Quantitative Research: Preparation of Incongruous Economic Data Sets for Archival Data Analysis

Gunnar Lucko, Ph.D., A.M.ASCE[1]; and Zane W. Mitchell Jr., Ph.D., M.ASCE[2]

Abstract: In the field of construction engineering and management, archival data sets are not always as correct and consistent as it would be desirable. Between different sources that are studied, e.g., companies, they may differ in format or content and within them, they may still be incongruous and require substantial preparation. This makes examining theories and extracting trends from historic data more difficult than it is for carefully controlled experimental studies or for collecting new data. The purpose of this paper is not to review the regression models that the writers developed during their research, but to focus on the data preparation that had to be applied before those analyses. The objective is to outline various techniques that can be applied to archival data that are related to construction engineering and management to give researchers a set of best practices on data preparation that can assist them in gleaning truths from them.

DOI: 10.1061/(ASCE)CO.1943-7862.0000078

CE Database subject headings: Construction companies; Construction equipment; Construction industry; Data analysis; Economic factors; History; Computer software; Statistics; Databases.

Author keywords: Construction companies; Construction equipment; Construction industry; Data analysis; Economic factors; History; Inflation, economic; Qualitative analysis; Spreadsheets; Statistical models.

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

Introduction

Construction engineering and management research is by its nature strongly related to field operations. Contracting companies routinely measure numerous aspects thereof, especially time, cost, and productivity, to fulfill the project management functions of planning, optimization, and control. This accumulates a valuable pool of data for researchers to explore. However, such data may be collected in a "home-grown" manner, lacking a standard format across different companies (Mitchell 1998). Data from a single source can still require substantial preparation to be usable, especially if they are paper based. Since quantifying economic phenomena is crucial to stay competitive in today's industry and years of archival data can be at the researcher's disposal, the challenge is to devise scientifically valid ways to compare "apples to apples."

Both writers have extensive experience in preparing archival economic data. Mitchell (1998) focused on repair costs of construction equipment. Data were accumulated in various media from four heavy/highway construction firms whose size, fleet, and computing infrastructure differed. The resolution ranged from detailed records of each individual repair to periodically aggregated data. These dissimilar data had to be integrated and prepared to derive predictive equation. Lucko (2003) focused on the residual value of equipment. Data on four major manufacturers were collected from two different auction record publishers that covered numerous types of equipment. While the data sets were in electronic form, their contents still had to be checked for errors, matched with other data that characterized e.g., the size or the macroeconomic context, corrected for inflation, normalized, and finally purged of incomplete or redundant entries.

For the purpose of this paper, *archival data* refers to data that were created before being used for research or that are collected concurrently but independent of the purpose of the research. The type or contents of the data and the method of collection can vary widely. Archival research can be employed for many research questions. For the purpose of illustration, the writers present practical examples of data preparation techniques from their research studies in construction equipment economics, but are hopeful that these subject-unspecific techniques will assist researchers on other topics as well. *Data preparation* refers to systematically collating and transforming unformatted, unconnected, or otherwise initially unusable data into a consistent and coherent data set that can be used for a specific analytical purpose. Yu (2007) cited earlier work that listed data cleaning, transformation, integration, reduction, and discretization as separate elements of data preparation. Rajagopalan and Isken (2001) additionally noted that the "preparation involves enhancing and enriching" the data and defined *data quality* to mean the "accuracy, relevance, completeness" of data with respect to their intended purpose, which was refined into dimensions including their accessibility, sample size, completeness, consistency, credibility, clarity, format, correctness, unbiasedness, and timeliness (Pipino et al. 2002). Poor quality data invalidate research studies (De Veaux and Hand 2005) and "compromises decision making" in the business environment, a major competitive disadvantage (Redman 1998).

[1]Assistant Professor, Director, Construction Engineering and Management Program, Dept. of Civil Engineering, The Catholic Univ. of America, 620 Michigan Ave., NE, Washington, D.C. 20064 (corresponding author). E-mail: lucko@cua.edu

[2]Associate Professor, Chair, Dept. of Engineering, Univ. of Southern Indian, Evansville, IN 47712. E-mail: zwmitchell@usi.edu

Note. This manuscript was submitted on July 24, 2008; approved on April 3, 2009; published online on April 30, 2009. Discussion period open until June 1, 2010; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Construction Engineering and Management***, Vol. 136, No. 1, January 1, 2010. ©ASCE, ISSN 0733-9364/2010/1-49–57/$25.00.

Importance

Zhang et al. (2003) considered data preparation to be one of four steps of knowledge discovery, which they described as an iterative process, alluding to it being part of the scientific method. They highlighted its fundamental importance by estimating "that data cleaning and preparation takes approximately 80% of the total data engineering effort. Data preparation is, therefore, a crucial research topic. However, much work in the field of data mining was built on the existence of quality data" (Zhang et al. 2003. Soibelman and Kim (2002) cited earlier work with a somewhat different estimated effort yet drew the same conclusion: "60% of the time goes into preparing the data for mining, thus highlighting the critical dependency on clean, relevant data. The actual mining step typically constitutes about 10% of the overall effort. Thus, the process of data preparation is one of the most important parts of the entire process and one of the most time consuming and difficult." It is further acknowledged that data preparation is in fact indispensable, as hardly any archival data are directly usable in an analysis. "One of the major obstacles to using organizational data for mining and knowledge discovery is that, in most cases, it [*sic*] is not amenable for mining in its natural form" (Rajagopalan and Isken 2001. Due to the essential nature of data preparation, research has not only sought to advance techniques, as is described in "Data Mining" and the remainder of this paper, but has also studied its positive impact on knowledge discovery in various domains (Zhang et al. 2003; Rajagopalan and Isken 2001).

