11-2019

# Numerical, Secondary Big Data Quality Issues, Quality Threshold Establishment, & Guidelines for Journal Policy Development

Anita Lee-Post
*University of Kentucky*, Anita.Lee-Post@uky.edu

Ram Pakath
*University of Kentucky*, ram.pakath@uky.edu

# Numerical, Secondary Big Data Quality Issues, Quality Threshold Establishment, & Guidelines for Journal Policy Development[1]

Anita Lee-Post

(Anita.Lee-Post@uky.edu)

& Ram Pakath*

(Ram.Pakath@uky.edu; * Corresponding Author)

C. M. Gatton College of Business & Economics

University of Kentucky

Lexington, KY 40506-0034

## Abstract

An IS researcher may obtain Big Data from primary or secondary data sources. Sometimes, acquiring primary Big Data is infeasible due to availability, accessibility, cost, time, and/or complexity considerations. In this paper, we focus on Big Data-based IS research and discuss ways in which one may, post hoc, establish quality thresholds for *numeric*al Big Data obtained from *secondary* sources. We also present guidelines for developing journal policies aimed at ensuring the veracity and verifiability of such data when used for research purposes.

## Key Words

Data Quality; Big Data; Secondary Data; Numerical Data; Quality Threshold.

## Acknowledgment

We thank Dr. Simon Sheather, Ph.D., University of Kentucky, for referring us to the Big Data sets used in this study.

[1] Forthcoming in Decision Support Systems

## 1. Introduction:

Marsden and Pingry (2018) draw attention to "major, persistent numerical data quality issues," in published Information Systems (IS) academic research that undermines the ability of researchers to replicate prior empirical and analytical IS research. For a forthcoming Decision Support Systems journal issue, the authors issued a Call for Papers (CFP), focusing on (a) a response to their paper, (b) detailing quality thresholds for "quantitative" data used in IS research, and/or (c) detailing and arguing for journal policies that emphasize data quality and research reproducibility.

Marsden and Pingry (2018) identify seven alternative ways to generate numerical data (termed, "data types") used in IS research: (i) interviews, (ii) surveys, (iii) field experiments, (iv) quasi-experiments, (v) controlled laboratory experiments, (vi) empirically observed with or without accuracy control, and (vii) purchased data from third-party vendors. They assess the accuracy of each of these data types by how clearly and precisely one can answer seven questions about the data (i.e., What, When, Where, Why, Who, Which, and How?), hereafter referred to in this paper as, "6W-1H." Because a single data set could serve multiple purposes over time, they argue that the "Why?" question is the least important but regard accurate answers to the remaining questions as *necessary* conditions for data accuracy, validity, and research reproducibility. The authors conclude that data generated using carefully controlled and documented laboratory experiments that use Vernon Smith's (1982) induced-value approach to laboratory experimentation represent the gold standard or upper threshold of data quality. At the other extreme, the authors regard survey data as often representing the lowest quality threshold and present their reasons for this view.

In this paper, we focus on Big Data-based research, given the increasing use of such data. A researcher may obtain Big Data from primary or secondary data sources. Sometimes, acquiring primary Big Data is infeasible due to availability, accessibility, cost, time, and/or complexity considerations. Given the availability of a vast number of secondary Big Data sources, we discuss ways in which one may post hoc

establish quality thresholds for *numerical* Big Data obtained from *secondary* sources. We also advocate guidelines for journal policy development to help ensure the veracity and verifiability of such data when used in academic research. Thus, this paper is an attempt at addressing the foci mentioned in (b) and (c) of Marsden and Pingry's (2018) Call for Papers in the context of secondary, numerical Big Data use in IS research.

We organize the remainder of this paper as follows: In Section 2, we briefly describe Big Data and sources of Numerical Big Data. Using a real-world, secondary Big Data set as an example, we identify several data quality issues that the data set suffers from in Section 3, followed by a discussion on how one may, post hoc, establish data quality thresholds for such data sets. In Section 4, we consider possible ways that academic research journals in the IS (and other) field(s) could assess to what extent the quality of secondary, numerical Big Data used in research meets quality thresholds. We present concluding remarks in Section 5.

