Int. J. Production Economics 154 (2014) 72–80

Contents lists available at ScienceDirect

Int. J. Production Economics

journal homepage: www.elsevier.com/locate/ijpe

Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications

Benjamin T. Hazen [a,*], Christopher A. Boone [b], Jeremy D. Ezell [c], L. Allison Jones-Farmer [c]

[a] Department of Marketing and Supply Chain Management, College of Business Administration, 310 Stokely Management Center, University of Tennessee, Knoxville, TN 37996-0530, USA
[b] Department of Marketing and Logistics, College of Business Administration, Georgia Southern University, 1332 Southern Drive, Statesboro, GA 30458, USA
[c] Department of Supply Chain and Information Systems Management, Harbert College of Business, 401 Lowder Building, 415 West Magnolia Avenue, Auburn University, Auburn, AL 36849, USA

ARTICLE INFO

Article history:
Received 17 October 2013
Accepted 12 April 2014
Available online 26 April 2014

Keywords:
Data quality
Statistical process control
Knowledge-based view
Organizational information processing view
Systems theory

ABSTRACT

Today's supply chain professionals are inundated with data, motivating new ways of thinking about how data are produced, organized, and analyzed. This has provided an impetus for organizations to adopt and perfect data analytic functions (e.g. data science, predictive analytics, and big data) in order to enhance supply chain processes and, ultimately, performance. However, management decisions informed by the use of these data analytic methods are only as good as the data on which they are based. In this paper, we introduce the data quality problem in the context of supply chain management (SCM) and propose methods for monitoring and controlling data quality. In addition to advocating for the importance of addressing data quality in supply chain research and practice, we also highlight interdisciplinary research topics based on complementary theory.

1. Introduction

Many have argued that the market focus of competition has evolved from that of competition between individual firms to competition between entire supply chains (Craighead et al., 2009; Ketchen and Hult, 2007; Slone, 2004; Whipple and Frankel, 2000). The resulting focus on supply chain management (SCM) has forced managers to rethink their competitive strategies (Zacharia et al., 2011), with many now seeking to "win with data" (Hopkins et al., 2010). Supply chain managers are increasingly reliant upon data to gain visibility into expenditures, identify trends in costs and performance, and support process control, inventory monitoring, production optimization, and process improvement efforts. In fact, many businesses are awash in data, with many seeking to capitalize on data analysis as a means for gaining a competitive advantage (Davenport, 2006). Data science, predictive analytics, and "big data" are each thought to be part of an emerging competitive area that will transform the way in which supply chains are managed and designed (Waller and Fawcett, 2013).

The emerging field of data science combines mathematical, statistical, computer science, and behavioral science expertise to tease insights from enterprise data, while predictive analytics describes the set of data science tools leveraged for future outcome prediction attempts (Barton and Court, 2012; Davenport and Patil, 2012). Big data is a more tenuous term, its definition and usage changing to include more than just the size or volume of the organization's data but the variety and velocity as well (Megahed and Jones-Farmer, 2013). As coined by Waller and Fawcett (2013), we collectively refer to these three related topics as data science, predictive analytics, and big data (DPB). Considering both the proliferation of DPB activities for supply chain management and the fact that the data upon which these DPB functions rely are often plagued with errors (Dey and Kumar, 2010), there is an important need to examine the data quality problem as it pertains to the field of SCM.

In the epic poem *Rime of the Ancient Mariner*, Samuel Taylor Coleridge states, "Water, water, everywhere, nor any a drop to drink." Data embodies the same degree of uselessness for consumption if it is of poor quality. Indeed, the degree to which data can be used is largely determined by their quality (O'Reilly, 1982). Poor quality data can have a direct impact on business decisions (Dyson and Foster, 1982; Warth et al., 2011) and have been shown to promote a number of tangible and intangible losses for businesses (Batini et al., 2009). The costs of poor data quality have

* Corresponding author. Tel.: +1 919 722 8218.
E-mail addresses: hazen@utk.edu (B.T. Hazen), caboone@georgiasouthern.edu (C.A. Boone), jde0009@tigermail.auburn.edu (J.D. Ezell), joneall@auburn.edu (L.A. Jones-Farmer).

http://dx.doi.org/10.1016/j.ijpe.2014.04.018
0925-5273/

been estimated to be as high as 8% to 12% of revenues for a typical organization and may generate up to 40% to 60% of a service organization's expenses (Redman, 1998); this translates into losses that are estimated to exceed billions of dollars per year (Batini and Scannapieco, 2006; Dey and Kumar, 2010). Poor data quality can be equally damaging to less tangible areas including job satisfaction, decision quality, and propagation of mistrust between and within organizations (Redman, 1996). Supply chain managers are seeing the problems and impacts attributed to poor data quality growing in importance. Although high quality data has always been a must-have for these managers, quality issues are increasing as firms' desires and capability for the analysis of ever larger amounts of acquired data similarly increase (Parssian et al., 2004). In fact, a recent survey of over 3000 business executives found that one in five executives consider data quality as a primary obstacle to adopting more robust data analytic-based strategies (Lavalle et al., 2011).

The goal of this paper is to introduce and stress the need for the monitoring and control of data quality in supply chain management processes and provide a starting point for future research and applications. The remainder of this paper is structured as follows. The paper begins with a short overview of the data production process. We then describe data quality and define its constituent dimensions. Then, we discuss methods for monitoring, controlling, and improving data quality and use a practical example of how one organization employed such a method to enhance data quality in its supply chain. We next examine how the data quality problem can be viewed through the lenses of systems theory, the knowledge-based view, and organizational information processing view to provide a series of theory-based topics to guide future research. Finally we discuss both managerial implications and additional research considerations.

2. Data production

Several scholars have drawn an analogy between product manufacturing and "data manufacturing" (Arnold, 1992; Ballou et al., 1998; Emery, 1969; Huh et al., 1990; March and Hevner, 2007; Ronen and Spiegler, 1991; Wang and Kon, 1993; Wang et al., 1995). For instance, Wang et al. (1995) proposed a simple framework, depicted in Fig. 1, to describe the similarities between the two manufacturing processes.

