PREPARATION OF INCONGRUOUS ECONOMIC DATASETS FOR
REGRESSION ANALYSIS

Gunnar Lucko\textsuperscript{1}, Ph.D., A.M.ASCE, Zane W. Mitchell Jr.\textsuperscript{2}, Ph.D., M.ASCE

\textsuperscript{1} Assistant Professor of Civil Engineering and Director of Construction Engineering and Management Program, Department of Civil Engineering, The Catholic University of America, 620 Michigan Avenue NE, Washington, DC 20064
Phone: 202-319-4381, fax: 202-319-6677, email: lucko@cua.edu

\textsuperscript{2} Assistant Professor and Deputy Department Head, Department of Civil and Environmental Engineering, United States Air Force Academy, Colorado Springs, CO 80840
Phone: 719-333-2054, fax: 719-333-2234, email: zane.mitchell@usafa.edu

ABSTRACT
Datasets in construction engineering and management are not always as clean and consistent as one would desire. They may not match in format or content between companies that are studied and within them they may require substantial preparation. This makes the examination of theories and the extrapolation of trends from historic data more difficult than it is for carefully controlled experimental studies or collection of 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 prior to those analyses. It outlines general techniques that can be applied to field data from construction companies or data repositories to give researchers various best practices that will help gleaning further truths from already existing data.

INTRODUCTION
Field operations are a fundamental component of the area of construction engineering and management. Contractors routinely measure various aspects, e.g. time, cost, and productivity, to fulfill their project management functions of planning, optimization, and control. This accumulates a valuable pool of data for construction researchers to explore. However, such data are often collected in a ‘home-grown’ manner, lacking a
standard format across different companies (Mitchell 1998). Data, especially paper-based, can require substantial preparation to be usable. Scholarly publications may contain separate sections on how their data were collected (Severson et al. 1994), but hardly describe what practical problems were encountered in preparing them. Literature specifically on the preparation of data for construction engineering and management studies is sparse, underlining the importance of this paper. A dilemma emerges; quantification of economic phenomena is crucial to stay competitive in today’s industry and years of existing data can be at the researcher’s disposal, if only scientifically valid ways to compare ‘apples to apples’ can be devised for them.

Central questions for the researcher at the onset of a study include the type, range, and source of the necessary data. In this paper it is assumed that the pros and cons of using existing data from one or several field sources instead of collecting new observations in the field or the laboratory have been evaluated, which is not an easy decision. The researcher has great control over the collection methodology for new data, but no influence on how the existing data were created. Data from field sources are more prone to containing irregularities (noise), which can be somewhat dampened by their potentially larger volume. Yet laboratory studies are often infeasible or impossible for examining economic phenomena in construction. The trade-off thus is between the challenges of extracting quality data versus savings in time, which can reach the order of years. This paper presents various techniques that the writers have successfully used in their research on heavy construction equipment. Mitchell’s (1998) objective was to devise a predictive model for repair costs based on data from four companies. Lucko (2003) built a model for residual value from auction records.

DATA SOURCES

Proprietary Data

The construction industry is very competitive; contracts are won or lost by small margins. While construction firms are often willing to cooperate with researchers who develop and test theories that will help them improve their business, they are disinclined to share their data in a forum where other firms might gain an advantage. In Mitchell’s (1998) study, some of the firms competed against each other in the same market. Proprietary corporate data therefore typically require written permission to be usable for non-commercial scholarly purposes. Confidentiality agreements regulate the disclosure of planned publications to the data supplying company and also their contents. They can establish that only significantly aggregated, averaged, or artificially created data that exhibit similar statistical properties to the original ones may be published and that descriptions and illustrations are created without making raw data discernable. The agreement should still permit discussion of general aspects with other researchers. Other precautions are anonymity of participating companies and protecting raw data by limiting access and destroying them after a specified time.

Multiple Sources

Ideally, data should be collected from several sources to allow validation of their order of magnitude, mean, and variance and enable cross-comparisons. However, each additional source also introduces new biases and possible sources of errors. It is essential to fully understand the mechanism by which the data were collected and
recorded, e.g. by interviewing the respective staff (Mitchell 1998). Data sources also vary in accuracy and reliability. Governmental sources are generally considered of high quality and typically reveal the data collection process while corporate sources often exhibit more incongruities. Sources must be coded, e.g. with integer numbers, in the final merged dataset to allow testing for statistical differences between them. After surveying the data available from different sources, it is important to determine what common data fields are available and whether or not those data fields suit the purpose of the study at hand. During the source selection phase of Mitchell’s (1998) research, it was found that the data from some companies, although extensive, did not have all of necessary attributes to contribute to the study. Such data of insufficient quality should be eliminated, even if it reduces the potential overall sample size.