Applicability

Types of construction engineering and management studies for which the data preparation techniques described in this paper can be useful include, but are not limited to, further research on equipment economics (Fan et al. 2008, 2007), productivity studies based on field records such as schedules, logs, and other routine records (Thomas and Horman 2006; Thomas 2000) or based on macroeconomic measures and industrywide cost records (Rojas and Aramvareekul 2003; Allmon et al. 2000), safety studies based on accident reports, databases of industry statistics, and hazard analyses (Carter and Smith 2006; Arboleda and Abraham 2004; Huang and Hinze 2003), metastudies that evaluate large quantities of previously published studies and possibly their data sets to explore the breadth and depth of knowledge on a topic (Abudayyeh et al. 2006, 2004; Pietroforte and Stefani 2004; Chevallier and Russell 1998), data mining and sociology-type analyses of correspondence, internal memoranda, site diaries, and other written records of construction projects (Zhu et al. 2007; Caldas and Soibelman 2006; Hajjar and AbouRizk 2000), and any other types of studies that may in part or fully rely on archival records. Additionally, studies that collect new data with paper records, especially survey questionnaires that, e.g., explore the productivity or site safety (Ezeldin and Sharara 2006; Zayed and Halpin 2004; Mohamed 2002) or the management or worker attitudes and motivations (Cox et al. 2006; Ling 2002; Chang 2001) can also benefit from exercising the cautious approach to ensuring data quality that is advocated and illustrated by this paper.

Scholarly publications typically contain a section on how data were collected (Rojas and Kell 2008), but not necessarily describe what practical problems were encountered. The particular contribution of this paper is therefore to assist future researchers at the critical interface between data collection and analysis, both of which are fundamental steps within the scientific method. It describes successful preparation techniques for a meaningful statistical analysis, e.g., hypothesis testing or regression modeling. Details on how to overcome problems in structure, format, labeling, frequency, and other properties are illustrated with specific examples of economic data. The time and cost intensive effort of developing and executing the data collection from scratch, which could potentially take years to yield a sufficient sample size for analysis, is thus avoided. While the writers sought to address the broad range of problems that they encountered in their own research, there is no guarantee that this list is exhaustive or applicable to all future studies. Considering that the characteristics of existing data and the manners in which they can be examined are limited only by the researchers' creativity, it is virtually impossible to provide one general approach of data preparation that would anticipate and mitigate all possible problems.

Data Mining

Recent years have seen a tremendous growth in archival data created through an increased use of information technologies in the business environment (Rajagopalan and Isken 2001), particularly due to less expensive hardware, e.g., sensors and storage media (Soibelman and Kim 2002). This vast pool of data holds valuable knowledge for whose extraction scientific principles must be applied. This process is known in the literature as *knowledge discovery in databases* (Ng and Soibelman 2003) as a process following knowledge storing and sharing (Rajagopalan and Isken 2001), or more specifically as *data mining* (Stegemann and Buenfeld 2004), which refers to the actual process of extracting trends or patterns (Zhang et al. 2003). The objective of data mining is to enable extracting rules on clustering, classification, association (e.g., linear regression), sequence, or time series (Lee et al. 2008). Its application can confirm, refine, or expand existing knowledge in a particular domain. Examples abound, e.g., in health care (Goodall 1999), market research (Hand et al. 2000), aviation safety (Jeske and Liu 2007), and prediction of hurricane damages (Nawari 2008). Engineering applications include pavement management (Attoh-Okine 1997), RC design (Roddis and Zhang 2000), water reservoirs (Bessler et al. 2003), contaminants in concrete (Stegemann and Buenfeld 2004), traffic flows (Amado and Virkler 2006), and many others (Fayyad and Smyth 1999). In construction engineering and management, knowledge discovery has been used, e.g., to classify construction documents (Caldas and Soibelman 2002) and to support facilities maintenance (Ng and Soibelman 2003) through mining text data. A recent study on the residual value of heavy equipment (Fan et al. 2008) continued the earlier work of the writers with an autoregressive tree approach. Current research continues to expand the analytical approaches of data mining, e.g., by using neural networks, fuzzy logic, and genetic algorithms (Yu 2007). All studies had in common that they analyzed very large data sets whose imperfections required manipulation for data preparation.

General Considerations

Central questions for the researcher at the onset of a study include the type, range, and source of the necessary data. It is assumed that the researcher has already evaluated the advantages and disadvantages of using archival data from one or several sources instead of collecting new observations in the field or the laboratory. This is not an easy decision. The researcher has great control over the collection methodology for new data, but no influence on

J. Constr. Eng. Manage., 2010, 136(1): 49-57

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

how the existing data were created. Data from field sources are more prone to containing irregularities or noise, which can be somewhat dampened by their potentially larger volume. Yet laboratory studies are often infeasible or impossible for examining economic phenomena due to their high cost, lengthy duration, and limited range. The challenge to extract quality data thus emerges. Several questions need to be answered satisfactorily regarding the existing data (Rajagopalan and Isken 2001):

- Which data are required and which are available?
- What are the type, range, and source(s) of the data?
- Was the collection methodology of the data adequate?
- How can incongruous data be matched consistently?
- How can various sources be merged into one data set?
- How are and should the data be formatted and labeled?
- How can errors in the data set be detected and corrected?

Variables and Sample Sizes

Even before existing data are compiled, the researcher must decide what will constitute the "input" and "output" of the planned model. The response, or dependent variable, is predicted from one or several explanatory, or independent variables. Variables that are not relevant to the study or that lack data points at hand should not be included (Soibelman and Kim 2002). Before the analysis it is unknown which ones of the possible explanatory variables will contribute to the final regression model in a statistically significant way. The number of explanatory variables should thus initially be kept as large as possible. Furthermore, the central limit theorem dictates that a larger sample size should yield a greater precision of the predictions that will be made with the model. The reason for this is that to accurately reflect the variability of the real phenomenon, the model needs to be built from data that contain a sufficient amount of such variability in all of their variables, as sparse data sets yield only insights of limited validity (Stegemann and Buenfeld 2004). Statistical equations are provided in the literature (e.g., Green 1991) to determine how many samples from the population are sufficient, depending on the desired confidence level and on the number of explanatory variables in the model. Acceptable and expected ranges of values for each variable should be determined a priori by the researcher, which defines the scope of the study. For example, the residual value study covered only common types of construction equipment of up to 15 years of age to ensure a broad applicability of its results. Sample sizes outside this range were too small to be valid. Within each data point, the type, make, model, serial number, year of manufacture, and condition rating were reported, along with the auction firm, location, date, price, and a brief description of the setup of the machine (Lucko 2003).