Wang (1998) extended this comparison between data and manufacturing processes by suggesting data quality should be addressed via a Total Data Quality Management (TDQM) cycle, which calls for continuously defining, measuring, analyzing, and improving data quality. This approach is similar to Deming's (2000) Total Quality Management cycle (Plan, Do, Check, and Act) and analogous to the Define, Measure, Analyze, Improve, Control (DMAIC) cycle, as ascribed by Six Sigma, for the data manufacturing process. However, Jones-Farmer et al. (2013) pointed out that, unlike the DMAIC cycle from Six Sigma, there is no control stage recommended in the TDQM cycle. With no means for controlling the quality of data, there is no framework for inducing continuous improvement in the data production process. Two important contributions of this paper are to introduce the need for continuous improvement in the SCM data production process, and to suggest a familiar framework for establishing a quality control mechanism regarding data quality.

The analogy of the data production process to a manufacturing process is, perhaps, one of the most widely accepted views in the literature. Although there are many similarities between the data production process and the manufacturing process, we will discuss what we feel are two of the most important differences. In a manufacturing process, raw materials are input into a process, the materials are transformed, and the resulting output is a manufactured product. The raw materials are generally depleted as the goods are produced. In the data production process, the data represent the input into the data production process, and a transformed data product is the output of the production process. The data are generally not depleted through production. A bad batch of data in the data production process will remain until it is actively cleaned up or removed. Perhaps the most pertinent, yet challenging difference between a manufacturing and data production process relates to the difficulty with regard to measuring the quality of intangible data. 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 most important quality characteristics.

3. Dimensions of data quality

Research suggests that data quality is comprised of several dimensions (Ballou and Pazer, 1985; Ballou et al., 1998; Pipino et al., 2002; Redman, 1996; Wand and Wang, 1996; Wang and Strong, 1996). Both Wang and Strong (1996) and Lee et al. (2002) organize data quality dimensions into two categories: *intrinsic*, referring to attributes that are objective and native to the data and *contextual*, referring to attributes that are dependent on the context in which the data are observed or used. Contextual dimensions include relevancy, value-added, quantity (Wang and Strong, 1996), believability, accessibility, and reputation of the data (Lee et al., 2004, 2002). Measures of these dimensions have 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). Contextual dimensions of data quality lend themselves more towards *information* as opposed to *data*, because these dimensions are formed by placing data within a situation or problem specific context (Batini et al., 2009; Davenport and Prusak, 2000; Haug et al., 2009; Watts et al., 2009). Because we

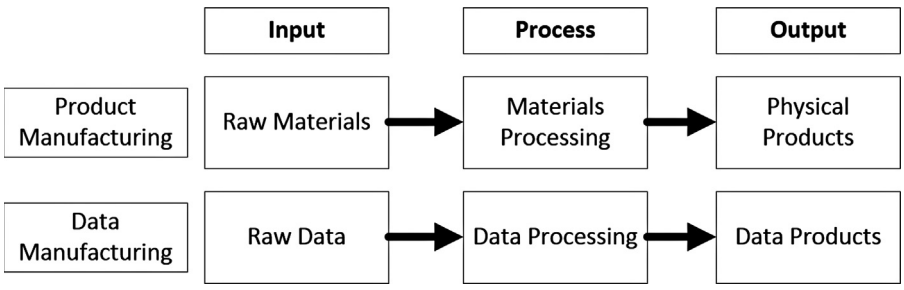

Fig. 1. An analogy between product and data manufacturing processes (Wang et al., 1995).

Table 1
Dimensions of data quality.

Data quality dimension	Description	Supply chain example
Accuracy	Are the data free of errors?	Customer shipping address in a customer relationship management system matches the address on the most recent customer order
Timeliness	Are the data up-to-date?	Inventory management system reflects real-time inventory levels at each retail location
Consistency	Are the data presented in the same format?	All requested delivery dates are entered in a DD/MM/YY format
Completeness	Are necessary data missing?	Customer shipping address includes all data points necessary to complete a shipment (i.e. name, street address, city, state, and zip code)

consider the quality of data, not information, as it moves through a production-like process, we limit our discussion of quality to consideration of the intrinsic measures of data quality.

The literature consistently describes intrinsic data quality along four dimensions: accuracy; timeliness; consistency; and completeness (Ballou and Pazer, 1985; Batini et al., 2009; Blake and Mangiameli, 2011; Haug and Arlbjørn, 2011; Haug et al., 2009; Kahn et al., 2002; Lee et al., 2002; Parssian, 2006; Scannapieco and Catarci, 2002; Wang and Strong, 1996; Zeithaml et al., 1990). Below, we explore and use the aforementioned literature to define and describe these four dimensions.

Accuracy refers to the degree to which data are equivalent to their corresponding "real" values (Ballou and Pazer, 1985). This dimension can be assessed via comparing values with external values that are known to be (or considered to be) correct (Redman, 1996). A simple example would be a data record in a customer relationship management system, where the street address for a customer in the system matches the street address where the customer currently resides. In this case, accuracy of the street address value in the system could be assessed via validating the shipping address on the most recent customer order. No problem context or value-judgment of the data is needed: it is either accurate or not. Its accuracy is entirely self-dependent.

Timeliness refers to the degree to which data are up-to-date. Research suggests that timeliness can be further decomposed into two dimensions: (1) currency, or length of time since the record's last update, and (2) volatility, which describes the frequency of updates (Blake and Mangiameli, 2011; Pipino et al., 2002; Wand and Wang, 1996). Data that are correct when assessed, but updated very infrequently, may still hamper efforts at effective managerial decision making (e.g., errors that occur in the data may be missed more often than not with infrequent record updating, preventing operational issues in the business from being detected early). A convenient example measure for calculating timeliness using values for currency and volatility can be found in Ballou et al. (1998), p. 468, where currency is calculated using the time of data delivery, the time it was entered into the system, and the age of the data at delivery (which can differ from input time). Together, currency and volatility measures are used to calculate timeliness.

Consistency refers to the degree to which related data records match in terms of format and structure. Ballou and Pazer (1985) define consistency as when the "representation of the data value is the same in all cases" (p. 153). Batini et al. (2009) develop the notion of both intra-relation and inter-relation constraints on the consistency of data. Intra-relation consistency assesses the adherence of the data to a range of possible values (Coronel et al., 2011), whereas inter-relation assesses how well data are presented using the same structure. An example of this would be that a person, currently alive, would have for "year of birth" a possible value range of 1900–2013 (intra-relation constraint), while that person's record in two different datasets would, in both cases, have a field for birth year, and both fields would intentionally represent the person's year of birth in the same format (inter-relation constraint).