## 2. Big Data and Big Data Sources:

The concept of Big Data has been defined in various ways for several years now with Gartner Analyst Laney (2001) describing Big Data as data that we distinguish by its volume, velocity, and variety attributes (i.e., the 3 V's). Over the years, other terms have been added to the three V's, with, e.g., veracity being popularly used as a fourth "V." Others, such as the SAS Institute, have developed their own extensions (Variability, Complexity). None of these terms, however, is unambiguously defined in precise terms and interpretation varies with context and time (e.g., what is "big" data in particle physics (or, today) is different from what it is in marketing (or, in particle physics, tomorrow)). As such, we adopt the context-dependent view that Big Data is data: (i) that is of extremely large quantity; (ii) that is possibly arriving at a very fast rate; (iii) that contains structured, alphanumeric data elements possibly along with unstructured text, voice, video, and/or audio data elements; (iv) whose accuracy is not always guaranteed and may have to be established;

and, (v) which is possibly captured or assembled from different kinds of primary and/or secondary data sources.

Baesens et al. (2016) identify the following sources of Big Data: (a) Large-scale enterprise systems (e.g., ERP, SCM, CRM systems); (b) Online social graphs (e.g., Facebook, Twitter, Weibo); (c) Mobile devices (e.g., cell phones, tablets, laptops, PDAs); (d) Internet of Things (i.e., a sensor interconnected communication network of living and non-living objects (e.g., gadgets, vehicles, people, animals); and, (e) Open/Public data (secondary data, gathered and made available freely for others to use).

Note that Baesens et al. (2016) regard all data other than open/public data as being "primary" data or data gathered by an entity interested in obtaining that data first hand. As such, data category "e" (i.e., open/public data) is data from any or all of the preceding categories ("a" through "d") made available for others to use (i.e., is secondary data). Here, we take "others" to mean those not associated in any way with acquiring, first-hand, any of the data in the open/public data source(s) in question. Our focus here is on category "e." Further, given our interest in *numeric*al data, we do not consider data source "b" (which is largely textual) and any non-numerical content generated by the devices mentioned in "c" further in this paper. We briefly elaborate on categories "a" and "d" next to understand their data characteristics.

Big Data generated using Large-scale Enterprise Systems is *largely* comprised of *structured* data. For example, CRM systems contain Customer, Employee, Contract, and Purchase History data stored in standalone, client-server, or cloud databases. Much enterprise data is drawn from multiple other source points (e.g., transactional databases) within the enterprise. Additionally, *unstructured* data in the form of text content could also be part of such systems (e.g., email/instant messaging interactions with customers and employees). Depending on the enterprise (e.g., Amazon) and point in time (e.g., Black Friday), the data could also be voluminous and streaming (i.e., arriving at a fast rate). Taken in conjunction with the seven data types defined by Marsden and Pingry (2018), one may conclude that a CRM system could contain data of any type except those involving experiments (i.e., types iii, iv, and v). With a well-designed CRM and underlying

transactional systems, an enterprise should be able to answer all seven of the data accuracy-related questions (i.e., the 6W-1H questions) satisfactorily. However, there is no such guarantee with any or all enterprise system data.

The Internet of Things (IoT) is yet an unfolding concept. IoT Data is comprised of many kinds of data depending on the "things" involved. Examples include, status data (e.g., is the equipment working normally?; how many parking spaces are open at which locations?; which signs of heart disease does a patient exhibit, if any?); location data (e.g., where is the forklift truck at present?; which warehouse shelf stocks Brand X shampoo?); and, automation data (e.g., climate control/electricity use data generated by "smart" (i.e., IoT-enabled) buildings; what speed is the driverless car travelling at?)). As these examples illustrate, much IoT data is *structured*, alphanumeric data. However, *unstructured* content is also possible (e.g., navigation instructions to a parking location with available spots). Given the use of automation and sensors in the IoT, it would seem that much Big Data generated by the IoT should also pass muster with regard to the 6W-1H data accuracy attributes posited by Marsden and Pingry (2018). There is no guarantee, however, as two sensors located side-by-side could each record data quite differently (e.g., because one or both malfunction, one is more/less sensitive than the other is).

