A Discussion on Data Reliability: Evaluating Qualitative and Quantitative Data

By Kevin DeStefano and Faye Kim

International Cost Estimating and Analysis Association 2019 Professional Development and Training Workshop

Tampa Bay, Florida

May 2019

Table of Contents

Author Biographies

Kevin DeStefano is currently a Consultant for Herren Associates working as a cost engineer supporting NAVSEA 05C. His time with Herren has provided him with experience in the cost estimation and data analytics. Prior to Herren, Mr. DeStefano received his Bachelor’s degree in Industrial Engineering at The Pennsylvania State University.

Faye Kim is a Senior Consultant at Herren Associates currently supporting the Naval Sea Systems Command Cost Estimating and Industrial Analysis Division and has previously supported the Naval Center for Cost Analysis (NCCA). Ms. Kim received her Bachelor and Master of Science from the Pennsylvania State University in Engineering Science and Mechanics.

Company

Founded in 1989, Herren Associates is an engineering and management consulting firm with a proven record of maximizing the value of every taxpayer dollar. As trusted advisors to federal executives, we partner with clients to drive operational improvements and manage performance - maximizing efficiency and cost effectiveness.

Abstract

Obtaining data is an integral part of any analysis and collecting good data that is accurate and robust can be challenging. The quality and validity of data can strongly influence the final cost position recommendation that is presented to accurately and effectively inform decision makers. This paper will discuss data reliability, how to evaluate qualitative and quantitative data, and how to create and evaluate metrics to track the quality of data to ensure accurate cost estimates.

Introduction: The Importance of Data

Cost analysts use data every day on the job. Because of its heavy use in the cost industry, it is crucial for cost analysts to have a good understanding of why it is important to have good data and what constitutes good data.

As a cost analyst, data is a major driver of client delivery and the backbone for most products submitted. Data is essential in delivering accurate, credible, and defensible estimates, models, briefs, and final recommendations which are used by decision makers to enact change. Having unreliable and bad data can lead to bad decisions being made. The estimates and models that are put together to present to clients do not just affect the initial clients but professionals both in and out of the cost community.

But what constitutes good data? Good data is defined by its ability to be defended and used. Since data is rarely presented in its raw form it must be able to be manipulated and extracted from and still hold up the qualities of the entire set. The ability for a cost estimator to defend their data and the subsequent analysis comes down to validity and reliability, so cost estimators should question and evaluate new data. The validity of data is crucial to defending it. What makes a data set valid is its accuracy and precision. An accurate sample data set has a mean close to the population mean; a precise data set has a small deviation from its mean. The accuracy and precision of data are equally important to creating defensible estimates.

This paper will discuss the different types of data, their advantages and pitfalls, the collection and evaluation of data, how to create and utilize metrics, and an all-encompassing case study done by the team.

Types of Data

Data comes in all shapes and sizes but can fundamentally be broken down into two major groups: qualitative and quantitative. Quantitative data is defined by the measurements of quantities with numerical values. Some examples of this include time, cost, height. Qualitative data is defined by the description of characteristics through non-numerical values, such as color, contract type, and basis of estimates.

Quantitative data is integral in cost estimation as it provides the hard-valued inputs used in estimates, models, and more. It allows a cost analyst to go beyond saying "it will cost more than it did last year" or "it should cost less" and provide an actual price point. Quantitative data is inherently unbiased. There is no room for misinterpretation of the values. In cost, one hundred dollars is one hundred dollars no matter who is looking at it. This is a huge benefit of using quantitative data. Quantitative data, and more specifically cost data, can be universally understood. Someone on the West coast of the United States could get data from a supplier in Maine, repackage the data for their own use and send it to Japan and the data would stay the same and carry the same weight. Another advantage to quantitative data is its ability to be manipulated. It can be added, multiplied, and divided, and with the help of software when using extremely large data, it is possible to easily evaluate data and run statistical analysis. The ability to complete statistical analysis on quantitative data helps in the validation of the data as a useful input that is both reliable and accurate.

