Software Engineering Institute

An Investigation of Techniques for Detecting Data Anomalies in Earned Value Management Data

Mark Kasunic
James McCurley
Dennis Goldenson
David Zubrow

December 2011

TECHNICAL REPORT

CMU/SEI-2011-TR-027
ESC-TR-2011-027

Software Engineering Measurement and Analysis (SEMA)

Carnegie Mellon

Copyright 2012 Carnegie Mellon University.

This material is based upon work funded and supported by the United States Department of Defense under Contract No. FA8721-05-C-0003 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Department of Defense.

This report was prepared for the

Contracting Officer
ESC/CAA
20 Shilling Circle
Building 1305, 3rd Floor
Hanscom AFB, MA 01731-2125

NO WARRANTY

THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

This material has been approved for public release and unlimited distribution except as restricted below.

Internal use:* Permission to reproduce this material and to prepare derivative works from this material for internal use is granted, provided the copyright and "No Warranty" statements are included with all reproductions and derivative works.

External use:* This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other external and/or commercial use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

® Carnegie Mellon is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

TM Carnegie Mellon Software Engineering Institute (stylized), Carnegie Mellon Software Engineering Institute (and design), Simplex, and the stylized hexagon are trademarks of Carnegie Mellon University.

* These restrictions do not apply to U.S. government entities.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The authors would like to thank Mr. Robert Flowe (OUSD (AT&L)/ARA/EI) for his support and collaboration throughout this research project. We are grateful to Dr. Cynthia Dion-Schwarz, Director, Information Systems and Cyber Security, DDR&E, OSD-ATL, who sponsored our data quality research. We also thank Mr. John McGahan (DCARC IT Program Manager, Tecolote Software Product /Services Group) who along with Mr. Flowe was instrumental in helping us obtain access to the earned value management data used in this study. We thank all of the people from the organizations in the Office of the Secretary of Defense that shared their expertise, experiences, and insights regarding data quality. Finally, we thank Deb Anderson and Erin Harper for coordinating the editing support for this document.

Abstract

Organizations rely on valid data to make informed decisions. When data integrity is compromised, the veracity of the decision-making process is likewise threatened. Detecting data anomalies and defects is an important step in understanding and improving data quality. The study described in this report investigated statistical anomaly detection techniques for identifying potential errors associated with the accuracy of quantitative earned value management (EVM) data values reported by government contractors to the Department of Defense.

This research demonstrated the effectiveness of various statistical techniques for discovering quantitative data anomalies. The following tests were found to be effective when used for EVM variables that represent cumulative values: Grubbs' test, Rosner test, box plot, autoregressive integrated moving average (ARIMA), and the control chart for individuals. For variables related to contract values, the moving range control chart, moving range technique, ARIMA, and Tukey box plot were equally effective for identifying anomalies in the data.

One or more of these techniques could be used to evaluate data at the point of entry to prevent data errors from being embedded and then propagated in downstream analyses. A number of recommendations regarding future work in this area are proposed in this report.

1 Introduction

1.1 The Problem of Poor Quality Data

Organizations rely on valid data. They use the data to manage programs, make decisions, prioritize opportunities, and guide strategy and planning. But how reliable are the data organizations collect and use? The problem with poor data quality is that it leads to poor decisions. In addition, the rework required to correct data errors can be quite costly.

Existing evidence suggests that poor data quality is a pervasive and rampant problem in both industry and government. According to a report released by Gartner in 2009, the average organization loses $8.2 million annually because of poor data quality. The annual cost of poor data to U.S. industry has been estimated to be $600 billion [Gartner 2009]. Research indicates the Pentagon has lost more than $13 billion due to poor data quality [English 2009].

1.2 Data Quality Defined

Data are of high quality if "they are fit for their intended uses in operations, decision making, and planning" [Juran 1951]. This definition implies that data quality is both a subjective perception of the individuals involved with the data and the quality associated with the objective measurements based on the data set in question. A number of studies have indeed confirmed that data quality is a multi-dimensional concept [Ballou 1985, Ballou 1998, Huang 1999, Redman 1996, Wand 1998, Wang 1996]. An international standard data quality model identifies 15 data quality characteristics: completeness, consistency, credibility, currentness, accessibility, compliance, confidentiality, efficiency, precision, traceability, understandability, availability, portability, and recoverability [ISO 2008].

1.3 Data Defects vs. Data Anomalies

A *data defect* is defined as a data value that does not conform to its quality requirements.[1] Larry English defines it similarly as an item that does not conform to its quality standard[2] or customer expectation [English 2011].

Data defects come about in a variety of different ways, including human errors and errors created by faulty processing of the data. Examples of data defects include missing data, errors caused by typos, incorrectly formatted data, data that are outside the range of acceptable values for an attribute, and other similar problems. English has developed a classification of data defects that is summarized in Appendix A.

Some data defects are easier to detect than others. For example, a missing data value can be readily identified through simple algorithms that check for null values within a data field. Likewise,

[1] A quality requirement is an application requirement that eliminates or prevents data errors, including requirements for domain control, referential integrity constraints, and edit and validation routines.

[2] A quality standard is a mandated or required quality goal, reliability level, or quality model to be met and maintained [English 2011].

values that are clearly out of range of acceptable values for a datum can be detected using simple value checking methods (e.g., a living person's birth date that is incorrectly entered so that it appears that the person is 300 years old). However, there is a class of defects that are more difficult to pinpoint. These are the data values that are referred to as *anomalies*.

A *data anomaly* is not the same as a data defect. A data anomaly *might* be a data defect, but it might also be accurate data caused by unusual, but actual, behavior of an attribute in a specific context. Data anomalies have also been referred to as outliers, exceptions, peculiarities, surprises, and novelties [Lazarevic 2008].

Chandola and colleagues refer to data anomalies as patterns in data that do not conform to a well-defined notion of normal behavior [Chandola 2009]. This is similar to how Hawkins defines an outlier as "an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" [Hawkins 1980]. Johnson defines an outlier as "an observation in a data set which appears to be inconsistent with the remainder of that set of data" [Johnson 1992]. In this report, we use the term "anomaly" to refer to outliers, exceptions, peculiarities, and similarly unusual values.

Anomaly detection techniques have been suggested for numerous applications such as credit card fraud detection, clinical trials, voting irregularity analysis, data cleansing, network intrusion, geographic information systems, athlete performance analysis, and other data-mining tasks [Hawkins 1980, Barnett 1998, Ruts 1996, Fawcett 1997, Johnson 1998, Penny 2001, Acuna 2004, Lu 2003].

1.4 Current State of Practice

A bourgeoning industry has developed to address the problem of data quality. Software applications are available that detect and correct a broad spectrum of data defects that exist in enterprise databases. Losses due to data quality issues would be higher than they are if not for the adoption of these data quality tools. According to Gartner, the data quality tools market grew by 26% in 2008, to $425 million [Gartner 2009]. These tools are geared toward customer relationship management (CRM), materials, and to a lesser degree, financial data. Of the companies that use data quality tools, the Gartner survey found that 50% of survey respondents said they are using data quality tools to support master data management (MDM) initiatives, and more than 40% are using data quality technologies to assist in systems and data migration projects.

According to Ted Friedman, an analyst with The Gartner Group, data quality tools have been most often used in an offline, batch mode to cleanse data outside the boundaries of operational applications and processes [Kelly 2009]. Figure 1 provides an example of a typical CRM data identification/correction algorithm.

DF_ALPHA
ADM BIO PRODUCTS-DECATUR
ADM BIO PRODUCTS-VELVA
ADM CLINTON-CLINTON
ADM CLINTON-CLINTON
ADM CORN PROCESSING-DECATUR
ADM-CEDAR RAPIDS

Deriving standardarized names from within the data

DF_NAME	DF_ALPHA	Derived Name	Location
ALPHA 1	AD MID –PARENT	ARCHER DANIELS MIDLAND	
ALPHA 2	ADM BIO PRODUCTS-DECATUR	ARCHER DANIELS MIDLAND	DECATUR
ALPHA 1	AD MID –PARENT	ARCHER DANIELS MIDLAND	
ALPHA 2	ADM BIO PRODUCTS-VELVA	ARCHER DANIELS MIDLAND	VELVA
ALPHA 1	AD MID –PARENT	ARCHER DANIELS MIDLAND	
ALPHA 2	ADM CLINTON-CLINTON	ARCHER DANIELS MIDLAND	CLINTON
ALPHA 1	AD MID –PARENT	ARCHER DANIELS MIDLAND	
ALPHA 2	ADM CLINTON-CLINTON	ARCHER DANIELS MIDLAND	CLINTON
ALPHA 1	AD MID –PARENT	ARCHER DANIELS MIDLAND	
ALPHA 2	ADM CORN PROCESSING-DECATUR	ARCHER DANIELS MIDLAND	DECATUR
ALPHA 1	AD MID –PARENT	ARCHER DANIELS MIDLAND	
ALPHA 2	ADM-CEDAR RAPIDS	ARCHER DANIELS MIDLAND	CEDAR RAPIDS

Derived Name	Location
ARCHER DANIELS MIDLAND	DECATUR
ARCHER DANIELS MIDLAND	VELVA
ARCHER DANIELS MIDLAND	CLINTON
ARCHER DANIELS MIDLAND	CLINTON
ARCHER DANIELS MIDLAND	DECATUR
ARCHER DANIELS MIDLAND	CEDAR RAPIDS

Final transformation

Figure 1: Example of Data Defect and Correction Algorithm for CRM Data[3]

Chen and colleagues state that "there is much prior work on improving the quality of data that already resides in a database. However, relatively little attention has been paid to improved techniques for data entry" [Chen 2009]. Friedman notes, "Gartner advises clients to consider pervasive data quality controls throughout their infrastructure, ensuring conformance of data to quality rules at the point of capture and maintenance, as well as downstream…Companies should invest in technology that applies data quality rules to data at the point of capture or creation, not just downstream" [Kelly 2009].

Much of the current work on data quality in the Department of Defense (DoD) is limited to identifying missing or duplicate data and discrepancies in recorded values from multiple sources. Other work at the DoD focuses on identifying business rules to screen for defects in repository data. Work is also ongoing in the DoD to apply automated data screening techniques to identify defects.

[3] Figure 1 was adapted from Rademacher and Harter [Rademacher 2009].

1.5 Our Research Focus

In our data quality research, SEMA is focusing on the *accuracy* characteristic of the International Organization for Standardization (ISO) 25012 quality model. Within the model, *accuracy* is defined as "the degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use" [ISO 2008].

Specifically, the objective of the study described in this report was to evaluate statistical techniques that could be used proactively to identify more and varied kinds of data anomalies than have thus far been recognized in the DoD.

1.6 Collaborators and Data Source for this Research

To accomplish our objectives, we collaborated with the Office of the Under Secretary of Defense for Acquisition, Technology, and Logistics (OUSD (AT&L)), Acquisition Visibility (AV). As part of this collaboration, our research group was given access to data that are reported to the Earned Value Management (EVM) Central Repository (CR) by government contractors [DCARC 2011]. The EVM–CR provides and supports the centralized reporting, collection, and distribution for key acquisition EVM data, such as

- contract performance reports (CPRs)
- contract funds status reports (CFSRs)
- integrated master schedules (IMSs) for acquisition category (ACAT) 1C & 1D major defense acquisition programs and ACAT 1A major automated information system programs

The data used in this study was extracted from monthly EVM progress reports that follow the CPR format [OUSD 2011].

1.7 What is Earned Value Management?

Earned value management is a program or project management method for measuring performance and progress in an objective manner. EVM combines measurements of scope, schedule, and cost in a single integrated system.

Figure 2 summarizes some of the key concepts and data items of the EVM system. A detailed discussion of EVM is beyond the scope of this paper. For a detailed description, see the resources available from the Defense Acquisition University [DAU 2011].

For our data anomaly detection research, we selected several EVM variables:

- budgeted cost of work scheduled (BCWS)
- budgeted cost of work performed (BCWP)
- actual cost of work performed (ACWP)
- negotiated contract cost (NCC)
- contract budget base (CBB)

BCSW, BCWP, and ACWP are shown in Figure 2 and are used together to measure performance in the EVM system. NCC and CBB are figures associated with government contracts that remain

constant unless formally changed and hence are not routinely part of the EVM system of measures.

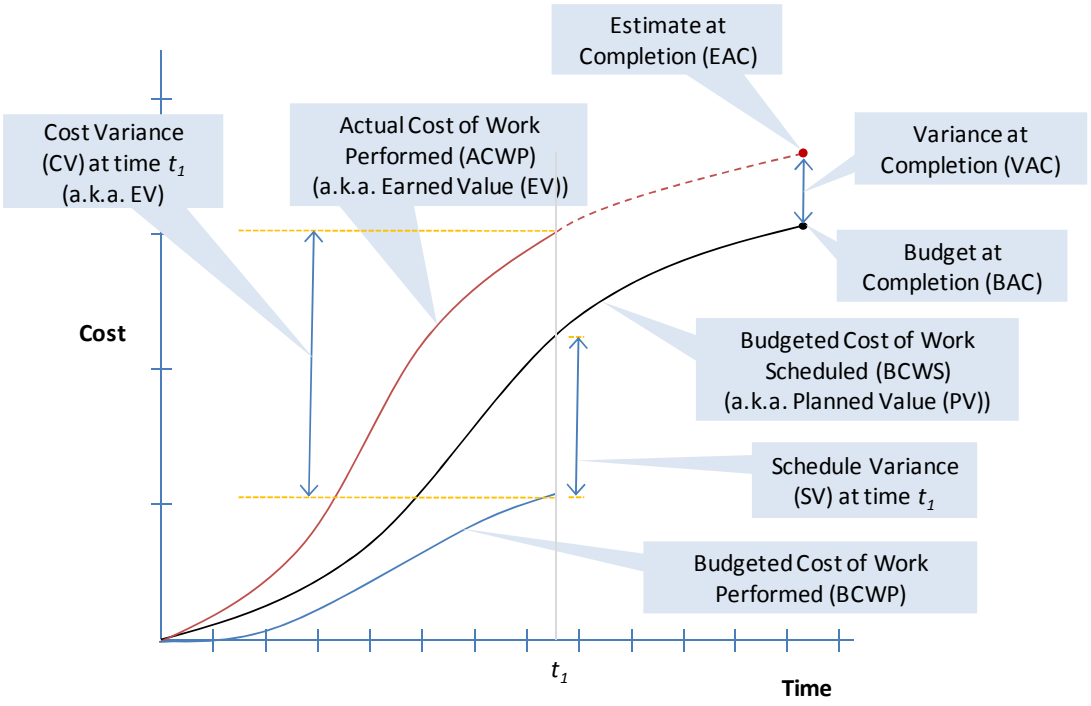

Figure 2: Key Concepts of Earned Value Management

2 Methodology

2.1 High-Level Approach

Our research approach used the following steps:

1. Conduct literature research.
2. Select data source.
3. Select test cases and establish anomalous data values.
4. Select anomaly detection techniques.
5. Analyze results.

Activities 1 through 4 are discussed in Sections 2.1.1 to 2.1.4. Section 2.2 describes each of the anomaly detection techniques and how they were applied to the EVM data.

2.1.1 Conduct Literature Search

Our literature research focused on the analytical strengths and limitations of existing anomaly detection techniques and their potential appropriateness for use in this research. Our team also reviewed the capabilities of some of the leading commercial data quality software tools to better understand the techniques that they incorporate. A brief summary of the review is presented in Appendix D.

Over 210 journal articles, web sites, and reference books were collected and catalogued for initial scanning by team members. The references were rated based on relevancy and items of high relevance were assigned to team members for in-depth review.

The techniques that were investigated can be grouped into the categories suggested by Chandola and his colleagues [Chandola 2009]. Their typology includes a wide-ranging set of techniques, including ones that are

- classification-based
- nearest-neighbor-based
- clustering-based
- statistical
- information theoretic
- spectral

Other useful sources reviewed include work by Kriegel and his colleagues [Kriegel 2010] and Hodge and Austin [Hodge 2004]. Kriegel and colleagues group the methods under the following categories:

- statistical tests
- depth-based approaches
- deviation-based approaches

- distance-based approaches
- density-based approaches
- high-dimensional approaches

Hodge and Austin partition their discussion of outlier detection methodologies under three overall categories: statistical models, neural networks, and machine learning. They also distinguish between clustering, classification, and recognition. There is no single definitive typology of anomaly detection techniques, and the techniques sometimes overlap several of these proposed categories. However, Chandola and colleagues provide a useful starter set to establish a high-level landscape of the techniques. All three papers, particularly the one by Chandola and colleagues, cite many references where these kinds of anomaly detection techniques have been used.

All of the techniques of anomaly detection that we describe in this document rely on the existence of patterns of "normal" behavior, from which the anomalies can be differentiated. Some of the techniques are limited to univariate data distributions while others consider anomalies based on atypical deviations from statistical relationships among two or more variables.

2.1.2 Select Data Source

During the latter part of 2010, our research team conducted two site visits to meet with data analysts from the DoD Acquisition Visibility (AV) organization. AV is responsible for providing accurate, authoritative, and reliable information supporting acquisition oversight, accountability, and decision making throughout the DoD. A key outcome of the meetings was the selection of the EVM-CR as the source of data for evaluating anomaly detection techniques. This repository source was selected based on several criteria, including: the ability to obtain access privilege to the data, the abundance and richness of the data, and existing reports of errors in the data submitted to the repository. This evidence was drawn from analyses conducted by AV analysts as they were preparing reports to support executive decision making.

Program performance information is reported to the EVM-CR on a monthly basis. The massive volume of EVM data reported each month is staggering. Using valuable analysts to do tedious, manual inspections of the data is impractical. For this reason, the development of an automated method for identifying potential data errors would be extremely beneficial since it would relieve the analyst from searching for needles in the proverbial haystack.

The EVM data was provided in MS-Excel workbook format. After receiving the data for this research study, the data set was organized for analysis and the contents characterized. It consisted of 6211 records associated with 359 program tasks. A program task is made up of multiple records in a time series. Each record in the data set contained 167 columns. Most of these columns were text fields containing descriptive and administrative details about the record, such as who submitted it, files that were submitted, when it was submitted, the contract under which it was being submitted, and so on. Given our focus on statistical techniques that apply to quantitative measures, most of the content in a record was not used.

2.1.3 Select Test Cases and Establish Anomalous Data Values

The research team decided it would be most efficient to focus on a sample of the data and chose to examine the time series profiles of the 359 program tasks. From these, the research team selected

four program tasks to use as test cases for evaluating the efficacy of the anomaly detection techniques. Criteria considered to select the cases included the number of records available and their completeness in terms of the variables of interest (i.e., BCWP, ACWP, BCWS, NCC, and CBB) As described further in Section 2.1.4, the nature of the data also influenced the techniques that could be used. The objective was to obtain an effective sample for evaluation purposes.

To establish the anomalous data values in the test cases, the team asked an OSD EVM subject matter expert (SME) to review them; this SME had extensive experience reviewing and analyzing data from the EVM-CR. This was necessary because the actual disposition of the data was unknown, and the research focus was on detecting anomalies that had a high probability of being defects.

We presented both the actual data values and graphical representations of the data and asked the SME to identify anomalies that should be investigated as possible data errors. One example of the results of the SME review is illustrated in Figure 3. The arrows indicate the values that the SME identified as data points that should be investigated as possible data errors. All test cases used in this research study are presented in Appendix B.

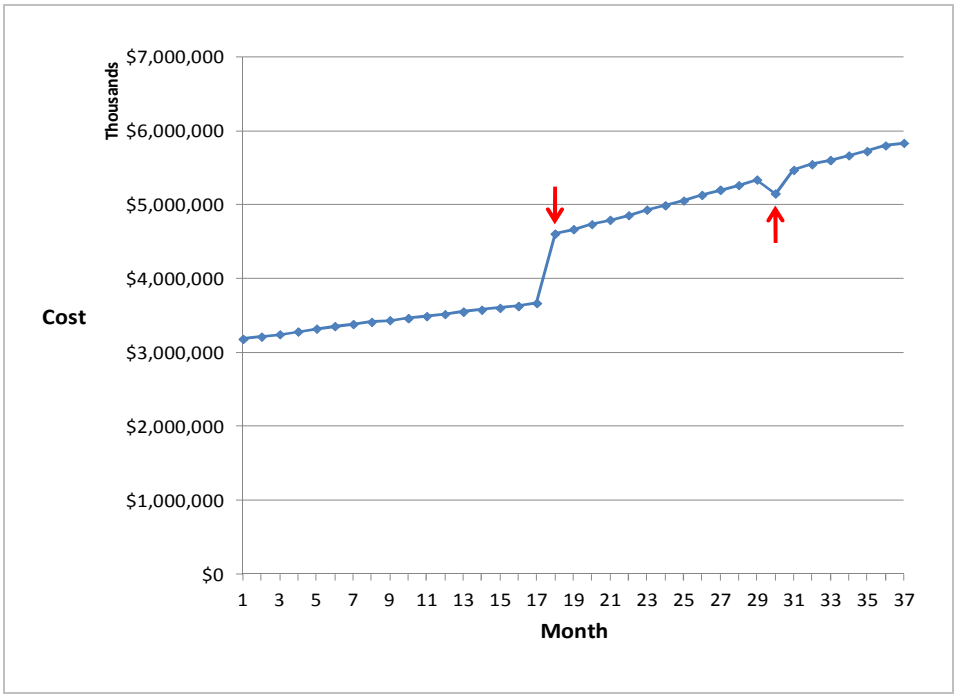

Figure 3: Example Test Case Used to Evaluate Effectiveness of Anomaly Detection Techniques

Figure 4 illustrates the evaluation scheme for the results of our analysis. Using the example illustrated in Figure 3, if an anomaly detection technique correctly identified values 18 and 30 as anomalies, then they would be tallied within the *True Positive* cell of Figure 4. For example, if a technique failed to identify 18 as an anomaly, that occurrence would be tallied as a *False Negative*. Similarly, if a technique identified a data point other than 18 and 30 as an anomaly, that value would be tallied as a *False Positive.* Values that are correctly not flagged as anomalies would be tallied as *True Negatives.*

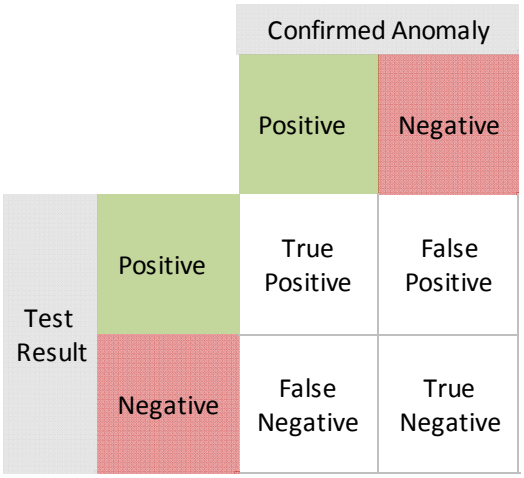

Figure 4: Scheme for Evaluating Effectiveness of Anomaly Detection Techniques

To evaluate the effectiveness of each anomaly detection technique, the team considered two key measures:

$$Detection\ Rate = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$False\ Positive\ Rate = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

The intent was to use the results of the SME review to determine the effectiveness of each anomaly detection technique that was to be investigated.

2.1.4 Select Anomaly Detection Techniques

To determine which anomaly detection technique is appropriate for a given situation, the nature of the data being assessed and the type of anomaly being searched for should be considered.