In the case of any of the three primary (numerical data) sources of interest, to the extent an academic researcher can obtain information related to the data accuracy attributes (i.e., 6W-1H) from the source providing the data for research, such as a business enterprise, the researcher will be in a position to furnish these to anyone who seeks clarification. However, being prepared to furnish such second-hand assurances in and of itself is not proof of the veracity of the data. The burden of proof rests with the researcher and herein lies the challenge.

Turning to the data source of interest to us, Open/Public Data, the data in such repositories is sourced from elsewhere. Secondary data sites are plentiful. As one example, Marr (2016) provides a listing of thirty-three open Big Data sites for public use. A popular site used in classrooms by students and teachers

is Kaggle.com that hosts over 14,000 open data sets at the time of writing. Another example is Data Planet from Sage Publishing that presently offers over 6 billion small and big data sets coupled with data crosscheck and visualization capabilities. Typically, however, such sites provide limited information in terms of the 6W-1H questions. In the following section, we make use of an example public Big Data set to both illustrate some of the data accuracy challenges such data could pose and propose ways in which a researcher could authenticate the data.

## 3. Quality Issues and Thresholds for Secondary, Numerical Big Data:

### 3.1. Illustrative Data Quality Issues:

To consider threshold establishment for secondary, numerical Big Data, we first illustrate some of the issues present in Public Big Data sets by considering the following data: 2013-2017 City of Chicago Taxi Trip (CCTT) Data https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/data). Table 1 provides a brief description of this data set (we have renamed some fields for clarity).

Of the 23 fields, 13 are numerical. As the site notes, the city's Department of Business Affairs and Consumer Protection (BACP), charged with assuring the quality and safety of taxi rides, makes this data available to the public. BACP gathered the above data through two "major payment processors" that process most of the taxi payments in the city. BACP takes some effort at ensuring data privacy by masking taxi medallion numbers, times (all times are rounded to the nearest 15 minutes), and locations (only census tracts and community areas are shown). BACP also undertook limited data *cleansing*. Specifically:

- Trip times less than 0 or greater than 86,400 seconds were removed. (Rationale: Trip times cannot be negative and trips cannot exceed 24 hours even accounting for stops, as city regulations do not allow working times that exceed a day).

- Trip lengths less than 0 or greater than 3,500 miles were removed. (Rationale: Travel distances cannot be negative and the farthest distance one can drive from Chicago and remain within the US territory is 3,500 miles.).
- If any component of the trip cost is less than $0 or greater than $10,000, all components of the trip cost were removed. (Rationale: trip cost cannot be negative; the choice of $10,000 is arbitrary).
- Some duplicate records (0.45%) were removed with duplicate records being identified using the following nine fields: <Taxi ID, Trip Start Time, Trip End Time, Trip Time, Trip Length, Trip Start Census Track, Trip End Census Track, Trip Start Community Area, Trip End Community Area>. (Rationale: A particular taxi cannot have two or more trips that start and end at the exact same times, on the same day, at the same start and end locations.)