DATA CHARACTERISTICS

Frequency and Resolution

The definition of what constitutes an individual data point depends on the internal practices of each source. Economic data from several different companies may deviate from each other in frequency and resolution. Such data often accumulate as events occur, e.g. an equipment repair. However, macroeconomic data typically have a monthly or quarterly frequency. Different frequencies, spanning from an individual machine to the economy at large, influence the time intervals of the response variable. The largest one generally becomes the valid frequency, e.g. the year of manufacture (Lucko 2003). Another consideration is the resolution of numerical data, i.e. their significant digits, which determines the possible accuracy of the response variable. Data can be averaged if their frequency is higher than needed, e.g. to match a monthly averaged annual producer price index with age, which was measured in years only. In Mitchell’s (1998) study of repair costs the time scale had to be determined to properly reflect how heavy construction equipment ages. The best way to account for aging of this type of asset was not calendar age, but age in terms of cumulative usage. This became the explanatory variable. Problematic in this study was one company without any dataset that included both cumulative usage and repair costs together. Repair costs were tracked in one database on a monthly basis. Cumulative hours of usage were available in a different database which recorded cumulative hours on a machine whenever an oil change took place. Calendar date was used as the common point to combine these two internal datasets. Costs from months in which oil changes took place were associated with the meter hour readings from those months. One problem was that cost data were reported in end-of-the-month increments but oil changes may not have anytime. Oil changes that took place on or prior to the 15\textsuperscript{th} day of a month were thus assumed to have taken place at its beginning, otherwise at its end. A certain amount of error is induced by this assumption. Repair costs that were associated with cumulative hours from oil changes early in a month are probably understated because not all costs that had taken place up to that oil change were necessarily included. By the same token, costs associated with oil changes that took place late in a month are probably overstated because they could include expenditures that occurred after the oil change. It was assumed that these errors were offsetting in the long run. It was critical that there was no more than one data pair per month (since

3
the repair cost data were tabulated on a monthly basis) and no more than one data pair per oil change. To allow more data pairs would have artificially fabricated data.

A problem encountered with data from other companies was that there were multiple data points for the same cumulative hours of use on some machines that occurred while those machines were idle for long periods. To make the data model work, all but one of the data pairs for multiple occurrences of the same cumulative hours of use were eliminated. The first data pair occurrence was kept because the repair costs associated with a specific cumulative meter hour reading for a machine should include all repair costs up to the point where those hours were reached. Costs incurred after the meter hour reading was reached were then ascribed to the next data pair.

**Storage Medium**

The medium in which data are recorded can range from informal paper records via ad-hoc spreadsheets to professional databases. Data might be readily accessible on personal computers (PC) or might need to be retrieved off of mainframe computers using data mining software. Paper records are more vulnerable to transcription errors, both by the individual who originally recorded the data and by the researcher wishing to use the data. Special care should be taken when scanning or typing data to an electronic format. Existing electronic data need to be at least skimmed to ascertain their quality. Find and sort functions of computer software assist with this task.

Spreadsheets are common for storing electronic data. Lucko (2003) subscribed to the online databases of Last Bid®, formerly Green Guide Auction Report™ and Top Bid, which covered most construction equipment auctions that were held in the North American market and obtained Microsoft® Excel spreadsheets for specific makes and models. All data had to be prepared by bringing them into the same medium and format, so that further verifying and ‘cleaning’ techniques can be applied evenly. Individual spreadsheets were time-stamped and consolidated into one workbook.

In Mitchell’s (1998) research, the data came in three different media: PC-formatted, mainframe, and manually recorded data. The data extracted from PCs were the easiest to manipulate and assimilate. These files came in a format that could be used directly by standard software packages. Data in spreadsheets (Excel or Corel® Quattro) required the least amount of preparation because they were already in the file format that would be used for analysis. Others were stored in databases (Microsoft® Access or Corel® Paradox). They were extracted from the database files into spreadsheets. In some cases, this was faster than preparing data that already came in spreadsheets, as queries to extract the data into the desired spreadsheet format could be generated.