The team's research focus was to identify techniques for finding specific types of data anomalies associated with accuracy. Data profiling methods and tools are already available for identifying and correcting the following:

- missing data
- incomplete data
- improper formats
- violations of business rules
- redundancy

Therefore, the team purposely ignored these basic types of data anomalies and focused on the accuracy attribute of five variables:

1. budgeted cost of work scheduled (BCWS)
2. budgeted cost of work performed (BCWP)
3. actual cost of work performed (ACWP)

4. negotiated contract cost (NCC)
5. contract budget base (CBB)

The first three variables are cumulative cost values that are reported on a monthly basis. NCC and CBB are not cumulative. Based on our initial profiling of these variables, we believed that statistical analysis approaches would be fruitful as a means of identifying anomalous data that could be caused by error. The assumption was that a normal datum belongs to a grouping or a patterned distribution of other data points. When the grouping or distribution is understood, a model can be developed that will establish the boundaries of what constitutes a region of normalcy outside of which a datum is considered as being anomalous.

2.2 Anomaly Detection Techniques Investigated

As part of the literature review, we identified a number of statistical anomaly detection approaches that looked promising. These techniques were specifically developed to identify anomalous data. They included the following:

- statistical control chart techniques, including the control chart for individuals, moving range (mR) control chart, exponentially weighted moving average chart, and moving average chart
- Grubbs', Rosner, and Dixon tests
- Tukey box plots
- auto regressive moving average (ARIMA) modeling

We also investigated the following control-chart-related techniques:

- 3-sigma outlier
- moving range
- SPI/CPI outlier

The approaches used for these techniques are summarized in Sections 2.2.1 to 2.2.5.

2.2.1 Statistical Control Chart Techniques

A control chart is a statistical device principally used for the study and control of repetitive processes. It is a line graph that displays variation in a time-ordered fashion. A center line and control limits (based on ± 3 standard deviations) are placed on the graph to help analyze the patterns in the data. Common cause variation occurs randomly and behaves like a constant system of chance causes that are predictable. While individual values are all different, as a group, they tend to form a pattern that can be described by a probability distribution. A process that experiences only common cause variation is said to be in statistical control. A process that experiences special cause variation is said to be out of statistical control. Special cause variation refers to any factors causing variation that cannot be adequately explained by any single probability distribution of the output.

Walter Shewhart introduced the first control chart system during the 1930s [Shewhart 1931]. Since then, a large number and wide variety of control chart schemes have been developed for specific applications and objectives. For example, some control chart schemes are effective for detecting anomalies in a data set, while others are effective for detecting a subtle shift in the aver-

age value of a key characteristic measure. Some control chart implementations assume continuous-scaled measurement data, while other chart schemes assume the use of discrete data (such as defect counts).

Based on our research, we selected several control charts that held potential for identifying anomalies in the data. These are listed in Table 1. While the appearance of the different control charts is similar, the parameters of the charts themselves are very different. Parameter calculations for each of the control charts are accessible in the references provided in Table 1.

Table 1: Statistical Control Chart Techniques Used

Name	References – Control Chart Technique
Control Chart for Individuals	[Florac 1999, NIST 2011b, Wheeler 2000,Wheeler 2010, Keen 1953]
Moving Range Control Chart	[Florac 1999, NIST 2011b, Wheeler 2000,Wheeler 2010, Keen 1953]
Exponentially Weighted Moving Average Chart	[NIST 2011b, Crowder 1989, Montgomery 2005]
Moving Average Chart	[StatSoft 2011, Roberts 1959]

We explored the efficacy of each control chart on each of the EVM variables under study. For the EVM variables BCWS, BCWP, and ACWP, the following approach was taken (both for control chart and other techniques described in the following sections):

1. Filter EVM data based on task name of interest.
2. Calculate the month-to-month difference for the time series:[4]

$$\text{BCWS-Diff} = \text{BCWS}_{(\text{Month i})} - \text{BCWS}_{(\text{Month i-1})}$$

$$\text{BCWP-Diff} = \text{BCWP}_{(\text{Month i})} - \text{BCWP}_{(\text{Month i-1})}$$

$$\text{ACWP-Diff} = \text{ACWP}_{(\text{Month i})} - \text{ACWP}_{(\text{Month i-1})}$$

3. Paste calculated values in Minitab[5] and generate a control chart (or run other tests).
4. Analyze the results by comparing the generated control chart to the relevant time series test case and compile results.

For the EVM variables NCC and CBB, the above steps were followed, with the exception of step 2, which was eliminated since NCC and CBB are non-cumulative variables.

An example of this type of control chart analysis is illustrated in Figure 5. The time series cumulative profile of ACWP is indicated in the chart at the right of the diagram. The control chart for the data is on the left. Two data anomalies are detected in the control chart as indicated by the values' positions above the upper control limit.

[4] BCWS, BCWP, and ACWP are cumulative values. The indicated calculations transform the data into monthly cost values.

[5] Minitab is a statistical software package developed at Pennsylvania State University. See the Minitab website for more information (http://www.minitab.com).

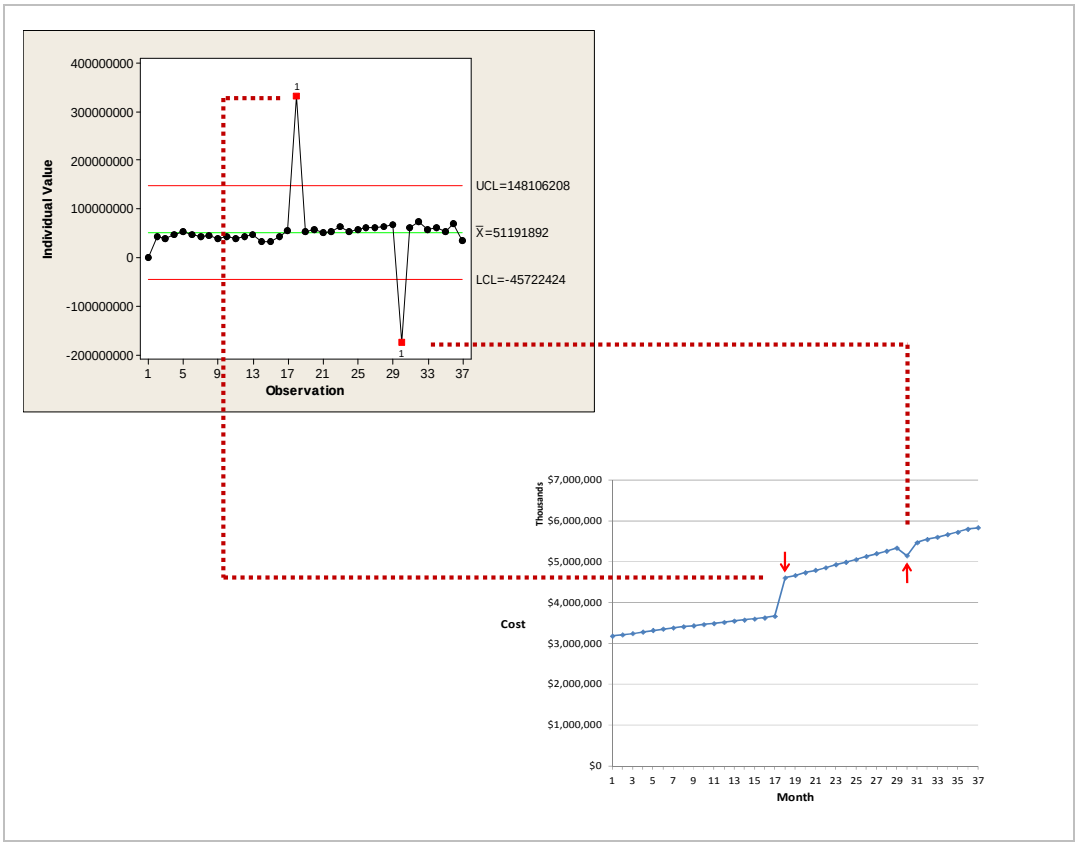

Figure 5: Example of Control Chart Comparison to Corresponding Test Case Profile

2.2.2 Grubbs' Test

Grubbs' test is a statistical test developed by Frank E. Grubbs to detect anomalies in a univariate data set [Grubbs 1969]. Grubbs' test is also known as the maximum normed residual test [Stefansky 1972]. Grubbs' test is defined for the statistical hypothesis:

H_0: The data set does not contain any anomalies.

H_a: There is at least one anomaly in the data set.

The test statistic is the largest absolute deviation from the data set mean in units of the data set standard deviation and is defined as

$$G = \frac{\max_{i=1,n} |X_i - \bar{X}|}{s}$$

where

$\bar{X}$ is the sample mean of the data set

s is the standard deviation of the data set

The hypothesis, $H_{0,}$ is rejected at the significance level, α, if

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/2n,n-2}}{n-2+t^2_{\alpha/2n,n-2}}}$$

where

$t^2_{\alpha/2n,n-2}$ denotes the upper critical value of the *t* distribution with (n-2) degrees of freedom and a significance level of $\alpha/2n$

Grubbs' test detects one anomaly at a time as illustrated in Figure 6. Multiple iterations are executed until no anomalies are discovered.

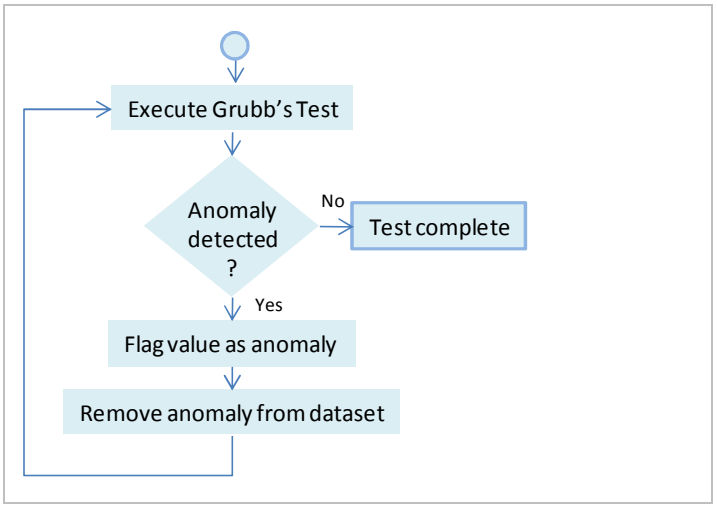

Figure 6: Grubbs' Test Algorithm

The approach described in Section 2.2.1 was used to implement Grubbs' test for the EVM variables BCWS, BCWP, and ACWP. Grubbs' test is a statistical test based on the assumption that the data are approximated by a normal distribution. Therefore, in our research, when suspected anomalies are removed from the transformed values of BCWS, BCWP, and the resultant data *reasonably* approximates a normal distribution. However, the NCC and CBB datasets are very different from BCWS, BCWP, and ACWP, making Grubbs' test ineffective for detecting anomalies for those variables.

2.2.3 Rosner Test

Rosner developed a parametric test designed to detect 2 to 10 anomalies in a sample composed of 25 or more cases [Rosner 1975, Rosner 1983]. The test assumes that the data are normally distributed after the suspected anomalies are removed. As described above for the other tests, the Rosner test is performed on the monthly non-cumulative values. The test requires that the suspected anomalies be identified by inspecting the data beforehand. Once the maximum number of possible anomalies is identified, then they are ordered from most extreme to least extreme.

Using ordered data, the following steps are performed for the Rosner test:

1. The sample mean and standard deviation are calculated based on the *n* sample values. *k* equals the number of suspected anomalies.
2. The sample value with the largest deviation from the mean is used to calculate the test statistic R_i as follows:

$$R_{(1)} = \frac{\left|x_1 - \bar{x}_{(1)}\right|}{s_{(1)}}$$

   X(i) is the value with the largest deviation from the mean but can be either the largest or smallest value in the sample.

3. The sample value X(1) is then removed from the sample, and the mean X(2), S(2), and R(2) are calculated from the *n*-1 values.
4. The previous steps are repeated until all *k* suspected anomalies have yielded corresponding R(*k*) test statistics.
5. Each R(*i*) is compared in sequentially reverse order to a table of critical values for Rosner's test [EPA 2000]. If the computed statistic R(*i*) is greater than the table value, then there are *I* number of anomalies.

The Rosner test is best illustrated with an example. For our example, 37 data entries were ordered by magnitude and used for the Rosner calculations.

Looking at the data, which represents a time series of month-to-month differences, the team hypothesized that there could be four anomalous entries. These are displayed in Table 2 as Ys. Choosing to test for four anomalies, the first iteration calculated the mean of the entire sample and the largest deviation from the mean to calculate the R value as described in the steps above. As the iterations progressed, the sample mean and the standard deviation were reduced as the entries with the largest deviations were dropped from each successive calculation. When four iterations were performed, the test of the R statistic failed for the fourth entry, but was positive for the third. This means that the Rosner test confirmed that there are three anomalies in this data set.

Table 2: Rosner Example

	Mean	Std Deviation	Y	R
R(1)	323,350	204,601	1,031,831	**3.46**
R(2)	303,108	167,056	878,047	**3.44**
R(3)	286,198	135,801	701,373	**3.06**
R(4)	273,617	116,054	540,379	2.30

The calculated R(i) is bolded where it exceeds the tabled critical value. For completeness, the three data records (Y) identified as anomalies in this example are shown.

2.2.4 Dixon Test

The Dixon test (sometimes referred to as Dixon's extreme test or Dixon's Q test) was designed for identifying anomalies when the sample size is less than or equal to 30 [Dixon 1951]. Recent research has extended its applicability to samples up to 100 and improved the precision and accuracy of the critical values for judging the test results [Verma 2006]. The test measures the ratio of

difference between an anomaly and its nearest neighbor to the range of the sample (see Table 2). The tests do not rely on the use of the mean or standard deviation of the sample.

Dixon initially posited a series of six calculations to test for anomalies. Which test to use depends on the sample size and whether the test is for a single anomaly or pairs of anomalies. The tests for pairs of anomalies were designed to account for masking effects when there is more than one extreme anomaly present in the data [Barnett 1998]. The data are tested for normality and, if necessary, transformed to fit a normal distribution. For our data, we used either a Box-Cox transformation or a Johnson transformation n the earned value data, the negotiated contract cost (NCC) and the contract budget base (CBB) variables could not be normalized, so the Dixon test was not appropriate and was not used.

The data are then ordered, and the largest or smallest extreme values are tested by calculating the following appropriate statistic.

Table 3: Dixon Calculations for Identification of Anomalies

Sample Size n	Test Statistic	To Test for Smallest	To Test for Largest
$3 \le n \le 7$	r_{10}	$\frac{x_2 - x_1}{x_n - x_1}$	$\frac{x_n - x_{n-1}}{x_n - x_1}$
$8 \le n \le 13$	r_{21}	$\frac{x_3 - x_1}{x_{n-1} - x_1}$	$\frac{x_n - x_{n-2}}{x_n - x_2}$
$n \ge 14$	r_{22}	$\frac{x_3 - x_1}{x_{n-2} - x_1}$	$\frac{x_n - x_{n-2}}{x_n - x_3}$

The value of the calculated r_n is then compared to a table of critical values. If the calculated r_n is greater than the corresponding critical value, the value can be characterized as an outlier. In this research, the critical values at the 95% confidence level were used.[6]

The Dixon test is meant to identify one extreme outlier, although the r_{21} and r_{22} statistics have been shown to be robust in the presence of more than one anomaly [Ermer 2005]. For our purposes, we were interested in the performance of the Dixon test compared to other anomaly detection techniques using the monthly differences for the earned value variables BCWS, BCWP, and ACWP.

To judge the efficacy of the Dixon test in identifying anomalies, a series of rolling brackets was imposed on the data for each of the three earned value variables. That is, when testing r_{10} for a large extreme datum, the statistic was calculated by using three consecutive data records at a time. For r_{21}, we used 8 consecutive cases and for r_{22}, we used 14 consecutive cases. Both the largest and smallest values were tested. The anomalous data records identified using this technique are shown in Appendix C.

[6] ISO 57255 suggests that if the test result is significant at the 95% level but not at 99%, the datum should be characterized as a straggler and requires further examination [Huah 2005].

2.2.5 Tukey Box Plot

Originally developed by John Tukey for exploratory data analysis, box plots have become widely used in many fields. As described earlier, the test is performed on the monthly non-cumulative values.

Figure 7 contains an image from the JMP statistical package's help file on outlier box plots.[7] The box runs from the 1st through the 3rd quartile (25th and 75th percentiles) of the entire data distribution; the distance between the two ends of the box is called the interquartile range. The whiskers stretch to the outermost data points both above and below the box, in each of which lie another 1.5*(interquartile range) of the data points. Any dots above or below the whiskers are classified as anomalies.

The bracket to the left of the box is the range in which the densest 50% of those data lie. The confidence diamond represents the confidence interval in which a sample's mean most likely lies, which may not be the same as the median as represented in the example.

Box plots as originally envisaged by Tukey make no assumption of statistical normality. They are simply based on distribution of the data by percentiles. The region between the ends of the whiskers contains 99.3% of the observations, which makes box plots equivalent to the 3σ technique for Gaussian data, although it is slightly more generous than Shewhart's rule of approximately 99.7% for identifying anomalies for statistical process control methods [Shewhart 1931]. As a result, a few more data points may be classified as anomalies using box plot techniques than in techniques using more stringent criteria.

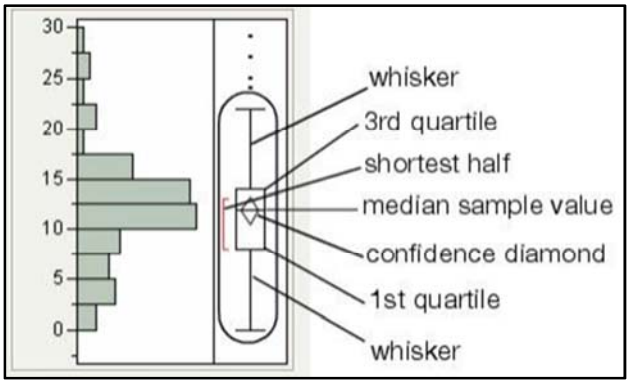

Figure 7: Interpreting Tukey Outlier Box Plots

2.2.6 Autoregressive Integrated Moving Average (ARIMA) Models

ARIMA models are widely used for both short- and long-term extrapolation of economic trends [Box 1970]. A particular strength of ARIMA is that it encompasses many related statistical time series methods in one general framework. While ARIMA models were originally intended (and continue to be used most widely) for modeling time series behavior and forecasting, they have

[7] A similar figure can be found in Release 8 of the JMP *Statistics and Graphics Guide* (http://www.jmp.com/support/downloads/pdf/jmp8/jmp_stat_graph_guide.pdf).

been used by others for anomaly detection as well [Bianco 2001, Chen 2005, Galeano 2004, Tsay 2000].

Many time series patterns can be modeled by ARIMA, but all such patterns amenable to ARIMA would have an autocorrelation or partial autocorrelation element to the model (that is, the value of any particular data record is related to earlier values). Differencing the data (calculating the differences between data values) is the step that simplifies the correlation pattern in the data.[8] Often a cyclic or seasonal pattern must be accounted for in the model. Once the proper order of differencing has been identified, the observations are integrated to characterize the overall trend in the original time series data (which accounts for the "I" in ARIMA). Autoregressive (AR) and/or moving average (MA) terms may be necessary to correct for any over- or under-differencing. An AR term or terms may be necessary if a pattern of positive autocorrelation still exists after the integration. An MA term or terms may be necessary if any negative autocorrelation has been introduced by the integration; this is likely to happen if there are step jumps where the original series mean increases or decreases at some thresholds over time. The goal of ARIMA is to account for all factors that determine the values in the time series so that any residual variation is attributable to "noise." Obviously, the best fit accurately models the values in the series while minimizing noise. Statistical software handles all the needed calculations and produces an array of visual outputs to guide the selection of an appropriate model.

Fortunately, the EVM time series that we analyzed tends to have much simpler best model fits than are sometimes required for more complex time series with seasonal cycles. ARIMA models can be quite varied in their construction; for our data, a nonseasonal ARIMA is appropriate. Such a model is classed as an ARIMA *(p,d,q)* model where

- p is the number of autoregressive terms
- d is the number of nonseasonal differences
- q is the number of lagged forecast errors in the prediction equation

Since ARIMA models often are nonlinear, the best fits are displayed by line and curve segments. An example is shown in Figure 8, which displays one of the 20 time series that we used to compare the anomaly detection methods described in this report. The actual data values are represented as dots, some of which are identified as anomalies using the Tukey box plots that are described below. The most extreme anomalies appear clearly outside of the confidence intervals displayed around the best fit in the figure. Using many of the existing statistical packages, any data point can be identified simply by mousing over it.

[8] The non-cumulative series are first-differenced series in mathematical terms. The transformation is done by subtracting the numerical value of its immediately preceding data point from the numerical value of each succeeding data point. The difference between the two will be positive if the prior value was smaller, negative if the succeeding value is smaller, and zero if they are the same. Statistical software packages do the same transformation automatically for as many time lags of integration as are necessary to find the best model fit (e.g., second differences, which are simply the differences between consecutive first differenced values).

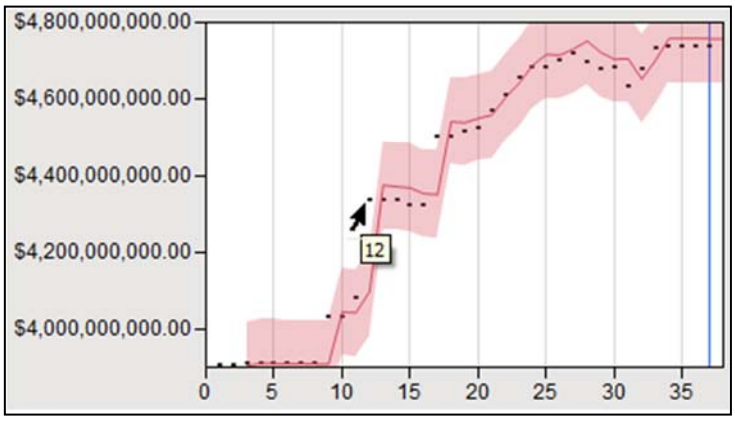

Figure 8: An Example ARIMA Best Fit of an EVM Distribution

Simple first difference (e.g., $X_t - X_{t-1}$) fits are sufficient in instances that we analyzed in doing our comparisons of anomaly detection methods. In addition, an ARIMA model can almost always be fit for variables that other anomaly detection methods do not handle well (e.g., for EVM management reserve).

The analysis of the model's residuals also plays a key role in determining the suitability of any particular model. We used Tukey box plots as part of the residual analysis to avoid making assumptions about normality of the data distributions as well as for their intuitive interpretation of what constitutes an anomaly (see Section 2.2.5).

2.2.7 3-Sigma Outlier

Many of the techniques discussed thus far operate on the data of a single program task and use all of the data within the program task data set as part of the anomaly detection technique. The 3-sigma outlier test is an automated algorithm that we developed as a way to evaluate the entire EVM data set, including all program tasks within the data set. The algorithm was implemented in a Microsoft Excel application. Rather than use the entire task data, the algorithm evaluated accumulated data beginning at month three (i.e., with three data values) and then carried out iterations for months four to n (where n is the total number of values in the program task). When a new program task ID was encountered, the calculations and counters were reset to initiation. A summary of the algorithm is depicted in Figure 9.