Completeness refers to the degree to which data are full and complete in content, with no missing data. This dimension can describe a data record that captures the minimally required amount of information needed (Wand and Wang, 1996), or data that have had all values captured (Gomes et al., 2007). Every field in the data record is needed in order to paint the complete picture of what the record is attempting to represent in the "real world." For example, if a particular customer's record includes a name and street address, but no state, city, and zip code, then that record is considered incomplete. The minimum amount of data needed for a correct address record is not present. A simple ratio of complete versus incomplete records can then form a potential measure of completeness.

A summary of the dimensions of data quality is presented in Table 1. Once data quality measures are understood, these quality measures can be monitored for improvement or adherence to standards. For example, data can be tagged as either accurate or not. Once tagged, there should be a method in place to monitor the long-term accuracy of the data. Combined with the measuring and monitoring the other three data quality dimensions, this helps to ensure that the records in the dataset are as accurate, timely, complete, and consistent as is practical.

Understanding the four intrinsic dimensions of data quality allows us to operationally define measures for these dimensions and apply tools to actively monitor for data quality problems. For instance, total quality management approaches (Porter and Rayner, 1992; Redman, 1992), process capability analyses (Veldman and Gaalman, 2013), statistical process control (SPC), and additional quality tools and theories might help inform data quality management techniques, and investigation into using these techniques in the context of the data quality problem is needed. To this end, tools from SPC have been suggested as a natural fit for monitoring and improving data quality over time (Jones-Farmer et al., 2013). In particular, control charts can be used to improve data quality, not batch-by-batch, but in the overall data production process. Although there are several quality methods that should be examined in future data quality research, we suggest that SPC control chart methods might be most useful as an illustrative example of controlling and monitoring data quality in a supply chain DPB setting. Thus, in the next section we describe details regarding how SPC control chart methods can be used to monitor and control data quality in a supply chain, and provide a brief example case study.

4. Controlling data quality with SPC

Both academic and practitioner literature has stated the need for improved data quality for effective management and decision making (Redman, 1998). To this end, much of the research has examined ways in which to assess the quality of data products after they are created (Dey and Kumar, 2010; Parssian et al., 2004). Although useful, this practice is akin to quality checking finished

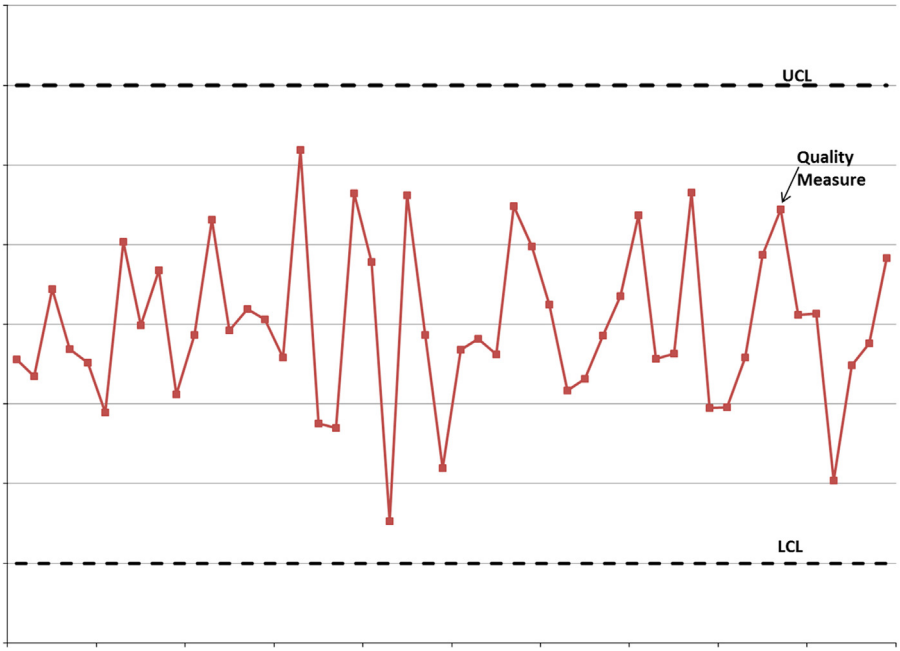

Fig. 2. An example Shewhart-type control chart showing an in-control process.

products at the end of a production line. Like with manufacturing, monitoring and controlling for quality throughout the duration of the data production process might be a more useful endeavor as deficiencies can be addressed in real time and corrected before they create cascading defects.

Redman (1992, 1996, 2001) emphasizes the importance of maintaining a process focus when considering data quality. He advocates the use of simple tools such as the histogram, fishbone diagram, and Pareto chart for cleaning up a data production process. For more information on the application of simple quality tools in general, the interested reader is referred to Montgomery (2013) or specific to data quality, Redman (1992). Once the initial quality efforts have improved the current state of data quality, bringing it into an in-control state, ongoing process monitoring should be used to maintain the required level of data quality. Unfortunately, as discussed above, there is no such control stage proposed in Wang's (1998) TDQM cycle.

Advanced control chart methodologies provide long established and commonly employed methods for monitoring and controlling production quality (Lieberman, 1965; Mitra, 2008; Woodall, 2000). Indeed, SPC, topics are often discussed in the SCM literature (e.g. Choi and Rungtusanatham, 1999; Fraisat and Sawalha, 2013; Laframboise and Reyes, 2005; Rahman, 2006; Sila et al., 2006). While SPC methods are common in SCM, these methods are not commonly applied in the production of data used in support of managing a supply chain. Therefore, we propose SPC methods as one means by which to monitor, control, and ultimately improve the quality of data used for SCM.

We suspect the lack of commonplace application of advanced SPC methods to monitoring data quality is due, in part, to the lack of awareness of the applicability of the methods on the part of practitioners, but also because SPC methods were developed based on assumptions relevant to the actual processes themselves, and not necessarily the data by which managers use to control these processes. We focus on the data production process, which includes data collection, storage, retrieval, and processing. We consider the output of this process, a data set, as a product, much like a product of a manufacturing process (Wang et al., 1995). Similar to those who examine Six Sigma in a manufacturing environment (e.g. Zu et al., 2010), we are motivated to examine how the data production process can be effectively monitored, controlled, and improved using control charts for the purpose of enhancing the quality of the data supply chain professionals use to manage their processes.