We ran several data quality tests involving some of the numerical fields of this partially cleansed data set using the visual and predictive analytics package JMP Pro, from SAS, and summarize our findings in Table 2. In assessing Entity Integrity violations, we used the same nine fields (mentioned earlier) that BACP used to check for duplicate entries. We have highlighted (in bold) the relatively larger values in the table, but our point is that despite the cleansing attempts by BACP, the data set yet has multiple issues. In fact, some of the cleansing steps BACP undertook has actually exacerbated the missing value problem (e.g., the BACP left blank Trip Length fields with values "< 0" or "> 86,400," but the records themselves remain in the data set). A thorough assessment would involve additional checks. For example: (a) Does the same taxi show up on multiple, overlapping trips in the same time window?; (b) Does the Taxi ID match with Taxi Company in every instance?; (c) Are the Trip Start and Trip End Census Tract, Community Area, and Centroid entries mutually consistent?; and, (d) Are the Trip Costs consistent with official Taxi Rates (found at, https://yellowcabchicago.com/rates/)? There also is the concern of how to go about treating missing/incorrect values, a topic too involved to discuss here.

Out of curiosity, we accessed the New York Taxi and Limousine Commission (NY TLC) Trip Record Data (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page). This data set spans January 2009 – June 2018 at present and contains over a billion records pertaining to Yellow Taxicab, Green Cab, and For Hire Vehicle rides. The site notes that TLC-authorized technology providers had gathered the data on their behalf. The TLC makes it clear that it cannot guarantee the accuracy of the data. The fields in these datasets are similar, though not identical, to those in the CCTT data set. We ran analogous checks on the Yellow and Green taxi sets for just the month of January 2018. Here, the most glaring observation was that the Trip Total did not equal the sum of the component costs in 1,467,402 of the 8,759,874 records (i.e., 16.75%) for Yellow Taxicabs in January 2018 and 75,437 of the 793,529 records (i.e., 9.51%) for the Green Cabs. These data sets also had small percentages of *negative* Trip Total entries (0.05% and 0.24%, respectively).

### 3.2. Establishing Data Quality Thresholds for Numerical, Secondary Big Data:

If a researcher is the primary data acquirer, then establishing and adhering to quality checks is largely in his/her control. Even here, there is potential for dirty data. One of the authors was involved, many years ago, in a carefully controlled and documented induced-value laboratory experiment, a setting described as establishing the gold standard for data quality thresholds by Marsden & Pingry (2018). A student researcher was further processing data generated by human participants in parallel, using spreadsheet software and accompanying data analysis programs running on multiple machines, in a networked research laboratory. One day, the student alerted the team to the fact that numerical, subject-generated data values in individual worksheets that should have remained static were arbitrarily changing. We verified this to be the case, ceased the attempt, and restarted the analysis using a small number of standalone machines despite the delays that this induced. The problem did not manifest itself in the new environment. We had no explanation for why the participant-generated data was morphing on the networked machines and chalked this up to, “some network error.” In this instance, the student chanced upon this behavior as the analysis would run for several hours in the lab and this person happened to stop by during a run. Had the run terminated before his/her arrival,

none of us may have detected this data quality problem. Analysis based on dirty data would have found its way into a publication and, perhaps, never discovered. This anecdote underscores Marsden and Pingry's (2018) observation, "… no data type is inherently of low quality and no data type guarantees high quality …."

As our discussions in Section 3.1 reveals, data provided with cleansing and documentation by a secondary provider, even with implicit or explicit quality assurances, could contain unaddressed quality issues that merit further attention. Establishing quality with secondary data is a more difficult task as one usually is doing this post hoc – i.e., quality control during a (data acquisition) process is usually "cheaper" than control after the process has generated output. In addition, the purpose for which a researcher uses the secondary data requires establishing quality checks tailored to the task that could be different from checks, if any, applied in originally gathering the data. Thus, one cannot establish hard thresholds for any secondary data set without cognizance of the application at hand. For instance, consider Age (in years) as a field. The upper and lower data value bounds set for a data set about Pre-School through Elementary School children would be different from those for Middle school through High school students. If the data is an extract from a Big Data set about all school students, it is a researcher's responsibility to first cleanse all of the data and then extract data rows of interest. It would be incorrect to extract data rows of interest first and then attempt cleansing because of the checks and crosschecks needed to establish whether the *seeming* data rows of interest are indeed the *correct* data rows of interest. For instance, a student taking AP Calculus shows up as a middle schooler. Was this actually a middle school student taking AP classes (e.g. as a gifted student) or was he/she actually a high schooler? If one is unable to get this clarified, one must drop this "seeming" middle schooler from further consideration. Below, we discuss some ways to conduct *post hoc* authentication.