While quantitative data has a lot of advantages and practical uses in cost estimating, it still has some serious pitfalls. Quantitative data often lacks the ability to accurately portray the nuances and complexities of what the data may be measuring. Because it is bound to only numerical values, numerical data lacks the interpretation of data that qualitative data has. Quantitative data can also be deceptive in its representation of a whole population. In data analytics a sample of the full data set is often the only thing used or available. Depending on the data source, a sample can be manipulated to misrepresent the population while still appearing valid.

Qualitative data is also integral in cost estimation as it often provides insights into the "how" or "why" of data. This allows cost analysts the ability to go a level below the numerical data and really delve into the origin of a data set. Unlike quantitative data, qualitative data is not explicit. You can continue to ask questions about it and receive new data and inputs that were not previously provided. Another advantage of qualitative data is the human aspect. Qualitative data allows for human input and experience to play a role in estimation. Qualitative data can give you the human aspect and to show you the nuances and complexities, or in other words the nonquantifiable inputs that must be considered. Qualitative data can provide a more descriptive picture when compared to quantitative as it is not bound to numerical form.

All the advantages to qualitative data can quickly become its greatest pitfalls. Because there is no discrete answer or input, there is room for human error and bias. While you can continue to delve deeper and deeper into the meaning of the data you can end down rabbit holes that mean nothing or never reach a finite end. This can become timely and in turn costly. With human input being the source of the data, one must seriously weigh the influence of bias into your analysis of the data. To put it simply, differing viewpoints can determine how the data is interpreted. Collecting accurate and defendable data can be difficult because of the human bias and underlying agendas. Qualitative data analysis can be a very taxing process that takes up a lot of time and resources because of the difficulties in its manipulation in large sets.

Through the combination of qualitative and quantitative data a defendable and valid data set can be created. This is something that should be kept in mind during the data collection process.

Data Collection

Collecting data can seem straight forward. However, there are many factors that impact collecting data. The importance of good collection of data can be clearly seen by a cost analyst when the desired product is fully understood. When a cost analyst understands what the data will be used for and where it is needed, they will be better prepared to ask the right questions and make the right requests that will lead to the reception of the best possible data. It is the foresight to see not just the data that will be needed in the immediate future, but also the data that will be used across the whole project as inputs, cross checks, and other supporting data that helps increase and prove the validity of the data set.

When collecting data, it is important to ask for as much of the relevant population data as possible. The larger a sample size of data collected, the more precise, accurate, and defendable the analysis on the data set can become. In other words, the more data you receive, the less likely you are to have an incomplete view of what's available. This includes asking for historical data. Historical data can be crucial to data defense as it can be used to accurately forecasting outward, cross check the values of new data, and give insight into the decision-making process of past recommendations. When asking for data it is also of best practice to ask for both the quantitative and qualitative data that supports the end goal to ensure you avoid the inherent pitfalls of the data itself.

It is very important that the data provider completely understands what is needed, not just the cost analyst. Especially when asking for qualitative data, cost analysts cannot rely on the provider to completely understand what is needed. There is a lot of room for interpretation. This begins with the scope of work. The cost analyst asking for the data should have a wealth of knowledge on their project, what they have, and what they still need. Failure to provide some sort of scope to the providers can lead to a lack of or overabundance of data. Cost analysis has a large time commitment on their hands either through trying to extrapolate full answers from no data or by stripping away the useless data from the large data. The more that the data provider understands about the project and its necessities, the more usable the data will be. Continuing with scope, it is important to layout the scope of the question. Cost analysts want to make sure that they are not getting too little or too much but also need to make sure they are getting all their questions answered. This means that cost analysts must ensure they lay out exactly what they need to be delivered to them

Knowing what data is needed is the first step for a thorough collection. Due to the inherent bias of the human input it is important to frame questions in an unbiased way. For example, if it is desired to have a lower cost in the estimate it is not appropriate to ask for the data in a way that frames the received data in a way that does not accurately represent the data. By trying to get a certain sample of data you can still have data with a credible source but lose the accuracy of the entire set. It is equally important to evaluate the bias of the data provider. Cost estimators should understand what stake the provider has in the end product to help determine how they might respond to certain request.

