Detecting Anomalous Cost Data in an Integrated Data Warehouse

Lawrence Brown, Northrop Grumman Information Systems
Lawrence.Brown@ngc.com

Overview

Data quality is always a key issue when using historical data to predict future costs. A small data quality issue can lead to a significant error in projected costs. This quality problem is particularly acute if:

- The data comes from several sources that have different characteristics
- Data from the source systems are being used for purposes beyond their original intent
- Some of the source systems are missing key data attributes that must be inferred by a complex set of business rules
- The source systems have evolved over time
- The data warehouse is used to support important decisions

This paper presents a systematic approach to finding the largest data quality issues in a warehouse that is challenged by these issues. The approach described in the paper was developed for the Air Force Total Ownership Cost (AFTOC) program. The goal of the program is to provide accurate and timely cost data for Air Force weapon systems to variety of Air Force, government personnel and contractors. AFTOC integrates data from the Air Force cost, budget, supply, personnel, fuel and ammunition systems to produce an integrated view of weapon system costs. This program is managed by the Air Force Cost Analysis Agency and supported by 309th Maintenance Wing at Hill AFB, Battelle and Northrop Grumman. The program operates a web site with standard reports, On-Line Analysis Processing (OLAP) and ad-hoc query of supply data. AFTOC data is used to prepare budgets, develop return on investment cases, and provide reports to senior Air Force and DoD officials.

The Challenges

All the AFTOC data sources have data quality issues and the process of integrating the sources can create additional issues. For example, the Air Force cost reporting system does not identify the weapon system, but it does provide some hints that allow the costs to be assigned to a weapon system or allocated among two or more. The supply system was designed to move parts to the right locations, but it also provides rich detail on the costs of those parts. Similarly, the cost system shows personnel costs, but it provides little information about the cost

categories of those personnel. AFTOC uses to the personnel system to identify costs by pilots, maintainers, support staff, etc. Because of the nature of the source systems, the data will not be perfect. The key challenge is to find the most inconsistent data so that it can be fixed or the consumers can be warned of the issues. AFTOC reports on 200 weapon systems, 14 years of history, 50 costs categories, 200 different bases, 1000 cost centers, 1000 budget codes, 1000 element of expense codes, 10 Appropriations, 12 commands, and 300 units. Thus, there are 10^{21} different possible combinations. Even though most of these combinations are empty, there are many possible ways to misplace data. It is too challenging to prove the data is correct. Our approach is to:

- Try very hard to prove the data is wrong
- Where we find issues, either fix them or notify the users
- When we cannot find problems, we build our confidence in the accuracy of the data

A Systematic Approach to Data Quality

While an ad-hoc approach may find some of the issues, this approach is time consuming and open-ended. AFTOC has developed a systematic approach which can be applied to other data warehouses. The approach looks for specific types of data anomalies. We use the term anomalies because we are looking for inconsistencies in the data. These anomalies may be data errors, processing errors or they may be related to changes in Air Force operations. The approach is to define a number of anomaly detection tests. Each test follows the steps shown in Figure 1.