Evidences of Value

Valuation is an important topic in economic studies and provides a key input for investment decisions. Value is a somewhat elusive concept that seeks to indicate how much a good or service could be sold for in the market at a particular time. The "fair market value" is the price could be realized in an open transaction in a competitive market between an equally informed and voluntarily acting buyer and seller (Lucko 2003). Valuation, or appraisal, is only useful if it can occur without having to actually sell the item. Therefore, the data on which it is based must be "evidences of value" (Cowles and Elfar 1977), i.e., actual sales. According to these criteria, public auctions with full information disclosure provide ideal realizations of economic value. Auction purchase prices are far superior to mere sales offers, which only reflect the expectation of a seller without ever having been substantiated in the market. In some cases, their use is unavoidable, e.g., for normalizing the residual value with the manufacturers' suggested retail price, or list price, due to unavailability of actual purchase prices as is described next.

Besides comparison with many similar sales, other valuation approaches exist and are used in various industries, e.g., considering the expected future income that can be generated by an asset. Some confusion is introduced by the concept of depreciation, which typically refers to a loss in book value for taxing a capital investment with simplified accounting methods, e.g., straight-line or declining balance, which may not reflect its actual physical deterioration (Perry et al. 1990). Mitchell (1998) found that companies were charged different amounts for the initial purchase of the same type of construction equipment. Trade-ins and lease allowances clouded the valuation. The list price was used as the initial value to mitigate this, regardless of what companies paid. This standardization was feasible because the study focused on predicting recurring repair costs.

Recurring Cost Data

Many different types of expenses can be included in economic models. They can range from straightforward and tangible expenses, e.g., fuel, to complicated and intangible expenses, e.g., the cost of obsolescence. Expenses can be broken down into three broad categories: Direct, provisional, and collateral costs. The researcher must decide which types of costs the economic study will address and remove any superfluous existing data from the data set before analysis.

Direct costs are quantifiable, clear, and directly related to owning, operating, and maintaining an asset. They occur regularly and predictably in a given accounting period and affect a company's operating budget. They are offset by the revenue stream generated by the asset. Examples of direct expenses for companies that operate heavy equipment include fuel, oil, tracks or tires, maintenance, repairs, financing interest and principal payments, taxes, licenses, and insurance.

Provisional costs are internal costs that are intended to cover the anticipated costs of continuous processes that are difficult to quantify or discrete events that only occur a limited number of times during the life of the asset. For example, the costs of major repairs or rebuilds of a unit of equipment are handled by charging a provisional hourly rate to build a repair reserve that is balanced across a particular unit, group, or fleet. Depreciation is another type of provisional cost, which allows for the fact that the value of an asset decreases with the passage of time alone.

Collateral costs are more difficult to quantify and not always part of economic models in the literature, nor used much by managers in the industry practice. They can include obsolescence costs, associated resource impact costs, lack of readiness costs, service level impact costs, and alternative method impact costs (Vorster and de la Garza 1990). Obsolescence costs are incurred as an asset "ages" technologically versus new technologies that provide increased productivity, reliability, or versatility. They can take the form of higher repair or production costs or may appear as bids that are lost because of the then-higher costs of assets with the old technology.

J. Constr. Eng. Manage., 2010, 136(1): 49-57

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

Data Sources

Proprietary Data

In the highly competitive environment of the construction industry, contracts can be won or lost by very small margins (Liu and Ling 2005). Although construction firms often are willing to cooperate with researchers who develop and test theories to help them improve their business, they are disinclined to share their economic data in a form by which competitors might gain an advantage. Proprietary corporate data therefore require written permission to be usable for noncommercial scholarly purposes. Such confidentiality agreement regulates disclosing planned publications to the source. Specifically, it should clearly establish what type and extent of raw, aggregated, or processed data may be published. If raw data are strongly protected, descriptions and illustrations can even use artificial data whose statistical properties are similar to the originals. The agreement should also permit discussing scientific questions of the project with other researchers. Anonymity of participating companies or individuals and protection of the raw data by limiting access and destroying them after a specified time are other typical precautions.

Multiple Sources

Ideally, economic data should be collected from several sources to allow validating the order of magnitude, mean, and variance of the data and enable crosscomparisons that can include demographic data about each source. However, data sources vary in accuracy and reliability depending on their type and internal procedures. Each additional source introduces new biases and possible errors. Governmental sources are generally considered of high quality are consist data in a consistent manner while corporate sources often exhibit more incongruities. Different sources must be coded in the final data set, e.g., with integers or time stamps, to allow statistical testing for differences. While surveying the extent of the available data sets, it is important to ascertain if pertinent data fields are common and complete across all sources. If an individual data set contains less than all of the attributes to address the research question it may need to be eliminated, however extensive, unless the missing entries can be reconstructed in a valid manner.

Data Editing

Frequency and Resolution

It is essential to understand the mechanism by which economic data were collected and recorded, e.g., from technical descriptions or interviews with the respective staff. The terminology and definitions of what constitutes a data point depends on the internal procedures of a source. Data may be either continuous or discrete with a restricted or unlimited range of possible values. Since existing data are not specifically collected for the purpose of the particular study, their frequency and resolution may be inconsistent between different sources. Corporate data may be measured ad hoc when a particular event occurs, e.g., an equipment repair. Macroeconomic data on the other hand often have a monthly or quarterly frequency. Different frequencies that can span from cost items of an individual unit of equipment to the economy at large must be matched in a valid response variable. The largest interval determines the frequency to which data of higher frequency must be adjusted, averaged, or interpolated, e.g., to create an average annual producer price index (PPI) from monthly values to match the years in which equipment age was measured. Another important consideration is the *data resolution*, i.e., the significant digits in their values, which analogously determines the possible numerical accuracy of the response variable.

The equipment age was an explanatory variable in the repair cost study. It could be measured in units ranging from calendar age in years to cumulative usage in hours, the latter of which was chosen (Mitchell 1998). One source presented the severe problem that the explanatory and response variables were recorded in separate databases at different frequencies; meter hour readings were recorded whenever an oil change occurred but repair costs were recorded with a lower frequency at the end of each month. These disparate data were matched via calendar date based on the assumption that events on or before the 15th day of a month could be grouped with its beginning and events after the 15th day with its end, which may have introduced small errors but made the overall study possible. Repair costs that were associated with cumulative hours from oil changes early in a month may have been understated and the costs associated with later oil changes may have been overstated. Fortunately, the errors would be offsetting in the long run if oil changes occurred approximately evenly distributed over time. The matching process ensured that exactly one data pair existed for the lowest data frequency, in this case months. Other sources provided multiple data points for the same cumulative hours of usage if equipment was idle for long periods. All but the first value of multiple occurrences of the same data were therefore eliminated, which was subsequently matched with repair costs as described previously.