When dealing with qualitative data specifically, there are some things to keep in mind for data collection. It is important to use neutral wording so that it does not influence the provider. For example, cost analysts should avoid positive and negative adjectives/verbs when referring to neutral options. When giving the data providers a scale to choose from it is important to define the end points and to also make sure that the end points encompass a full scale of options. If scaling is from 1-10 it is important to make sure that you are not pushing people towards a response by making all the options positive or negative. For example, if the questions is asked, “how negative of an impact has the new initiative had on you?” all the responses will be negative. It is best practice to keep the responses neutral like, “how has the new initiative impacted you?” to reduce the potential bias in the results. Another thing to keep in mind when collecting qualitative data is the way you are asking for the response. Certain questions should be yes and no answer and some should be multiple choice, but it is important to make sure that you leave room for open ended responses when appropriate to allow for human input of these qualitative data responses. By giving the wrong response type data providers can be pigeon-holed into how much or little they can give.

While there are some best practices that cost analyst can follow to try and improve the quality of data received, there is no cookie-cutter way to get the best usable data. There are often unavoidable and unfortunate roadblocks that arise. Whether its unresponsive data sources, poor formatting of data sets, or biased data providers, some level of reliability may be missing from a data set. Cost analysts often must make the best of what they receive through thoughtful evaluation and analysis.

The Power of Metrics

The process of collecting and evaluating data can seem like a daunting task. As large data sets are received it is often expected that cost analysts have a quick turnaround from the moment they receive data to its use for a project. One way to help simplify the process is to create metrics that track the quality of the data received. Metrics are a way to track and evaluate a given data set by creating a new measurement. This new measurement is derived from the given data set and helps prove the reliability and validity of data. The power of metrics comes in that there are no limits to what you can evaluate with them.

Metrics can be made for anything a cost analyst wants, and they do not have to be fool proof. With the massive amount of data that a cost analyst can receive, metrics can be a good way to find a starting point for data analysis. Metrics in the cost estimating world can vary immensely and still have a useful place in the data validation process. Metrics can be simple like "days between request and reception" or complex like "quantity independence error by line item", both providing insights that will help in the implementation of the data into a model and the future collection of data.

Metrics can be created during the data collection process to ease the burden of tracking what has been received and what is still needed. Simple metrics can simply have a yes or no output but help track a large variety of things. Whether the data been received, contains qualitative data, or contains sources, are all examples of quickly created metrics that can be used to evaluate what is coming in and out. More complex metrics can also be used to evaluate data such as percent hardware, inflation rates, quantity scaling that can be used to check on the general validity and accuracy of new data.

When creating metrics, there are some practices that can yield the best results. First, create metrics that will propel you towards the overall goal. It is easy to get hung up creating metrics for anything and everything. Metrics should be used to help ensure that the data is reliable and valid. Second, not all metrics have to have a direct tie to reliability and validity. If you can use the metric to compare the data source with old sources or analogous sources, which in turn helps prove the validity of the data source, then it is a good metric. Third, don't over complicate metrics. Certain metrics are created using complex equations. However, some of the best metrics are simple created using simple arithmetic. Finally, understand the limitations of metrics. Metrics are a very powerful tool in the data evaluation process, but they are not perfect. Cost analysts create the formulas, bounds, and restraints of a metric and there can be mistakes that lead to misrepresentation and misinterpretation. Cost analysts must be sure that the metrics they create actual track and evaluate what they claim to.

Using Metrics

In cost estimation, metrics can be used to flag actions that need to be taken in the collection process, make informed decisions regarding the use of the data, and eventually refine cost estimates. If metrics are created well, they should be able to make the data evaluation process quicker and easier while increasing the overall validity of the data set and making the end product more defensible.

Using metrics to track incoming data is an easy way to ensure that all the requested data is received and of good quality. Metrics for tracking can be as simple as checking that the data arrived, that it arrived in proper form, and that it passes a visual eye test of quality. With new data sets it is easy for data providers to misinterpret a request. Metrics are a good way of ensuring that the data received answers all the questions you had. With such simple 'yes or no' answered metrics, if the metric comes up negative it immediately flags the cost analyst and tells them that they need to circle back with the data provider to have discussion about the missing material or the format in which it was sent. Some more creative metrics can be used to check the general quality of what was submitted. Creating metrics to ensure that the estimates have a level of effort in their completion can be another way to flag the need for a circle back with data providers. In this process of checking level of effort, metrics can be used to look at the use of inflation, check to what level of detail the estimate went, and the quantity scaling that was used. These metrics give actual leverage when coming back to a data provider and asking them to resubmit better data.