This technique simulates the real-world situation of monitoring data as it is being recorded in a database, rather than the retrospective inspection of data once the entire data set is available.

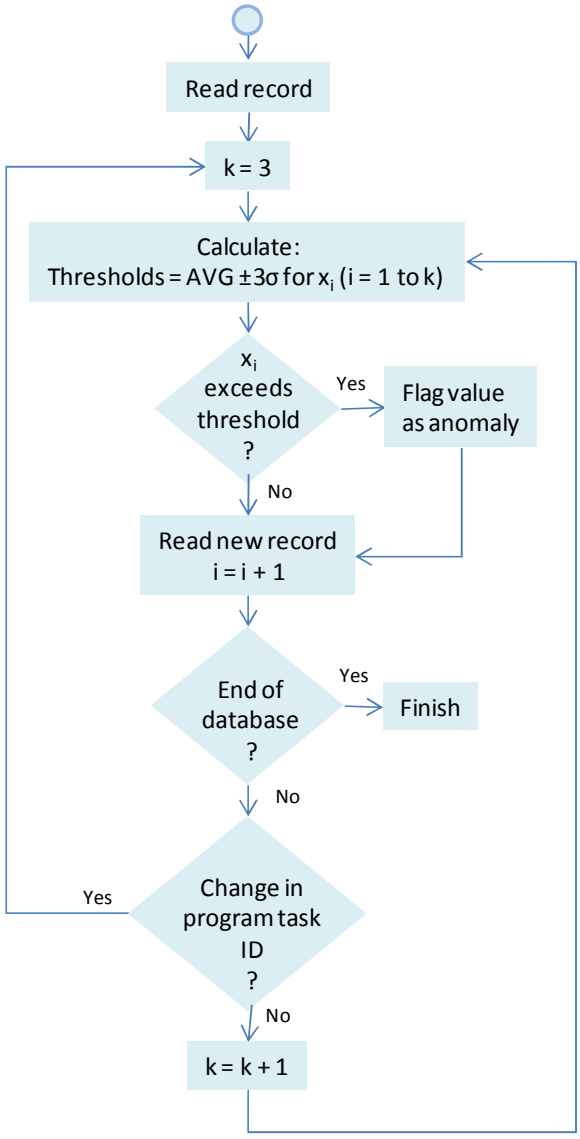

Figure 9: 3-Sigma Outlier Algorithm

2.2.8 Moving Range Technique

We developed the moving range technique following the control chart analyses listed in Table 1. Based on the efficacy of the mR control chart for detecting anomalies in NCC and BCC data, the moving range technique is an adaptation of this particular control chart scheme.

As in the 3-sigma outlier test, we used a Microsoft Excel application that evaluates accumulated data beginning at month three (i.e., with three data values) and then carries out iterations for months four to n (where n is the total number of values in the program task). When a new program task ID was encountered, the calculations and counters were reset to initiation.

The general flow of the algorithm is the same as that shown in Figure 9 except for the anomaly detection test, which is depicted in the third box from the top of the diagram. The anomaly detection test for the moving range technique is as follows:

$$mR_i = |X_i - X_{i-1}|$$

$$\bar{R} = \frac{\sum_{i=2}^{i=k} mR_i}{k-1}$$

$$UCL = D_4\bar{R}$$

where

X_i is the value of NCC or CBB for record i

k is the number of data values in the program task; for k individual values, there are k-1 ranges

D_4 is the sample-size-specific anti-biasing constant for n=2 observations that are used in calculating mR_i [Montgomery 2005]

A value is flagged as an anomaly if

$$mR_i > UCL$$

2.2.9 SPI/CPI Outlier

In the earned value management system, the schedule performance index (SPI) and cost performance index (CPI) are defined as

$$SPI = \frac{BCWP}{BCWS}$$

$$CPI = \frac{BCWP}{ACWP}$$

Our research team explored the use of these variables as a way to normalize the entire data set (i.e., the multiple program data available in the data set) so that anomaly detection analysis was not constrained to a program task by program task evaluation. This approach was explored because there was a possibility that anomalous SPI and CPI values could be detected across the entire data set (that is, across multiple program tasks).

The SPI/CPI outlier technique was implemented as follows for SPI:

1. Calculate $SPI_{Diff} = SPI_i - SPI_{i-1}$ for i=2 to n where n is the total number of records in the EVM data set.
2. Calculate the average value, $\bar{X}$, of the SPI_{Diff} values.
3. Calculate the standard deviation (sd) of the SPI_{Diff} values.
4. Calculate $T_U = \bar{X} + (3 * sd)$ and $T_L = \bar{X} - (3 * sd)$.
5. If $SPI_{Diff} > T_U$, flag the value as an anomaly; if $SPI_{Diff} < T_L$, flag the value as anomaly and investigate the corresponding EVM measures.

The SPI/CPI outlier technique was implemented as follows for CPI:

1. Calculate $CPI_{Diff} = CPI_i - CPI_{i-1}$ for i=2 to n where n is the total number of records in the EVM data set.
2. Calculate the average value, $\bar{X}$, of the CPI_{Diff} values.
3. Calculate the standard deviation (sd) of the CPI_{Diff} values.
4. Calculate $T_U = \bar{X} + (3 * sd)$ and $T_L = \bar{X} - (3 * sd)$.
5. If $CPI_{Diff} > T_U$ then flag the value as an anomaly; if $CPI_{Diff} < T_L$then flag the value as an anomaly and investigate the corresponding EVM measures.

3 Results and Discussion

3.1 Comparison of Techniques

In this research study, we evaluated the following anomaly detection techniques:

- control chart for individuals
- moving range (mR) control chart
- exponentially weighted moving average (EWMA) control chart*
- moving average control chart*
- Grubbs' test
- Rosner test
- Dixon test
- autoregressive integrated moving average (ARIMA)
- Tukey box plot
- 3-sigma outlier
- moving range technique
- SPI/CPI outlier*

* *These techniques were found to be completely ineffective for detecting anomalies in the EVM data. Therefore, they are not discussed further in this report.*

We found that some techniques were effective for discovering anomalies in the variables BCWS, BCWP, and ACWP, but proved ineffective for detecting anomalies in the NCC and CBB variables. This is because the variables behave in fundamentally different manners. BCWS, BCWP, and ACWP are cumulative variables whose typical time series profile is curvilinear. NCC and CBB are reported values tied to the contract and do not typically change on a month-to-month basis. When these variables do change over time, the resultant time series appears as a step function. We partitioned the analysis results into two sections to reflect the different character of these two groups of variables and the techniques that were used to detect anomalies within the variables.

3.2 Performance of Techniques Applied to BCWS, BCWP, and ACWP

Figure 10 provides a graphical summary of the performance of the techniques that were found to be effective for BCWS, BCWP, and ACWP when the results of all test cases were combined. Table 4 shows the same results in tabular format, and a further breakdown of the results is presented in Appendix C.

With respect to detection rate, it may appear that Grubbs' test outperformed all other tests (with the highest detection rate of 85.4%). However, the differences in detection rates among the five top performers (i.e., Grubbs' test, Rosner test, box plot, ARIMA, and control chart for individu-

als) are not statistically significant. These techniques as a group did perform better than the four remaining techniques, and this outcome is statistically significant.[9] A probable explanation for this difference is that the top performers benefited from the use of the entire set of data in each test case data set to construct the statistical parameters of the anomaly detection technique. However, the four techniques represented on the right of Figure 10 were implemented in such a way as to simulate the monthly accumulation of EVM data over time. These techniques evaluated the existence of anomalies sequentially, without using all information in the data set to evaluate whether the new incoming data was anomalous. For example, at month six, the variable value was tested using only the available six data values. Then, the next record was read and value seven was evaluated using n=7. But, for the five techniques on the left of the graph, the entire data set (e.g., 73 values in some cases) was used to evaluate the month six value to determine whether it was anomalous. Having the benefit of all the information in the data set likely led to the detection rate effectiveness of the five top performing techniques.

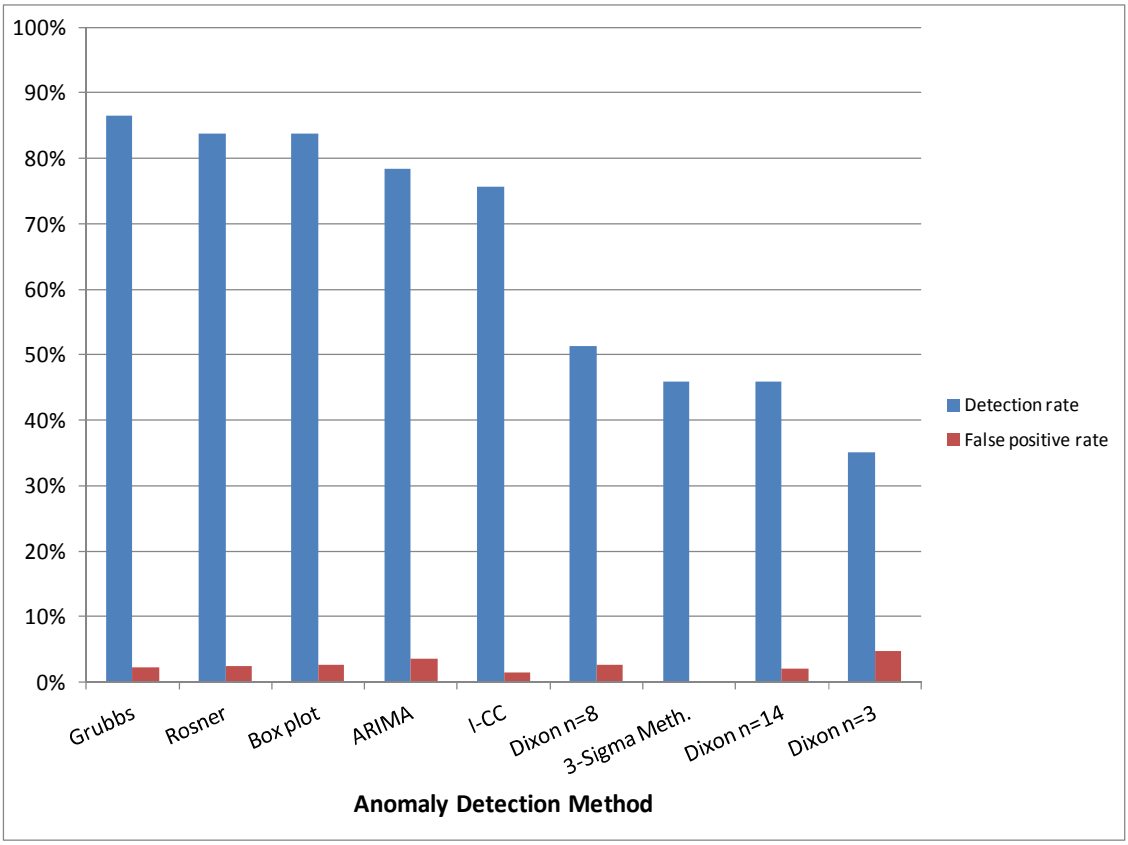

Figure 10: Anomaly Detection Effectiveness for EVM Variables BCWS, BCWP, and ACWP Across All Four Test Cases

[9] A Chi-Square test for equalities of proportions and significance tests for two proportions were used to establish statistical significance of the differences between techniques. See Appendix C for the details of these tests.

Table 4: Anomaly Detection Effectiveness for EVM Variables BCWS, BCWP, and ACWP Across All Four Test Cases

	Grubbs	Rosner	Box plot	ARIMA	I-CC	Dixon n=8	3-Sigma Meth.	Dixon n=14	Dixon n=3
Detection rate	86.5%	83.8%	83.8%	78.4%	75.7%	51.4%	45.9%	45.9%	35.1%
False positive rate	2.2%	2.4%	2.6%	3.6%	1.5%	2.7%	0.2%	2.0%	4.8%

Having established that these five techniques perform similarly with respect to anomaly detection rate (i.e., those that appear in the top row, starting in the left-most column of Table 4), the false alarm rates were compared among them. These appear in the bottom row of Table 4. The differences in false alarm rate among the five top performers were statistically insignificant, based on the outcome of the Chi-square test for equalities of proportions.[10] Therefore, based on our two measures of effectiveness (that is, detection rate and false alarm rate), our analysis suggests that Grubbs' test, Rosner test, box plot, ARIMA, and the control chart for individuals (I-CC) all performed at the same level.

Sections 3.2.2 through 3.3.4 describe some of the qualitative factors associated with each of the five techniques that performed well with respect to detection rate. These qualitative factors are summarized in Table 5.

Table 5: Qualitative Criteria Used to Evaluate High Performance Anomaly Detection Techniques

Qualitative Criterion	Definition	Indicators
Efficiency	The extent to which time is well used for the intended task.	The number of times human intervention is required (for the purpose of decision-making) before the technique can execute to completion.
		The amount of human intervention time required by a technique to complete the evaluation of the data set.
Flexibility	Susceptible to modification or adaptation. The ability of a technique to respond to potential changes affecting its value delivery in a timely and cost-effective manner.	The validity of results when data are from a non-Gaussian distribution.
		Effectiveness of technique for small and large sample sizes.
		Ease with which the sensitivity of the anomaly detection technique can be adjusted.
Simplicity	Freedom from complexity, intricacy, or division into parts.	Amount of burden put on someone to understand the technique or to try to explain it to a measurement and analysis novice.
Extensibility	The ability of a technique to be operationalized in a production environment with minimal effort or disruption of the existing system.	The level of effort required to extend technique to implementation in production environment.

[10] See Appendix C for the details of the significance test results.

In the following sections, each of the high-performing anomaly detection techniques are discussed with respect to the qualitative criteria listed in Table 5.

3.2.1 Control Chart for Individuals

The control chart for individuals was a top performer as determined by the two measures of effectiveness used in this study: it had a detection rate of 75.7% and a false alarm rate of 1.5%. This control chart is a popular tool for understanding variation in a process and system and is particularly well-suited for identifying anomalies in a data set. Anomalies are identified by their appearance above the upper control limit or below the lower control limit of the control chart. The centerline and control limits of the chart are calculated using the available data. Therefore, the control chart operates best when there is sufficient data to generate an accurate portrayal of the average and standard deviation of the data set. For small data sets (n<10), the control limits ($\bar{X} \pm 3\sigma$) may generate additional false alarms due to an inflated standard deviation caused by the contribution of even a single deviant data value. As n increases, the calculations of the control limits become more reliable in terms of representing the true standard deviation of the data set. In practice, control limits based on n<10 are typically referred to as trial limits until additional data become available.

With respect to sensitivity, the upper and lower control limits can be adjusted for increased or decreased sensitivity. While the typical application is based on 3σ limits, these can be adjusted up or down to change the sensitivity of the detection scheme. For example, under certain conditions, this adjustment is implemented when control charts are used to monitor industrial processes. Here, they are sometimes referred to as *warning limits* [Montgomery 2005].

Incorporating the control chart for individuals scheme into a data collection stream of activities would not be difficult. The implementation of the tool in a production environment would be relatively straightforward and practical to accomplish.

3.2.2 Grubbs' Test

Grubbs' test was also a top performer with a high detection rate of 86.5% and a relatively low false alarm rate of 2.2%. With regard to efficiency, Grubbs' test is not difficult to apply when using a statistical package such as Minitab for the analysis. A macro has been developed for Minitab that implements Grubbs' test for a specific alpha [Griffith 2007]. When the test is performed manually, the calculations are compared to a look-up table that provides critical values. Grubbs' test works well for small n as well as large n. Tables of Grubbs' test critical values are available for n=3 to 100 [Lohninger 2011].The computations are straightforward, as is the comparison of the Grubbs' test statistic to a table of critical values that are available in many statistics books and reference sources [Lohninger 2011, Dunkl 2007].

While empirical results using Grubbs' test were impressive, the test assumes an approximate normal distribution. In cases where there is a large departure from normality, false alarms may be generated due to non-normality, rather than the presence of anomalies.

The sensitivity of Grubbs' test can be adjusted by changing the value of alpha. The alpha used in this research study was set to $\alpha = 0.05$.

3.2.3 Rosner Test

The Rosner test detected 83.8% of the anomalies with a false positive rate of 2.4%, making it the second best performer among the techniques presented here. Unfortunately, the Rosner test suffers several unique drawbacks that make its implementation problematic. First, the test is not generally available in statistical software packages. Although the algorithm involved is not complex, the iterative nature of the technique complicates the programming requirements such that it may be beyond the skills of a normal user. Any organization seeking to implement the Rosner test would need to devote resources to develop such software. A second major drawback is the maximum limitation of 10 anomalies and a minimum of 25 data records. Analysis of a long-term data series might exceed the limit of 10 anomalies, particularly when investigating programs with life cycles that span decades. The minimum of 25 data records also means that a program task would have to produce more than two years of data before the Rosner test could be used. Third, the Rosner test requires the analyst to identify the suspected anomalies *before* initiating the test. Although this might be done visually, it means additional time and effort on the part of an analyst in order to implement the test.

Like the Grubbs' and Dixon tests, the Rosner test assumes an approximate normal distribution of the non-anomalous data (those data records remaining after the anomalies are removed). This makes it susceptible to false positives when there is a departure from normality. The Rosner test also produces a test statistic, which is compared to a table of critical values that is widely available [U.S. Army 2008, Barnett 1998, EPA 2000, Gibbons 2001, Rosner 1983]. The sensitivity of the Rosner test is also adjustable; the alpha used in this research study was set to $\alpha = 0.05$ and critical values are available for $\alpha = 0.01$ and $\alpha = 0.005$.

3.2.4 Tukey Box Plot

The Tukey box plot technique for non-cumulative distributions was also a top performer, with a high detection rate of 83.8% and a relatively low false alarm rate of 2.6%. Box plots can be generated easily and efficiently using many readily available statistical packages. Transformation of the time series into non-cumulative format is easily done in a spreadsheet and can be done with a single mouse click in many statistical packages. Box plots make no assumptions about normality or other statistical properties, and the results are easy to interpret and describe intuitively. The cut-off points for determining what constitutes an anomaly can be easily adjusted based on historical experience and the judgment of domain experts in validating the statistical results. The anomalies can be identified for validation by domain experts with a simple copy and paste from the data tables in any good statistical package. The necessary procedures could be easily automated for use in a production environment.

3.2.5 ARIMA

The ARIMA technique was also a top performer, with a high detection rate of 78.4% and a relatively low false alarm rate of 3.6%. For someone experienced with statistical packages, ARIMA techniques are relatively straightforward to use for anomaly detection in relatively simple univariate time series, such as the EVM data that we analyzed. There is no need to transform the time series data into non-cumulative series, which saves time and may be helpful for EVM analysts who are accustomed to visualizations of cumulative time series. Semi-automated software tools and relatively painless guidance for finding the best ARIMA model fit can be made available to

EVM domain experts in a production environment. Anomalies can be easily determined by importing the residuals from an ARIMA model into existing box plot software.

A particular strength of ARIMA is that it subsumes many related statistical time series techniques into one general framework, and it may prove to be more widely applicable for EVM and other time series data that are more complex than those we used to compare statistical anomaly detection techniques thus far. A potential drawback is over-fitting to the data, potentially causing the number of false negatives to increase.

3.3 Performance of Techniques Applied to NCC and CBB

Selection of an anomaly detection scheme is dependent on the characteristics of the data. The time series behavior of the EVM variables NCC and CBB is fundamentally different than the behavior of the variables BCWS, BCWP, and ACWP. NCC and CBB are non-cumulative variables whose time series profiles typically (but not always) appear as step functions (see Appendix B). Techniques that performed well for detecting anomalies in BCWS, BCWP, and ACWP did not necessarily work well for NCC and CBB.

The following four techniques effectively identified anomalies in NCC and CBB:

- mR control chart (CC)
- moving range technique
- ARIMA
- Tukey box plot

Figure 11 summarizes the ability of these techniques to discover anomalies in the NCC and CBB variables of the four test cases. Table 6 presents the results in tabular format.

All four proved to be 100% effective in discovering data anomalies in the test cases. With respect to false alarm rates, some techniques performed better than others; however the differences were statistically insignificant (see Appendix C).

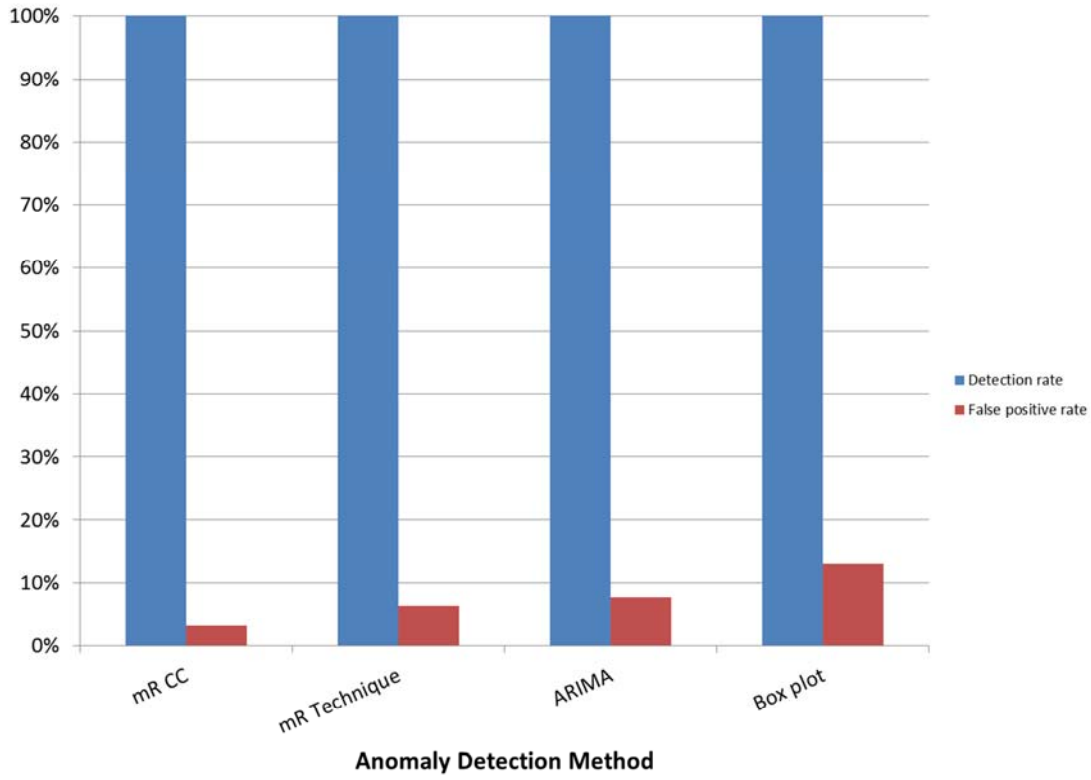

Figure 11: Anomaly Detection Effectiveness for EVM Variables NCC and CBB Across All Four Test Cases

Table 6: Anomaly Detection Effectiveness for EVM Variables NCC and CBB Across All Test Cases

	mR CC	mR Technique	ARIMA	Box Plot
Detection rate	100.0%	100.0%	100.0%	100.0%
False positive rate	3.0%	6.3%	7.6%	12.9%

In Sections 3.3.1 through 3.3.4 we discuss some of the qualitative factors associated with each of the four techniques that performed well with respect to detection rate. These qualitative factors are summarized in Table 5 on page 25.