Shewhart (1931) first introduced control charts as a method for monitoring the output quality of manufacturing processes. The methods were popularized following World War II as Deming (2000) used control charts to aid the Japanese in rebuilding their manufacturing base. Fig. 2 gives an example of a simple Shewhart-type control chart. The dotted lines labeled UCL and LCL represent upper and lower control limits, respectively, and are based on the statistical nature of the process under consideration. Each data series plotted over time represents a measure of a process characteristic. Values that fall between the UCL and LCL are considered subject to only usual or common-cause process variation. When all points in a process fall between the control limits, the process is considered to be in-control. Points that plot either above the UCL or below the LCL are considered signals to a potential out-of-control scenario, or affected by some force that is not expected within the usual operating confines of the process. When a control chart signals a potential out-of-control event, the process operators investigate for possible root causes of the signal. For more information about control charts, we refer the reader to Montgomery (2013).

Researchers in SPC have advanced the field of control charts far beyond the traditional Shewhart charts like the one illustrated above. More advanced control charts include the Cumulative Sum (CUSUM) (Page, 1961), the Exponentially Weighted Moving Average (EWMA) (Roberts, 1959), multivariate Shewhart-type control charts such as Hotelling's (1947) T^2 chart , multivariate versions of the CUSUM (Crosier, 1998), and EWMA (Lowry et al., 1992), and many others (Ho and Quinino, 2013; Ou et al., 2012; Wu et al., 2009). In addition, many control charts have been developed to monitor categorical and discrete process characteristics (see, e.g., Woodall (1997) and Topalidou and Psarakis (2009)). Like most statistical methods, the different control charts are designed to work in different scenarios with different types of data. Jones-Farmer et al. (2013) give an overview of the advanced control charting tools that are applicable to monitoring the intrinsic measures of data quality.

We found evidence of some use of control charts for monitoring data quality (Pierchala et al., 2009; Redman, 2001; Sparks and OkuGami, 2010); however, the applications have not been widely adopted by practitioners. Indeed, there are many opportunities for advancement in the practical application of control charts to data quality in the SCM context. With the proliferation of DPB and the increasingly important role of the supply chain in the success of today's businesses, we suggest the use of control charts for monitoring supply chain data quality. Next, we demonstrate how such an approach might be employed to enhance data quality in a supply chain setting.

4.1. Use of SPC to monitor and control supply chain data: An example case study

As part of a field-based research effort, we examined the data management program of an organization that remanufactures jet engines and related components for military aircraft and introduced the use of control chart methods to enhance data quality. This particular data management system is used for real-time monitoring of the closed-loop jet engine inventory for one particular cargo aircraft. Engine location and repair status are among some of the important information tracked in this database. Data products derived from this system are used by line, middle-, and senior-level managers at several locations for a wide variety of decision-making purposes (e.g. to determine if a particular aircraft is currently capable to deploy overseas in the sense that none of its engines require extensive maintenance or inspections that deployed locations are incapable of conducting).

For the purpose of this illustration, we limit our examination to data records regarding jet engine compressors (which is one subcomponent of a jet engine). As described below, records for eight different compressors were captured in real-time. To maintain brevity in our example, we focus here on one of the four intrinsic data quality dimensions: completeness. Completeness was measured at the record level and defined as

$$C_{i.k} = \begin{cases} 0 \text{ if record is complete} \\ 1 \text{ if record is incomplete} \end{cases}$$

for $i = 1, \ldots, 8$, compressors and $k = 1, \ldots, N_R$, part records. Thus, we have eight binary variables describing completeness.

The first 400 observations taken can be used as the reference sample. Table 2 shows the phi coefficients (Cohen et al., 2003) estimating the correlation among the eight completeness variables. The values given along the diagonal are the estimated proportion of incomplete records.

Table 2
Correlation matrix.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8
Comp1	.515							
Comp2	−.074	.488						
Comp3	.123**	.083*	.553					
Comp4	−.067	.153***	.107**	.600				
Comp5	−.030	.097*	.099**	.029	.418			
Comp6	.054	.097*	.069	−.004	.092*	.373		
Comp7	−.047	.072	.333	−.080	.041	.100**	.533	
Comp8	−.062	−.003	−.049	.039	.001	−.010	.038	.443

Note: All table values are estimated from 400 reference sample observations. The values along the diagonal estimate the percent of incomplete records for each of eight compressors. The lower diagonal observations represent the correlation between each row and column variable. The notations, *, **, and *** indicate significance at the .10, .05, and .01 level of significance, respectfully.

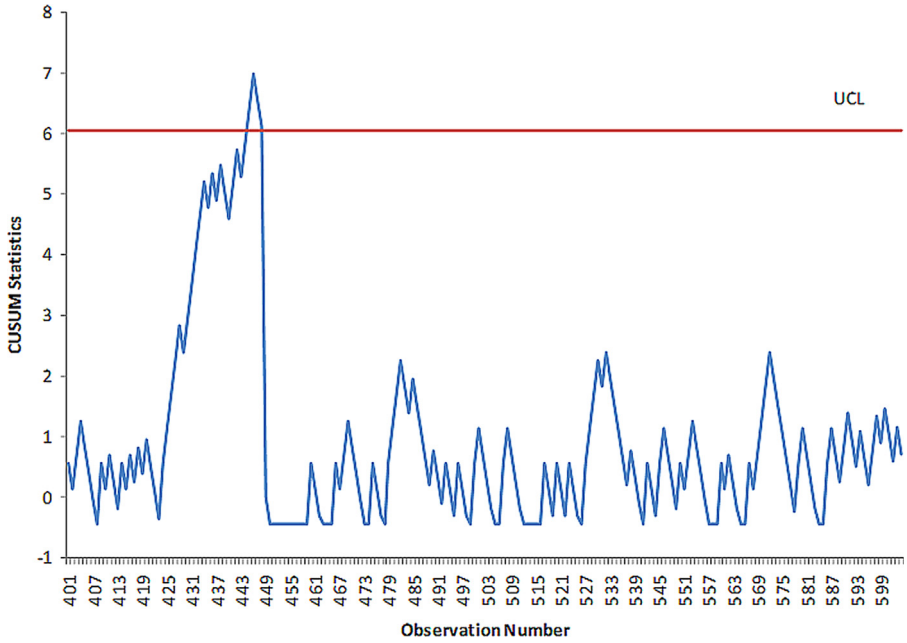

Fig. 3. Bernoulli CUSUM chart of completeness of component 6 for aircraft maintenance database.

