), "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, 12, 5–33. [29,31,32]
- Warth, J., Kaiser, G., and Kügler, M. (2011), "The Impact of Data Quality and Analytical Capabilities on Planning Performance: Insights From the Automotive Industry," in *10th International Conference on Wirtschaftsinformatik*, Zurich, Switzerland, pp. 322–331. [29,32]
- Watts, S., Shankaranarayanan, G., and Even, A. (2009), "Data Quality Assessment in Context: A Cognitive Perspective," *Decision Support Systems*, 48, 202–211. [31]
- Wixom, B. H., and Watson, H. J. (2001), "An Empirical Investigation of the Factors Affecting Data Warehousing Success," *Management Information Systems Quarterly*, 25, 17–41. [30]
- Yang, C.-H., and Hillier, F. S. (1970), "Mean and Variance Control Chart Limits Based on a Small Number of Subgroups," *Journal of Quality Technology*, 2, 9–16. [36]
- Yang, Z., Xie, M., Kuralmani, V., and Tsui, K.-L. (2002), "On the Performance of Geometric Charts With Estimated Control Limits," *Journal of Quality Technology*, 34, 448–458. [34]
- Zhang, H., Albin, S. L., Wagner, S. R., Nolet, D. A., and Gupta, S. (2010), "Determining Statistical Process Control Baseline Periods in Long Historical Data Streams," *Journal of Quality Technology*, 42, 21–35. [36]
- Zhang, M., Peng, Y., Schuh, A., Megahed, F. M., and Woodall, W. H. (2012), "Geometric Charts With Estimated Control Limits," *Quality and Reliability Engineering International*, 29, 209–223. [34,35,36]