We used JMP Pro for our quick analyses to spot some readily-discernable quality issues. The process of data "wrangling" (i.e., data acquisition, data unification, and data cleansing) can be accomplished in different ways based on circumstances. For instance, one may use "traditional" spreadsheet and database management tools like MS Excel and MS Access/MS SQL Server/Oracle Database. One may also

accomplish wrangling by writing code in an environment like R or using a language like Python. Today, sophisticated wrangling capabilities are available via standalone and/or integrated products from certain Data Cleansing, Business Intelligence, Data Visualization, and Data Analytics vendors such as Quadient (Data Cleaner), IBM (Cognos Analytics), Alteryx, and SAS (Visual Analytics). Modern products also embody advanced statistical and machine learning capabilities to help ease the cleansing task but none *fully automates* the task under all circumstances and one cannot *assume* any is thorough in cleansing. An extensive list of about 80 contemporary products with data cleansing capabilities is available at softwareadvice.com. Choices include products ranging from the inexpensive to the very expensive; niche products (i.e., tailored to particular industries) to general-purpose products; on-premise or cloud-based products; products targeted at Small, Medium, or Large businesses; and products that run on Windows, Mac, or Linus platforms.

Data cleansing should help prepare a secondary Big Data set user better respond to some of the 6W-1H questions. How well, he/she is able to answer each question is context dependent. Some sites help answer some of the questions unambiguously. For example, in the CCTT and NY TLC data sets, we unequivocally know *what* data was gathered, *when* it was gathered, *where* it is located, and *why* it was gathered by the primary sources. We also know, somewhat less clearly, *who* gathered the data (the BACP through two payment processors for the CCTT data and via two TLC-authorized technology providers for the NY TLC data). We have little information on the "how" question (i.e., what instruments and/or artifacts where involved). Through a researcher's own data wrangling efforts, we also can determine *who* the researcher is, *why* he/she acquired the data, *what* data he/she is using, *where* his/her data extract is located, and to what extent the data quality was improved. Both taxi data sites also provide contact information (dataportal@cityofchicago.org or @ChicagoCDO; FOIL@tlc.nyc.gov) so that one can seek further clarifications on the source data, how it was gathered, and/or request additional data (if available). Such features make data from these secondary sources more amenable for academic research use.

We note that, without such secondary sources, these data would be impossible to gather in entirety even if one had the required resources (time, money, and expertise). This is also the case with other confidential data like healthcare data, pharmaceutical data, diseases data, and corporate data, for instance, which are usually gathered, used, and sometimes released later for public use with suitable masking. Examples are numerous and include FBI Crime data, CDC Cause of Death data, Medicare Hospital Quality data, Bureau of Labor Statistics data, Dow Jones Weekly Returns data, and Walmart Historical Sales data. In our view, restricting all Big Data used in academic research to be only *primary* data, would severely curtail academic research given non-availability, lack of access, time or cost constraints, and/or complexity considerations. Academic IS research must allow for the use of carefully selected and vetted secondary Big Data. Consider, e.g., the errors in the SEC's EDGAR system whose data is the basis for many top-tier Finance journal publications. It is only as of October 2018 that the SEC began emphasizing data quality, beginning with EDGAR Release 18.3 (https://xbrl.us/news/sec-efm-release48/).