Mainframe data posed another challenge. Some companies ran reports and generated all printouts from their mainframe computer using individually programmed applications. The data were converted by transferring them to ASCII print files, which could then be opened on a PC, but were not parsed into discrete components. They could not be directly used in spreadsheets as all of the data in a given row of the ASCII file would be imported into one cell. Although converters exist in spreadsheet programs that permit the parsing of data, they only work with data that originally were in tabular format, which was not always the case. Mainframe reports also contained cost data that were not a part of this study. Usually more than one report was needed to accumulate all data on a machine. The process of extracting the useful data was simplified by building templates in DataWatch® Monarch™ to recognize row
formats. Cost data were extracted, associated with specific machines, and subtotaled by cost code for each month. Unwanted cost data were then filtered out by cost code and a spreadsheet containing only the desired data was exported. Paper-only records were the hardest to prepare for analysis. In some cases, folders containing records and receipts for each machine had to be examined page-by-page to enter them into a spreadsheet. Due to the amount of labor involved in transcribing these raw data, paper-only data sources were used only when electronic sources were not available.

**Formatting and Labeling**

Since terminology strongly depends on the data source, it is recommended that each column of data points receive an intuitive label for future reference. Spreadsheet software allows formatting cells as text, numbers, dates, or prices according to their type of data. Editing functions in text processing software can be used to find and replace superfluous blank spaces, inconsistent capitalization, and discrepancies among individual entries. To remove superfluous formatting from electronic data it can be useful to copy and paste datasets first into a simple text editor and then into the spreadsheet software. Sorting columns of data alphabetically and numerically brought cells with symbols, e.g. “N/A” or “-”, to the top to be filled with “.” This symbol is commonly used in statistics software to represent missing data (Lucko 2003). Column header cells with minimum, maximum, mean, and variance were added to numerical data in the spreadsheet to gauge their range, centricity, and spread. Calculated data columns were added, e.g. determining age as the difference between the manufacturing date and the date of the auction at which the machine was sold. Age had an accuracy of years, since only the year of manufacturing was known, not its exact day or month. Sorting by numeric values, e.g. by age, revealed unreasonable extremes caused by incomplete entries, e.g. negative values (199 - 1996) or large values (2002 - 199 _ ), that were corrected or deleted. Setup descriptions was scanned for negations, e.g. “no”, “not”, and “inoperable,” to delete entries of machines that had missing parts or were otherwise unsuitable. Finally, simple IF formulas in the spreadsheet converted postal ZIP codes and state name abbreviations into auction regions to be used as an explanatory variable in the statistical model. For large datasets it is recommended to create a control column that provides an independent checksum of a conversion formula and verifies its cell references and commands.

In Mitchell's (1998) research, two types of accounting errors occurred that had to be corrected prior to analysis. The most common error was that of negative repair costs for a given month. When the question of negative costs was posed to the companies involved, the answer obtained was that the negative charges were due to either overcharges or mistaken charges that occurred in an earlier month. In order to fix this type of error, the mistaken charges were removed in the months that they occurred and the negative charges were eliminated. The mistaken charges were moved to the months in which they actually occurred. The reason for doing this was to eliminate the false fluctuations in cumulative repair cost induced by adding and subtracting charges that should not have been there in the first place. Another error, though not as common as negative charges, was that of replaced hour meters. This was obvious in that either the cumulative cost at time zero was not zero or the cumulative hours associated with a given machine went down with the passage of calendar time. The fix for this problem was to first confirm that the meter had been replaced. After this
confirmation, the cumulative hours at replacement of the meter were used as a baseline to correct later readings. In cases where the cumulative hours at time of replacement were not available, the machine was eliminated from the dataset.

**Typographical Errors**

Skimming the entire sorted dataset for deviations from the regular ‘pattern’ among the data can help identifying and eliminating typographical errors. Lucko (2003) was able to make very detailed corrections, some of which may have stemmed from converting the original handwritten record into electronic form, as they typically occurred between similar looking letters and digits:

- **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 varied the spelling in alpha-numeric entries, e.g. the model of the machine:

- **Incomplete model name:** PC300, not PC300HD-5;
- **Incomplete series name:** PC100 II, not PC100C Series II.

It was found that long numbers contained superfluous zeros that were eliminated. Serial numbers are sequential for newly manufactured equipment, which allowed correcting sudden deviations in the year of manufacture in otherwise constant entries.

**DATA CONVERSION PROCESS**

**Verbal Descriptors**

A problem that is frequently encountered in existing datasets is that they contain explanatory variables with categorical verbal descriptors. All explanatory variables must be expressed numerically to be usable in a statistical analysis. In Lucko’s (2003) study, the manufacturer name, condition rating, and auction region were categorical. Some were hierarchical, e.g. condition (excellent, very good, good, fair, poor); others were not, e.g. region (Northeast, South, Midwest, West). Both had to be transformed into numerical form. Categorical variables often serve as somewhat subjective proxies for complex phenomena. Condition e.g. 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. However, simply assigning numerical values, i.e. integers, is problematic in the statistical model. For non-hierarchical variables this would fabricate a hierarchy. For hierarchical variables it would assume regular distances within such hierarchy.