Fig. 3 shows a control chart of the completeness scores for the next 204 observations of compressor 6. Here, we chose to use a Bernoulli CUSUM control chart to monitor the incomplete records. Because each record was determined either incomplete (1) or complete (0), the data may be well-modeled by a Bernouli distribution. The use of the CUSUM control chart for individual (as opposed to subgrouped) Bernoulli random variables was originally proposed by Reynolds and Stoumbos (1999) as a fast-detection alternative to the Shewhart *p*-chart for continuous monitoring of dichotomous variables. The use of the Bernoulli CUSUM for monitoring dichotomous quality characteristics can lead to a faster detection of process changes because it eliminates the need to accumulate a large subgroup of observations prior to plotting a control chart statistic. As can be seen from the chart, out-of-control signals were given on observations 445–448 following a series of incomplete records. Using this chart, managers were able to detect a data quality problem. Following observation 448, corrective action was taken that included retraining the data entry workforce. At that point, the CUSUM chart was reset, and the process continued in an in-control state.

The example described above is brief in nature for the purpose of description in this paper. It should be noted that this method can be implemented on a large scale to examine data products generated from a data warehouse on all four dimensions of data quality. Although we do not go into specific detail here, the interested reader is referred to Jones-Farmer et al. (2013) for a deeper discussion and specific details on applying control chart methodology to monitor and improve the data production process.

Now that we have outlined the data production process, described the dimensions of data quality, and demonstrated a practical means for controlling data quality using an SPC framework, we turn to a discussion of future research needs in the area of enhancing data quality for DPB in the supply chain. In the next section, we highlight applicable theory that might help inform future research in this area and propose topics in need of further exploration.

5. Theory-based research opportunities

We propose that considering data quality should be common practice in future supply chain DPB research. At a minimum, data quality should be acknowledged, measured, monitored, and controlled. Ideally, the goal of any measurement or monitoring activity would be to improve the quality of the process and product. We believe that simply acknowledging, measuring, and monitoring the quality of a firm's SCM data will lead to inevitable improvement in the quality of the data. Further, establishing an ongoing monitoring scheme will allow for future data acquisitions to be controlled, with the goal of improving the quality of the data and ultimately the decisions that are based on the data.

Although our consideration thus far has been on the technical nature of data quality, there are also specific and theoretically based research questions pertinent to data quality and SCM that can and should also be addressed in future research. While there are surely several theories that can be used as a basis to study emerging problems regarding data quality in the context of DPB and SCM, as a starting point we highlight three. More specifically, we frame some example research questions in the context of the knowledge-based view (KBV), systems theory, and the organizational information processing view (OIPV).

5.1. Knowledge-based view

The resource-based view (RBV) suggests that organizations create competitive advantage via employment of the resources at their disposal (Barney, 1991). The theory is often used to frame research in the SCM domain to describe competition between firms and supply chains (Defee et al., 2010; Fawcett and Waller, 2011). Most importantly, to translate short-term competitive in to sustained competitive advantages, resources must be valuable (the resource enables the firm to create value for customers), rare (the resource is not widely available), inimitable (the resource cannot be easily replicated or procured by other firms), and non-substitutable (other firms cannot employ an alternative resource that offers similar utility or drives the resource into obsolescence) (Barney, 1991).

As an extension of RBV, the KBV considers knowledge as one such resource that can be valuable, rare, inimitable, and non-substitutable (Grant, 1996). The value of DPB as a knowledge resource and its impact on the firm's competitive advantage in the market are both dependent on data quality. This highlights the importance of the consideration of a firm's data quality levels and processes. For instance, a DPB effort might not necessarily create value for a firm if a determined baseline level of data quality is not attained. Conversely, a high level of data quality might enable DPB efforts that are rare, inimitable, and perhaps non-substitutable among competitors. Considering the KBV, we suggest the following research questions:

- Is there a relationship between perceived or known levels of data quality and DPB usage in supply chain applications?
- Does data quality play an intervening role in the relationship between DPB activities and measures of supply chain performance?
- If DPB activities can be considered a knowledge resource, then when, how, and how often should firms conduct an analysis of the quality of their data in order to maintain this knowledge resource?
- Similar to the above, precisely how can data quality analyses affect the value of both (a) a firm's data and (b) a firm's DPB activities as knowledge resources capable of enhancing supply chain performance?
- How does the quality of the data affect perceived value placed on DPB efforts when making strategic decisions regarding competitive actions?

5.2. Systems theory

Systems theory is another commonly used theory in the SCM literature (Chicksand et al., 2012; Ketchen and Hult, 2007) that might provide a useful lens through which to view the data quality problem. System theory suggests that organizations are open and porous systems that interact with their surrounding environment, and thus are continually evolving (Von Bertalanffy, 1951). From this perspective, it is easy to envision a supply chain as a system of connected nodes (Towill et al., 1992) that are interacting with and relying on inputs from the external environment and from each other (Blackhurst et al., 2011). Similarly, information systems that support DPB can be viewed as SCM sub-systems operating within a real-world feedback control system (Orr, 1998). This is because such systems are by nature intra-organizational and rely upon interactions with and inputs from a variety of both inter- and intra-firm actors.

Orr (1998) suggests that there must be a mechanism to synchronize system data with changes in the environment. As the system absorbs data at all node points, the quality of that data could change as rapidly as the volume being gathered. As such, the measurement and control of data quality might be especially important from a systems theory perspective when investigating

the impact of DPB on SCM performance. Considering systems theory, we suggest the following research questions:

- How can organizations integrate data quality and control initiatives into their existing or emerging supply chain DPB programs?
- What are the costs of poor data quality to specific supply chain DPB initiatives? What are the costs of maintaining "acceptable" data quality, and do these costs result in an adequate return on investment? How is "acceptable" data quality defined for supply chain operations?
- How does one firm's data quality affect the DPB efforts of partner firms within the supply chain?
- Considering the boundary-spanning role of logistics and supply chain operations, how does the level of data quality affect firm processes outside of logistics and supply chain operations?

5.3. Organizational information processing view

The OIPV suggests that organizations are imperfect decision making systems due to incomplete information, which is a function of uncertainty and equivocality (March and Simon, 1958). While uncertainty refers to incomplete knowledge, equivocality results from conflicting interpretations about a decision-making situation (Daft and Lengel, 1986; Galbraith, 1974). Levels of data quality might play a role in regard to the completeness of knowledge and interpretations of information.