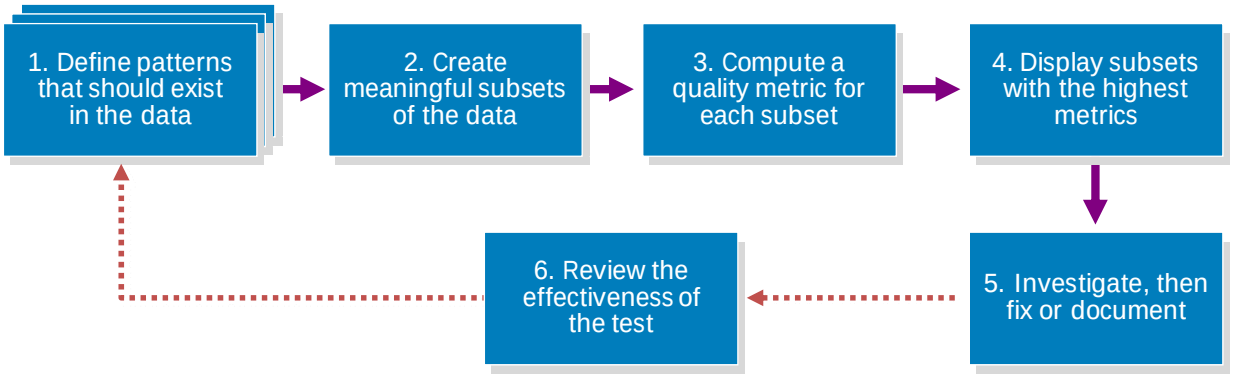

Figure 1 - A Systematic Approach

The steps for each test are:

1. Define a model that should exist in the data. For example, fuel consumption should be proportional flying hours. The cost of replacement parts, adjusted for inflation, should not change much from month to month
2. Create meaningful subsets of the data. If the subsets are too large, some problems may be masked. If they are too small the natural randomness of the data may appear to be problems.

3. Compute a quality metric for each subset that measures how well the subset of the data matches the model. This metric is often a simple statistic such as a correlation or the residual from a regression.

4. Display the subsets with the highest metrics in a manner so that the largest misfits are obvious. This may be a simple sorted list, a scatter chart, or a scorecard format

5. Investigate the largest anomalies, then fix or document.

6. Review the effectiveness of each test. Find ways to improve the process over time. Which tests are no longer finding issues? What issues were found by users that were missed during QA?

Examples

There are two basic types of anomaly tests. The first compares data from two or more source systems and the second looks at consistency over time. Because AFTOC data is categorized as "For Official Use Only", these examples do not show specific numbers and systems.

Comparing Source Systems Tests

Since AFTOC data is integrated from several sources, there is an opportunity to compare the data from the systems for consistency.

AFTOC receives fuel consumption and flying hours by specific tail number from two different systems. The model is that all the aircraft of a given type at a given base should consume about the same amount of fuel for each hour flown. The subsets are defined by Aircraft type, Base, and year. Most of the subsets consist of 10 to 30 pairs of hours and gallons. The metric used is the correlation of these pairs. The display used is a simple scorecard format shown in figure 2 with color to highlight low correlations. Low correlations can be due to incomplete or duplicated data sources or processing errors.

		FY2003	FY2004	FY2005	FY2006
First Aircraft Type	Base 1	0.184	0.999	0.976	0.999
	Base 2	0.995	0.978	0.986	0.998
	Base 3	0.961	0.996	0.976	0.985
	Base 4	0.956	0.999	0.991	0.999
Second Aircraft Type	Base 5	0.995	0.988	0.895	0.996
	Base 6	0.950	0.992	0.974	0.994
	Base 7	0.969	0.961	0.977	0.945
	Base 8	0.975	0.985	0.978	0.995

Figure 2 - Fuel / Flying Hour Correlations

A couple of recent results are shown in figures 3 and 4. These scatter plots show gallons consumed on the vertical axis and flying hours plotted on the horizontal access. Each point represents one tail number for one year.

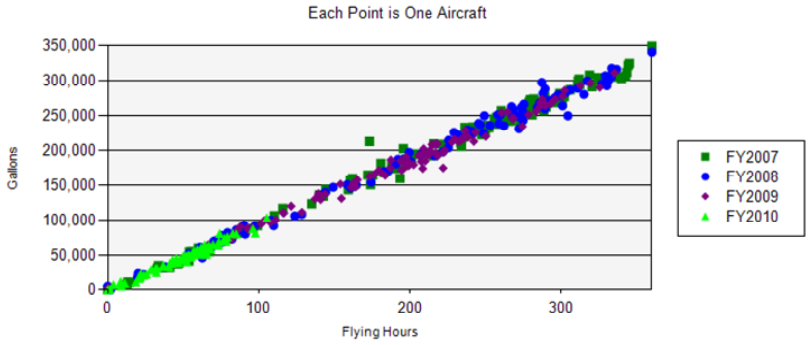

Figure 3 - Example of a High Correlation

Figure 3 shows a base and aircraft with a very high correlation for the last several years. In the figure, each point is one aircraft in one year. This chart shows that the reported fuel and flying hours follow the expected pattern. Almost all the point fall along a diagonal that matches the fuel consumption factor for this aircraft.