Storage Medium

The media in which economic data are recorded range from informal papers via electronic office files to commercial data repositories. Actual paper records are vulnerable to transcription errors that may occur during their initial recording or during extraction for research use. To be usable for analysis, data must be carefully converted into an electronic format by manual transcription or by employing optical character recognition software. Existing electronic data may be readily accessible on PC or may be retrieved off central mainframe computers or distributed networks with data mining software. Their quality must still be verified, e.g., using various sorting functions to identify any fields that contain missing or impossible entries.

Spreadsheets are common for storing economic data and can often be exported directly from local database software or online databases such as Last Bid, formerly Green Guide Auction Report, and Top Bid that cover most heavy equipment auctions but require a subscription fee and various governmental sources of macroeconomic indicators that generally are free of charge. Manufacturers and distributors provide additional valuable data, e.g., price lists, on paper, in text or spreadsheet files, Adobe portable document format, or embedded in Web sites. These must be converted into a common medium and format with separators between each entry, so that data cleaning techniques can subsequently be applied evenly to the consolidated final data set.

The data in the repair cost study initially were in paper records, various PC files, and mainframe computers. *PC and database files* required only little preparation by converting them into formats such as Microsoft Access or Corel Paradox into spreadsheets, or by only adjusting the layout if data were already in Microsoft Excel or Corel Quattro format. Converting database

J. Constr. Eng. Manage., 2010, 136(1): 49-57

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

files was faster than PC files because queries extracted the data into the layout as desired for analysis.

Preparing *mainframe data* can be more challenging because they may be housed in applications that were programmed specifically for a company. Transferring them to unformatted American Standard Code for Information Interchange (ASCII) files is a suitable solution (Rajagopalan and Isken 2001) but adds the data preparation step of separating one row of unformatted entries. Converters within spreadsheet programs allow parsing data but work only if they were originally in tabular format. Moreover, pertinent data may be disjointed across several mainframe reports, e.g., for the same unit of equipment. Data mining software such as, e.g., DataWatch Monarch can help structuring data through creating custom templates and provides graphical reports. For example, cost data were extracted, filtered to include only repair costs, associated with a specific unit of equipment, subtotaled by cost code for each month, and then exported.

Paper records prove to be the hardest to prepare for a meaningful analysis in the experience of the writers. Making them usable is a labor-intensive process that is susceptible to transcription, transposition, omission, and structural errors if files are illegible, unorganized, or incomplete. When the preferred electronic data did not exist for the repair cost study, manual transcription of loose leaf repair receipts into a spreadsheet and subsequent proofreading were required.

Formatting and Labeling

Individual data points typically consist of several columns. Since terminology for economic data are determined by the source, it is recommended that a consistent and intuitive labeling is applied to entire final data set during data preparation. Spreadsheet software allows designating cells as text, numbers, dates, or prices according to their contents. Superfluous formatting, e.g., printer symbols (Rajagopalan and Isken 2001), blank spaces, and tabulator stops can be removed with the editing functions of text processing software. An efficient method of removing formatting is pasting data as unformatted text into spreadsheet software, which unifies all of their fonts.

Sorting the columns of Lucko (2003) residual value data alphabetically brought blank cells and those containing symbols, e.g., "N/A" for not applicable or "-," to the top, which were filled with "." This symbol is commonly used in statistics software to represent missing values. Some data had to be calculated in additional columns, e.g., age as the difference between the year of manufacture and the date of the auction at which the equipment was sold. Age only had an accuracy of years, because the day or month of manufacture was not known. The column of setup descriptions was scanned for negations, e.g., "no," "not," and "inoperable," to isolate entries of equipment that was lacking parts. Location-related data, e.g., postal ZIP codes and state name abbreviations, were converted by simple IF commands in the spreadsheet into geographic regions that served as an explanatory variable in the statistical model. For large data sets it is recommended to add a control column that automatically verifies the correct conversion of each entry with an adequate checksum, which can additionally be formatted to be easily visible.

Dimension and Sign

Sorting data numerically can reveal unreasonable extremes caused by partially complete entries, e.g., negative (199_–1996) or large ages (2002–199_) that must be corrected or deleted. For numerical data it is useful to add column headers that calculate minimum, maximum, mean, and variance in the spreadsheet to quickly gauge their range, centricity, and spread. Two types of accounting errors had to be corrected in the repair cost study. Most common were negative repair costs. Closer investigation with the respective companies showed that negative charges were due to either overcharges or mistakes that had occurred in an earlier month and were removed. Another error occurred due to replaced hour meters in equipment, which caused conspicuous sudden decreases in cumulative hours of usage over time or nonzero cost at time zero. After confirming that a meter had been exchanged, the cumulative hours after replacement were corrected to being continuous or the data point was removed entirely if the value directly before replacement was not known. Whatever the exact nature of the economic data, introducing checks and balances that verify the numerical integrity in terms of dimension, sign, sequence, intervals, and other parameters within or between entries can significantly enhance the data quality.

Typographical Errors

Visually skimming the entire sorted data set for deviations can help identifying and eliminating typographical errors. Lucko (2003) was able to identify and correct errors that occurred between similar looking letters and digits while transcribing the handwritten records into electronic form:

- Letter or digit switches: 7XM instead of 7MX within a serial number.
- Letter or digit changes: 1=I, 2=Z, 3=6=8=B=S, 7=F=T.

Others were spelling variations in alphanumeric entries, e.g., the model name of the equipment:

- Incomplete model name: Excavator PC300 instead of PC300HD-5.
- Incomplete series name: Excavator PC100 II instead of PC100C Series II.
- Missing hyphen: Excavator PW301 instead of PW30-1.

Long numbers may contain superfluous initial zeros that can be removed. Finally, since serial numbers are awarded consecutively to newly manufactured equipment, sudden deviations in the year of manufacture between otherwise constant or sequential entries were corrected to preserve the correct time sequence (Rajagopalan and Isken 2001). Correct spelling and grammatical variations need to be ascertained especially for studies that mine text data (Jeske and Liu 2007).