The OIPV considers three primary components: information processing needs, information processing capabilities, and the fit between needs and capabilities (Tushman and Nadler, 1978). Information processing needs indicate the information required by the organization to enable effective decision-making, whereas information processing capabilities indicate the organization's actual capacity to structure and utilize information to support decision-making (Tushman and Nadler, 1978). Fit is the degree to which a firm's information processing capabilities satisfy its information processing needs (Tushman and Nadler, 1978). It is via fit that firms can reduce uncertainty and equivocality, enhance decision-making capabilities and, subsequently, enhance performance (Trautmann et al., 2009; Wu et al., 2013). As such, OIPV can be a useful lens through which to examine how efforts to enhance data quality might lead to increased levels of performance realized via DPB. Considering the OIPV, we suggest the following research questions:

- Does the degree of data quality play an intervening role in the relationship between supply chain information processing needs and capabilities?
- Does the degree of data quality impact the range of information processing needs? Does enhanced data quality cause a re-scope of processing needs as processing capabilities are enhanced?
- Does enhancing data quality help to reduce uncertainties surrounding DPB activities used for SCM?

6. Implications and concluding remarks

As alluded to above, the study of DPB in general and the data quality problem in particular require interdisciplinary collaboration in order to advance. For instance, information systems experts are needed to provide insight into how data are collected, stored, processed, and retrieved. SCM domain experts are needed to ensure that the right problems are the analysis being performed and results derived thereof are relevant (Waller and Fawcett, 2013). Additionally, emerging data quality research suggests the need for statistical and analytical experts who are knowledgeable in methods required to measure, monitor, and control data quality (Jones-Farmer et al., 2013). Working together, scholars from these and other disciplines can employ the right technologies and techniques to solve the right problems.

Even though the field-based effort described herein provides support for the application of data quality control methods to data products in a supply chain environment, more study is needed. For example, an audit released by the United States Air Force (USAF) (Air Force Audit Agency, 2013) reviewed the data entered by both organic and contract maintenance personnel into one of two field level maintenance information systems (determined by the aircraft being repaired). Data from these two field level systems are automatically transferred to an enterprise level system that serves as the USAF's single source of aircraft maintenance data. The enterprise level system is used by the USAF in determining the need for weapon system inspections and provides information on weapon system status, utilization, and configuration. Unfortunately, the audit revealed errors in the reporting of over 30% of the observed maintenance actions. Auditors found that many of the maintenance actions entered into the system were incomplete or inaccurate, which resulted in the improper grounding of aircraft and an increase in maintenance manpower costs. Chief among the causes of the errors identified in the audit was a failure to monitor the data transfer between the systems and a failure to establish effective monitoring and control processes for notifying and correcting data errors. Similar to the field-based example presented earlier, this scenario represents an opportunity to examine methods to improve the quality of data products.

However, data quality issues and the opportunities for the application of data quality control methods are not isolated to military or maintenance intensive operations. For example, the proliferation of radio frequency identification provides firms with a large amount of data to enhance visibility throughout the supply chain; unfortunately, the data generated via such technology is often rife with errors (Delen et al., 2007). The lack of common data format standards and the transfer of data between dissimilar systems also plague supply chain firms that rely upon DPB to drive growth and innovation. Errors in customers' addresses for example can result in advertisements and shipments being sent to the wrong customers, which could lead to the loss of potential sales and poor customer service. Because customer service is shown to be a key antecedent to firm performance in the supply chain (Leuschner et al., 2013), such poor service can indeed be detrimental. Thus, researchers should seek to study the application of data quality control methods in a variety of supply chain environments.

The increasing importance of data to supply chain managers should lead to an amplified awareness and sensitivity to their need for high quality data products. The results of decisions based on poor quality data could be costly. Thus, supply chain managers should begin to view the quality of the data products they depend upon for decisions in much the same way they view the quality of the products their supply chain delivers. Managers who appreciate the value of data products that are accurate, consistent, complete, and timely should consider the potential of using control methods to improve the quality of data products much like these methods improved the quality of manufactured products.

In this paper, we have presented a review of the literature on data quality from the perspective of DPB in the supply chain. This review includes literature that frames data production as a process and defines measures of data quality. We introduce application of SPC methods as a means to control data quality in the supply chain, and suggest theory-based topics for future research. We hope that our introduction to the data quality problem in the context of DPB in the supply chain will encourage interdisciplinary collaboration to further develop tangible methods for controlling data, and examining the effects thereof.

References

Air Force Audit Agency, 2013. Audit Report: Serialized Parts Configuration Management, Washington, D.C. Accessed January 24, 2014. ⟨http://www.afaa.af.mil/⟩.

Arnold, S.E., 1992. Information manufacturing: the road to database quality. Database 15 (5), 32–39.

Ballou, D.P., Pazer, H.L., 1985. Modeling data and process quality in multi-input, multi-output information systems. Manage. Sci. 31 (2), 150–162.

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

Barney, J., 1991. Firm resources and sustained competitive advantage. J. Manage. 17 (1), 99–120.

Barton, D., Court, D., 2012. Making advanced analytics work for you. Harvard Bus. Rev. 90, 79–83.

Batini, C., Cappiello, C., Francalanci, C., Maurino, A., 2009. Methodologies for data quality assessment and improvement. Assoc. Comput. Mach. Comput. Surv. 41 (3), 1–52.

Batini, C., Scannapieco, M., 2006. Data Quality: Concepts, Methodologies and Techniques. Springer-Verlag, New York.

Blackhurst, J., Dunn, K.S., Craighead, C.W., 2011. An empirically derived framework of global supply resiliency. J. Bus. Logist. 32 (4), 374–391.

Blake, R., Mangiameli, P., 2011. The effects and interactions of data quality and problem complexity on classification. Assoc. Comput. Mach. J. Data Inf. Qual. 2 (2), 1–28.

Chicksand, D., Watson, G., Walker, H., Radnor, Z., Johnston, R., 2012. Theoretical perspectives in purchasing and supply chain management: an analysis of the literature. Supply Chain Manage.: Int. J. 17 (4), 454–472.

Choi, T.Y., Rungtusanatham, M., 1999. Comparison of quality management practices: across the supply chain and industries. J. Supply Chain Manage. 35 (1), 20–27.

Cohen, J., Cohen, P., West, S.G., Aiken, L., 2003. Applied Multiple Regresstion/Correlation Analysis for the Behavioral Sciences, third ed.. Lawrence Earlbaum Associates, Manway, NJ.

Coronel, C., Morris, S., Rob, P., 2011. Database Systems: Design, Implementation, and Management. Cengage Learning, Boston, MA.

Craighead, C.W., Hult, G.T.M., Ketchen Jr., D.J., 2009. The effects of innovation—cost strategy, knowledge, and action in the supply chain on firm performance. J. Oper. Manage. 27 (5), 405–421.

Crosier, R.B, 1998. Multivariate generalizations of cumulative sum quality-control schemes. Technometrics 30 (3), 291–303.