Metrics should be used to help make informed decision about the use of data. Creating metrics can help in the cross-checking process. Metrics should be used as a way to flag any data that is outside of a reasonable threshold of its expected value. In the cost world it can be powerful to check a future years cost to a known historical cost. Creating metrics to see the difference between the expected value and the new data value can help flag data that may be unreliable and require further investigation. Metrics should be used in the data validation process. Sometimes the data received is not perfect. Metrics can help defend the decision to include or exclude the data from estimates, even if it was the desired source. Another powerful way to use metrics to track data is during post-delivery. Creating metrics to compare the final decided data value to those which you received can help create a repository of data reliability information. These metrics can help understand the bias, uncertainty, and validity of data sources to help with future decisions.

Through metrics, cost estimators have an extra tool in the defense of their estimates. Metrics can seem like a lot of work for a large data set but a large majority of them can be automated. The quicker the analysis of incoming data the more time there is to do real analysis to the data and thus add value to their estimates. Metrics should be used by all analysts to help improve the way the handle data.

Case Study

The team has been working diligently to improve the process in which data is received and evaluated. These estimates rely heavily on the inputs of different data suppliers. Each year a data call is put out to the data suppliers requesting the cost data for the different inputs that fall under the estimates umbrella. With multiple estimates and almost one hundred inputs each, the data can pile up very quickly. In order to submit the best estimates the team must intake all the data, evaluate it, analyze it, and implement it in very short time periods. Proper analysis of all the input data requires deep dives into them, which can be a very long process.

To improve the data collection and analysis process, the team began reviewing the process in which data is requested, received, and evaluated for use. First, we evaluated the status of the preliminary data that we received to determine the best metrics to use to track the upcoming data call. We created a scorecard to facilitate the tracking. A scorecard is a visual way of seeing all your data sources and how they compare in accuracy, validity, and precision based off self-created metrics. The scorecard was used to track what data was included in the data sets, its level of effort, and how it compared to previous estimates/data calls for the same data points. The team wanted to create these metrics to evaluate whether the inputs were reliable enough to be defended if they were used in the cost estimate. A snippet of the scorecard can be seen below in Figure 1.

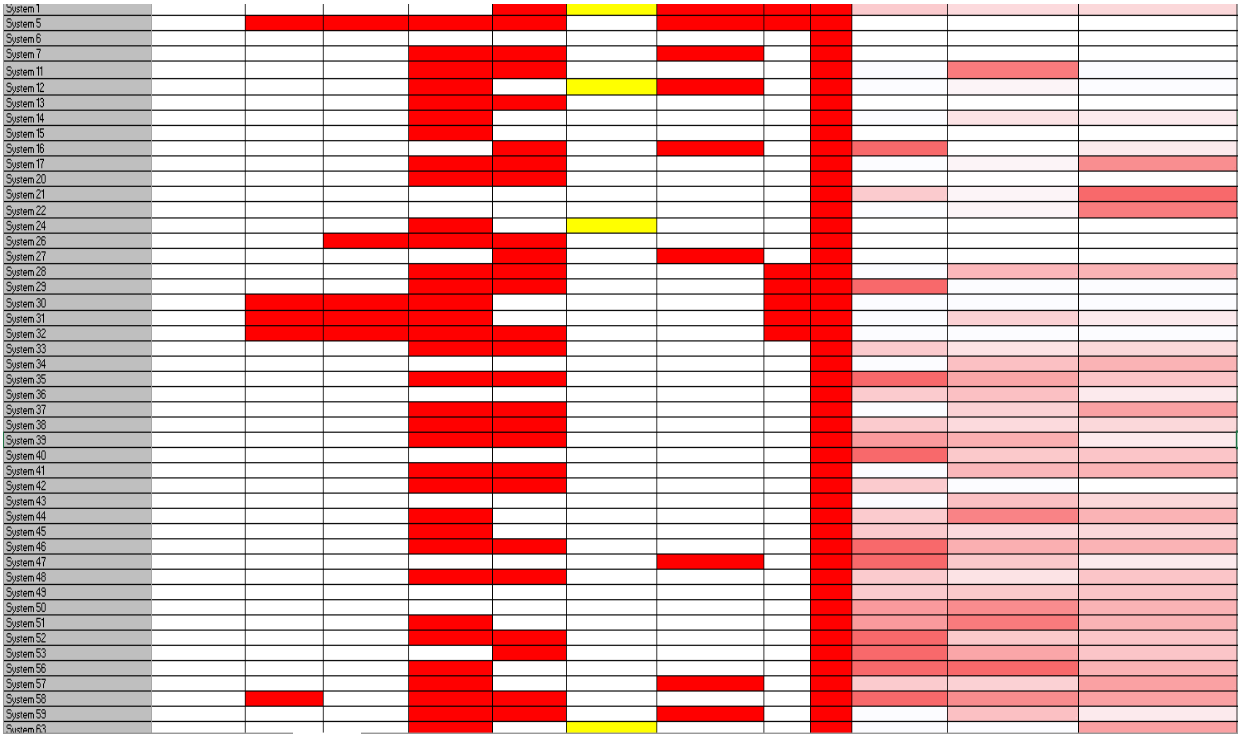
System 1
System 5
System 6
System 7
System 11
System 12
System 13
System 14
System 15
System 16
System 17
System 20
System 21
System 22
System 24
System 26
System 27
System 28
System 29
System 30
System 31
System 32
System 33
System 34
System 35
System 36
System 37
System 38
System 39
System 40
System 41
System 42
System 43
System 44
System 45
System 46
System 47
System 48
System 49
System 50
System 51
System 52
System 53
System 56
System 57
System 58
System 59
System 63