3.3.1 Moving Range Control Chart

The mR control chart performed well for detecting anomalies in NCC and CBB variables, with a detection rate of 100% and a false positive rate of 3%. When used in the industrial domain, the mR control chart is paired with the control chart for individuals to monitor the variation of a process [Montgomery 2005]. However, for our purposes, the mR control chart was used solely for detecting anomalies for these variables.

This technique can be easily automated and does not require human judgments or interaction to execute the sequence of steps required for anomaly identification. The approach is straightfor-

ward. As with all control charts, anomalies are indicated by the appearance of a data point above the upper control limit or below the lower control limit.

3.3.2 Moving Range Technique

The moving range technique was essentially a direct implementation of the moving range chart within a Microsoft Excel spreadsheet application. The difference between the two was that the moving range chart relied on the entire data set for analysis of anomalies, while the moving range technique considered only the available subset of data available when the EVM data was reported. Using only a subset of the data for anomaly evaluation led to additional false alarms as compared to the mR control chart.

Implementing this technique confirmed that it would not be difficult to automate the mR control chart within a production environment.

3.3.3 ARIMA

The ARIMA technique performed well for the NCC and CBB variables, with a high detection rate of 100% and a relatively low false alarm rate of 7.6%. ARIMA is equally applicable to cumulative and non-cumulative series, including series with step jumps such as NCC and CBB. The same semi-automated software tools and relatively painless guidance for finding the best ARIMA model fit could be available to EVM domain experts in a production environment, and the anomalies could be easily determined by importing the residuals from an ARIMA model into existing box plot software.

3.3.4 Tukey Box Plot

The Tukey box plot technique did well, with a detection rate of 100%. The false alarm rate of 12.9% is relatively higher for the NCC and CBB series, although not statistically significantly so. As noted for the comparisons of the BCWS, BCWP, and ACWP time series, box plots are equally easy to use and interpret for any time series, and the cut-off points for determining what constitutes an anomaly can be easily adjusted based on experience. The necessary procedures could be easily automated for use in a production environment.

4 Conclusion

4.1 Summary of Results

In this research study, we investigated the efficacy of anomaly detection techniques on earned value management data submitted on a monthly basis by government contractors to the EVM-CR. Five variables from the data set were analyzed for anomalies. Based on their time series behavior (see Appendix B), the variables fell into two categories as shown in Table 7.

Table 7: Summary of EVM variables investigated in this study.

Group 1	Group 2
Budgeted Cost of Work Scheduled (BCWS)	Negotiated Contract Cost (NCC)
Budgeted Cost of Work Performed (BCWP)	Contract Budget Base (CBB)
Actual Cost of Work Performed (ACWP)	

4.1.1 Summary of Results – BCWS, BCWP, ACWP

Of the various techniques we analyzed in this study, we found that five techniques were equally effective for identifying anomalies in BCWS, BCWP, and ACWP. These techniques were:

- Grubbs' test
- Rosner test
- Tukey box plot
- ARIMA
- control chart for individuals

The Grubbs' and Rosner tests are better suited for addressing anomaly detection in small sample sizes as well as large sample sizes. On the other hand, the effectiveness of Tukey box plot, ARIMA, and the control chart for individuals rely on the existence of larger sample sizes (approximately $n>10$).

The Grubbs' and Rosner tests assume that the data are from an approximate normal distribution. In cases of non-normal data, there is a chance that anomalies will escape detection. However, Tukey box plot, ARIMA, and control chart for individuals are more robust in that they are not as sensitive to departures from normality.

In production environments, some techniques will require more human judgments than others. We believe that Grubbs, Rosner, Tukey box plot, and control chart for individuals could all be implemented in an automated environment without significant effort or disruption. However, ARIMA would require significant software programming to address the logic required to implement the technique in a fully automated way.

Therefore, when choosing among the top performers in this group, the conditions and trade-offs must be considered. Given the simplicity and robustness in situations of non-normality, the Tukey box plot appears to be a stand-out performer when sample sizes are greater than 10, while either Grubbs' or Rosner tests should be used when the sample size is small.

4.1.2 Summary of Results – NCC, CBB

Three techniques were found to be effective for discovering anomalies in the NCC and CBB variables. The techniques are

- ARIMA
- mR control chart
- moving range technique

These techniques performed at 100% effectiveness for identifying data anomalies in our test cases. The differences in the false alarm rate among the techniques were insignificant.

The moving range technique is an adaptation of the mR control chart. The techniques are essentially the same except that the moving range technique evaluated the data one record at a time (for n>3), while the mR control chart used the entire data set of values.

As stated in Section 4.1.1, ARIMA is somewhat complex because it requires human judgment as part of the method. Implementing a fully automated ARIMA method would be more costly than implementing a method based on the moving range of the data. The calculations and anomaly detection rules associated with the moving range technique are simple and would be easy to implement as an automated stand-alone anomaly detection system. Therefore, moving range is recommended as the technique of choice for the detection of anomalies in variables whose time series behave similarly to NCC or CBB.

4.2 Challenges Encountered During This Research

We encountered a number of challenges during the course of this research project. First, we were not able to test our techniques against data that had been previously verified as error free. We dealt with this issue by involving an EVM subject matter expert to identify probable defects that we used as test cases in our analysis.

A second challenge involved distinguishing data errors from accurate data that depicted anomalous program behavior. Data anomalies are detected by measuring the departure of values from what they are expected to be. The expectation in this research was based on statistical and probabilistic models and distributions. When a value is within an expected range, it is treated as valid and accurate. However, when it is a measurable departure from what is expected, it is treated as anomalous. Defining a normal region that minimizes the number of false positive and false negative anomalies can be difficult. The boundary between valid and anomalous values is often imprecise. Thus, an anomalous observation that lies close to the boundary distinguishing valid and anomalous values can actually be valid, and vice-versa [Chandola 2009].

A third challenge was the nature of EVM-type data, as it represents actual performance and is not from a stochastic process that can be modeled. Human intervention is at play as program managers make adjustments to the allocation of resources based on the current state of the program. This redistribution of resources throughout the program causes the performance indicator to change in ways that may not be predictable.

Finally, an additional concern associated with this factor is the process for resolving whether a defect is caused by an error or by actual program performance. In all cases, when an anomaly is discovered, the only reliable way to determine its true nature is to trace the data value back to the

source to conduct root cause analysis. In this study, we were unable to obtain traceability back to the source (individual or authoritative record) that could resolve the nature of the anomaly. As in the previously identified challenge, we mitigated this issue by consulting with EVM subject matter experts to distinguish anomalies (identified in our test cases; see Appendix B) resulting from probable data defects vs. anomalies attributable to actual program performance.

4.3 Implications of This Research

Because the cost of poor data quality is a significant problem in government and commercial industry, the National Resource Council (NRC) report, *Critical Code: Software Producibility for Defense*, Recommendation 2-2 states: "The DoD should take steps to accumulate high-quality data regarding project management experience and technology choices" [NRC 2010]. But committing errors is part of the human condition. We all do it, no matter how careful we are. We rely on quality checks, peer reviews, and inspections to weed out errors in the final product. Without these safeguards defects are injected into the product and processes and remain there.

Information is the product of the data life cycle. As noted in Figure 12, the potential for errors is significant because errors can be injected whenever human beings *touch* the data through processing and analysis activities as the data are transformed into information that supports decision making. Correcting the data errors represents costly rework to determine the source of the error and fix it. When errors go undetected, flawed analysis leads to potentially flawed decisions that are based on the derived information. Also, since many information systems involve multiple shared repositories, data errors are replicated and propagate uncontrollably. This is why it is important to focus on correcting data errors at the time of entry rather than downstream in the data life cycle where the errors become embedded.

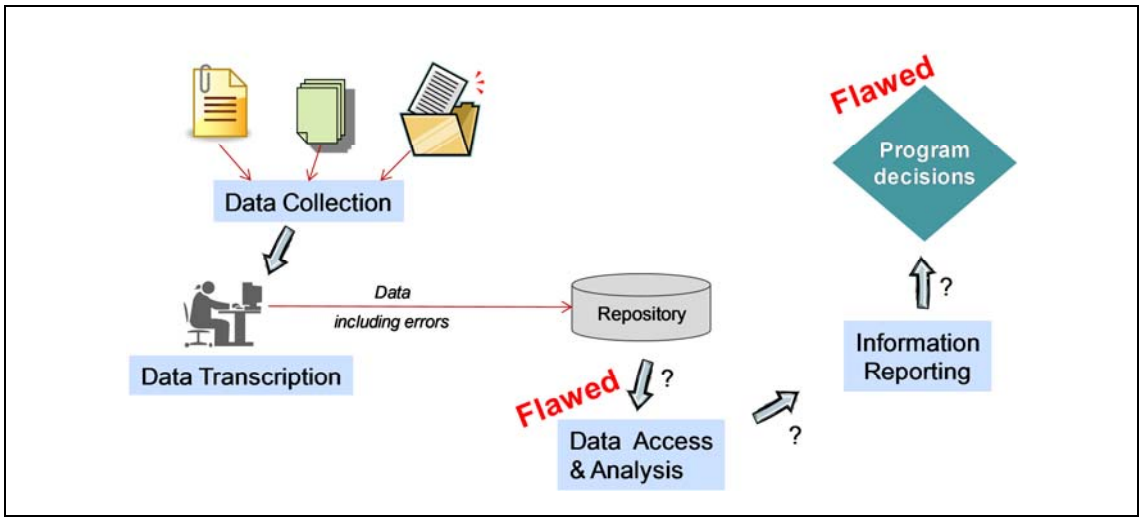

Figure 12: The Data Life Cycle

Many organizations are flooded with data, and error detection methods are ad hoc or non-existent. While some errors are detected through manual "sanity" checks of the data, many types of errors escape detection due to the volume of data and the difficulty and tediousness of manual inspection.

The purpose of this research study was to investigate the efficacy of methods that detect potential data errors through automated algorithms. The development of automated support would improve data quality by reducing data defects and release the analysts from the tedious and repetitive task of manual inspection so they can focus their efforts more productively.

4.4 Recommendations

This research demonstrates that statistical techniques can be implemented to discover potential data anomalies that would have otherwise gone undetected. We believe that it would be technically feasible and potentially very practical to codify the high performing statistical techniques into automated procedures that would scan and screen data anomalies when data are being entered into a repository. Such a capability could be coupled to and preceded by more basic types of error checking that would initially screen basic types of errors from the data based on business rules. There also may be significant potential for improving anomaly detection based on multivariate approaches.

Future research should focus on the cost/benefit analysis to determine the economic advantages of automating a data anomaly detection capability that could serve as the front end of a data collection system. While it appears there will be a need for back-end checks that use all of the available records for a program, it may be that highly effective front-end checking would eventually eliminate the need for such a process.

Appendix A Data Defect Taxonomy

This table was adapted from the work of Larry English [English 2009].

Table 8: Data Defect Taxonomy

Data Defect	Description
Definition conformance	Data values are consistent with the attribute definition.
Existence	Each process has *all* the information it requires.
Record existence	A *record* exists for every real-world object or event the enterprise needs to know about.
Value existence	A given data element has a full value stored for all records that *should* have a value.
Completeness	Each process or decision has *all* the information it requires.
Value completeness	A given data element (fact) has a full value stored for all records that *should* have a value.
Validity	Data values conform to the information product specifications.
Value validity	A data value is a valid value or is within a specified range of valid values for this data element.
Business rule validity	Data values conform to the specified business rules.
Derivation validity	A derived or calculated data value is produced correctly according to a specified calculation formula or set of derivation rules.
Accuracy	The data value *correctly* represents the characteristic of the real-world object or event it describes.
Accuracy to reality	The data *correctly* reflects the characteristics of a real-world object or event being described. Accuracy and precision represent the highest degree of *inherent* information quality possible.
Accuracy to surrogate source	The data agree with an original, corroborative source record of data, such as a notarized birth certificate, document, or unaltered electronic data received from a party outside the control of the organization that is demonstrated to be a reliable source.
Precision	Data values are correct to the right level of detail or granularity, such as price to the penny or weight to the nearest tenth of a gram.
Non-duplication	There is *only one* record in a given data store that represents a single real-world object or event.
Source quality and security warranties or certifications	The source of information (1) guarantees the *quality* of information it provides *with remedies for non-compliance;* (2) documents its certification in its Information Quality Management capabilities to capture, maintain, and deliver Quality Information; (3) provides objective and verifiable measures of the quality of information it provides in agree-upon quality characteristics; and (4) guarantees that the information has been protected from unauthorized access or modification.
Equivalence of redundant or distributed data	Data about an object or event in one data store is *semantically* equivalent to data about the same object or event in another data store.

Data Defect	Description
Concurrency of redundant of distributed data	The information float or lag time is acceptable between (a) when data are knowable (created or changed) in one data store to (b) when it is knowable in a redundant or distributed data store, and concurrent queries to each data store produce the same result.
Currency	The “age” of the data are correct for the knowledge workers’ purpose or purposes.

Appendix B Test Cases: Earned Value Management Data

This appendix presents the test cases we used to evaluate the effectiveness of the anomaly detection methods investigated as part of this research study. The arrows on each graph indicate values that were identified as possible data errors by and OSD subject matter expert.

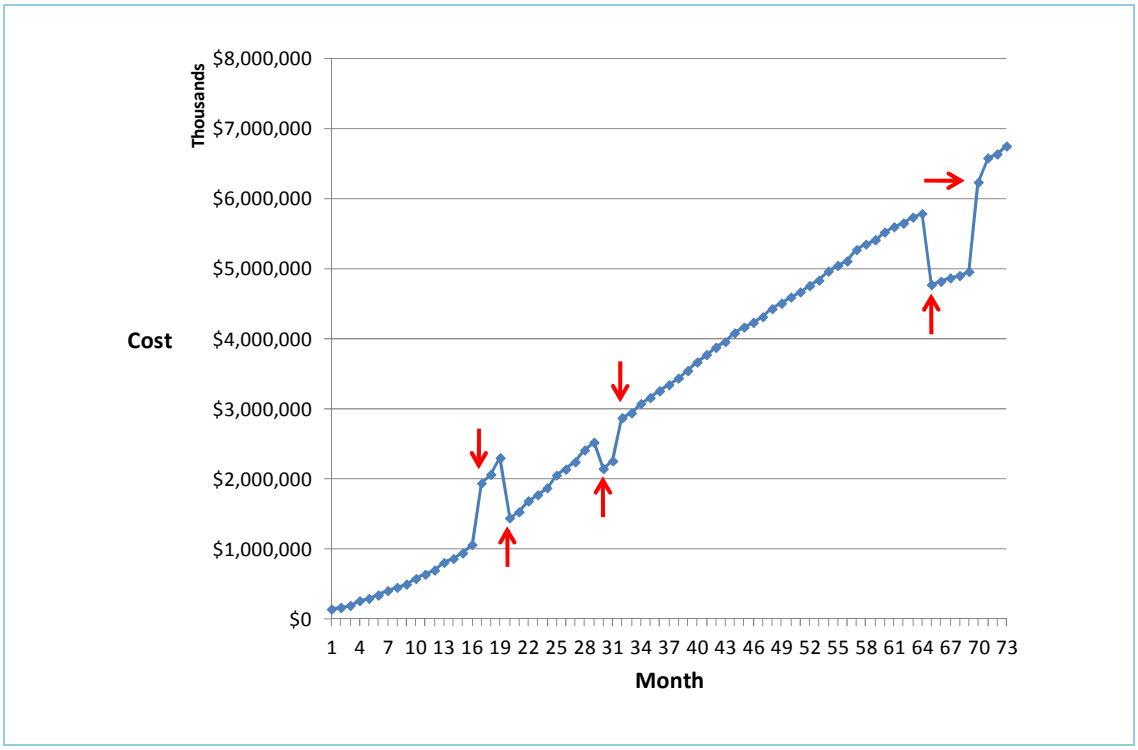

Figure 13: Time Series Plots of Case #1 BCWS Data

Table 9: Date and Error Values for Case #1 BCWS Data

Month ID	Possible Error Value
17	1,940,676,000
20	1,444,025,000
30	2,148,585,000
32	2,873,670,000
65	4,775,742,000
70	6,238,964,000

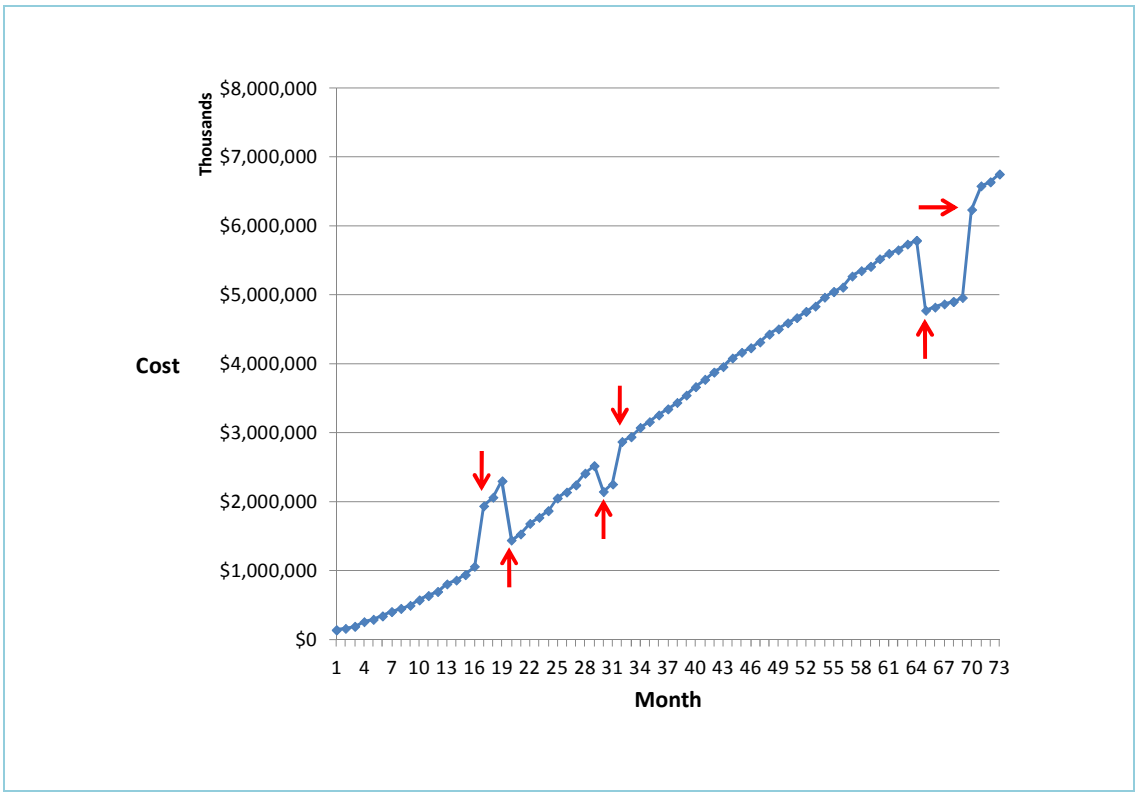

Figure 14: Time Series Plots of Case #1 BCWP Data

Table 10: Date and Error Values for Case #1 BCWP Data

Month ID	Possible Error Value
17	1,909,818,000
20	1,423,075,000
30	2,091,860,000
32	2,761,025,000
65	4,745,235,000
70	6,171,406,000

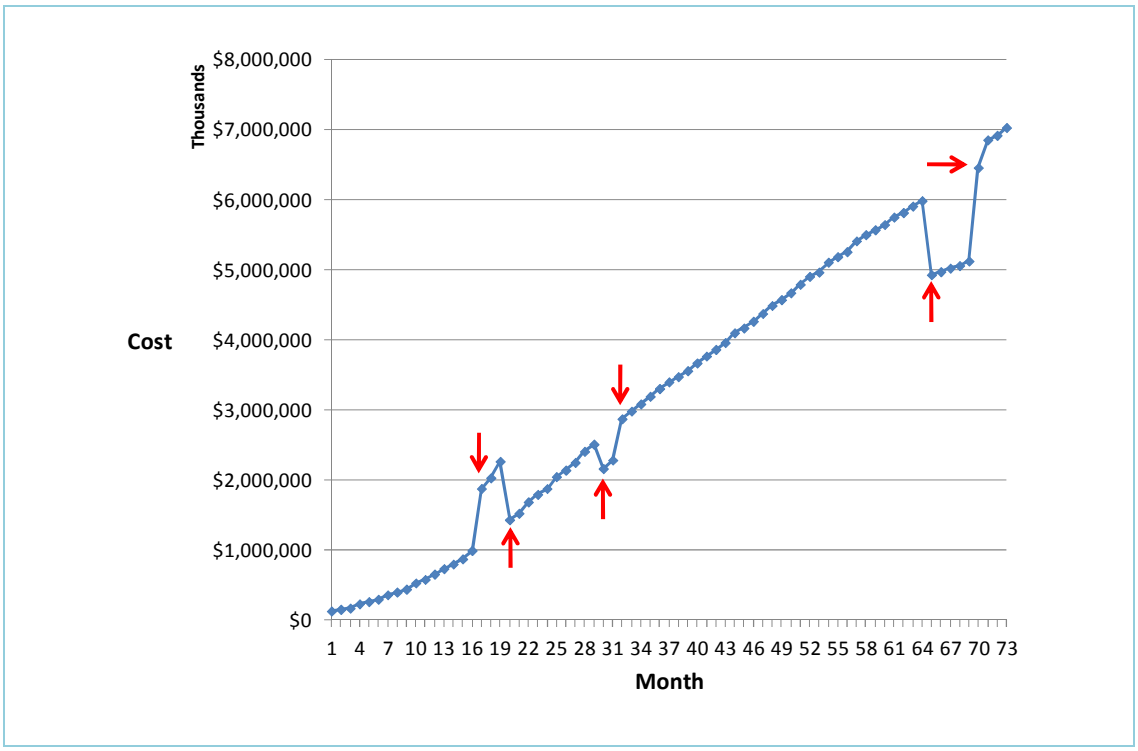

Figure 15: Time Series Plots of Case #1 ACWP Data

Table 11: Date and Error Values for Case #1 ACWP Data

Month ID	Possible Error Value
17	1,879,200,000
20	1,432,706,000
30	2,163,096,000
32	2,873,672,000
65	4,931,343,000
70	6,459,977,000