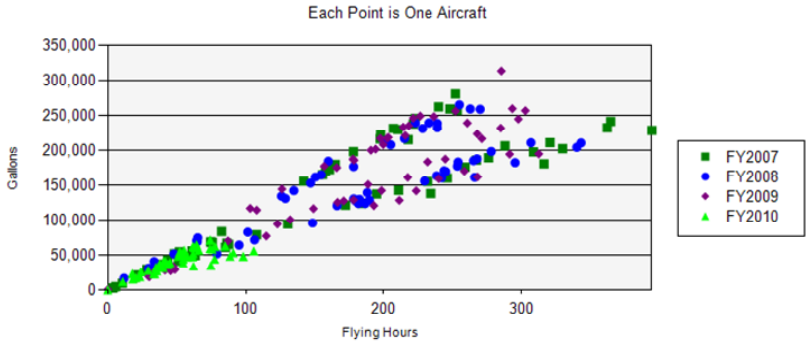

Figure 4 - Example of a Low Correlation

A base / Aircraft combination with a low correlation is shown In Figure 4. The bimodal appearance of the distribution is clearly an anomaly. In this example, the anomaly was due to different versions (block numbers) at the base involved. The newer versions of this aircraft are larger, have more powerful engines and consume more fuel. The model for this test was based on the assumption that all the aircraft of a given type at a base would have similar configurations and perform similar missions. For this particular base and aircraft, this assumption was not valid. This is a case where the data is correct, but many AFTOC users would not understand the issue. This is the type of situation that we document for our users.

Other tests that compare source systems are:

- Cost v. Budget – the data sources arrive in different formats and levels of detail and they are processed by different business rules.
- Cost v. Supply – Prior to 2008, the supply system fed the cost system and this test insured that this process was working correctly. In 2008, the cost system is populated with estimates of supply costs based on historical data.

- Cost v. Fuel – do the two systems, after processing, show the same costs
- Supply v. Flying Hours – do the bases with the most flying hours for a given type of aircraft also have the largest supply costs?

Time Series Tests

These tests look at data over time. The Air Force manages costs in four major categories: Procurement; Research Development Test and Evaluation (RDT&E); Operations and Maintenance (O&M); and Military Personnel (Milpers). Procurement and RDT&E funds do not show any year to year consistency at the weapon system level. For example, a major modification to an aircraft is funded under the Procurement appropriation. In this case there would be a large increase on Procurement funds for one or two years. On the other hand, O&M and Milpers costs should be stable over time – especially after adjusting for changes in inventory, flying hours and inflation. Figure 5 shows four types of time series curves.