Conversion Process

Reconstruction

Data sets need to be checked for completeness, as missing data severely reduce the data quality and curtail the attainability of research objectives (Stegemann and Buenfeld 2004). If the overall data set contains redundancy or a predictable pattern, it is possible to fill gaps by carefully comparing the incomplete entry with neighboring ones, provided that no data are fabricated a posteriori. Only in rare cases do data sets overlap to a significant degree, such as for the equipment auction data from Last Bid and Top Bid. They provided two independent records of the same auctions at which a particular unit of heavy equipment had been sold. Redundancy between the two data sets allowed verification by comparing pairs of entries via unique identifiers such as date, location, and serial number. Merging nonredundant entries can

J. Constr. Eng. Manage., 2010, 136(1): 49-57

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

increase the sample size, but should be verified to match the quality of redundant ones. Machines appeared in more than one data point if they were sold several times during their economic life. It was assumed that such multiple entries were independent and could therefore be retained in the overall data set. Reconstruction required that all data points in the merged data set were sorted in the hierarchical order of model, serial number, and auction date. A computer macro *FillGaps* automated this process. If a cell contained "." then the macro matched several identifiers with preceding or succeeding entries to find redundancy, i.e., two records of the same auction sale. The placeholder was then overwritten with the correct value from the adjacent entry as per the pseudocode:

```
//FillGaps [Array EqData (Row, Column)]
for each Row in EqData
do if (Row,Column1)="."
then if (Row, Column2).
Date=(PreviousRow,Column2).Date
and (Row,Column3)=(PreviousRow,Column3)
and (Row,Column4)=(PreviousRow,Column4)
then (Row,Column1)←(PreviousRow,Column1)
else if (Row, Column2).
Date=(NextRow,Column2).Date
and (Row,Column3)=(NextRow,Column3)
and (Row,Column4)=(NextRow,Column4)
then (Row,Column1)←(NextRow,Column1)
```

A variation of this technique also identified the aforementioned sudden changes in an otherwise consecutive sequence. Remaining redundant entries had to be deleted once gaps were filled. A macro *DeleteDoubles*—similar to the previous one—compared identifiers, allowing for a small variability between prices due to currency conversion, and deleted the first entry of each pair.

Verbal Descriptors

A problem that is often encountered in existing data sets is that they contain explanatory variables that consist of verbal descriptors. All explanatory variables must be expressed in numerical terms to be usable in a statistical analysis. For example, the manufacturer name, condition rating, and auction region were categorical variables in the residual value study. Some are hierarchical, e.g., condition (excellent, very good, good, fair, poor); others are not, e.g., region (northeast, south, midwest, west). Categorical variables often serve as somewhat subjective proxies for modeling complex phenomena. Condition is typically appraised visually using a checklist and aggregates the result of wear and tear on different parts of a machine, e.g., tires or tracks, undercarriage, and engine, which is offset by maintenance and repairs. If at all possible, verbal descriptors should be systematically replaced with meaningful numbers (Soibelman and Kim 2002). However, simply assigning integers would only be permissible for a truly hierarchical phenomenon and yet would still be statistically problematic, as it would assume a constant distance between categories.

A valid solution is to convert the integer into binary numbers that serve as sets of "indicator variables." Table 1 shows how four verbal descriptors are represented by an equivalent triplet of zeros and ones. Numbering begins with one, because starting at zero would slightly favor the first category and de facto consider it the benchmark for any further variability in the particular explanatory variable. Each binary provides only part of the distinction between categories, so that either all or none should be included in a statistical model. A drawback is that the conversion removes any actual hierarchical information that may have been implied in the integers.

Table 1. Conversion of Verbal Descriptors to Binary Numbers

		Binary number		
Descriptor	Number	n_1	n_2	n_3
A	1	0	0	1
B	2	0	1	0
C	3	0	1	1
D	4	1	0	0

Economic Indicators

Macroeconomic phenomena are created by complex interactions of many participants. Indicators attempt to capture one specific aspect of a local, regional, national, or global economic situation in a numeric measure. Vast amounts of data are available from government agencies, but also from financial news services, corporate publications, and independent research organizations, e.g., for business cycle indicators. They can provide useful explanatory variables of the economic context within which construction companies acted. But it is probable that they are correlated if they measure similar aspects of the economy (Perry et al. 1990), e.g., inflation indicators, or the cost indices by *Engineering News Record* magazine. It is necessary to determine which ones contribute significantly to the predictive power of the statistical model. Numerous pairs of economic indicators have very high Pearson coefficients of correlation R_{corr}, e.g., the building and construction cost indices, gross domestic product and total retail sales, consumer price index and PPI for machinery and equipment, and others for values from January 1980 to August 2002. Lower frequencies were matched with higher ones by assuming them to be piecewise constant. This slightly decreased the correlation but allowed using all data points. Some indicators are available in regular and seasonally adjusted form to remove recurring swings from actual economic trends. The Bureau of the Census provides, e.g., seasonally adjusted annual rates, which allows direct comparisons between monthly, quarterly, and annual economic values. Consistency in selecting seasonally adjusted or unadjusted indicators is recommended.

Inflation Correction

The phenomenon of inflation is an imbalance between the supply of money and the supply of goods and services (Bodie et al. 2002). Manufacturers' suggested retail prices, auction prices, maintenance and repair costs, and any other cost data must be corrected to an arbitrary common date before they can be analyzed. Inflation adjustments can be made with an annual percentage rate, which creates an exponential growth. If data are insufficient or unavailable to estimate such rate accurately, using governmental indices as per Eq. (1) is a widely accepted technique

$$\text{Price}_2 = \text{Price}_1 \cdot \frac{\text{Index}_2}{\text{Index}_1} \tag{1}$$

where Price=any price in U.S. dollars and Index=price index. They are ratios of current to past prices for standardized baskets of goods or services. Indices with different levels of detail and

J. Constr. Eng. Manage., 2010, 136(1): 49-57

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

composition are available. Some apply to the entire economy; others are regional or cover only specific industries, commodities, and stages of processing. The PPI for finished goods was used to adjust costs related to heavy equipment (Cross and Perry 1995, 1996; Kastens 1997). Once an index has been used to correct for inflation, it cannot be an explanatory variable. Otherwise, the model would suffer from multicollinearity, i.e., a very high correlation between them, which precludes obtaining a closed-form solution. Ideally, all explanatory variables, including economic indicators, should thus be only weakly correlated. A more complex method for calculating a composite price index was proposed by Douglas (1975). If specific indices can be applied to specific categories within the data, such weighted index can be beneficial. For example, the not seasonally adjusted annual PPI for all finished goods grew from 2002 to 2007 by 19.9%, 27.2% for materials and components for construction, and even 74.5% for steel mill products (Bureau of Labor Statistics 2008). A researcher focusing on steel-related construction would be ill advised to use the general PPI under this situation.