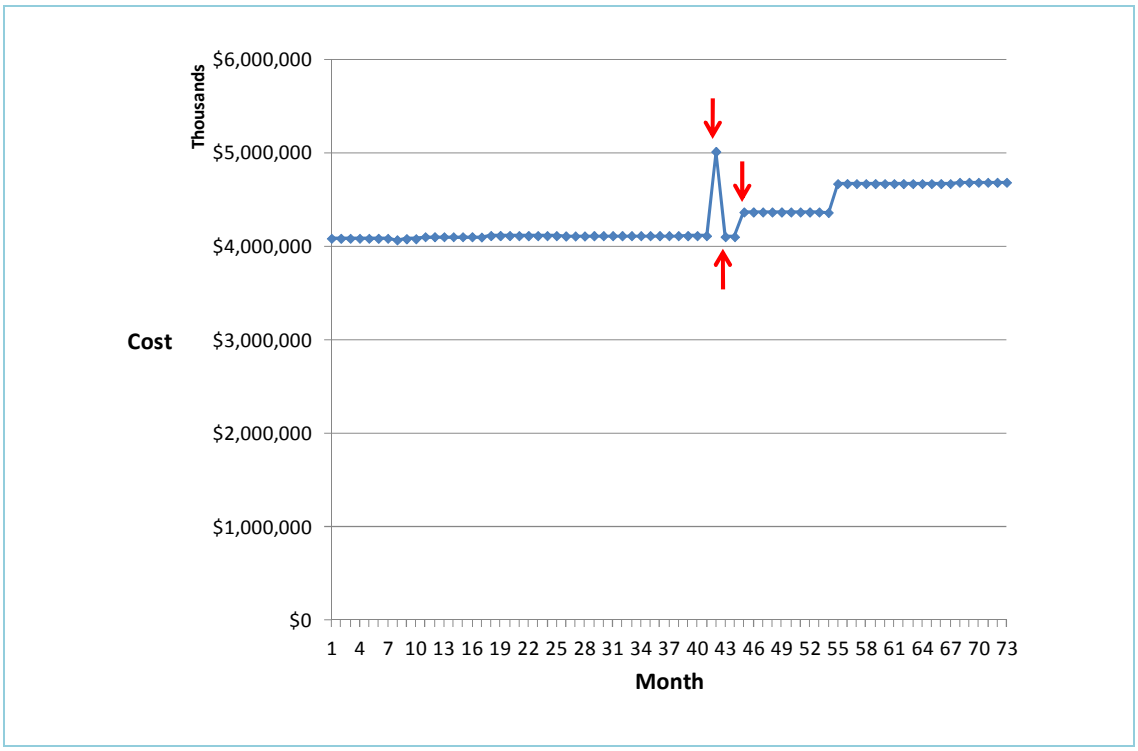

Figure 16: Time Series Plots of Case #1 NCC Data

Table 12: Date and Error Values for Case #1 NCC Data

Month ID	Poss. Error Value
42	5,017,373,000
43	4,105,475,900
45	4,369,780,800

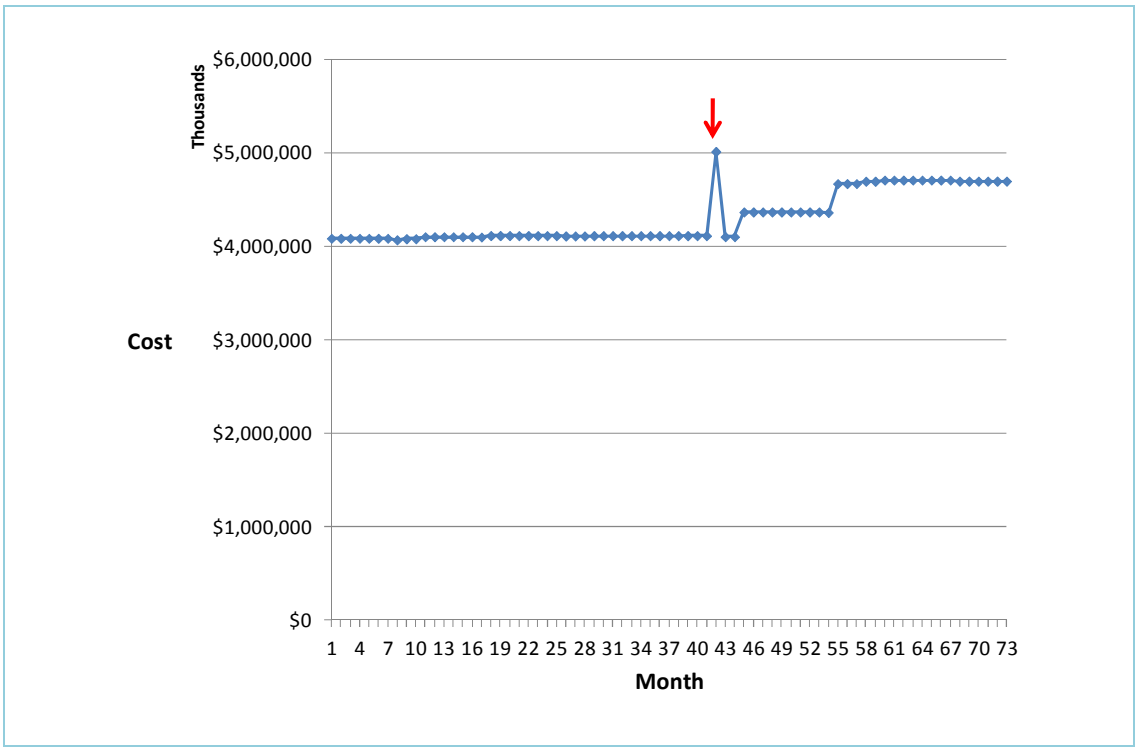

Figure 17: Time Series Plots of Case #1 CBB Data

Table 13: Date and Error Values for Case #1 CBB Data

Month ID	Poss. Error Value
42	5,017,373,000

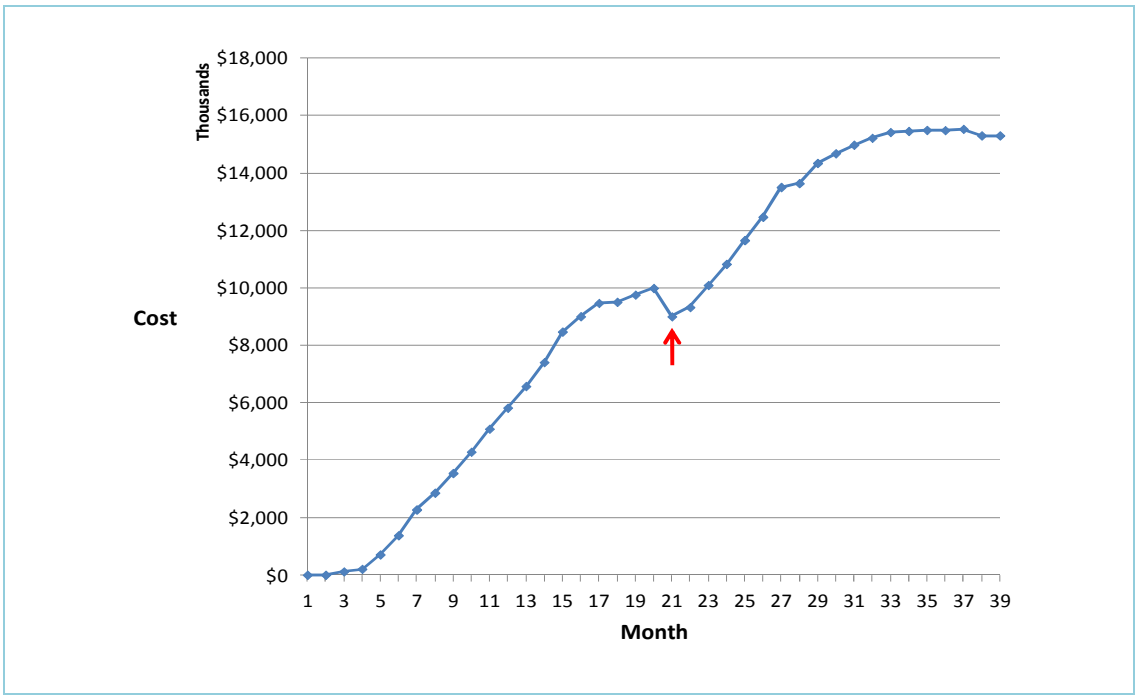

Figure 18: Time Series Plots of Case #2 BCWS Data

Table 14: Date and Error Values for Case #2 BCWS Data

Month ID	Poss. Error Value
21	8,999,568

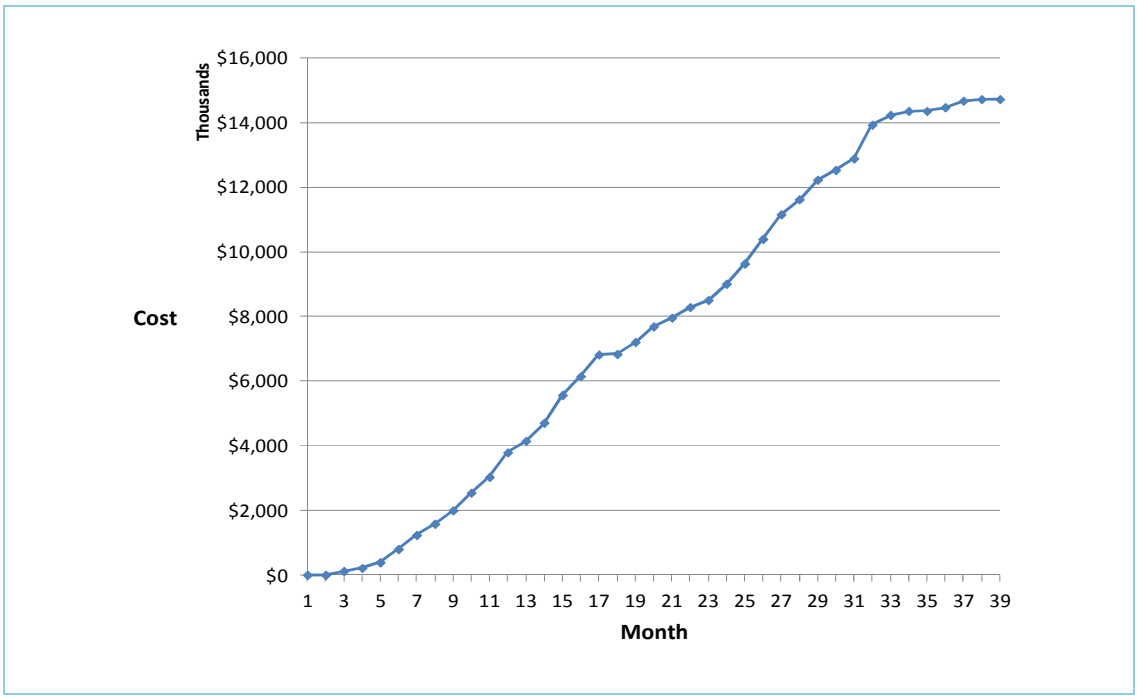

Figure 19: Time Series Plots of Case #2 BCWP Data

Table 15: Date and Error Values for Case #2 BCWP Data

Month ID	Poss. Error Value
n/a	n/a

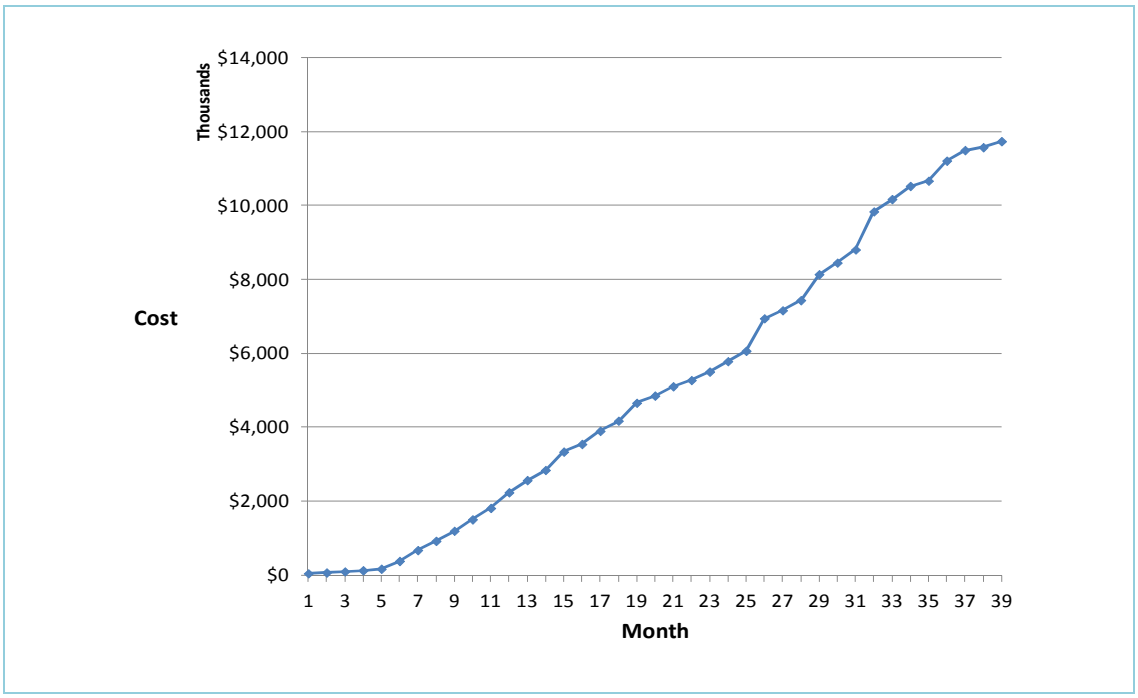

Figure 20: Time Series Plots of Case #2 ACWP Data

Table 16: Date and Error Values for Case #2 ACWP Data

Month ID	Poss. Error Value
n/a	n/a

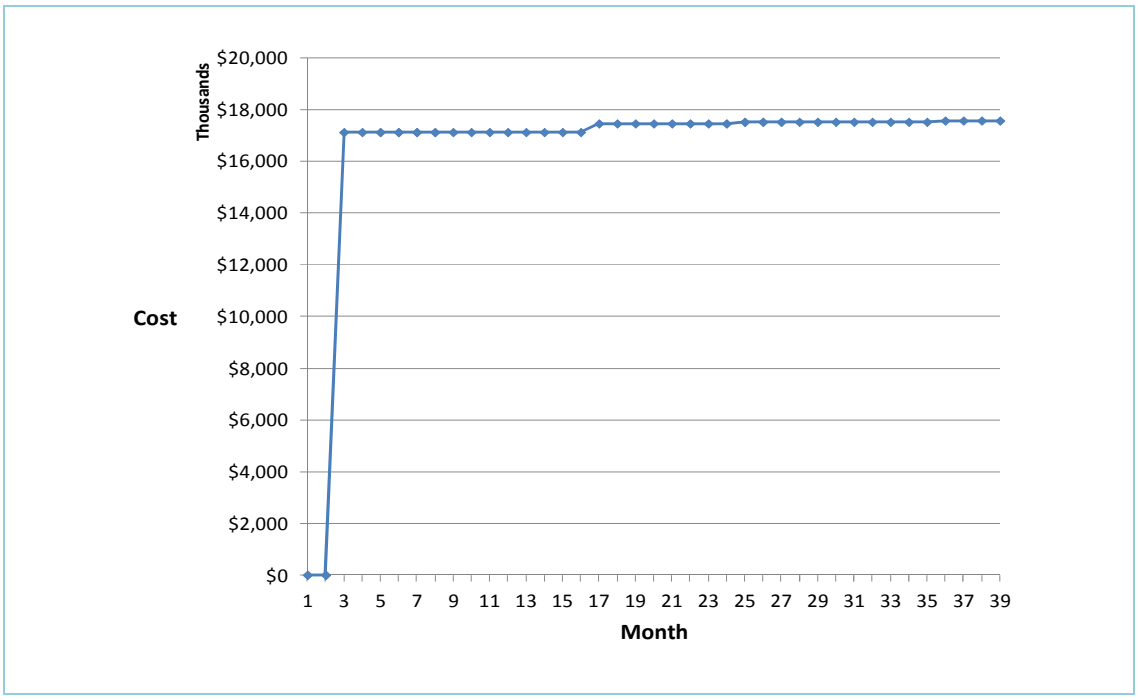

Figure 21: Time Series Plots of Case #2 NCC Data

Table 17: Date and Error Values for Case #2 NCC Data

Month ID	Poss. Error Value
n/a	n/a

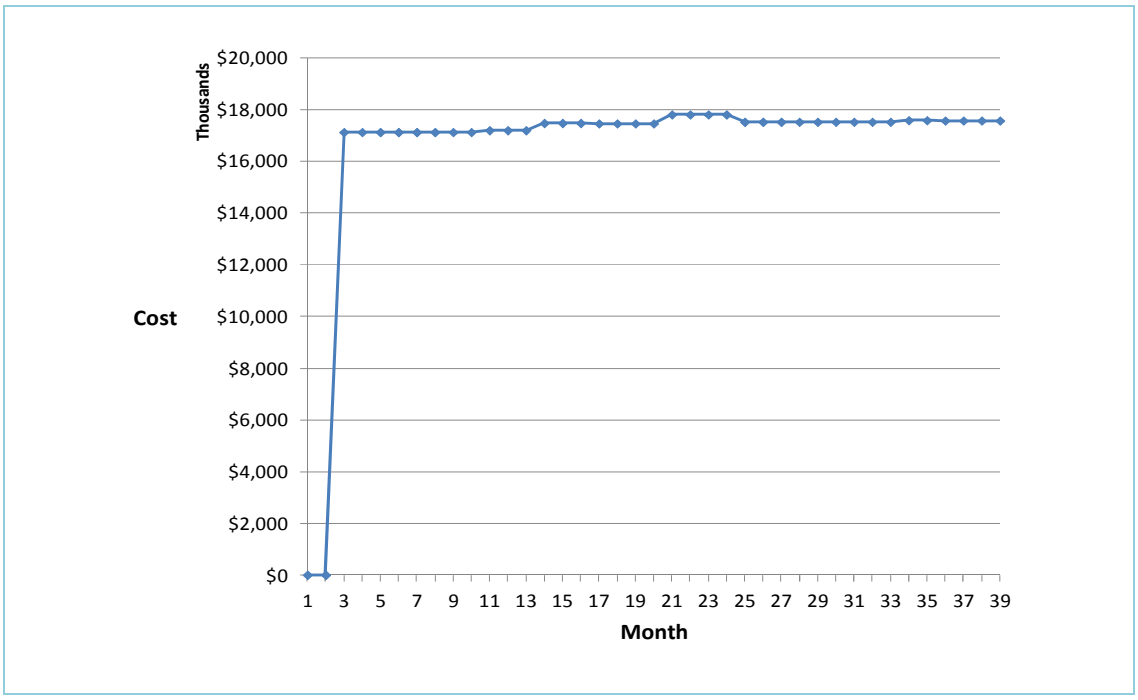

Figure 22: Time Series Plots of Case #2 CBB Data

Table 18: Date and Error Values for Case #2 CBB Data

Month ID	Poss. Error Value
n/a	n/a

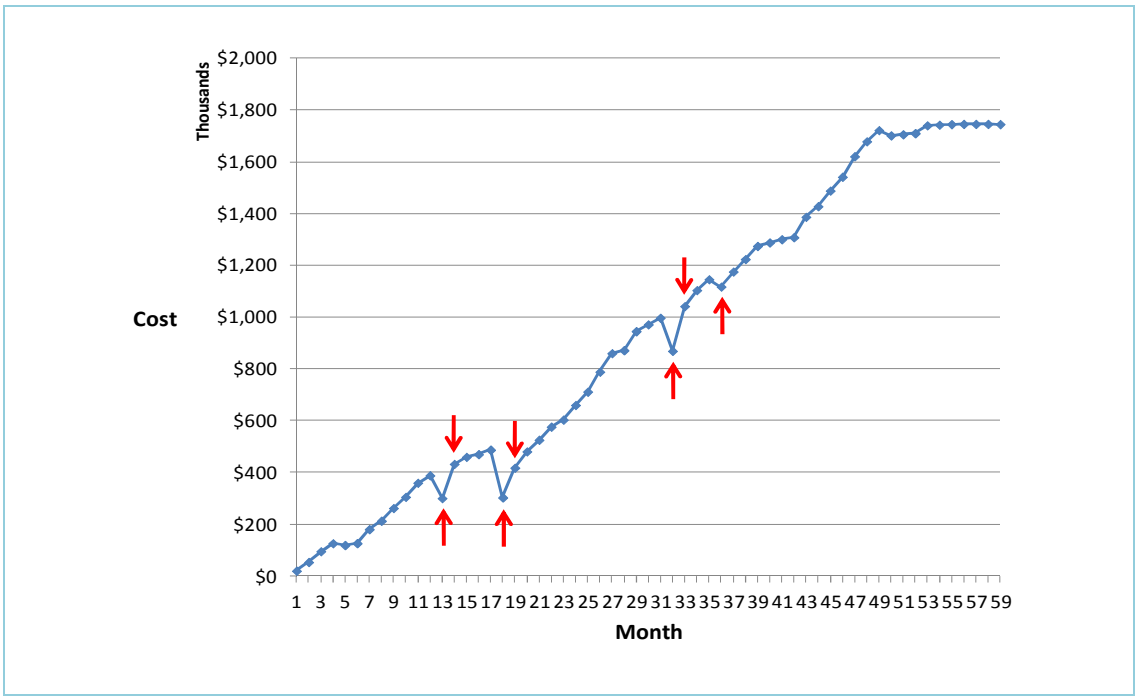

Figure 23: Time Series Plots of Case #3 BCWS Data

Table 19: Date and Error Values for Case #3 BCWS Data

Month ID	Poss. Error Value
13	299,778
14	432,306
18	302,882
19	417,285
32	869,116
33	1,041,546
36	1,117,395