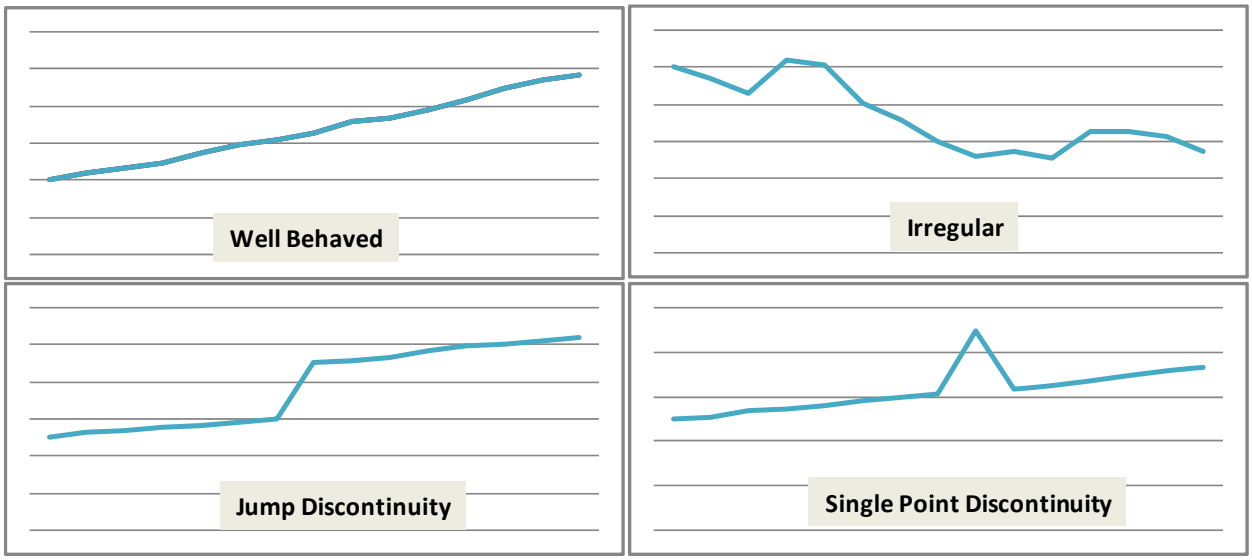

Figure 5 - Four Types of Time Series

The well behaved is the ideal. The next curve, Irregular, is interesting, but our experience shows this is not a data quality problem. These time series often represent processes such as Depot Maintenance that is less predictable. The next two curves, Jump Discontinuity and Single Point Discontinuity, are certainly anomalies to be investigated. AFTOC has developed specific tests to find these types of curves.

Single Point Discontinuities

To find the single point discontinuities, the model is that O&M and Milpers costs, after adjusting for inflation[1], should fit a regression with time and flying hours as independent variables. The subsets used are based on one aircraft type and one of the 7 major cost categories. This results in approximately 700 subsets with each one currently consisting of 14 time points. The metric used is based on computing the residuals from the entire time interval. Then, one at a time, a point is removed to see how much the residuals are reduced. More precisely, the metric for each subset is:

[1] Inflation in this context is the published Air Force cost year over year cost changes by category. There are separate factors for fuel, personnel, supplies, etc.

$$M = 1 - \sqrt{\frac{\min(\square[Ri)}{R}}$$

Where *R* is the sum of the squares of residuals for the regression over the entire interval and *Ri* is sum of the squares of the residuals for the regression that is computed omitting the i^{th} point. M is almost 1.0 if all the data points except one fit the model well. M is near 0.0 if omitting any single data point does not improve the fit. This is true of the well behaved and irregular curves shown in Figure 5. The results of the computation are displayed as a list of subsets sorted by the metric. An analyst can then focus on the subsets with the highest metrics.

Jump Discontinuities

The jump discontinuities are found using a similar approach. The regression for each subset is computed over the whole interval and then the interval is partitioned into two parts. The regression and residuals are computed for each part. In this case,

$$M = 1 - \sqrt{\frac{\min(R(1,i) + R(i+1,n))}{R(1,n)}} \quad where\ 4 \leq i \leq n-4$$

Where *R(i,j)* is the sum of the squares of the residuals from i^{th} data point to the j^{th} data point. In other words, consider all the partitions for each subset with no partition smaller than 4 points. For each subset, the algorithm finds the partitioning that minimizes the residuals to compute the quality metric for this subset.

Unusual Monthly Supply Costs

Another test that is applied to supply data looks for unusual monthly costs for a given aircraft type. Our experience has shown that unusual months are often due to bad cost information in the supply system or a mistake in associating the right account information with a transaction that caused us to report the costs to the wrong aircraft type. The model for this test is that the supply costs for a month should be similar to costs in the previous and subsequent months. A subset is defined for each weapon system and month. The metric is the difference between monthly costs and a moving average smoother which is defined as:

$$M_i = Abs(S_i - (0.15*S_{i-2} + 0.20*S_{i-1} + 0.30*S_i + 0.02*S_{i+1} + 0.15*S_{i+2}))$$

Where S_i is the supply costs in month i for each weapon system. Figure 6 shows a sample of a time history for one aircraft. The point in red was one of the largest anomalies.