Figure 1: Metric Scorecard

To evaluate what was received, the team created simple yes or no style metrics. This include "was there a BOE?", "Was the Input signed?", and "was a contract number given?". This flagged what was still need from the data suppliers and allowed the team to submit rework data requests in a professional manner, ensuring that the chances of multiple where minimal.

To evaluate the level of effort on estimates submitted, the team created metrics to check the escalation, inflation, line item logic errors, quantity independence errors, and quantity scaling. This helped provide the team insight into the validity of the data before its introduction into the estimate. These metrics could be used to adjust the uncertainty applied to the estimate that may not have been added if not previous tracked.

To evaluate the new inputs against old data, the team utilized well documented values to cross check the data within a threshold. This was then used to prioritize which data sets needed more attention in the deep dive process and again helped determine how good the data would be for immediate use. Part of the scorecard method used with the metrics was color formatting. By setting threshold values and adding a visual color scale to show how far the data is from the desired location the team was able to identify which data sets were the biggest offenders easily.

The overarching issue that had been revealed through the scorecard was the general disconnect between what the team was receiving versus what they needed to be able to put together the best, defensible estimate through the easy visualization of the scorecard. The team compared what was sent out in the previous request to find out if the disconnect was coming from the analyst side or data supplier side for each issue. The implementation of the scorecard allowed a revamp of the data call to be efficient and effective. The team added all the data they had not requested in the past to the data call and made sure to follow good practices to layout the requests in an easily understood and interpreted way. The team helped conduct a meeting with the data suppliers to present the new request and help clarify any requests in person. This helped eliminate the interpretation pitfall that comes with data calls.

After the data call the team began to refine the scorecard to be used for the future data review. The team has added new metrics to be tracked to help increase the effectiveness of the review process and expedite any rework data call requests. As data has started to slowly be sent in there is a clear difference in what was received in previous years and the new data call, which can be tracked by a metric of its own. As the new data comes in the team will continue to evaluate the data using the scorecard and metrics to continue and improve the process.

Closing

Data is such a vital part of being a cost estimator that its collection and analysis should be done with the best practices. Through a knowledge and understanding of the importance of data, the different types of data, and how to properly collect data, cost analysts can best equip themselves to fulfill client needs in a fashion that is professional, timely, and defensible. Even with the best collection practices, analysts still must scrutinize all data that comes their way to ensure that it is the best they can get. One great way to evaluate and track data is through the implementation of metrics. Metrics can be used to determine what is still need in a data set and the reliability of the source and its subsequent data. Through good practice and experience a cost analyst can progress forward how data is handled in and around their company, client space and the cost community.