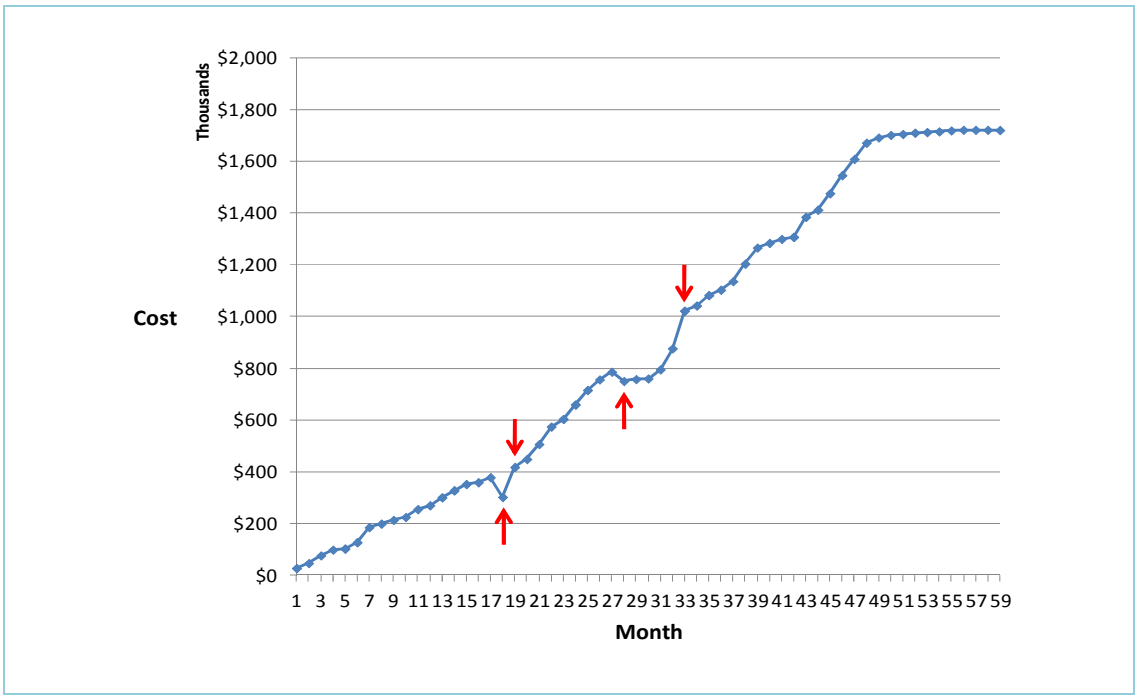

Figure 24: Time Series Plots of Case #3 BCWP Data

Table 20: Date and Error Values for Case #3 BCWP Data

Month ID	Poss. Error Value
18	301,159
19	417,285
28	749,861
33	1,022,003

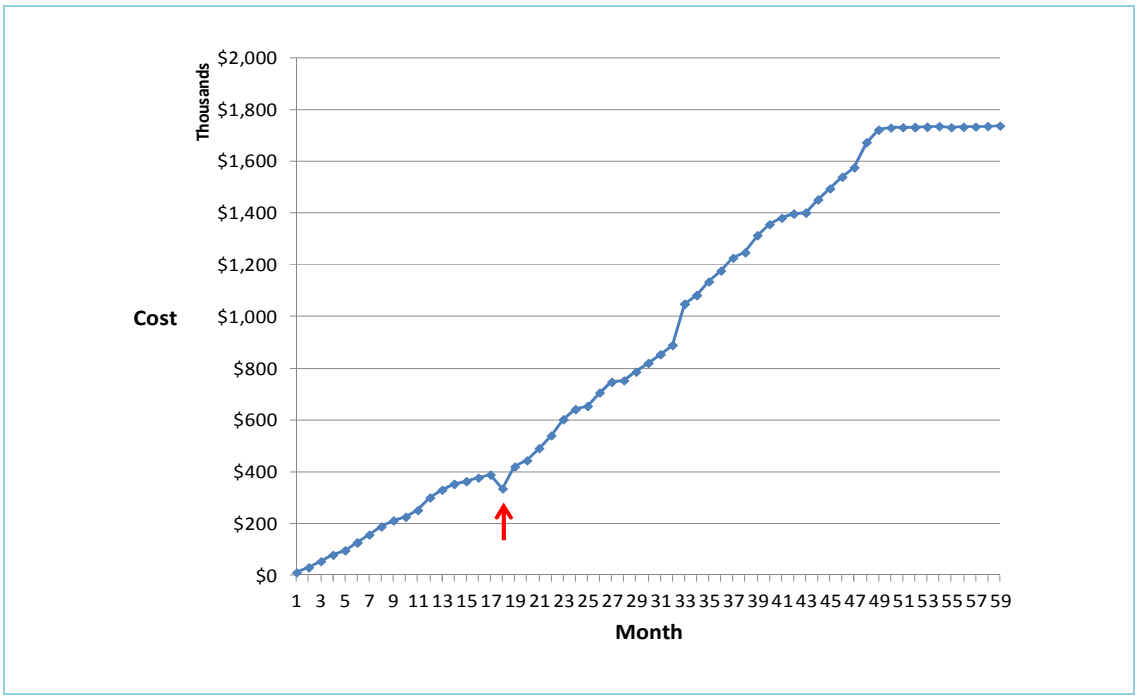

Figure 25: Time Series Plots of Case #3 ACWP Data

Table 21: Date and Error Values for Case #3 ACWP Data

Month ID	Poss. Error Value
18	333,775

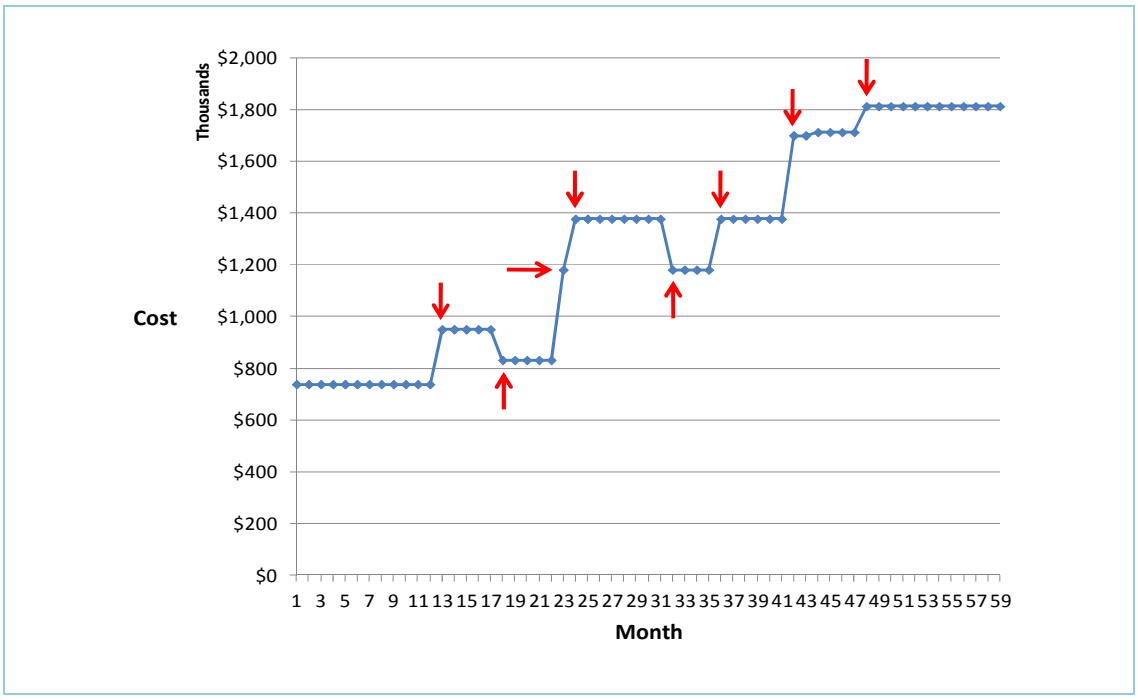

Figure 26: Time Series Plots of Case #3 NCC Data

Table 22: Date and Error Values for Case #3 NCC Data

Month ID	Poss. Error Value
13	950,311
18	831,202
23	1,181,208
24	1,377,971
32	1,181,208
36	1,378,253
42	1,700,921
48	1,813,922

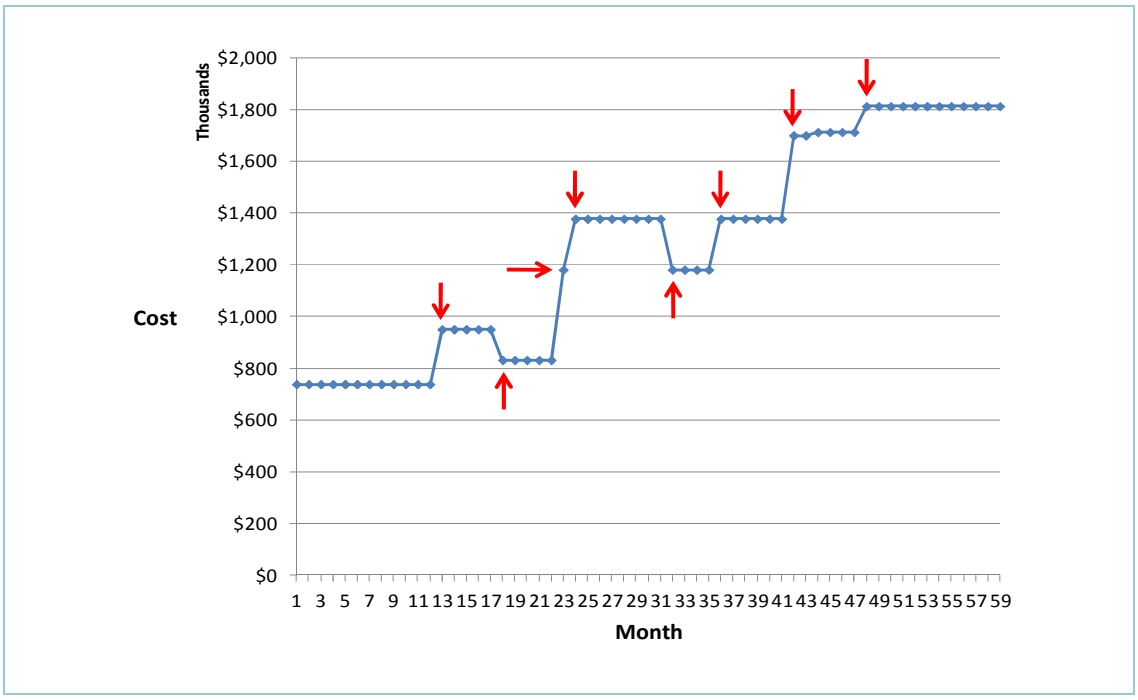

Figure 27: Time Series Plots of Case #3 CBB Data

Table 23: Date and Error Values for Case #3 CBB Data

Month ID	Poss. Error Value
13	950,311
18	831,202
23	1,181,208
24	1,377,971
32	1,181,208
36	1,378,253
42	1,700,921
48	1,813,922

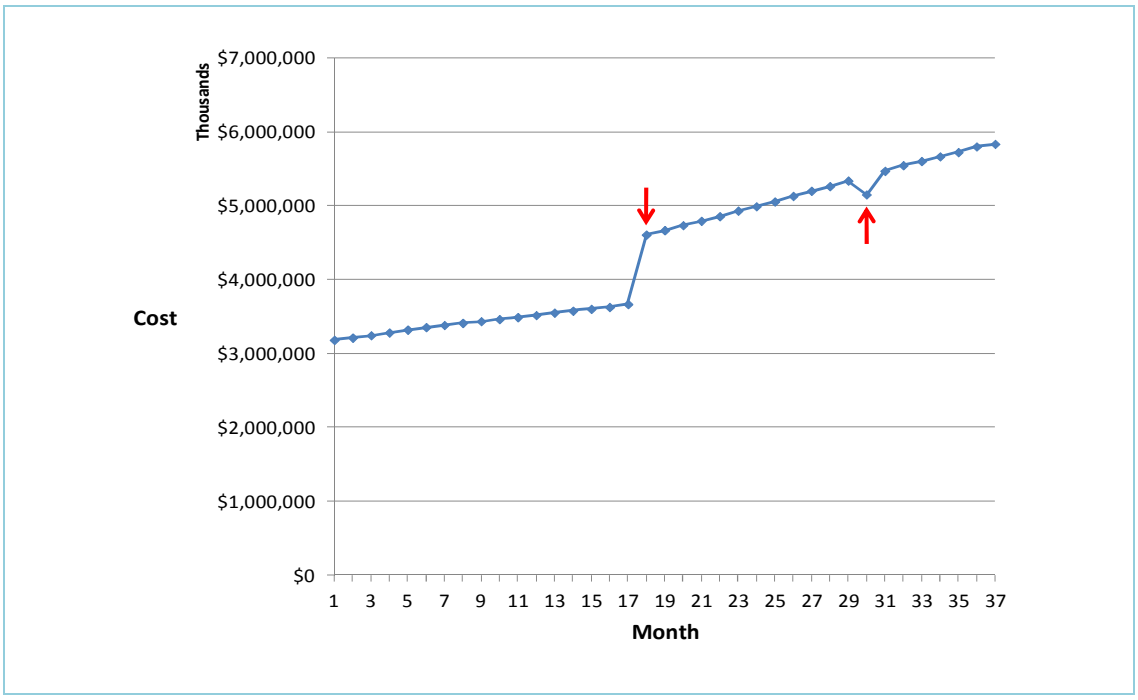

Figure 28: Time Series Plots of Case #4 BCWS Data

Table 24: Date and Error Values for Case #4 BCWS Data

Month ID	Poss. Error Value
18	4,595,154,254
30	5,145,641,276

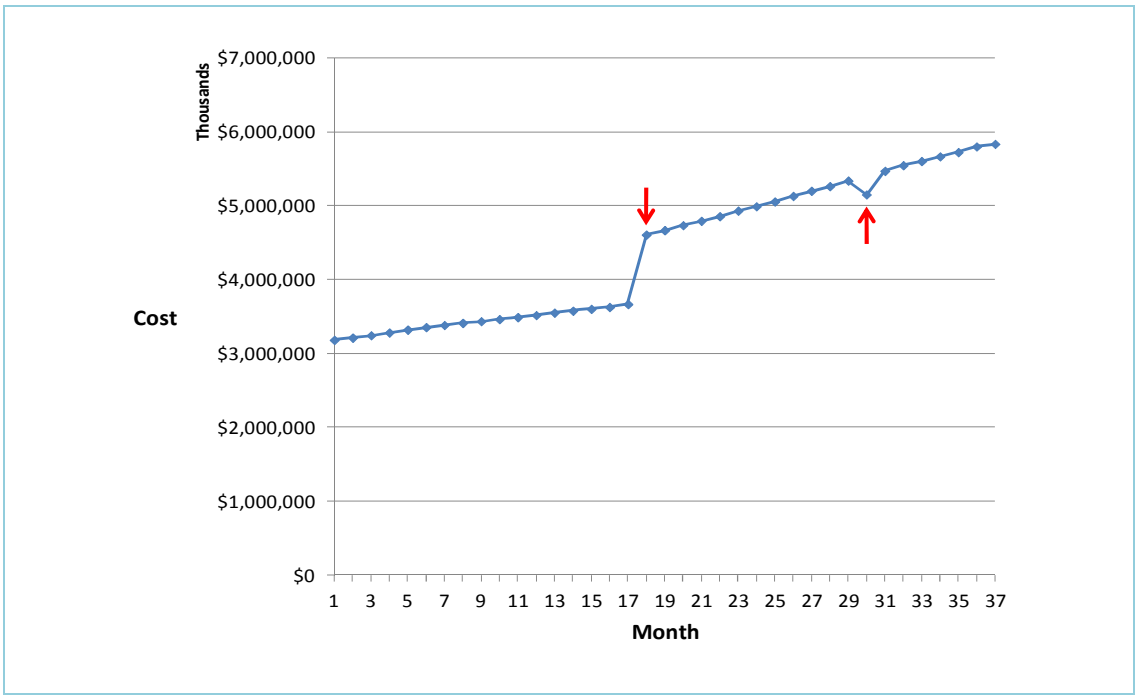

Figure 29: Time Series Plots of Case #4 BCWP Data

Table 25: Date and Error Values for Case #4 BCWP Data

Month ID	Poss. Error Value
18	4,608,395,985
30	**5,151,418,521**

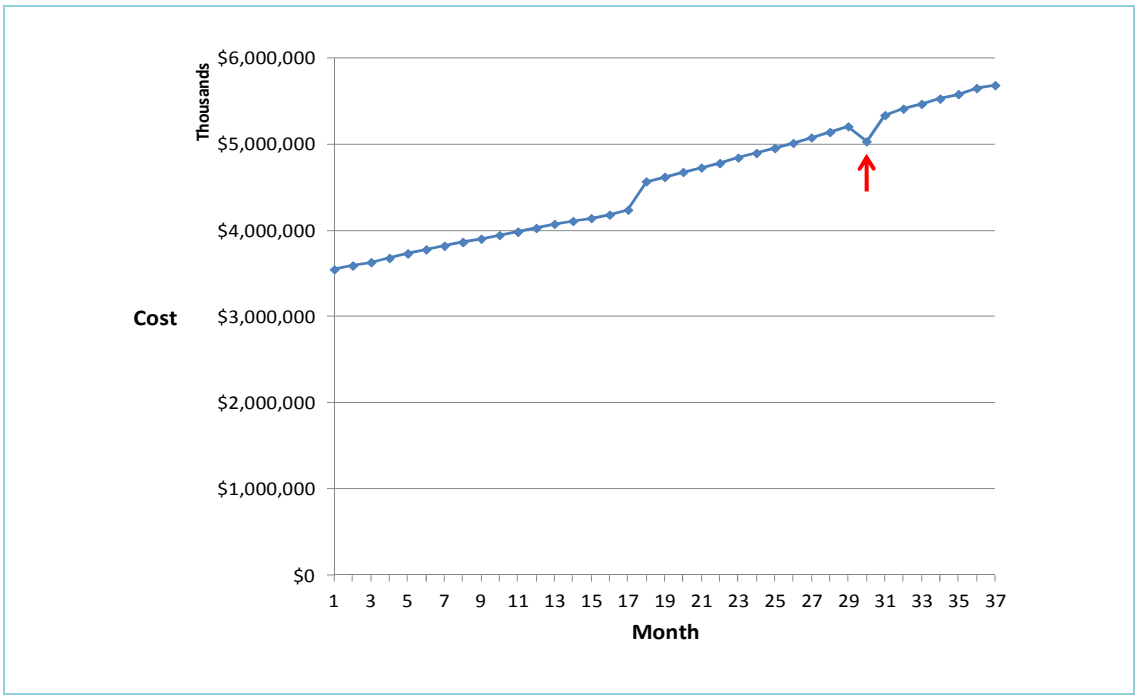

Figure 30: Time Series Plots of Case #4 ACWP Data

Table 26: Date and Error Values for Case #4 ACWP Data

Month ID	Poss. Error Value
29	5,033,355,639

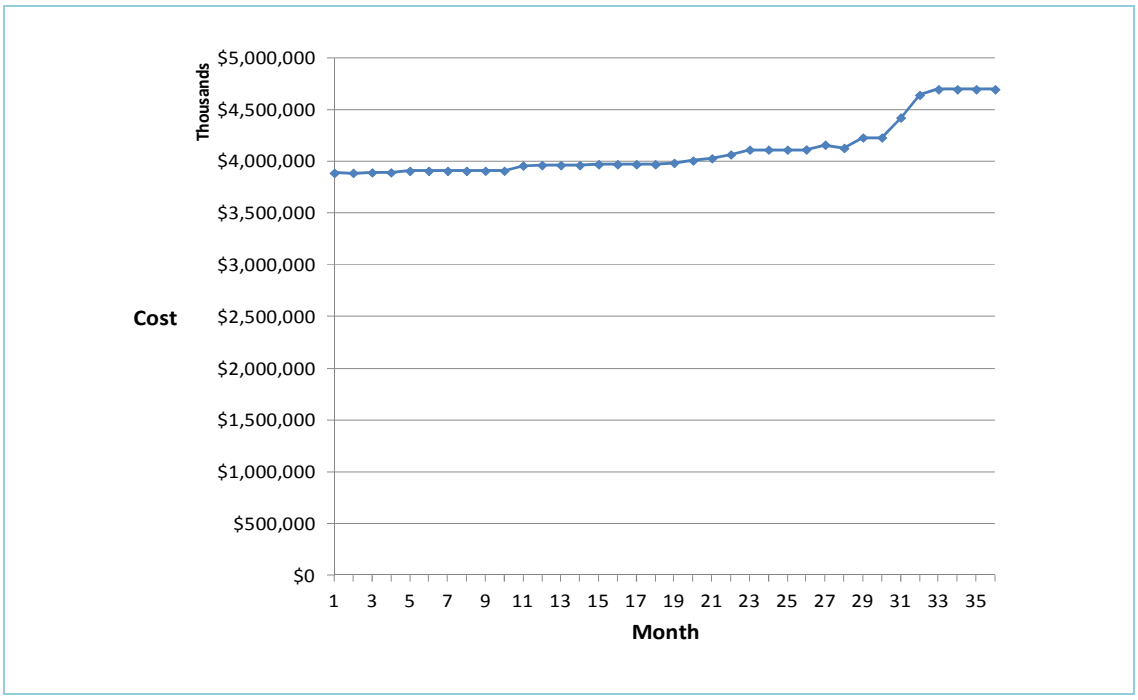

Figure 31: Time Series Plots of Case #4 NCC Data

Table 27: Date and Error Values for Case #4 NCC Data

Month ID	Poss. Error Value
n/a	n/a

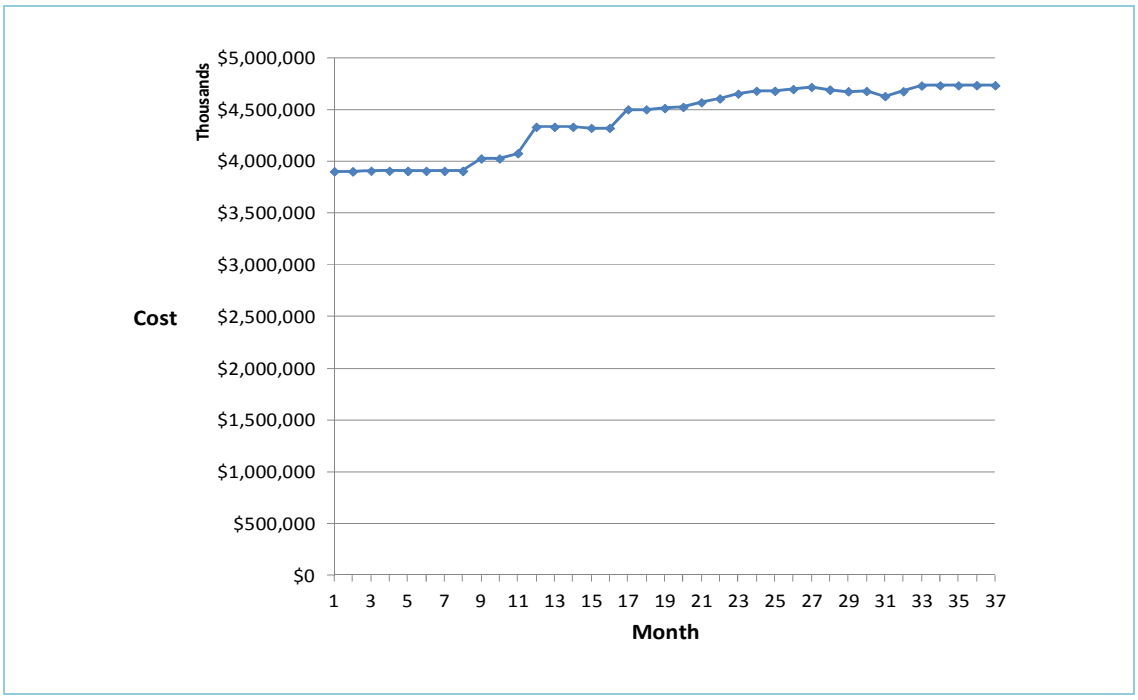

Figure 32: Time Series Plots of Case #4 CBB Data

Table 28: Date and Error Values for Case #4 CBB Data

Month ID	Poss. Error Value
n/a	n/a

Appendix C Detailed Tabular Results

This appendix presents a detailed summary of results by EVM Variable for each of the four test cases illustrated in Appendix B.