Matching Data

If existing data are merged with other data, the question of how to efficiently match them arises. For example, the auction records in the residual value study needed to be categorized by adding size parameters to the data set, normalized by adding list prices, corrected for inflation, and matched with economic indicators. If data need classification, the categories should reflect common features of the item of interest, e.g., footprint area or number of floors for residential construction. Equipment can be characterized based on its performance or capacity, e.g., standard operating weight (empty), general purpose bucket volume, and net horse power (flywheel). A catalog of size parameters was assembled from data by manufacturers and their distributors. Minor conversions and rounding between English and metric units were performed. Rather than introducing an explanatory variable for size, a separate model was created for each smaller, more consistent category, which significantly improved the goodness-of-fit. A computer macro *MatchData* read unique identifiers in each row of the existing data, looped through a given block of new data, and in case of a match wrote the new data next to the existing. For list prices and sizes, the identifier was the model name; for economic indicators it was the auction date.

Normalization

Normalizing variables by dividing them by a baseline value makes them comparable across categories. Dollar values in the residual value study were divided by manufacturers' suggested retail prices, i.e., list prices. Actual initial sales prices would have been ideal, but in their absence for proprietary reasons, list prices were assumed to be generated in a consistent manner across manufacturers. They are devoid of discounts that specific companies may receive and unbiased, unlike purchase prices. Collected in analogy to the size parameters, list prices were interpolated across gaps and, if necessary, extrapolated with an inflation correction before the matching.

Some economic research may investigate assets without identical attributes, but costs can vary strongly depending on these attributes. For example, units of construction equipment can vary in many regards. Differences between the machines may occur in physical setup and in their usage, the latter of which are harder to discern and may be realized in a proxy measure, i.e., a condition rating. It is recommended to create data sets that are as homogenous as possible as long as the required sample size suffices. Predictive equations then seek to quantify the relationship between the response variable and the attributes. For example, Mitchell (1998) compared unlike machines by indexing their repair costs to list prices with a cumulative cost index (CCI) as per Eq. (2)

$$\mathrm{CCI}_t = \frac{PP_0 + \sum_0^t (P_t + L_t + O_t)}{PP_0} \tag{2}$$

where CCI=cumulative cost index; P=cost of parts; L=cost of labor; O=other maintenance costs at time t; and PP_0=list price. All of these cost items were cumulative.

Statistical Preparation

Outliers

Data may take on extreme values due to measurement or recording errors. It is therefore prudent to purge the final data set of such *outliers* (Soibelman and Kim 2002). They are defined as data points that differ significantly from the basic relationship captured by the other data, either in their sign or magnitude, and can substantially distort regression models. A manual technique for identifying them is to graph all data in one or several scatterplots, but various statistical techniques also exist, e.g., residuals that measure how much a data point deviates from the model. It is recommended to use scaled residuals due to their constant variance (Montgomery et al. 2001). Their absolute value then determines if a data point is earmarked for removal, which can be implemented with an IF command in an additional column of the spreadsheet. The threshold should be set depending on the characteristics of the data set in consultation with a statistician. After outliers are identified, it is best to try to identify the reason for the deviation. The statistical model then contains only coefficients that were calculated from the "cleaned" final data set.

Relative Dominance

A statistical issue that manifested itself in the repair cost data is an uneven distribution of data points between individual economic assets that together compose a data set. Some machines were represented with significantly more data points than others. Reasons for this deviation included different usages but also the data collection style of the individual sources. Dominant machines would have had more of an influence on the regression analysis than those with fewer data points relative dominance can be addressed by interpolating data to discrete, evenly spaced intervals, which are chosen to match the data at hand, e.g., 500 h of usage in the repair cost study. It is important to emphasize is that only one data point should be interpolated between any two actual ones, as otherwise the statistical integrity of the data set would suffer from fabricating data.

Variable Selection

Note that for observational studies, including archival research that uses previously existing data from construction companies, no direct causality can be established statistically. Nonetheless, carefully selecting the components and structure allows creating

J. Constr. Eng. Manage., 2010, 136(1): 49-57

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

predictive models with high confidence levels. Techniques for variable selection start the analysis phase after data of possible explanatory variables have been collected with a sufficient sample size and prepared. Pairs of variables (Stegemann and Buenfeld 2004), e.g., two explanatory variables or one explanatory and one response variable, are visually inspected in scatterplots to identify multicollinearity and any trends or patterns. Multiple explanatory variables yield a triangular matrix of scatterplots. These observations assist in determining the model composition, i.e., what mathematical function(s) to use, what order of variable terms, and what possible interaction terms between them, if any. Plotting data or derivatives thereof, e.g., the normal probability plot of the residuals, also yields information whether important regression assumptions are fulfilled by the data set at hand.

Both overfitting and underfitting a model with explanatory variables reduces its predictive power. Three techniques are therefore commonly used to select appropriate variables for being included in a regression model, forward selection, backward elimination, and stepwise selection (Montgomery et al. 2001). *Forward selection* starts with an "empty" model, tests which variable is most significant if added, adds it, and proceeds with testing the remaining variables. *Backward elimination* is the reverse process that sequentially removes the least significant variables from a "full" model and can assist in initially quickly discarding unwanted variables. *Stepwise selection* extends forward selection and alternates between adding and removing to arrive at an overall best-fitted model. While the statistical modeling can be aided additionally with an autoregression algorithm that creates and tests many possible regressions models (Fan et al. 2008), it ultimately still depends on the conscious, experienced researcher and remains both an art and a science.