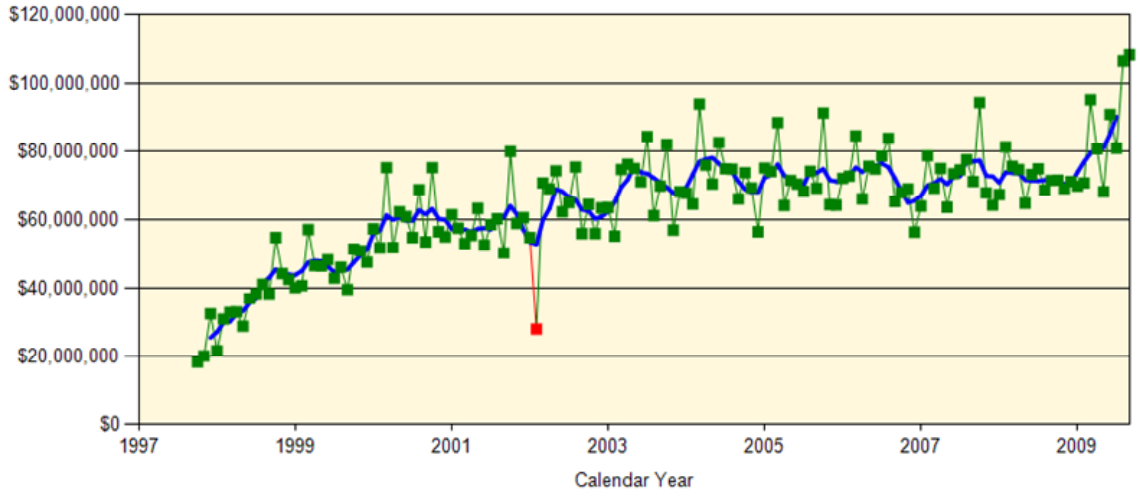

Figure 6 - Monthly History of Supply Costs

Significant Revisions to Historical Data

Another time series test is the comparison of quarterly versions of data. Each quarter, AFTOC receives new data feeds for the last several years. Also, we sometimes improve the business rules that are used to assign and allocate costs to weapon systems. Our philosophy is to apply the same business rules to all years of data. The model for this test is that costs in completed years should not change very much. The costs do change because some types of appropriations can be spent over several years (AFTOC reports costs based on the year of congress appropriated the funds). Even for funds that have to be obligated in the year of appropriation (one year money) there are adjustments to completed years. This test uses a subset of weapon system and fiscal year which results in about 2,500 subsets. The metric for each subset is the root sum square (RSS) of the differences in costs between the costs reported in the previous quarter of data for completed years. The RSS is performed based on costs by command and two different costs breakdowns. One of cost break downs is reported in the source data and one determined during AFTOC processing. The RSS metric is high if the reported costs changed substantially or if the costs were reported to a different command or in a different cost category. This test has found several cases where mistakes in cost data were corrected in the source and it also shows when costs were reported after the close of the quarter. This test also helps us to understand the impact of changes to the source systems and the AFTOC cost assignment/allocation business rules.

Current Quarter Data that Does not Match Prediction

Another time series detection test compares a prediction of O&M and Milpers for the current quarter using historical data. The model assumption is that these costs should be very predictable from previous cost and flying hour data. The subsets used are aircraft and seven major cost categories (about 700 subsets). For each subset, historical costs are adjusted for inflation and a linear regression is performed on the historical data to predict the costs for the current quarter. The regression uses time and flying hours as independent variables. The metric used is the difference between the predicted and reported costs for the subset.