4. Quality Threshold Policy Guidelines for Academia:

We advocate the following guidelines for developing journal policies for all submissions using secondary, numerical Big Data:

i. The authors must clearly identify all secondary Big Data sources, provide access information for these sources, and provide contact information of entities who may be reached for clarifications about source data.

ii. The sources that provide the secondary data must be credible and well known (journal-specific examples of admissible and inadmissible sources may be provided, as a guide to authors).

iii. The authors must extract relevant data from the sources accessed, further process the data as necessary (i.e., to improve quality and/or to customize the data for the application of interest), and be ready to make the processed data accessible to the journal upon request. In particular, authors must carefully document

quality issues addressed and changes made to the data to render it usable. In so doing, the authors must also clearly describe the steps (tools and techniques) used for post hoc data processing.

iv. Authors must provide clear answers, to whatever extent is possible, to the 6W-1H questions. These answers should help a journal decide if it must solicit further information, or if it could send the article out for further review, or reject the submission outright.

v. The article itself must provide links to the original secondary data sources, the extracted data, and to the data as processed for use, for the benefit of readers.

vi. Individual journals, groups of journals, or IS academic groups (e.g., departments at universities, IS academic societies), could choose to establish repositories listing credible, well-known secondary data sources or hosting/pointing to data from such sources. These may also include vetted, primary data gathered by prior researchers that the repository makes available for others to use.

We developed a checklist (see Figure 1) that academic journal reviewers/editors could use in authenticating the quality of Big Data sources and extracts used when reviewing research articles.

## 5. Concluding Remarks:

In this paper, we posit that it is not always possible to acquire primary Big Data due to availability, accessibility, cost, time, and/or complexity considerations. Therefore, we consider secondary sources of Big Data as viable sources for exploitation for academic research. A compelling reason for this is the vast number of open secondary data sources that are available today containing governmental, financial, economic, health, sports, societal, corporate or other data made available for public use. Using two city-government data sources, we first identify some of the numerical data quality issues present in these sets as examples of such issues in general. We then present different ways in which a researcher could cleanse such data, ranging from largely manually guided efforts to highly automated ones. We draw attention to the growing body of sophisticated data wrangling tools that a researcher could exploit to help ease the task of wrangling. We then move to answer the 6W-1H question that Marsden and Pingry (2018) recommend for verifying and

establishing the quality level of a given data set. We provide responses to these questions in the case of the two data sets used in this study. The level of data quality finally assigned to a given secondary data set is context-dependent (i.e., dependent on the extent to which one can clearly and completely answer each of the 6W-1H and other questions shown in Figure 1).

We narrate a past personal experience that depicts that even data obtained from a carefully controlled, induced-value laboratory experiment could be susceptible to quality issues during or after it was gathered. We conclude by advancing a set of guidelines and a checklist aimed at helping IS journals enhance the veracity and verifiability of Big Data obtained from secondary sources. Taking into consideration Marsden and Pingry's (2018) observation that no data type can be of inherent "low" quality or of guaranteed "high" quality, we take the view that academic IS journals should consider credible sources of *secondary* Big Data as viable sources of Big Data for academic IS research, provided a researcher can answer the questions in the checklist (Figure 1) to the journal's satisfaction (i.e., we expect, journals in the IS field would vary in their quality expectations, just as in any field). While our focus has been on numerical data, our arguments would largely carry over to non-numerical data as well.

## Bibliography


Baesens, B., Bapna, R., Marsden, J. R., Vanthienen, J., Zhao, J. L., 2016, "Transformational Issues in Big Data and Analytics in Networked Business," Management Information Systems Quarterly, 40(4), pp. 807-818.

Laney, D., 2001, "3D Data Management: Controlling Data Volume, Velocity, and Variety," Application Delivery Strategies, File 949, Meta Group Inc.

Marr, B., 2016, "Big Data: 33 Brilliant and Free Data Sources Anyone Can Use," Forbes.com, Feb 12, 2016, accessed 01/31/2019.

Marsden, J. R., Pingry, D. E., 2018, Numerical Data Quality in IS Research and the Implications for Replication, Decision Support Systems, 115, pp. A1-A7.