Outlook

Messner (2003) suggested organizing corporate data based on the organization, its processes and commitments, its facilities, and the environment. An accessible and well-organized data structure that considers these recommendations facilitates their analysis at any point in time, including a posteriori research studies even and especially on to glean insights on aspects that are not part of the routine daily business functions and their reporting requirements. While researchers may have the time and training to transform unsuitable data sets with the data preparation techniques, employees of construction companies may find it challenging to perform the necessary steps on initially unusable archival data. Additionally, pending the adoption of new data management systems, e.g., accounting software, companies are likely to continue collecting data in the same way as they have before. Awareness is therefore called to the fact that data can be a singularly valuable pool of knowledge that is yet to be extracted by data mining techniques. The effort of setting up a data collection and storage system is therefore an investment whose future payoff in terms of savings or productivity improvements is still unknown. In the meantime, the various data preparation techniques that this paper has described are hoped to serve as useful guidelines.

Conclusions

This paper has described how different types of existing archival data in the area of construction engineering and management can be prepared in a valid manner while preserving their integrity. Preparations included validating the sample size, handling confidential data, unifying the format, reconstructing missing values as far as possible, checking the data set for various types of inconsistencies, matching incongruous data with each other, performing an inflation adjustment on economic data, removing outliers, and selecting what explanatory variables are included in the model. While each new study of archival data will present the researcher with new and somewhat unique problems, these general techniques are hoped to serve as guidelines and as a source of inspiration in surmounting the crucial and often underestimated phase of data preparation, which alone enables the subsequent analysis and determines its quality and success.

Acknowledgments

The first writer thanks Joseph D. Lombardo of Learning Seed and Justin P. Molineaux of Computech for their advice on creating effective pseudocode.

References

Abudayyeh, O., Dibert-DeYoung, A., and Jaselskis, E. J. (2004). "Analysis of trends in construction research: 1985–2002." *J. Constr. Eng. Manage.*, 130(3), 433–439.

Abudayyeh, O., Dibert-DeYoung, A., Rasdorf, W. J., and Melhem, H. (2006). "Research publication trends and topics in computing in civil engineering." *J. Comput. Civ. Eng.*, 20(1), 2–12.

Allmon, E., Haas, C. T., Borcherding, J. D., and Goodrum, P. M. (2000). "U.S. construction labor productivity trends, 1970–1998." *J. Constr. Eng. Manage.*, 126(2), 97–104.

Amado, V., and Virkler, M. R. (2006). "Using data mining to analyze archived traffic related data." *Proc., 2006 9th Int. Conf. on Applications of Advanced Technology in Transportation*, K. C. P. Wang, B. L. Smith, D. R. Uzarski, and S. C. Wong, eds., ASCE, Reston, Va., 310–318.

Arboleda, C. A., and Abraham, D. M. (2004). "Fatalities in trenching operations—Analysis using models of accident causation." *J. Constr. Eng. Manage.*, 130(2), 273–280.

Attoh-Okine, N. O. (1997). "Rough set application to data mining principles in pavement management database." *J. Comput. Civ. Eng.*, 11(4), 231–237.

Bessler, F. T., Savic, D. A., and Walters, G. A. (2003). "Water reservoir control with data mining." *J. Water Resour. Plann. Manage.*, 129(1), 26–34.

Bodie, Z., Kane, A., and Marcus, A. J. (2002). *Investments*, 5th Ed., McGraw-Hill, New York.

Bureau of Labor Statistics. (2008). "Producer price indexes: Databases, tables & calculators by subject." *U.S. Dept. of Labor*, ⟨http://www.bls.gov/data⟩ (July 24, 2008).

Caldas, C. H., and Soibelman, L. (2002). "Automated classification of construction project documents." *J. Comput. Civ. Eng.*, 16(4), 234–243.

Caldas, C. H., and Soibelman, L. (2006). "A combined text mining method to improve document management in construction projects." *Proc., 2006 Int. Conf. on Computing in Civil Engineering of ASCE*, H. Rivard, H. Melhem, and E. Miresco, eds., ASCE, Reston, Va., 2912–2918.

Carter, G., and Smith, S. D. (2006). "Safety hazard identification on construction projects." *J. Constr. Eng. Manage.*, 132(2), 197–205.

Chang, S.-T. (2001). "Work-time model for engineers." *J. Constr. Eng. Manage.*, 127(2), 163–172.

Chevallier, N., and Russell, A. D. (1998). "Automated schedule generation." *Can. J. Civ. Eng.*, 25(6), 1059–1077.

Cowles, H. A., and Elfar, A. A. (1977). "Valuation of industrial property: A proposed model." *Eng. Econ.*, 23(3), 141–161.

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.

Cox, R. F., Issa, R. A., and Frey, A. (2006). "Proposed subcontractor-based employee motivational model." *J. Constr. Eng. Manage.*, 132(2), 152–163.

Cross, T. L., and Perry, G. M. (1995). "Depreciation patterns for agricultural machinery." *Am. J. Agric. Econom.*, 77(1), 194–204.

Cross, T. L., and Perry, G. M. (1996). "Remaining value functions for farm equipment." *Appl. Eng. Agric.*, 12(5), 547–553.

De Veaux, R. D., and Hand, D. J. (2005). "How to lie with bad data." *Stat. Sci.*, 20(3), 231–238.

Douglas, J. (1975). *Construction equipment policy*, McGraw-Hill, New York.

Ezeldin, A. S., and Sharara, L. M. (2006). "Neural networks for estimating the productivity of concreting activities." *J. Constr. Eng. Manage.*, 132(6), 650–656.

Fan, H., AbouRizk, S. M., and Kim, H. (2007). "Building intelligent applications for construction equipment management." *Proc., 2007 ASCE Int. Workshop on Computing in Civil Engineering*, L. Soibelman and B. Akinci, eds., ASCE, Reston, Va., 192–199.

Fan, H., AbouRizk, S. M., Kim, H., and Zaïane, O. (2008). "Assessing residual value of heavy construction equipment using predictive data mining model." *J. Comput. Civ. Eng.*, 22(3), 181–191.

Fayyad, U. M., and Smyth, P. (1999). "Cataloging and mining massive datasets for science data analysis." *J. Comput. Graph. Stat.*, 8(3), 589–610.