Other Time Series Tests

In addition the time series tests described above, the following tests are also run:

- Constant Dollar Fuel – is the cost for a gallon of aviation fuel, adjusted for inflation, constant over time?
- Constant Dollar Personnel – is the average cost for an officer, enlisted or civilian, adjusted for inflation, nearly constant over time?
- Compare Two Years of Engine Costs – are current year aircraft engine costs, by engine type and cost category similar to previous years?
- Indirect Predictions – AFTOC also reports infrastructure costs by base. This test looks for a consistent time history in various cost categories.

Integration of these Techniques into the AFTOC Processing Cycle

AFTOC publishes most of its data quarterly, normally two months after the end of the quarter. During the last week before publication, the tests described here are to find the largest anomalies. In a typical quarter a dozen anomalies are detected. The list is prioritized based on the approximate magnitude of the issue. Next the subject matter experts examine the anomalies to determine if they were due to:

- A bad data feed (e.g. missing or duplicate data)
- An unusual occurrence in the data (e.g. $40M in the net cost field for a pair of gloves)
- AFTOC business rules that did not work well for a particular situation (e.g. costs that must allocated among a mix of very different aircraft)
- An error in processing (e.g. a missing a step in the processing)
- A change in Air Force operations (a new maintenance strategy for an aircraft)

When possible, the data is corrected prior to publication by requesting new data feeds or reprocessing to fix mistakes. In some cases we contact the commands or weapon systems program offices to improve the understanding of anomalous data. If it is not feasible to fix the data prior to publication, or if the data is really correct, just unusual, then the users are notified of the issues.

Implementation tools

The techniques described in this paper are implemented via database queries (MS SQL Server) imbedded within a reporting tool (MS Reporting Services). These tools are used for the production database and data delivery, so they were the natural choices for implementing the anomaly detection. Other databases or reporting tools would work as well. Reporting tools also have the advantage of “drill down” to more detailed reports. For example, a particular anomaly test might display its results as a scatter chart. In a reporting environment, the analyst can click on a point on the scatter chart and see a report showing detail behind that point. Database queries using the Structured Query Language (SQL) are very effective at computing simple statistics. These statistics are defined as expressions based on sums, sums of squares, sums of products, etc. SQL queries can filter, group, and sort the data very well. The performance of this type of data retrieval and computation is surprisingly fast on modern computers. For example, a query that involves tens of thousands of regressions runs in a few seconds.

All of the techniques could be implemented in a spreadsheet and most of them were prototyped with Excel. The problem with a spreadsheet solution is that it tends to require many manual steps involving cutting and pasting data, which is time consuming and error prone.

There are many powerful statistical and data mining tools, but they are focused on a careful and focused analysis of one issue. The anomaly detection applies simple models to thousands of cases to find the most interesting ones.

There are good commercial tools data quality tools, but their emphasis is different. They look more at data attributes and record level quality, while the focus here is on measures (data values). These tools will take a representation of a mailing address and convert it into a standard so that each distinct address is stored only once in the warehouse. For AFTOC, this attribute problem has already been solved.

Application of these Techniques into Other Data Warehouses

While the models and anomaly detection tests described in this paper are unique to AFTOC, the systematic approach should apply across a wide range of warehouses. For example, in a data warehouse of retail data, the models might have to include adjustments for seasonality. With AFTOC, regression has proven to be a simple and reliable predictive method. For some time series, more advanced techniques may be appropriate. With AFTOC, we can test the data before we publish it. In other warehouses, the data may have to be loaded nightly and the tests may have to be done while the data is "live". The warehouses that will benefit the most from this structure approach have the following characteristics

- The data is used for important decisions.
- The data is integrated from several sources, none of which are perfect.
- Source systems have evolved over time, but the users need a consistent view.
- There is ample anecdotal evidence of data quality issues.
- Data is often used a summary level where significant data quality issues may not be obvious.

Every owner of a data warehouse should have an answer for the question "How accurate is data?". While the question is simple, often the answer is very complex. The techniques presented here can provide part of that answer.