A solution is to convert the integer into binary sets of indicator variables. Table 1 converts four verbal descriptors into a triplet of zeros and ones that act like switches. Each binary provides only part of the distinction. The triplet was chosen over an equally possible binary pair to not assign all zeros to one ‘benchmark’ category.

<table>
<thead>
<tr>
<th>Descriptor</th>
<th>Number</th>
<th>Binary 1</th>
<th>Binary 2</th>
<th>Binary 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Either all or none should be included in a statistical model. A drawback is that this conversion removes any hierarchical information that may have existed. The numbering begins with one, since zero would favor the first category and consider it the *de facto* benchmark for further variability in the particular explanatory variable.

**Reconstruction**

Datasets need to be checked for completeness and consistency. If each data point contains several pieces of information and the dataset contains redundancy, it is possible to carefully fill gaps by sorting and comparing neighboring entries, provided that no data are fabricated *a posteriori*. Only in rare cases do datasets directly overlap, such as the data from Last Bid® and Top Bid used by Lucko (2003). They provided independent records of the same heavy equipment auction events. Redundancy of the two datasets allowed verification by comparing pairs of entries, i.e. one machine being sold at a particular auction, via unique identifiers in the data point such as date, location, and serial number. Machines could appear in several data points if they were sold several times during their life. It was assumed that they were independent.

Reconstruction required that all data points in the merged dataset were sorted in the hierarchical order of model, serial number, and auction date. A computer macro called *FillGaps* automated the otherwise time-intensive process. For several columns that contained cells with “.” the macro first matched several identifiers with preceding or succeeding entries to find redundant data points, i.e. two records of the same auction sales event. The placeholder “.” was then replaced with the value of the directly adjacent cell through the use of the *FillGaps* macro. A variation of this technique was applied to the aforementioned changes in an otherwise sequential list of numerical values. Once gaps had been filled, any redundant entries had to be identified and deleted. A second macro called *DeleteDoubles* was programmed that – similar to the previous one – compared identifiers, allowing for a small variability between auction prices due to currency conversion, and deleted the first entry of each pair. Non-redundant entries still increased the sample size, but had to be verified by other means to ensure they achieved the same quality as the redundant data points.

**Inflation Correction**

The phenomenon of inflation is an imbalance between the supply of money on the one side and the supply of goods and services on the other side (Bodie *et al.* 2002). Manufacturers’ suggested retail prices, auction prices, maintenance and repair costs, and any other items that are modeled should be corrected for inflation 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 said rate accurately for the application of interest, using governmental indices as per Eq. 1 is a widely accepted technique. They are ratios of current to past prices for standardized baskets that contain a mixture of goods and services. Indices of different levels of detail and composition are available. Some apply to the entire economy; others are regional and just cover specific industries, commodities, and stages of processing. The producer price index for finished goods (PPI) was considered to be applicable for construction projects and has been used in several other studies to adjust for inflation the costs related to heavy equipment (Cross and Perry 1995, Cross and Perry 1996, Kastens 1997) as per Eq. 1:
\[ \text{Price}_i = \text{Price}_1 \cdot \frac{\text{Index}_i}{\text{Index}_1} \]  

\[ \text{Eq. 1} \]

where \( \text{Price} \) is any price in U.S. dollars and \( \text{Index} \) is a price index. Once an index, e.g. the PPI, has been used for inflation correction, it cannot to be used as an explanatory variable. Otherwise, the model would suffer from multicollinearity, i.e. a very high Pearson coefficient of correlation between them, which would make obtaining a unique closed-form solution impossible. Ideally, all explanatory variables, including economic indicators, should therefore be only weakly correlated with each other.

### Matching Data

If existing data are augmented by other data, the question arises of how to efficiently match them. In Lucko’s (2003) study, the auction records were categorized with size parameters, normalized by list prices, corrected for inflation, and matched with economic indicators for the respective dates. A size parameter catalog was assembled from manufacturers and their distributors in analogy to the list price catalog. Size parameters characterized the equipment by its performance or capacity, e.g. standard operating weight (empty), general purpose bucket volume, and net horse power (flywheel). 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, but more consistent category to improve the goodness-of-fit. A computer macro called \textit{MatchData} was reading 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 size parameters, the identifier was the model name; for economic indicators it was the auction date.