Daft, R.L., Lengel, R.H., 1986. Organizational information requirements, media richness and structural design. Manage. Sci. 32 (5), 554–571.

Davenport, T.H., 2006. Competing on analytics. Harvard Business Review 84 (1), 84–93.

Davenport, T., Patil, D., 2012. Data scientist: the sexiest job of the 21st century. Harvard Bus. Rev. 90, 70–76.

Davenport, T.H., Prusak, L., 2000. Working Knowledge: How Organizations Manage What They Know. Harvard Business Press, Boston, MA.

Defee, C.C., Williams, B., Randall, W.S., Thomas, R., 2010. An inventory of theory in logistics and SCM research. Int. J. Logist. Manage. 21 (3), 404–489.

Delen, D., Hardgrave, B.C., Sharda, R., 2007. RFID for better supply-chain management through enhanced information visibility. Prod. Oper. Manage. 16 (5), 613–624.

Deming, W.E., 2000. Out of the Crisis. The MIT Press, Boston, MA.

Dey, D., Kumar, S., 2010. Reassessing data quality for information products. Manage. Sci. 56 (12), 2316–2322.

Dyson, R.G., Foster, M.J., 1982. The relationship of participation and effectiveness in strategic planning. Strategic Manage. J. 3 (1), 77–88.

Emery, J.C., 1969. Organizational Planning and Control Systems: Theory and Management. Macmillian, New York, NY.

Fawcett, S.E., Waller, M.A., 2011. Making sense out of chaos: why theory is relevant to supply chain research. J. Bus. Logist. 32 (1), 1–5.

Fraisat, L.E., Sawalha, I.H., 2013. Quality control and supply chain management: a contextual perspective and a case study. Supply Chain Manage.: Int. J. 18 (2), 194–207.

Galbraith, J.R., 1974. Organization design: an information processing view. Interfaces 4 (3), 28–36.

Garvin, D.A., 1984. What does "product quality" really mean? Sloan Manage. Rev. 26 (1), 25–43.

Garvin, D.A., 1987. Competing on the eight dimensions of quality. Harvard Bus. Rev. 65 (6), 101–109.

Gomes, P., Farinha, J., Trigueiros, M.J., 2007. A data quality metamodel extension to CWM. In: Fourth Asia-Pacific Conference on Conceptual Modeling. Australian Computer Society, Inc., pp. 17–26.

Grant, R.M., 1996. Toward a knowledge-based theory of the firm. Strategic Manage. J. 17 (1), 109–122.

Haug, A., Arlbjørn, J.S., 2011. Barriers to master data quality. J. Enterp. Inf. Manage. 24 (3), 288–303.

Haug, A., Arlbjørn, J.S., Pedersen, A., 2009. A classification model of ERP system data quality. Ind. Manage. Data Syst. 109 (8), 1053–1068.

Ho, L.L., Quinino, R.d.C., 2013. An attribute control chart for monitoring the variability of a process. Int. J. Prod. Econ. 145 (1), 263–267.

Hopkins, M.S., Lavalle, S., Balboni, F., 2010. The new intelligent enterprise: 10 insights: a first look at the new intelligent enterprise survey. MIT Sloan Manage. Rev. 52 (1), 22.

Hotelling, H., 1947. Multivariate quality control. In: Eisenhart, O. (Ed.), Selected Techniques of Statistical Analysis. McGraw-Hill, New York, pp. 113–184.

Huh, Y.U., Keller, F.R., Redman, T.C., Watkins, A.R., 1990. Data quality. Inf. Softw. Technol. 32 (8), 559–565.

Jones-Farmer, L.A., Ezell, J.D., Hazen, B.T., 2013. Applying control chart methods to enhance data quality. Technometrics , http://dx.doi.org/10.1080/00401706.2013.804437.

Kahn, B.K., Strong, D.M., Wang, R.Y., 2002. Information quality benchmarks: product and service performance. Commun. ACM 45 (4), 184–192.

Ketchen Jr, D.J., Hult, G.T.M., 2007. Bridging organization theory and supply chain management: the case of best value supply chains. J. Oper. Manage. 25 (2), 573–580.

Laframboise, K., Reyes, F., 2005. Gaining competitive advantage from integrating enterprise resource planning and total quality management. J. Supply Chain Manage. 41 (3), 49–64.

Lavalle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N., 2011. Big data: analytics and the path from insights to value. MIT Sloan Manage. Rev. 52 (2), 21–32.

Lee, Y.W., Pipino, L., Strong, D.M., Wang, R.Y., 2004. Process-embedded data integrity. J. Database Manage. 15 (1), 17 (87).

Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, R.Y., 2002. AIMQ: a methodology for information quality assessment. Inf. Manage. 40 (2), 133–146.

Leuschner, R., Charvet, F., Rogers, D.S., 2013. A meta-analysis of logistics customer service. J. Supply Chain Manage. 49 (1), 47–63.

Lieberman, G.J., 1965. Statistical process control and the impact of automatic process control. Technometrics 7 (3), 283–292.

Lowry, C.A., Woodall, W.H., Champ, C.W., Rigdon, S.E., 1992. A multivariate exponentially weighted moving average control chart. Technometrics 30 (1), 46–53.

March, J., Simon, H., 1958. Organizations. John Wiley & Sons, New York.

March, S.T., Hevner, A.R., 2007. Integrated decision support systems: a data warehousing perspective. Decis. Support Syst. 43 (3), 1031–1043.

Megahed, F.M., Jones-Farmer, L.A., 2013. A statistical process monitoring perspective on big data. In: Lens, H.-J., Schmid, W., Wilrich, P.-T. (Eds.), Frontiers in Statistical Quality Control, 11th ed. Springer, New York.

Mitra, A., 2008. Fundamentals of Quality Control and Improvement, third ed.. John Wiley & Sons, Hoboken, NJ.

Montgomery, D.C., 2013. Introduction to Statistical Quality Control, seventh ed.. John Wiley & Sons, Hoboken, NJ.

O'Reilly, C.A., 1982. Variations in decision makers' use of information sources: the impact of quality and accessability of information. Acad. Manage. J. 25 (4), 756–771.

Orr, K., 1998. Data quality and systems theory. Commun. ACM 41 (2), 66–71.

Ou, Y., Wu, Z., Tsung, F., 2012. A comparison study of effectiveness and robustness of control charts for monitoring process mean. Int. J. Prod. Econ. 135 (1), 479–490.

Page, E.S., 1961. Cumulative sum charts. Technometrics 3 (1), 1–9.