Goodall, C. R. (1999). "Data mining of massive datasets in healthcare." *J. Comput. Graph. Stat.*, 8(3), 620–634.

Green, S. B. (1991). "How many subjects does it take to do a regression analysis?" *Multivar. Behav. Res.*, 26(3), 499–510.

Hajjar, D., and AbouRizk, S. M. (2000). "Integrating document management with project and company data." *J. Comput. Civ. Eng.*, 14(1), 70–77.

Hand, D. J., Blunt, G., Kelly, M. G., and Adams, N. M. (2000). "Data mining for fun and profit." *Stat. Sci.*, 15(2), 111–126.

Huang, X., and Hinze, J. (2003). "Analysis of construction worker fall accidents." *J. Constr. Eng. Manage.*, 129(3), 262–271.

Jeske, D. R., and Liu, R. Y. (2007). "Mining and tracking massive text data: Classification, construction of tracking statistics, and inference under misclassification." *Technometrics*, 49(2), 116–128.

Kastens, T. (1997). "Farm machinery operation cost calculations." *Kansas State University Farm Management Guide Rep. No. MF-2244*, Kansas State Univ. Agricultural Experiment Station and Cooperative Extension Service, Manhattan, Kan.

Lee, J.-R., Hsueh, S.-L., and Tseng, H.-P. (2008). "Utilizing data mining to discover knowledge in construction enterprise performance records." *Journal of Civil Engineering and Management*, 14(2), 79–84.

Ling, Y. Y. (2002). "Model for predicting performance of architects and engineers." *J. Constr. Eng. Manage.*, 128(5), 446–455.

Liu, M., and Ling, Y. Y. (2005). "Modeling a contractor's markup estimation." *J. Constr. Eng. Manage.*, 131(4), 391–399.

Lucko, G. (2003). "A statistical analysis and model of the residual value of different types of heavy construction equipment." Ph.D. dissertation, Virginia Polytechnic Institute and State Univ., Blacksburg, Va.

Messner, J. I. (2003). "An architecture for knowledge management in the AEC industry." *Proc., 2003 Construction Research Congress*, K. R. Molenaar and P. S. Chinowsky, eds., ASCE, Reston, Va.

Mitchell, Z. W. (1998). "A statistical analysis of construction equipment repair costs using field data & the cumulative cost model." Ph.D. dissertation, Virginia Polytechnic Institute and State Univ., Blacksburg, Va.

Mohamed, S. (2002). "Safety climate in construction site environments." *J. Constr. Eng. Manage.*, 128(5), 375–384.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2001). *Introduction to linear regression analysis*, 3rd Ed., Wiley, New York.

Nawari, N. O. (2008). "The role of data mining techniques in the prediction of hurricane damages." *Proc., 2008 Structures Congress*, D. Anderson, C. Ventura, D. Harvey, and M. Hoit, eds., ASCE, Reston, Va., 1–10.

Ng, H. S., and Soibelman, L. (2003). "Knowledge discovery in maintenance databases: Enhancing the maintainability in higher education facilities." *Proc., 2003 Construction Research Congress*, K. R. Molenaar and P. S. Chinowsky, eds., ASCE, Reston, Va.

Perry, G. M., Bayaner, A., and Nixon, C. J. (1990). "The effect of usage and size on tractor depreciation." *Am. J. Agr. Econ.*, 72(2), 317–325.

Pietroforte, R., and Stefani, T. P. (2004). "ASCE Journal of Construction Engineering and Management: Review of the years 1983–2000." *J. Constr. Eng. Manage.*, 130(3), 440–448.

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

Rajagopalan, B., and Isken, M. W. (2001). "Exploiting data preparation to enhance mining and knowledge discovery." *IEEE Trans. Syst. Man Cybern., Part C Appl. Rev.*, 31(4), 460–467.

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

Roddis, W. M. K., and Zhang, L. (2000). "Equation discovery in databases from engineering." *Proc., 2000 8th Int. Conf. on Computing in Civil and Building Engineering*, R. Fruchter, F. Peña-Mora, and W. M. K. Roddis, eds., ASCE, Reston, Va., 890–897.

Rojas, E. M., and Aramvareekul, P. (2003). "Is construction labor productivity really declining?" *J. Constr. Eng. Manage.*, 129(1), 41–46.

Rojas, E. M., and Kell, I. (2008). "Comparative analysis of project delivery systems cost performance in Pacific Northwest public schools." *J. Constr. Eng. Manage.*, 134(6), 387–397.

Soibelman, L., and Kim, H. (2002). "Data preparation process for construction knowledge generation through knowledge discovery in databases." *J. Comput. Civ. Eng.*, 16(1), 39–48.

Stegemann, J., and Buenfeld, N. (2004). "Mining of existing data for cement-solidified wastes using neural networks." *J. Environ. Eng.*, 130(5), 508–515.

Thomas, H. R. (2000). "Schedule acceleration, work flow, and labor productivity." *J. Constr. Eng. Manage.*, 126(4), 261–267.

Thomas, H. R., and Horman, M. J. (2006). "Fundamental principles of workforce management." *J. Constr. Eng. Manage.*, 132(1), 97–104.

Vorster, M. C., and de la Garza, J. M. (1990). "Consequential equipment costs associated with lack of availability and downtime." *J. Constr. Eng. Manage.*, 116(4), 656–669.

Yu, W.-D. (2007). "Hybrid soft computing approach for mining of complex construction databases." *J. Comput. Civ. Eng.*, 21(5), 343–352.

Zayed, T. M., and Halpin, D. W. (2004). "Process versus data oriented techniques in pile construction productivity assessment." *J. Constr. Eng. Manage.*, 130(4), 490–499.

Zhang, S., Zhang, C., and Yang, Q. (2003). "Data preparation for data mining." *Applied Artificial Intelligence*, 17(5–6), 375–381.

Zhu, Y., Mao, W., and Ahmad, I. (2007). "Capturing implicit structures in unstructured content of construction documents." *J. Comput. Civ. Eng.*, 21(3), 220–227.

Downloaded from ascelibrary.org by Mississippi State Univ Lib on 07/05/19. Copyright ASCE. For personal use only; all rights reserved.