Table 29: Anomaly Detection Method Performance for EVM Variable BCWS

		I-CC	Grubbs	3-Sigma	Dixon n=3	Dixon n=8	Dixon n=14	Rosner	ARIMA	Box plot
Case #1	Total # records	73	73	73	73	73	73	73	73	73
	# Total Defects	6	6	6	6	6	6	6	6	6
	# True Positives	6	6	4	4	6	6	6	6	6
	# False Negatives	0	0	2	2	0	0	0	0	0
	# True Negatives	67	65	67	60	65	66	65	64	65
	# False Positives	0	2	0	7	2	1	2	3	2
	Detection rate	100.0%	100.0%	66.7%	66.7%	100.0%	100.0%	100.0%	100.0%	100.0%
	False positive rate	0.0%	3.0%	0.0%	10.4%	3.0%	1.5%	3.0%	4.5%	3.0%
Case #2	Total # records	39	39	39	39	39	39	39	39	39
	# Total Defects	1	1	1	1	1	1	1	1	1
	# True Positives	1	1	1	1	1	1	1	1	1
	# False Negatives	0	0	0	0	0	0	0	0	0
	# True Negatives	38	38	38	36	36	36	38	37	38
	# False Positives	0	0	0	2	2	2	0	1	0
	Detection rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	False positive rate	0.0%	0.0%	0.0%	5.3%	5.3%	5.3%	0.0%	2.6%	0.0%
Case #3	Total # records	59	59	59	59	59	59	59	59	59
	# Total Defects	7	7	7	7	7	7	7	7	7
	# True Positives	3	4	0	2	6	5	4	3	5
	# False Negatives	4	3	7	5	1	2	3	4	2
	# True Negatives	52	52	52	45	50	52	52	52	52
	# False Positives	0	0	0	7	2	0	0	0	0
	Detection rate	42.9%	57.1%	0.0%	28.6%	85.7%	71.4%	57.1%	42.9%	71.4%
	False positive rate	0.0%	0.0%	0.0%	13.5%	3.8%	0.0%	0.0%	0.0%	0.0%

Table continues on next page

Table 29, continued

		I-CC	Grubbs	3-Sigma	Dixon n=3	Dixon n=8	Dixon n=14	Rosner	ARIMA	Box plot
Case #4	Total # records	37	37	37	37	37	37	37	37	37
	# Total Defects	2	2	2	2	2	2	2	2	2
	# True Positives	2	2	0	2	1	2	2	2	2
	# False Negatives	0	0	2	0	1	0	0	0	0
	# True Negatives	34	34	35	33	33	33	34	34	34
	# False Positives	1	1	0	2	2	2	1	1	1
	Detection rate	100.0%	100.0%	0	100.0%	50.0%	100.0%	100.0%	100.0%	100.0%
	False positive rate	2.9%	2.9%	0	5.7%	5.7%	5.7%	2.9%	2.9%	2.9%
Totals	Total # records	208	208	208	208	208	208	208	208	208
	# Total Defects	16	16	16	16	16	16	16	16	16
	# True Positives	12	13	5	9	14	14	13	12	14
	# False Negatives	4	3	11	7	2	2	3	4	2
	# True Negatives	191	189	192	174	184	187	189	187	189
	# False Positives	1	3	0	18	8	5	3	5	3
	Detection rate	75.0%	81.3%	31.3%	56.3%	87.5%	87.5%	81.3%	75.0%	87.5%
	False positive rate	0.5%	1.6%	0.0%	9.4%	4.2%	2.6%	1.6%	2.6%	1.6%

Table 30: *Anomaly Detection Method Performance for EVM Variable BCWP*

		I-CC	Grubbs	3-Sigma	Dixon n=3	Dixon n=8	Dixon n=14	Rosner	ARIMA	Box plot
Case #1	Total # records	73	73	73	73	73	73	73	73	73
	# Total Defects	6	6	6	6	6	6	6	6	6
	# True Positives	5	6	4	2	2	1	6	6	6
	# False Negatives	1	0	2	4	4	5	0	0	0
	# True Negatives	67	65	67	66	67	66	65	65	65
	# False Positives	0	2	0	1	0	1	2	2	2
	Detection rate	83.3%	100.0%	66.7%	33.3%	33.3%	16.7%	100.0%	100.0%	100.0%
	False positive rate	0.0%	3.0%	0.0%	1.5%	0.0%	1.5%	3.0%	3.0%	3.0%
Case #2	Total # records	39	39	39	39	39	39	39	39	39
	# Total Defects	0	0	0	0	0	0	0	0	0
	# True Positives	0	0	0	0	0	0	0	0	0
	# False Negatives	0	0	0	0	0	0	0	0	0
	# True Negatives	38	39	39	34	38	39	39	37	39
	# False Positives	1	0	0	5	1	0	0	2	0
	Detection rate	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	False positive rate	2.6%	0.0%	0.0%	12.8%	2.6%	0.0%	0.0%	5.1%	0.0%
Case #3	Total # records	59	59	59	59	59	59	59	59	59
	# Total Defects	4	4	4	4	4	4	4	4	4
	# True Positives	3	3	3	0	0	0	2	1	1
	# False Negatives	1	1	1	4	4	4	2	3	3
	# True Negatives	55	55	55	52	53	54	55	55	55
	# False Positives	0	0	0	3	2	1	0	0	0
	Detection rate	75.0%	75.0%	75.0%	0.0%	0.0%	0.0%	50.0%	25.0%	25.0%
	False positive rate	0.0%	0.0%	0.0%	5.5%	3.6%	1.8%	0.0%	0.0%	0.0%

Table continues on next page

Table 30, continued

		I-CC	Grubbs	3-Sigma	Dixon n=3	Dixon n=8	Dixon n=14	Rosner	ARIMA	Box plot
Case #4	Total # records	37	37	37	37	37	37	37	37	37
	# Total Defects	2	2	2	2	2	2	2	2	2
	# True Positives	2	2	0	1	1	1	2	2	2
	# False Negatives	0	0	2	1	1	1	0	0	0
	# True Negatives	34	34	35	35	34	32	34	34	34
	# False Positives	1	1	0	0	1	3	1	1	1
	Detection rate	100.0%	100.0%	0.0%	50.0%	50.0%	50.0%	100.0%	100.0%	100.0%
	False positive rate	2.9%	2.9%	0.0%	0.0%	2.9%	8.6%	2.9%	2.9%	2.9%
Totals	Total # records	208	208	208	208	208	208	208	208	208
	# Total Defects	12	12	12	12	12	12	12	12	12
	# True Positives	10	11	7	3	3	2	10	9	9
	# False Negatives	2	1	5	9	9	10	2	3	3
	# True Negatives	194	193	196	187	192	191	193	191	193
	# False Positives	2	3	0	9	4	5	3	5	3
	Detection rate	83.3%	91.7%	58.3%	25.0%	25.0%	16.7%	83.3%	75.0%	75.0%
	False positive rate	1.0%	1.5%	0.0%	4.6%	2.0%	2.6%	1.5%	2.6%	1.5%

Table 31: Anomaly Detection Method Performance for EVM Variable ACWP

		I-CC	Grubbs	3-Sigma	Dixon n=3	Dixon n=8	Dixon n=14	Rosner	ARIMA	Box plot
Case #1	Total # records	73	73	73	73	73	73	73	73	73
	# Total Defects	7	7	7	7	7	7	7	7	7
	# True Positives	5	7	4	1	2	1	7	7	7
	# False Negatives	2	0	3	6	5	6	0	0	0
	# True Negatives	66	65	66	65	66	65	65	63	65
	# False Positives	0	1	0	1	0	1	1	3	1
	Detection rate	71.4%	100.0%	57.1%	14.3%	28.6%	14.3%	100.0%	100.0%	100.0%
	False positive rate	0.0%	1.5%	0.0%	1.5%	0.0%	1.5%	1.5%	4.5%	1.5%
Case #2	Total # records	39	39	39	39	39	39	39	39	39
	# Total Defects	0	0	0	0	0	0	0	0	0
	# True Positives	0	0	0	0	0	0	0	0	0
	# False Negatives	0	0	0	0	0	0	0	0	0
	# True Negatives	38	37	39	39	39	39	36	36	36
	# False Positives	1	2	0	0	0	0	3	3	3
	Detection rate	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	False positive rate	2.6%	5.1%	0.0%	0.0%	0.0%	0.0%	7.7%	7.7%	7.7%
Case #3	Total # records	59	59	59	59	59	59	59	59	59
	# Total Defects	1	1	1	1	1	1	1	1	1
	# True Positives	1	1	1	0	0	0	1	1	1
	# False Negatives	0	0	0	1	1	1	0	0	0
	# True Negatives	56	57	57	58	55	57	57	56	56
	# False Positives	2	1	1	0	3	1	1	2	2
	Detection rate	100.0%	100.0%	100.0%	0.0%	0.0%	0.0%	100.0%	100.0%	100.0%
	False positive rate	3.4%	1.7%	1.7%	0.0%	5.2%	1.7%	1.7%	3.4%	3.4%

Table continues on next page

Table 31, continued

		I-CC	Grubbs	3-Sigma	Dixon n=3	Dixon n=8	Dixon n=14	Rosner	ARIMA	Box plot
Case #4	Total # records	37	37	37	37	37	37	37	37	37
	# Total Defects	1	1	1	1	1	1	1	1	1
	# True Positives	0	0	0	0	0	0	0	0	0
	# False Negatives	1	1	1	1	1	1	1	1	1
	# True Negatives	33	33	36	36	35	36	33	33	33
	# False Positives	3	3	0	0	1	0	3	3	3
	Detection rate	0.0%	0.0%	0	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	False positive rate	8.3%	8.3%	0	0.0%	2.8%	0.0%	8.3%	8.3%	8.3%
Totals	Total # records	208	208	208	208	208	208	208	208	208
	# Total Defects	9	9	9	9	9	9	9	9	9
	# True Positives	6	8	5	1	2	1	8	8	8
	# False Negatives	3	1	4	8	7	8	1	1	1
	# True Negatives	193	192	198	198	195	197	191	188	190
	# False Positives	6	7	1	1	4	2	8	11	9
	Detection rate	66.7%	88.9%	55.6%	11.1%	22.2%	11.1%	88.9%	88.9%	88.9%
	False positive rate	3.0%	3.5%	0.5%	0.5%	2.0%	1.0%	4.0%	5.5%	4.5%

Table 32: Anomaly Detection Method Performance for EVM Variable NCC

		mR-CC	Moving Range	ARIMA	Box plot
Case #1	Total # records	73	73	73	73
	# Total Defects	3	3	3	3
	# True Positives	0	3	3	3
	# False Negatives	3	0	0	0
	# True Negatives	70	69	69	66
	# False Positives	0	1	1	4
	Detection rate	0.0%	100.0%	100.0%	100.0%
	False positive rate	0.0%	1.4%	1.4%	5.7%
Case #2	Total # records	39	39	39	39
	# Total Defects	0	0	0	0
	# True Positives	0	0	0	0
	# False Negatives	0	0	0	0
	# True Negatives	39	36	36	36
	# False Positives	0	3	3	3
	Detection rate	n/a	n/a	n/a	n/a
	False positive rate	0.0%	7.7%	7.7%	7.7%
Case #3	Total # records	59	59	59	59
	# Total Defects	8	8	8	8
	# True Positives	0	8	8	8
	# False Negatives	8	0	0	0
	# True Negatives	51	51	51	48
	# False Positives	0	0	0	3
	Detection rate	0.0%	100.0%	100.0%	100.0%
	False positive rate	0.0%	0.0%	0.0%	5.9%

Table continues on next page

Table 32, continued

		mR-CC	Moving Range	ARIMA	Box plot
Case #4	Total # records	37	37	37	37
	# Total Defects	0	0	0	0
	# True Positives	0	0	0	0
	# False Negatives	0	0	0	0
	# True Negatives	37	34	29	32
	# False Positives	0	3	8	5
	Detection rate	n/a	n/a	n/a	n/a
	False positive rate	0.0%	8.1%	21.6%	13.5%
Totals	Total # records	208	208	208	208
	# Total Defects	11	11	11	11
	# True Positives	0	11	11	11
	# False Negatives	11	0	0	0
	# True Negatives	197	193	185	182
	# False Positives	0	4	12	15
	Detection rate	0.0%	100.0%	100.0%	100.0%
	False positive rate	0.0%	2.0%	6.1%	7.6%

Table 33: Anomaly Detection Method Performance for EVM Variable CBB

		mR-CC	Moving Range	ARIMA	Box plot
Case #1	Total # records	73	73	73	73
	# Total Defects	3	3	3	3
	# True Positives	3	3	3	3
	# False Negatives	0	0	0	0
	# True Negatives	69	65	65	53
	# False Positives	1	5	5	17
	Detection rate	100.0%	100.0%	100.0%	100.0%
	False positive rate	1.4%	7.1%	7.1%	24.3%
Case #2	Total # records	39	39	39	39
	# Total Defects	0	0	0	0
	# True Positives	0	0	0	0
	# False Negatives	0	0	0	0
	# True Negatives	39	34	32	32
	# False Positives	0	5	7	7
	Detection rate	n/a	n/a	n/a	n/a
	False positive rate	0.0%	12.8%	17.9%	17.9%
Case #3	Total # records	59	59	59	59
	# Total Defects	8	8	8	8
	# True Positives	8	8	8	8
	# False Negatives	0	0	0	0
	# True Negatives	51	51	51	51
	# False Positives	0	0	0	0
	Detection rate	100.0%	100.0%	100.0%	100.0%
	False positive rate	0.0%	0.0%	0.0%	0.0%

Table continues on next page

Table 33, continued

		mR-CC	Moving Range	ARIMA	Box plot
Case #4	Total # records	37	37	37	37
	# Total Defects	0	0	0	0
	# True Positives	0	0	0	0
	# False Negatives	0	0	0	0
	# True Negatives	34	34	34	34
	# False Positives	3	3	3	3
	Detection rate	n/a	n/a	n/a	n/a
	False positive rate	8.1%	8.1%	8.1%	8.1%
Totals	Total # records	208	208	208	208
	# Total Defects	11	11	11	11
	# True Positives	11	11	11	11
	# False Negatives	0	0	0	0
	# True Negatives	193	184	182	170
	# False Positives	4	13	15	27
	Detection rate	100.0%	100.0%	100.0%	100.0%
	False positive rate	2.0%	6.6%	7.6%	13.7%

Appendix D Analysis Results – Significance Tests

This section presents the significance tests that are referred to in the Results and Discussion Section of the document.

Table 34 presents the anomaly detection results for the combined variables BCWS, BCWP, and ACWP for all four tests cases.

Table 34: Anomaly Detection Effectiveness Results for BCWS, BCWP, and ACWP (n = 208)

	True Positives	False Negatives	True Negatives	False Positives	Detection Rate	False Positive Rate
I-CC	28	9	578	9	75.7%	1.5%
Grubbs	32	5	574	13	86.5%	2.2%
3-Sigma Meth.	17	20	586	1	45.9%	0.2%
Dixon (n=3)	13	24	559	28	35.1%	4.8%
Dixon (n=8)	19	18	571	16	51.4%	2.7%
Dixon (n=14)	17	20	575	12	45.9%	2.0%
Rosner	31	6	573	14	83.8%	2.4%
ARIMA	31	6	573	14	83.8%	2.4%
Box plot	31	6	572	15	83.8%	2.6%

Table 35: Chi-Square Goodness-of-Fit Test for Observed Counts in True Positives

	True Positives	Test Proportion	Expected	Contrib. to Chi-Square	N	DF	Chi-Squ.	p Value
I-CC	28	0.11	24.33	0.55251	219	8	19.32	0.013
Grubbs	32	0.11	24.33	2.41553				
3-Sigma Meth.	17	0.11	24.33	2.21005				
Dixon (n=3)	13	0.11	24.33	5.27854				
Dixon (n=8)	19	0.11	24.33	1.16895				
Dixon (n=14)	17	0.11	24.33	2.21005				
Rosner	31	0.11	24.33	1.82648				
ARIMA	31	0.11	24.33	1.82648				
Box plot	31	0.11	24.33	1.82648				

A p value of 0.013 demonstrates that there is a significant difference in the effectiveness of the listed techniques in Table 35.

Table 36: Test of Two Proportions (Dixon (n=8) and I-CC)

Technique	True Positives	Sample Percent	Estimate for Difference	95% Upper Bound for Diff.	Z	p Value (Fisher's Exact Text)
Dixon (n=8)	19	0.514	-0.243	-0.065	-2.25	0.026
I-CC	28	0.757				

The test of two proportions demonstrates a significant difference between the Dixon (n=8) technique and the I-CC technique for detecting anomalies (p value = 0.025). This implies that I-CC, Grubbs, Rosner, Box plot, and ARIMA are all significantly different in effectiveness from the techniques Dixon (n=8), 3-sigma Meth., Dixon (n=3), and Dixon (n=14) techniques.

Table 37: Chi-Square Goodness-of-Fit Test for Observed Counts in False Positives

	True Positives	Test Proportion	Expected	Contrib. to Chi-Square	N	DF	Chi-Squ.	p Value
I-CC	9	0.11	13.56	1.5310	122	8	29.377	0.000
Grubbs	13	0.11	13.56	0.0228				
3-Sigma Meth.	1	0.11	13.56	11.6293				
Dixon (n=3)	28	0.11	13.56	15.3916				
Dixon (n=8)	16	0.11	13.56	0.4408				
Dixon (n=14)	12	0.11	13.56	0.1785				
Rosner	14	0.11	13.56	0.0146				
ARIMA	14	0.11	13.56	0.0146				
Box plot	15	0.11	13.56	0.1539				

A p-value of 0.000 (listed in Table 37) demonstrates a significant difference in false positives generated by the techniques.

Table 38: Test of Two Proportions (3-Sigma and I-CC)

Technique	False Positives	Sample Percent	Estimate for Difference	95% Upper Bound for Diff.	Z	p Value (Fisher's Exact Text)
3-Sigma	1	0.002	-0.014	-0.005	-2.55	0.005
I-CC	9	0.015				

The test summarized in Table 38 demonstrates that the 3-sigma method generates fewer false positives than all other techniques, and this difference is statistically significant. Other test of proportions for false positives did not show significant differences.

Table 39 presents the anomaly detection results for the combined variables NCC and CBB for all four tests cases.

Table 39: Anomaly Detection Effectiveness Results for NCC and CBB (n = 208)

	True Positives	False Negatives	True Negatives	False Positives	Detection Rate	False Positive Rate
Moving Range	22	0	369	25	100%	6.3%
mR-CC	22	0	382	12	100%	3.0%
ARIMA	22	0	364	30	100%	7.6%
Box plot	22	0	343	51	100%	12.9%

Table 40: Chi-Square Goodness-of-Fit Test for Observed Counts in False Positives (NCC and CBB)

	True Positives	Test Proportion	Expected	Contrib. to Chi-Square	N	DF	Chi-Squ.	p Value
Moving Range	25	0.25	29.5	0.6864	118	3	26.746	0.000
mR-CC	12	0.25	29.5	10.3814				
ARIMA	30	0.25	29.5	0.0085				
Box plot	51	0.25	29.5	15.6695				

A p value of 0.000 demonstrates a significant difference in the generation of false positives by the techniques listed in Table 40.

Table 41: Test of Two Proportions (mR CC and Moving Range)

Technique	True Positives	Sample Percent	Estimate for Difference	95% Upper Bound for Diff.	Z	p Value (Fisher's Exact Text)
mR CC	30	0.076	-0.053	-0.017	-2.47	0.009
Moving Range	51	0.129				

The test of two proportions demonstrates a significant difference between the mR CC method and the moving range technique. The mR CC method generates fewer false positives and the difference is significant (p value = 0.009).

Other tests to two proportions failed to show significant differences in performance.

Appendix E Summary of Leading Enterprise Data Quality Platforms

This appendix provides summaries of six enterprise data quality platforms that were highlighted in *The Forrester Wave: Enterprise Data Quality Platforms, Q4 2010* [Karel 2010]. The focus of the summary for each tool is on data profiling capabilities associated with the platform.

Vendor	Tool
DataFlux	Data Management Studio
Harte Hanks Trillium Software	Trillium Software System
IBM	Info Sphere Foundation Tools
Informatica	Data Explorer Informatica Analyst Informatica Developer Informatica Administrator
Pitney Bowes Business Insight	Spectrum
SAP BusinessObjects	Data Quality Management

DataFlux

Tool: Data Management Studio

http://www.dataflux.com/Products/Data-Management-Studio.aspx

The DataFlux software product, Data Management Studio, is an integrated framework that allows users to plan, implement, and monitor data across multiple processes and technology components from a single user interface.

Data Management Studio addresses the data quality management features listed in this table below.

Tool feature	Description
data profiling	To execute a complete assessment of organization's data, examining the structure, completeness, and suitability of information assets
metadata analysis	To understand what data resources exist and extract and organize metadata from any source throughout the enterprise
data quality	To correct data problems, standardize data across sources and create an integrated view of corporate information
data integration	Create workflows to consolidate and migrate data from multiple data sources
data monitoring	Enforce business rules for quality, providing a foundation for an ongoing, highly-customized data governance program
data enrichment	Enrich address data with geographic, demographic, or other details, as well as standardize and augment data on products, materials, and services
entity resolution	Identify and resolve disparate data on customers and products

Data Profiling

The data profiling feature is used to analyze the structure, completeness, and suitability of information assets by

- developing a complete assessment of the scope and nature of data quality issues
- creating an inventory of data assets
- inspecting data for errors, inconsistencies, redundancies, and incomplete information

This table below lists the capabilities provided in the data profiling feature.

Data profiling capability	Description
Business rule validation	Ensure data meets organizational standards for data quality and business processes by validating data against standard statistical measures as well as customized business rules
Relationship discovery	Uncover relationships across tables and databases – and across different source applications
Anomaly detection	Detect data that falls outside of predetermined limits and gain insight into source data integrity
Data validation	Verify that data in tables matches its appropriate description
Pattern analysis	Ensure that your data follows standardized patterns to analyze underlying data and build validation rules
Statistical analysis	Establish trends and commonalities in corporate information and examine numerical trends via mean, median, mode, and standard deviation

Harte-Hanks Trillium Software

Software Tool: Trillium Software System

http://www.trilliumsoftware.com/home/products/TrilliumSoftware.aspx

The Trillium Software System is an integrated data quality suite that delivers a single user experience for complete, global, data quality life-cycle management across the enterprise. Built for seamless movement to and from views related to each phase of the data quality management life cycle, the Trillium Software System emphasizes upfront and ongoing investigation and improvement during process design, development, and test phases.

Tool component	Purpose	Description
TS Discovery	Automated data discover and data profiling	Provides a complete view of enterprise data assets; uncovers the true content, structure, rules, relationships, and quality of the data; and reveals issues that otherwise might remain hidden
TS Quality	Parsing, standardizing, and cleansing global data	Cleanses, matches, and unifies data across multiple data sources and data domains including customer, product, sales, and financial. TS Quality delivers data parsing, standardization, and cleansing solutions and the ability to implement data quality processes in high-performance, real-time environments
TS Insight	For scoring and tracking enterprise data quality	A data quality dashboard that provides visibility into the status of data quality enabling analysts to monitor, manage, and view trends of data quality metrics through intuitive scorecards, charts, and graphs

Director	For deploying and managing real-time data quality processes	A complete data quality application server, Director delivers data cleansing and matching services across multiple platforms, servers, and applications. Using Director, organizations can integrate and deploy batch and real-time data quality projects in multiple environments.

Data Profiling

TS Discovery is the automated data profiling and data discovery component of the Trillium Software System. Automated profiling capabilities provide valuable insight about the current state of the data. Early discovery highlights issues, anomalies, and previously unknown data problems by exposing:

- data content and context
- data structure and patterns
- data integrity and business rules
- relationships and referential integrity

TS Discovery routinely assesses data to ensure that high quality is maintained and monitors production systems for anomalies.