### Normalization

It is often useful to normalize variables to make them comparable across categories. All values are divided by their respective values from a baseline variable. In the residual value study, the dollar values were divided by their manufacturers’ suggested retail prices, also known as list price. Actual initial sales prices would have been ideal, but in their absence for proprietary reasons, list prices had to be assumed to be generated in a sufficiently consistent manner across manufacturers (Lucko 2003). List prices are also devoid of discounts that specific companies may receive and thus less biased than purchase prices. They were collected from manufacturers and their distributors, interpolated across gaps and extrapolated to the necessary range with the inflation correction, and matched with the other data on auction sales and sizes. Occasionally, a researcher may want to perform economic research on assets that do not have identical attributes. Units of construction equipment can vary in many regards. Mitchell used a cumulative cost index (CCI) instead of raw dollar figures to help normalize data for differing construction equipment. The differences between the machines could be characterized as physical and usage differences. Physical differences can be seen just by looking at the machine, e.g. type (bulldozer, scraper, etc.), size, and make (Caterpillar, Volvo, etc.). Usage differences are less apparent. This category relates to the application the machine normally performs and to the company that owns the machine. Repair costs, as well as initial purchase price, can differ considerably between the different categories. Mitchell (1998) was able to
compare unlike machines by indexing their repair costs to their initial list price to develop prediction equations. Eq. 2 was used to calculate the final response variable:

\[ CCI_t = \frac{\sum_t (P_t + L_t + O_t) + PP_o}{PP_o} \]  
Eq. 2

where \( CCI_t \) is the cumulative cost index at time \( t \), \( P_t \) is the cost of parts at time \( t \), \( L_t \) is the cost of labor at time \( t \), \( O_t \) are other maintenance costs at time \( t \), and \( PP_o \) is the list price of the machine. Parts, labor, and other maintenance costs were cumulative. The CCI provided the common ground for comparisons between non-identical machines.

PREPARATION FOR STATISTICAL ANALYSIS

Outliers

Considering that data may be faulty without being detectable, e.g. from measurement errors, it is prudent to prepare a dataset that has been purged of these so-called outliers. Such extreme data differ significantly from the basic relationship captured by the other data, either in their sign or magnitude and can often be visually identified in a scatter plot. They can substantially distort regression models. Statistical techniques can identify them. Residuals measure how much a data point deviates from the value that is predicted by the model. Scaled residuals are recommended for their constant variance (Montgomery et al. 2001). An IF formula in a new column marked data points for removal if the absolute value of their residual exceeded 3 (Montgomery et al. 2001). The final model coefficients were then based on the ‘cleaned’ dataset.

Relative Dominance

A statistical issue that manifests itself in the data collected by Mitchell (1998) was that of relative dominance. Some machines had more data pairs than other machines. This was due to differences in usage, dates of purchase, or data collection styles of the companies involved. Machines that had more data pairs could have had more of an influence on the regression analysis than those with fewer pairs had. The way that relative dominance was addressed was to use interpolated values of the CCI at discrete, evenly spaced intervals. What is important to emphasize is that there should only be one interpolated data pair between any two actual data pairs. The interpolation of more than one data pair between actual data pairs would be the creation of more data than actually exist. The cumulative usage interval selected must support the interpolation rule. Mitchell (1998) found 500-hour intervals to work well.

Variable Selection

Note that for observational studies, including research that uses existing data from construction companies, no causality can be established statistically. Nonetheless, predictive models with high confidence levels can still be created by carefully selecting their components and structure. Various techniques exist for researchers to identify and understand the possible contributions of their explanatory variables to a statistical model. A scatter plot of the dataset should be visually inspected to identify any trends or patterns, e.g. linear dependencies. If several explanatory variables exist, their possible pairs yield a triangular matrix of scatter plots, which can identify
pairwise interactions that may be included as separate terms. Such initial observations assist in determining the model composition, i.e. which mathematical function to use, what order of variable terms, and what interaction terms between them, if any. Three techniques can be used to statistically select appropriate explanatory 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 tests the remaining variables. Backward elimination is the inverse, sequentially removing least significant variables from a ‘full’ model, and assists in quickly discarding unwanted variables. Stepwise selection extends forward selection by alternating between adding and removing to arrive at the best model. Despite these techniques, statistical modeling still depends on the conscious researcher and remains both art and science.

CONCLUSIONS
This paper has described how different types of existing data in the area of construction engineering and management can be prepared in a valid manner while preserving their integrity. It included validating the sample size, handling confidential data, unifying the format, filling missing values as far as possible, checking the dataset for various types of inconsistencies, matching data with each other, performing an inflation adjustment on financial data, removing outliers and selecting what explanatory variables to ultimately include in the model. While each study of existing data will present the researcher with unique problems, these guidelines are hoped to serve as a reference and source of inspiration in surmounting the crucial, often underestimated phase between collecting data and being able to analyze them.

REFERENCES