Parssian, A., 2006. Managerial decision support with knowledge of accuracy and completeness of the relational aggregate functions. Decis. Support Syst. 42 (3), 1494–1502.

Parssian, A., Sarkar, S., Jacob, V.S., 2004. Assessing data quality for information products: impact of selection, projection, and Cartesian product. Manage. Sci. 50 (7), 967–982.

Pierchala, C.E., Surti, J., Peytcheva, E., Groves, R.M., Kreuter, F., Kohler, U., Chipperfield, J.O., Steel, D.G., Graham, P., Young, J., 2009. Control charts as a tool for data quality control. J. Off. Stat. 25 (2), 167–191.

Pipino, L.L., Lee, Y.W., Wang, R.Y., 2002. Data quality assessment. Commun. ACM 45 (4), 211–218.

Porter, L.J., Rayner, P., 1992. Quality costing for total quality management. Int. J. Prod. Econ. 27 (1), 69–81.

Rahman, S.-u., 2006. Quality management in logistics: an examination of industry practices. Supply Chain Manage.: Int. J. 11 (3), 233–240.

Redman, T.C., 1992. Data Quality: Management and Technology. Bantam Books, New York.

Redman, T.C., 1996. Data Quality for the Information Age. Artech House Publishers, Norwood, MA.

Redman, T.C., 1998. The impact of poor data quality on the typical enterprise. Commun. ACM 41 (2), 79–82.

Redman, T.C., 2001. Data Quality: The Field Guide. Digital Press, Boston, MA.

Reynolds Jr, M.R., Stoumbos, Z.G., 1999. A CUSUM chart for monitoring a proportion when inspecting continuously. J. Qual. Technol. 31 (1), 87–108.

Roberts, S.W., 1959. Control chart tests based on geometric moving averages. Technometrics 1 (3), 239–250.

Ronen, B., Spiegler, I., 1991. Information as inventory: a new conceptual view. Inf. Manage. 21 (4), 239–247.

Scannapieco, M., Catarci, T., 2002. Data quality under a computer science perspective. Arch. Comput. 2, 1–15.

Shewhart, W.A., 1931. Economic control of quality of manufactured product. Am. Soc. Qual..

Sila, I., Ebrahimpour, M., Birkholz, C., 2006. Quality in supply chains: an empirical analysis. Supply Chain Manage.: Int. J. 11 (6), 491–502.

Slone, R.E., 2004. Leading a supply chain turnaround. Harvard Bus. Rev. 82 (10), 114–121.

Sparks, R., OkuGami, C., 2010. Data quality: algorithims for automatic detection of unusual measurements. In: Tenth International Workshop on Intelligent Statistical Process Control, Seattle, WA.
Topalidou, E., Psarakis, S., 2009. Review of multinomial and multiattribute quality control charts. Qual. Reliab. Eng. Int. 25 (7), 773–804.
Towill, D.R., Naim, M.M., Wikner, J., 1992. Industrial dynamics simulation models in the design of supply chains. Int. J. Phys. Distrib. Logist. Manage. 22 (5), 3–13.
Trautmann, G., Turkulainen, V., Hartmann, E., Bals, L., 2009. Integration in the global sourcing organization—an information processing perspective. J. Supply Chain Manage. 45 (2), 57–74.
Tushman, M.L., Nadler, D.A., 1978. Information processing as an integrating concept in organizational design. Acad. Manage. Rev. 3 (3), 613–624.
Veldman, J., Gaalman, G., 2013. A model of strategic product quality and process improvement incentives. Int. J. Prod. Econ. , http://dx.doi.org/10.1016/j.ijpe.2013.03.002.
Von Bertalanffy, L., 1951. General system theory; a new approach to unity of science. Problems of general system theory. Hum. Biol. 23 (4), 302.
Waller, M.A., Fawcett, S.E., 2013. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. J. Bus. Logist. 34 (2), 77–84.
Wand, Y., Wang, R.Y., 1996. Anchoring data quality dimensions in ontological foundations. Commun. ACM 39 (11), 86–95.
Wang, R.Y., 1998. A product perspective on total data quality management. Commun. ACM 41 (2), 58–65.
Wang, R.Y., Kon, H.B., 1993. Towards total data quality management (TDQM). In: Wang, R.Y. (Ed.), Information Technology in Action: Trends and Perspectives. Prentice-Hall, Englewood Cliffs, NJ.
Wang, R.Y., Storey, V.C., Firth, C.P., 1995. A framework for analysis of data quality research. IEEE Trans. Knowl. Data Eng. 7 (4), 623–640.
Wang, R.Y., Strong, D.M., 1996. Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. 12 (4), 5–33.
Warth, J., Kaiser, G., Kugler, M., 2011. The impact of data quality and analytical capabilities on planning performance: In: Insights from the Automotive Industry, Tenth International Conference on Wirtschaftsinformatik. Association for Information Systems Electronic Library, Zurich, Switzerland.
Watts, S., Shankaranarayanan, G., Even, A., 2009. Data quality assessment in context: a cognitive perspective. Decis. Support Syst. 48 (1), 202–211.
Whipple, J.M., Frankel, R., 2000. Strategic alliance success factors. J. Supply Chain Manage. 36 (3), 21–28.
Woodall, W.H., 1997. Control charts based on attribute data: bibliography and review. J. Qual. Technol. 29 (2), 172–183.
Woodall, W.H., 2000. Controversies and contradictions in statistical process control. J. Qual. Technol. 32 (4), 341–350.
Wu, Y., Cegielski, C.G., Hazen, B.T., Hall, D., 2013. Cloud computing in support of supply chain information infrastructure: understanding when to go to the cloud. J. Supply Chain Manage. 49 (3), 25–41.
Wu, Z., Jiao, J., He, Z., 2009. A single control chart for monitoring the frequency and magnitude of an event. Int. J. Prod. Econ. 119 (1), 24–33.
Zacharia, Z.G., Sanders, N.R., Nix, N.W., 2011. The emerging role of the third-party logistics provider (3PL) as an orchestrator. J. Bus. Logist. 32 (1), 40–54.
Zeithaml, V.A., Berry, L.L., Parasuraman, A., 1990. Delivering Quality Service: Balancing Customer Perceptions and Expectations. Free Press, New York, NY.
Zu, X., Robbins, T.L., Fredendall, L.D., 2010. Mapping the critical links between organizational culture and TQM/Six Sigma practices. Int. J. Prod. Econ. 123 (1), 86–106.