IBM

Software Tool: InfoSphere Foundation Tools

http://www-01.ibm.com/software/data/integration/info_server/

http://www-01.ibm.com/software/data/infosphere/foundation-tools/index.html

Tool Component	Purpose	Description
Information Analyzer	To assess data quality	Deep profiling capabilities—provides a comprehensive understanding of data at the column, key, source, and cross domain levels
Data Architect	To design enterprise models	A collaborative data design solution used to discover, model, visualize, relate, and standardize diverse and distributed data assets
Fast Track	To capture design specifications	Streamlines collaboration between business analysts, data modelers, and developers by capturing and defining business requirements in a common, familiar format
Business Glossary	To manage business terms	Enables creating and managing an enterprise vocabulary and classification system, with ready-to-use industry standard terms and definitions
Metadata Workbench	To monitor business flows	Provides a window to a unified data integration platform, with insight into data source analysis, ETL (extract, transformation, load) processes, data quality rules, business terminology, data models, and business intelligence reports.
Discovery	To understand data relationships	Identifies and documents existing data, where it is located, and how it is linked across systems by intelligently capturing relationships and determining applied transformations and business rules

Data Profiling

IBM InfoSphere Information Analyzer is intended to help analysts understand data by offering data quality assessments, flexible data rules design and analysis, and quality monitoring capabilities. Capabilities include the following:

- deep profiling capabilities—provide a comprehensive understanding of data at the column, key, source, and cross domain levels
- multi-level rules analysis (by rule, record, or pattern) unique to the data quality space—provides the ability to evaluate, analyze, and address multiple data issues by record rather than in isolation
- shared metadata foundation—integrates the modules across IBM InfoSphere Information Server and IBM InfoSphere Information Server in support of the enterprise
- native parallel execution for enterprise scalability—enables high performance against massive volumes of data
- supports data governance initiatives through auditing, tracking, and monitoring of data quality conditions over time
- enhanced data classification capabilities help to focus attention on common personal identification information to build a foundation for data governance,
- used to proactively identify data quality issues, find patterns, and set up baselines for implementing quality monitoring efforts and tracking data quality improvements

Informatica

http://www.informatica.com/products_services/Pages/index.aspx#page=page-8

Informatica offers a number of products that share various capabilities related to data profiling and data quality. The software product *Data Explorer* is its primary product for data profiling. However, additional products are available that provide data profiling capabilities but are geared for specific roles within the organization, as shown in the table below.

Software Tool	Purpose
Data Explorer	Business analysts, data stewards, IT developers
Informatica Analyst	Line-of-business managers, data stewards, and business analysts
Informatica Developer	IT developers
Informatica Administrator	IT administrator

Software Tool: Data Explorer

Data profiling capabilities include the following:

- analyze data to automatically profile the content, structure, and quality of highly complex data structures
- discover hidden inconsistencies and incompatibilities between data sources and target applications
- easily customize new rules to automatically profile new data entries

Data mapping capabilities include the following:

- generate accurate source-to-target mapping between different data structures and define the necessary transformation specifications

- compare actual data and metadata sources to target application requirements
- find data gaps, redundancies, and inaccuracies to resolve before moving data
- identify data anomalies and create a normalized schema capable of supporting data

Connectivity capabilities include the following:

- profile all data types in a wide variety of applications, systems, and databases, including
  - .csv and flat files
  - RDBMS files (Oracle, DB2, UDB, Informix, Sybase)
  - ODBC data sources
  - VSAM and IMS, plus nonrelational data structures including COBOL copybooks
- extend support beyond basic customer data, such as names, addresses, and telephone numbers, to include product, financial, asset, pricing, and other data

Software Tool: Informatica Analyst

This easy-to-use, browser-based tool is designed to empower the business to proactively participate in data quality processes without IT intervention. It enables line-of-business managers, data stewards, and business analysts to

- profile, analyze, and create data quality scorecards
- drill down to specific records with poor data quality to determine their impact on the business and how to fix them
- monitor and share data quality metrics and reports by emailing a URL to colleagues
- define data quality targets and valid reference data sets
- specify, validate, configure, and test data quality rules
- collaborate efficiently with IT developers to share profiles and implement data quality rules
- identify anomalies and manage data quality exception records
- track data quality targets on an ongoing basis

Software Tool: Informatica Developer

This Eclipse-based data quality development environment is designed to enhance IT productivity. It enables IT developers to

- discover and access all data sources—whether they are on premise, with partners, or in the cloud
- analyze, profile, and cleanse data
- define and model logical data objects
- combine data quality rules with sophisticated data transformation logic
- conduct midstream profiling to validate and debug logic as it’s developed
- configure data quality services, provisioning data physically or virtually and at any latency

- reuse all profiling and rule specifications from business analysts and data stewards across all applications and projects

Software Tool: Informatica Administrator

This easy-to-use, browser-based tool with centralized configuration and deployment capabilities for managing the data integration environment enables IT administrators to

- manage services and nodes, including configurations that support grid and high availability
- oversee security and user management, including users, groups, roles, privileges, and permissions
- perform advanced monitoring and logging

Pitney Bowes Business Insight

Software Tool: Spectrum

http://www.pbinsight.com/

Pitney Bowes incorporates its data profiling feature in the Enterprise Data Governance product. The product is composed of two modules, Profiler Plus and Monitor Plus.

Profiler Plus is intended to help the user discover and understand the quality of data. Its data profiling features and integrated analysis management framework enhances, accelerates, and reduces risks in data analysis activities. Key benefits described by Pitney Bowes include the ability to

- decrease analysis timeframes by up to 90%
- reduce IT project risks with automated, up-front analysis
- simplify analysis activities through integrated management platform
- gain confidence in the quality of your data

Monitor Plus allows you to create rules to provide a proven way of checking and validating the data used in your business systems and applications, including the ability to

- run regular data checks using an external scheduler
- integrate monitoring into your existing operational data environment
- receive automatic alerts after every execution, with data reports sent directly to your inbox

Pitney Bowes also provides specific products related to other and more specific aspects of data quality. They include

- **Address New Module**: Capture, validate, and correct addresses for the U.S., Canada, and over 220 countries worldwide with the Address Now Module for the Spectrum Technology Platform

- **Advanced Matching Module**: Marketing and business processes rely on accurate data to identify and understand the relationships between records. The Advanced Matching Module recognizes customers, products, duplicates, and households across data sources.
- **Data Normalization Module**: Creates a uniform customer experience by standardizing terms in your data base with the Data Normalization Module for the Spectrum Technology Platform
- **Universal Addressing Module**: Address data exists throughout your enterprise in customer databases, call centers, web sites, and marketing systems. Reliable address information is required to communicate effectively with your customers, develop an accurate single customer view, and leverage your customer-facing technology investments. The Universal Addressing Module provides address validation, correction, and standardization technologies for more than 220 countries
- **Universal Name Module**: Provides flexible and global name knowledge to better segment and target your customer base while matching, standardizing, analyzing, and consolidating complex records with confidence

SAP BusinessObjects

Software Tool: Data Quality Management

http://www.sap.com/solutions/sapbusinessobjects/large/eim/data-quality-management/index.epx

SAP BusinessObjects Data Quality Management—which is also available in versions for Informatica, SAP solutions, and Siebel—delivers a solution to help analyze, cleanse, and match customer, supplier, product, or material data (structured or unstructured) to ensure highly accurate and complete information anywhere in the enterprise.

SAP BusinessObjects Data Quality Management includes the following features and functionality:

- data quality dashboards that show the impact of data quality problems on all downstream systems or applications
- the ability to apply data quality transformations to all types of data, regardless of industry or data domain, such as structured to unstructured data as well as customer, product, supplier, and material information
- intuitive business user interfaces and data quality blueprints to guide you through the process of standardizing, correcting, and matching data to reduce duplicates and identify relationships
- comprehensive global data quality coverage with support for over 230 countries
- comprehensive reference data
- broad, heterogeneous application and system support for both SAP and non-SAP sources and targets
- prepackaged native integration of data quality best practices for SAP, Siebel, and Informatica PowerCenter environments
- optimized developer productivity and application maintenance through intuitive transformations, a centralized business rule repository, and object reuse

- high performance and scalability with software that can meet high volume needs through parallel processing, grid computing, and bulk data loading support
- flexible technology deployment options, from an enterprise platform to intuitive APIs that allow developers quick data quality deployment and functionality

References/Bibliography

URLs are valid as of the publication date of this document.

[Acuna 2004]
Acuna, E. & Rodriquez, C. "Meta Analysis Study of Outlier Detection Methods in Classification." Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez. *Proceedings IPSI.* Venice 2004. http://academic.uprm.edu/eacuna/paperout.pdf

[Ballou 1985]
Ballou, D. P. & Pazer, H. L. "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems." *Management Science 31*, 2 (1985): 150-162.

[Ballou 1998]
Ballou, D. P.; Wang, R. Y.; Pazer, H.; & Tayi, G. K. "Modeling Information Manufacturing Systems to Determine Information Product Quality." *Management Science 40*, 4 (1998): 462-484.

[Barnett 1998]
Barnett, V. & Lewis, T. *Outliers in Statistical Data*, 3rd ed. John Wiley & Sons, Inc., 1998.

[Bianco 2001]
Bianco, A. M.; Ben, M. G.; Martinez, E. J.; & Yohai, V. J. "Outlier Detection in Regression Models with ARIMA Errors Using Robust Estimates." *J. Forecast. 20*, 8 (2001): 565–579.

[Böhrer 2008]
Böhrer, Armin. "One-sided and Two-sided Critical Values for Dixon's Outlier Test for Sample Sizes up to n = 30." *Economic Quality Control 23*, 1 (2008): 5–13.

[Bonadonna 2006]
Bonadonna, Costanza; Scollo, Simona; Cioni, Raffaello; Pioli, Laura; & Pistolesi, Marco. *Determination of the Largest Clasts of Tephra Deposits for the Characterization of Explosive Volcanic Eruption*. IAVCEI Commission on Tephra Hazard Modeling, Salcedo, Ecuador, January 16-18, 2006.

[Box 1970]
Box, George & Jenkins, Gwilym. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1970.

[Brockwell 1987]
Brockwell, Peter J. & Davis, Richard A. *Time Series: Theory and Methods*. Springer-Verlag, 1987.

[Burke 2006]
Burke, Shaun. "Missing Values, Outliers, Robust Statistics & Non-parametric Methods." C*hromatographyonline, LC•GC Europe Online Supplement* (January 2006): 19-24.

[Chandola 2009]
Varun Chandola, V.; Banerjee, A.; & Kumar V. "Anomaly Detection—A Survey." *ACM Computing Surveys 41*, 3 (July 2009): Article 15.

[Chen 2005]
Chen, D.; Shao, X.; Hu, B.; & Su, Q. "Simultaneous Wavelength Selection and Outlier Detection in Multivariate Regression of Near-Infrared Spectra." *Analytical Science 21*, 2 (2005): 161–167.

[Chen 2009]
Kuang C.; Chen, H.; Conway, N.; Dolan, H.; Hellerstein, J. M.; & Parikh, T. S. "Improving Data Quality With Dynamic Forms." *International Conference on Information and Communication Technologies and Development* (2009). http://db.cs.berkeley.edu/papers/icde10-usher.pdf

[Crowder 1989]
Crowder, S. V. "Design of Exponentially Weighted Moving Average Schemes." *Journal of Quality Technology 21* (1989):155-162.

[DAU 2011]
Earned Value Management. Defense Acquisition University, 2011. https://acc.dau.mil/CommunityBrowser.aspx?id=17609&lang=en-US

[DCARC 2011]
Defense Cost and Resource Center. http://dcarc.pae.osd.mil/.

[Dixon 1950]
Dixon, W. J. "Analysis of Extreme Values." *The Annals of Mathematical Statistics 21*, 4 (December 1950): 488-506.

[Dixon 1951]
Dixon, W. J. "Ratios Involving Extreme Values." *The Annals of Mathematical Statistics 22*, 1 (March 1951): 68-78.

[DOA 2008]
Department of the Army, U.S. Army Corps of Engineers. *Environmental Quality—Environmental Statistics*, Engineer Manual 1110-1-4014, January 2008.

[Dunkl 2007]
Dunkl, István. *Grubbs' Test*. http://www.sediment.uni-goettingen.de/staff/dunkl/software/pep-grubbs.pdf (2007).

[English 2009]
English, L. *Information Quality Applied*. John Wiley & Sons, 2009.

[English 2011]
English, L. *Glossary*. Information Impact International, Inc. http://www.infoimpact.com/resglossary.cfm

[EPA 2000]
U.S. Environmental Protection Agency, Office of Environmental Information. *Quality Guidance for Data Quality Assessment.* EPA/600/R-96/084, July, 2000.

[Ermer 2005]
Ermer, Joachim & Miller, John H., eds. *Method Validation in Pharmaceutical Analysis: A Guide to Best Practice*. John Wiley & Sons, 2005.

[Fawcett 1997]
Fawcett T. & Provost F. "Adaptive Fraud Detection." *Data-mining and Knowledge Discovery 1*, 3 (1997): 291-316.

[Florac 1999]
Florac, William A. & Carleton, Anita D. *Measuring the Software Process: Statistical Process Control for Software Process Improvement.* Addison-Wesley Professional, 1999.

[Galeano 2004]
Galeano, P.; Pea, D.; & Tsay, R. S. "Outlier Detection in Multivariate Time Series Via Projection Pursuit." Statistics and Econometrics Working Articles ws044211, Departamento de Estadistica y Econometrica, Universidad Carlos III, 2004.

Gartner 2009
Gartner Research Group. "Findings from Primary Research Study: Organizations Perceive Significant Cost Impact from Data Quality Issues." August 14, 2009.
http://www.gartner.com/id=1131012

[Gibbons 2001]
Gibbons, Robert D. & Coleman, David D. *Statistical Methods for Detection and Quantification of Environmental Contamination*. Wiley-Interscience, 2001.

[Griffith 2007]
Griffith, Daniel. *Outlier Test Using Grubbs' Test Code*. http://www.minitab.com/en-TW/support/macros/default.aspx?action=code&id=120 (2007).

[Grubbs 1969]
Grubbs, Frank. "Procedures for Detecting Outlying Observations in Samples." *Technometrics 11*, 1 (1969): 1-21.

[Hawkins 1980]
Hawkins D. *Identification of Outliers*. Chapman and Hall, 1980.

[Hodge 2004]
Hodge, Victoria and Austin, Jim. "A Survey of Outlier Detection Methodologies." Artificial Intelligence Review 22, 2 (85-126), 2009.

[Huah 2005]
Huah, Yeoh Guan. "Basic Statistical Methods for Chemical Analysis." *Proceedings of the Workshop on Basic Statistics for Chemical Analysis*. Malaysian Institute of Chemistry and the Universiti Sains Malaysia, September, 2005.

[Huang 1999]
Huang, K.; Lee, Y.; & Wang, R. Y. *Quality Information and Knowledge*. Prentice Hall, 1999.

[ISO 2008]
International Organization of Standardization. Software Engineering – Software Product Quality Requirements and Evaluation (SQuaRE) – Data Quality Model. http://www.iso.org/iso/catalogue_detail.htm?csnumber=35736

[Johnson 1992]
Johnson, R. *Applied Multivariate Statistical Analysis*. Prentice Hall, 1992.

[Johnson 1998]
Johnson, T.; Kwok, I.; & Ng, R. "Fast Computation of 2-Dimensional Depth Contours." *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1998.

[Juran 1951]
Juran, J. *Quality Control Handbook*. McGraw-Hill, 1951.

[Kadota 2003]
Kadota, K.; Tominaga, D.; Akiyama, Y.; & Takahashi, K. "Detecting Outlying Samples in Microarray Data: A Critical Assessment of the Effect of Outliers on Sample Classification." *Chem-Bio Informatics 3*, 1 (2003): 30–45.

[Karel 2010]
Karel, Rob. *The Forrester Wave: Enterprise Data Quality Platforms, Q4 2010*, October 29, 2010. http://www.forrester.com/rb/Research/wave%26trade%3B_enterprise_data_quality_platforms%2C_q4_2010/q/id/48300/t/2

[Keen 1953]
Keen, John & Page, Denys J. "Estimating Variability from the Differences Between Successive Readings." *Applied Statistics 2*,1 (1953): 13-23.

[Kelly 2009]
Kelly, Jeff. *Poor Data Quality Costing Companies Millions of Dollars Annually*. http://www.melissadata.com/enews/articles/09092010/3.htm (2009).

[Kitagawa 1979]
Kitagawa, G. "On the Use of AIC for the Detection of Outliers." *Technometrics, 21*, 2 (1979): 193–199.

[Kriegel 2010]
Kriegel, H.P.; Kroger, P.; and Zimek, A. "Outlier Detection Techniques." Tutorial at the 16th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, 2010.

[Lazarevic 2008]
Lazarevic, A.; Banerjee, A.' Chandola, V.; Kumar, V.; & Srivastava, J. "Data Mining for Anomaly Detection." *European Conference on Principles and Practice of Knowledge Discovery in Databases*. Antwerp, Belgium, September 2008.

[Lohninger 2011]
Lohninger, H. *Grubbs' Outlier Test.*
http://www.statistics4u.com/fundstat_eng/ee_grubbs_outliertest.html

[Lu 2003]
Lu C.; Chen, D.; & Kou, Y. "Algorithms for Spatial Outlier Detection." *Proceedings of the 3rd IEEE International Conference on Data-mining*. Melbourne, FL, 2003.

[Mélard 2000]
G. Mélard, G. & Pasteels, J. M. "Automatic ARIMA Modeling Including Interventions, Using Time Series Expert Software." *International Journal of Forecasting 16* (2000): 497 –508.

[Minitab 2010]
Minitab 16 Statistical Software. State College, PA: Minitab, Inc., 2010. http://www.minitab.com

[Montgomery 2005]
Montgomery, Douglas. *Introduction to Statistical Quality Control*. John Wiley & Sons, 2005.

[Nelson 1973]
Nelson, Charles. *Applied Time Series Analysis for Managerial Forecasting.* Holden-Day, 1973.

[NIST 2011a]
NIST/SEMATECH. *e-Handbook of Statistical Methods.*
http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc322.htm

[NIST 2011b]
NIST/SEMATECH. *e-Handbook of Statistical Methods.*
http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc324.htm

[NIST 2011c]
NIST/SEMATECH. e-Handbook of Statistical Methods.
http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm

[NRC 2010]
National Research Council of the National Academies. *Critical Code: Software Producibility for Defense.* The National Academic Press, 2010.

[OUSD 2011]
Office of the Undersecretary for Defense. *Data Item Description: Contract Performance Report (CPR).* http://www.marcorsyscom.usmc.mil/sites/scatt/DIDs/DI-MGMT-81466A.pdf (2005).

[Penny 2001]
Penny, K. & Jollifee, I. "A Comparison of Multivariate Outlier Detection Methods of Clinical Laboratory Safety Data." *The Statistician 50*, 3 (2001): 295-308.

[Rademacher 2009]
Rademacher, Joe & Harter, Clay. *DataFlux DataVision Technology Demonstration.* http://www.openspirit.com/filebin/pdfbin/Webcast_DataFlux_OpenSpirit_Sept_2009.pdf (2009).

[Redman 1996]
Redman, T. C. *Quality for the Information Age*. Artech House, 1996.

[Roberts 1959]
Roberts, S.W. "Control Chart Tests Based on Geometric Moving Averages." *Technometrics 1* (1959).

[Rorabacher 1991]
Rorabacher, David B. "Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon's *'Q'* Parameter and Related Subrange Ratios at the 95% Confidence Level." *Analytical Chemistry 63*, 2 (January 15, 1991): 139 – 146.

[Rosner 1975]
Rosner, B. "On the Detection of Many Outliers." *Technometrics 17* (1975): 221–227.

[Rosner 1983]
Rosner, B. "Percentage Points for a Generalized ESD Many-Outlier Procedure." *Technometrics 25* (1983): 165–172.

[Ruts 1996]
Ruts, Il & Rousseeuw, P. "Computing Depth Contours of Bivariate Point Clouds." *Computational Statistics and Data Analysis, 23* (1996): 153-168.

[Shewhart 1931]
Shewhart, Walter A. *Economic Control of Quality of Manufactured Product*. Van Nostrand, 1931.

[StatSoft 2011]
StatSoft. *What is Quality Control and Quality Control Charts*. http://www.statsoft.com/textbook/quality-control-charts/

[Stefansky 1972]
Stefansky, W. "Rejecting Outliers in Factorial Designs." *Technometrics 14*, 2 (1972): 469-479.

[Tsay 2000]
Tsay, R. S.; Pea, D.; & Pankratz, A. E. "Outliers in Multi-variate Time Series." *Biometrika 87*, 4 (2000): 789–804.

[Tukey 1977]
Tukey, John W. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[U.S. Army 2008]
Department of the Army, U.S. Army Corps of Engineers. *Environmental Quality—Environmental Statistics*, Engineer Manual 1110-1-4014. January, 2008.

[Valenzuela 2004]
Valenzuela, O.; Márquez, L.; Pasadas, M.; & Rojas, I. "Automatic Identification of ARIMATime Series By Expert Systems Using Paradigms of Artificial Intelligence." *Monografías del Seminario Mathematic Garcia de Galdeano 31* (2004): 425–435.

[Valenzuela 2008]
Valenzuela, O.; Rojas, I.; Rojas, F.; Pomares, H.; Herrera, L. J.; Guillen, A.; Marquez, L.; & Pasadas, M. "Hybridization of Intelligent Techniques and ARIMA Models For Time Series Prediction." *Fuzzy Sets and Systems, 159*, 7 (April 2008): 821-845.

[Verma 2006]
Verma, Surendra P. & Quiroz-Ruiz, Alfredo. "Critical Values for Six Dixon Tests for Outliers in Normal Samples up to Sizes 100, and Applications in Science and Engineering." *Revista Mexicana de Ciencias Geológicas 23*, 2 (2006): 133-161.

[Wand 1998]
Wand, Y. & Wang, R.Y. "A Product Perspective on Total Data Quality Management." *Communications of the ACM 41*, 2 (1998): 58-65.

[Wang 1996]
Wang, R. Y. & Strong, D. M. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems 12*, 4 (1996): 5-34.

[Wheeler 2000]
Wheeler, Donald J. *Understanding Variation: The Key to Managing Chaos.* SPC Press, 2000.

[Wheeler 2010]
Wheeler, Donald J. & David S. Chambers. *Understanding Statistical Process Control*, 3rd ed. SPC Press, 2010.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
(Leave Blank)	December 2011	Final

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
An Investigation of Techniques for Detecting Data Anomalies in Earned Value Management Data	FA8721-05-C-0003

6. AUTHOR(S)

Mark Kasunic, James McCurley, Dennis Goldenson, David Zubrow

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213	CMU/SEI-2011-TR-027

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
HQ ESC/XPK 5 Eglin Street Hanscom AFB, MA 01731-2116	ESC-TR-2011-027

11. SUPPLEMENTARY NOTES

12A DISTRIBUTION/AVAILABILITY STATEMENT	12B DISTRIBUTION CODE
Unclassified/Unlimited, DTIC, NTIS	

13. ABSTRACT (MAXIMUM 200 WORDS)

Organizations rely on valid data to make informed decisions. When data integrity is compromised, the veracity of the decision-making process is likewise threatened. Detecting data anomalies and defects is an important step in understanding and improving data quality.

The study described in this report investigated statistical anomaly detection techniques for identifying potential errors associated with the accuracy of quantitative earned value management (EVM) data values reported by government contractors to the Department of Defense.

This research demonstrated the effectiveness of various statistical techniques for discovering quantitative data anomalies. The following tests were found to be effective when used for EVM variables that represent cumulative values: Grubbs' test, Rosner test, box plot, autoregressive integrated moving average (ARIMA), and the control chart for individuals. For variables related to contract values, the moving range control chart, moving range technique, ARIMA, and Tukey box plot were equally effective for identifying anomalies in the data.

One or more of these techniques could be used to evaluate data at the point of entry to prevent data errors from being embedded and then propagated in downstream analyses. A number of recommendations regarding future work in this area are proposed in this report.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Data quality, data anomaly detection, automated anomaly detection, data integrity	103

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL