NORTHROP GRUMMAN

Detecting Anomalous Cost Data in an Integrated Data Warehouse

Society of Cost Estimation and Analysis Conference

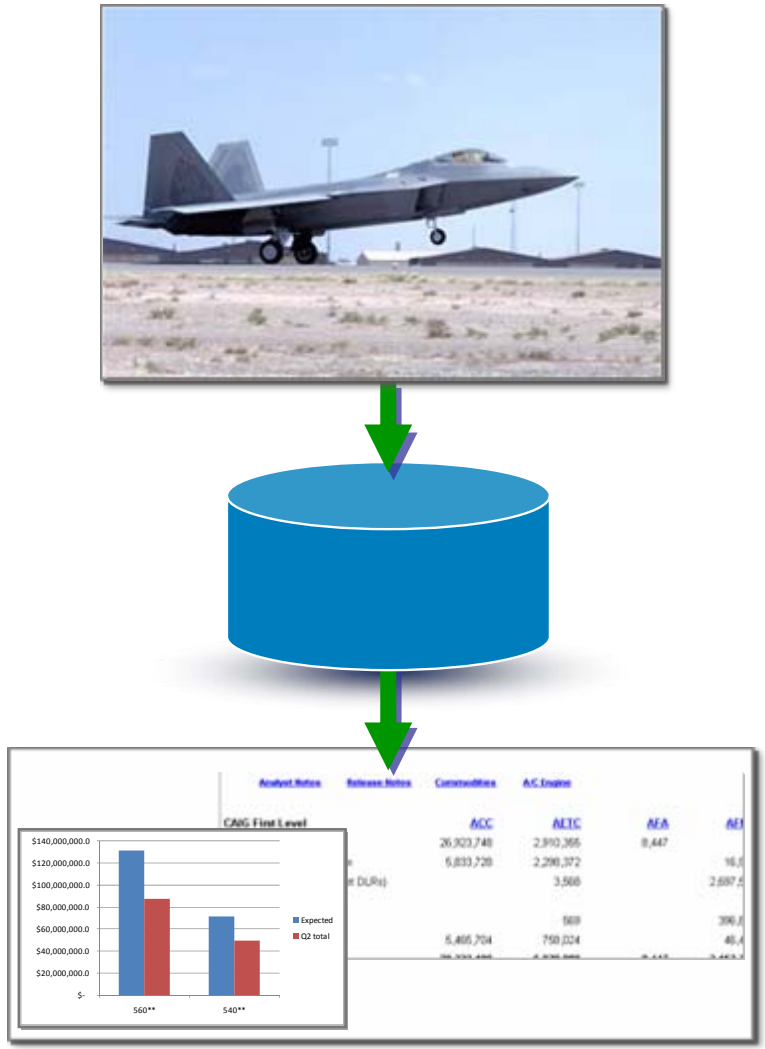
June 2010

Larry Brown

Session Overview

NORTHROP GRUMMAN

- Air Force Total Ownership Cost (AFTOC) program
- Impact of bad data
- A systematic approach to identifying anomalies
- Application of the approach to AFTOC data
- Application to other warehouses
- Recommendations
- Summary



Air Force Total Ownership Cost (AFTOC) Overview

NORTHROP GRUMMAN

- AFTOC provides detailed cost reporting for Air Force weapon systems and other activities
- AFTOC integrates data from many systems
  - Cost reporting – primary source
  - Budget – programmed costs and programmatics
  - Personnel – identify costs by job function (AFSC)
  - Supply – base detail, separate engine costs
  - Maintenance – Flying hour and inventory data
  - Fuel – more detailed than cost data
  - Others – ammunition, inflation, pay rates, etc.
- AFTOC stores the costs in data marts and provides
  - Standard reports
  - Online Analytic Processing (OLAP)
  - Ad-Hoc reporting
  - Special data requests

Who is AFTOC?

NORTHROP GRUMMAN

The AFTOC program team consists of four elements

Air Force Cost Analysis Agency – Program management, requirements, user support

Battelle
The Business of Innovation

Battelle, Prime Contractor – Cost analysis, database administration, data ingestion and processing

NORTHROP GRUMMAN

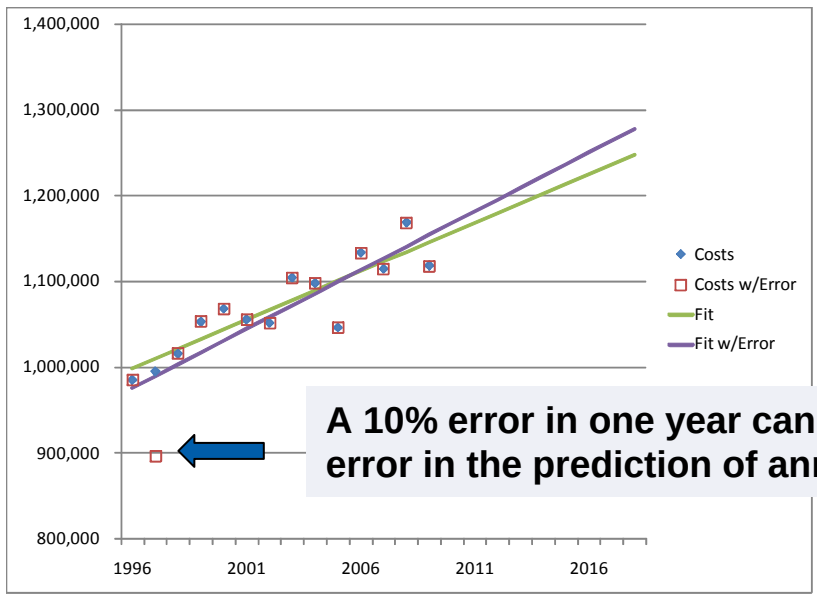
Northrop Grumman IS – Data mart design, reporting and analysis tools, data quality assessment

309 MXW (Hill AFB) – System administration and operation, user support

Impact of Bad Data

NORTHROP GRUMMAN

- Knowledge is the result of applying reasoning to data. Bad data will stress the reasoning and possibly degrade the knowledge
- More specifically bad data in a warehouse can
  - Lead to bad decisions. A small mistake in the perception of history may lead to a large error in the prediction of the future.
  - Increase the amount of effort to understand the data or corroborate it with other sources
  - Create the perception that the warehouse is so flawed that it is not helpful



A 10% error in one year can lead to a 20% error in the prediction of annual cost growth.

Typical Data Quality Approach

NORTHROP GRUMMAN

- Assess every data feed for number of records, total costs, etc.
- Look for duplicate records, leading or trailing blanks, etc.
- Validate every field in every input record before data is accepted for the warehouse
  - Is the state abbreviation valid?
  - Are the State, Area Code and Zip Code consistent?
- Package the extract, translate and load routines to ensure they run correctly and consistently

These are all good ideas, and there are good commercial tools to implement them.

The problem is that these techniques do not address overall consistency of the data.

If I fill 300 pages with correctly spelled words organized into grammatical sentences, have I written the great American novel?

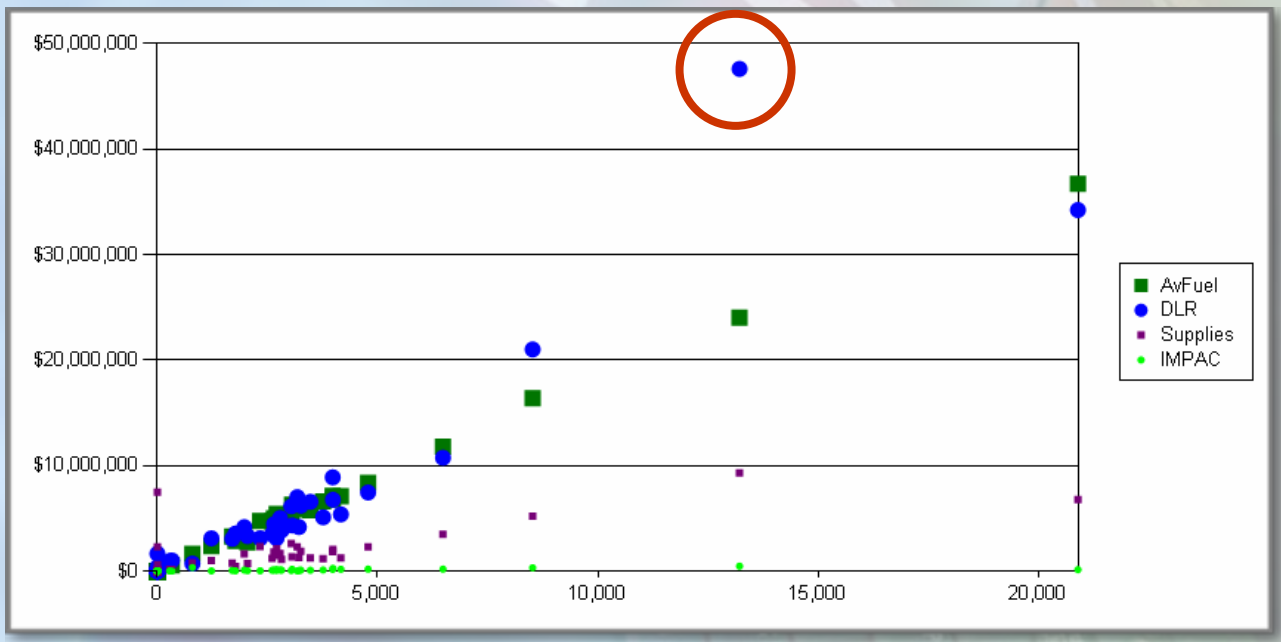
A Systematic Approach to Identifying Anomalies

NORTHROP GRUMMAN

- An anomaly is an unusual combination of data that most users could not easily explain
- Causes of anomalies
  - Bad, incomplete or duplicated source data
  - Processing mistakes
  - Bad logic for categorizing data
  - Changes in operations (the data is right, but surprising)
- The techniques described here can identify the anomalies. Determining the cause requires subject matter experts (often a scarce resource).

Examples of Anomalies

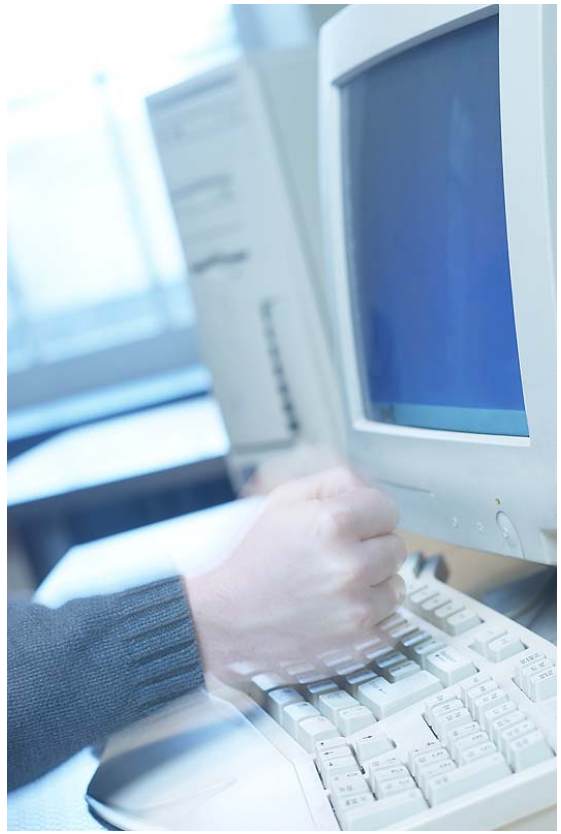
NORTHROP GRUMMAN



An Unsystematic Approach to Finding Data Anomalies

NORTHROP GRUMMAN

- Wait for the users to complain and try to fix their problems quickly
- Conduct Ad-Hoc investigation – "Fool around with the data" and report anything you find to be unusual. Problems
  - How do you know if you have "fooled around" long enough?
  - How do you know if your focused on the biggest issues

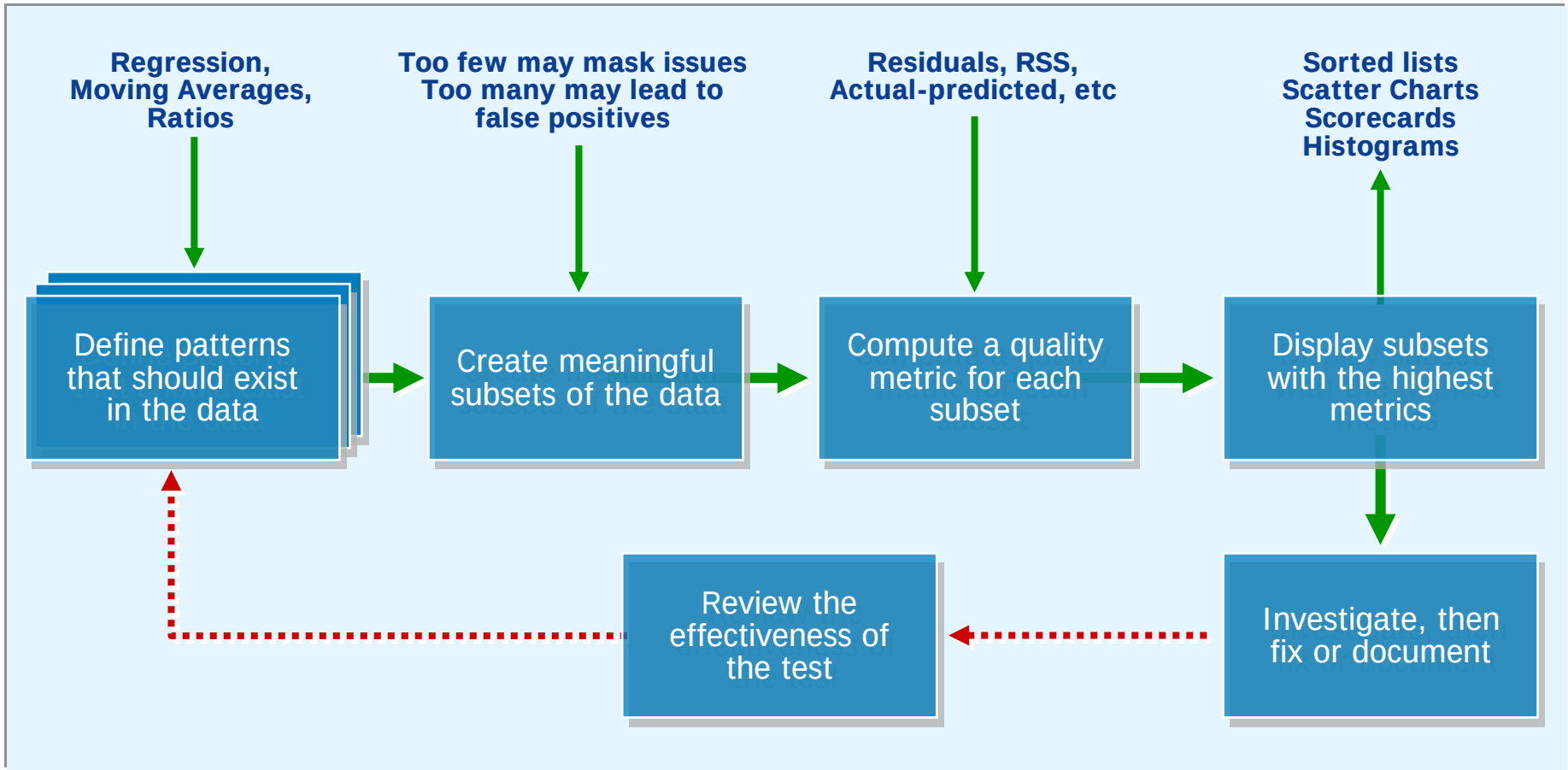


These approaches are costly and ineffective.

A Systematic Approach to Anomaly Detection

NORTHROP GRUMMAN

Define a series of tests using the this process



The Data Quality Challenge for AFTOC

NORTHROP GRUMMAN

- A lot of data
- Data entry errors
- Many different data sources
- Data sources change
- The Air Force changes
- Business rules and cross reference tables are used to derrive data attributes or allocate costs. For example, the cost reporting system does not
  - Identify the weapon system
  - Categorize military personnel costs into operations, maintenance, and support

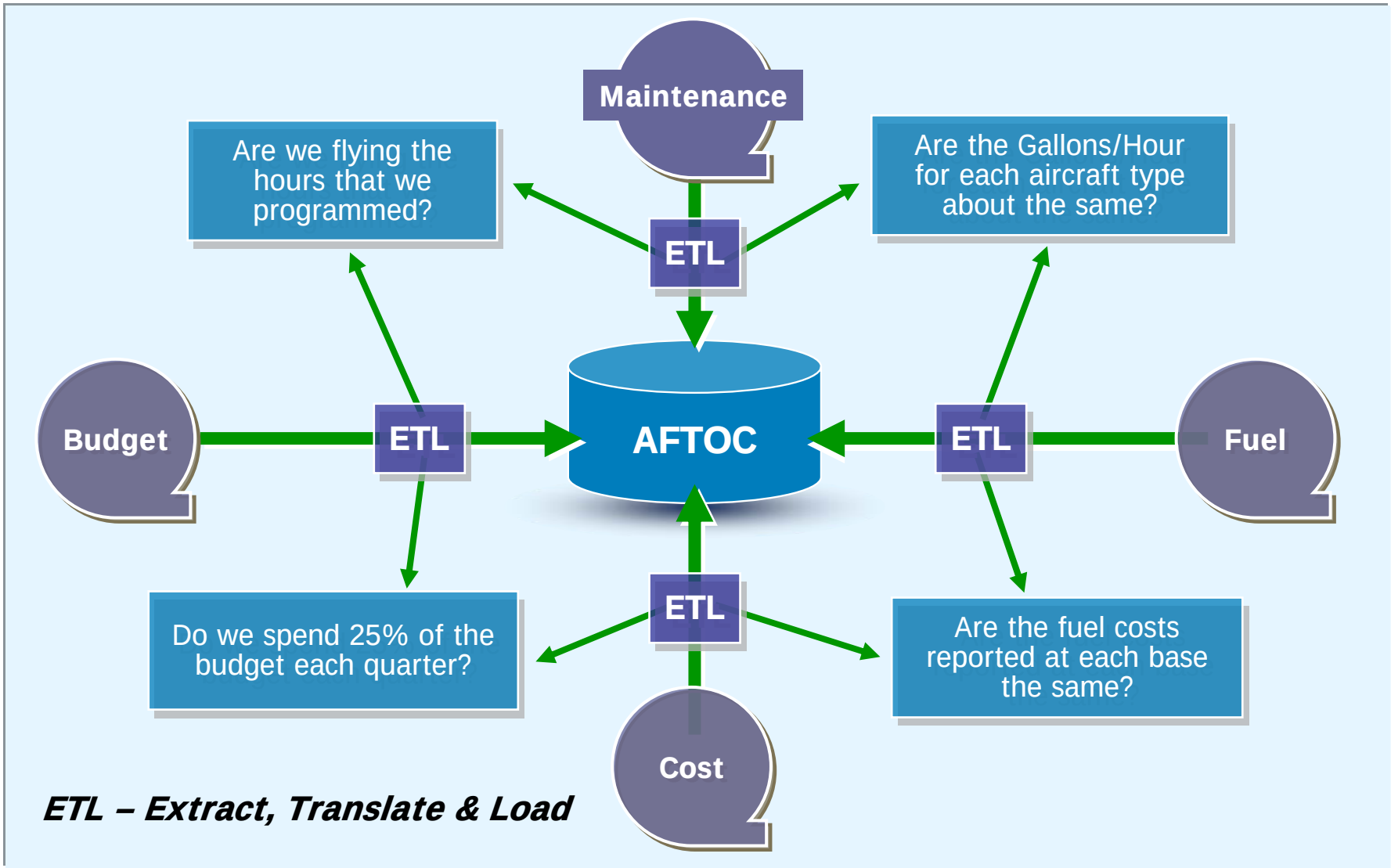
Application of Anomaly Detection in AFTOC

NORTHROP GRUMMAN

- AFTOC publishes most of its data quarterly (but the method can be applied to warehouses that publish data much more frequently)
- Before publication each quarter, 20 anomaly detection tests are run and the largest anomalies are investigated
  - When possible, errors are corrected or cost allocation rules are updated before publication
  - Unaddressed anomalies are documented for users
- The 20 tests fall into two categories
  - Are the data sources consistent with each other?
  - Is the processed cost data, viewed over time, consistent with itself?
- The tests are packaged SQL scripts and reports so they are documented and repeatable

Some Examples of Data Source Consistency

NORTHROP GRUMMAN



Assessing Consistency of Fuel and Flying Hours

NORTHROP GRUMMAN

Pattern – all aircraft of a given type at a given base should burn the same amount of fuel per hour

Subsets – All the tail numbers of a given type, base and year (3,000 subsets of about 15 points each)

Metric – the statistical correlation of gallons and hours for each subset.

Display – a scorecard format with color used to highlight low correlations with a link more detail

Air Force Total Ownership

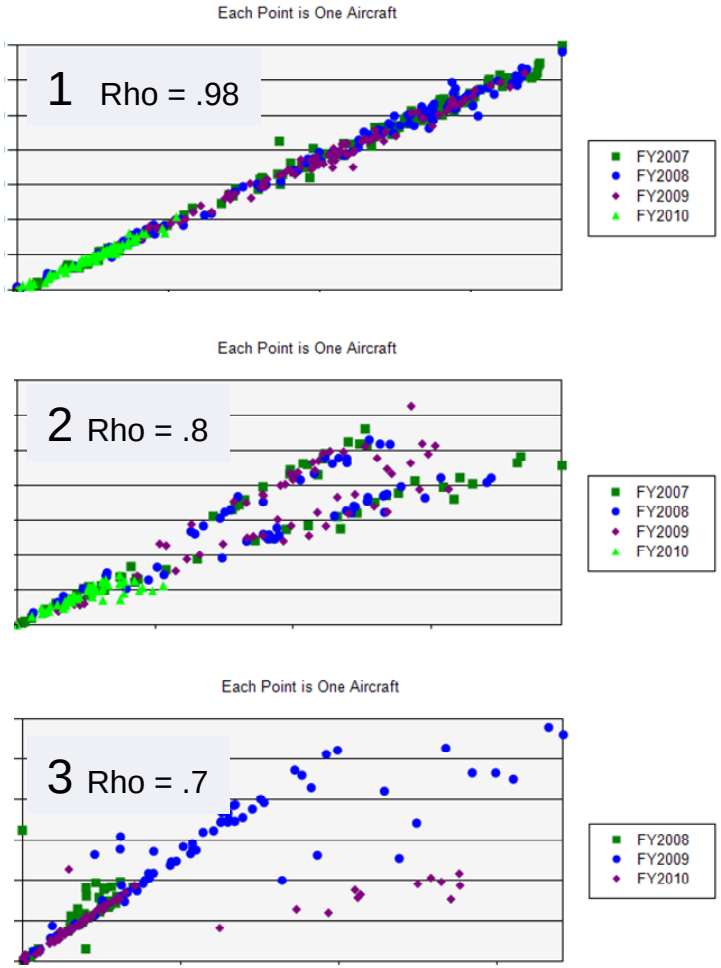
Fuel and Flying Hour Correlation

This scorecard contains the correlation between Fuel consumption and Flying Hours for each Tail Number. Click on a Base to see detail for that

		FY2003	FY2004	FY2005
Type 1	Base 1	0.079	0.998	0.976
	Base 2	0.991	0.975	0.976
	Base 3	0.961	0.996	0.969
	Base 4	0.957	0.999	0.990
	Base 5	0.981	0.972	0.882
	Base 6	0.753	0.957	0.956
	Base 7	0.993	0.942	0.969

Since AFTOC data is FOUO, specifics of aircraft type and base are not shown

Fuel and Hour Correlation Examples



These charts plot fuel gallons v. flying hours for one base and aircraft type each point corresponds to one tail number for one year. Rho is the correlation between Gallons and Hours. The value shown is an average for four years.

1.This base, with a high correlation in all years, shows the expected pattern. Almost all the points fall along line with a slope close to the published fuel factor.

2.This base, with a low correlation, shows an interesting bi-modal pattern. At this base, some of the aircraft were a newer model (block) with a higher consumption.

3.This base, with a low correlation, shows several tail numbers with low fuel reported. Aircraft from this base were involved in a series of air shows and refueled at commercial airports. Some of this fuel consumption has not (yet) been reported.

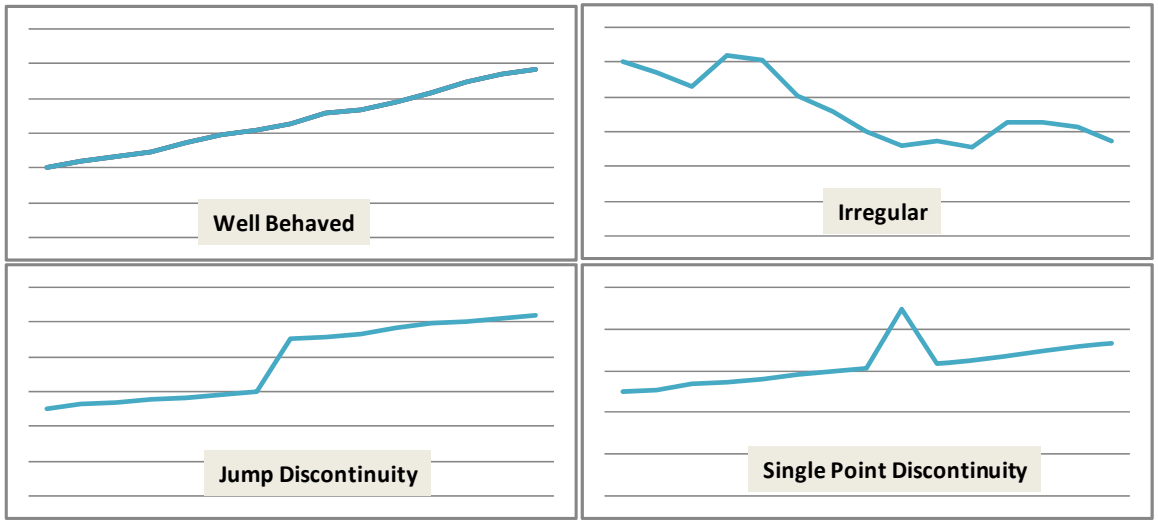
A systematic approach allows the analyst to focus on the most interesting cases.

Examples of Consistency Over Time

NORTHROP GRUMMAN

- Each quarter compare the reported costs with a prediction based on history
- Compare each release with the previous release. (AFTOC restates costs for completed years due to late cost reporting, adjustments and improvements in our business rules.)
- Look for single time points that are inconsistent with those around it. (details to follow)
- Look for “jump discontinuities” in time series

Types of Time Series



- The “Well Behaved” is the ideal.
- The “Irregular” is usually not a data quality problem. This pattern occurs for small fleets of aircraft or certain cost categories such as depot maintenance.
- The “Jump Discontinuity” is possibly
  - A change in source systems, use of codes, or AF policy
  - Different processing of the cost and budget data sources
- The “Single Point Discontinuity” is possibly
  - A bad attribute in the data
  - A bad cost record

"Single Point Discontinuities"

NORTHROP GRUMMAN

Pattern – Operating and support costs, adjusted for inflation, for an aircraft type and cost category should be continuous linear function of time and flying hours

Subsets – approximately 700 - one for each aircraft type (100 types) and cost category (7 categories). Most subsets have 14 years of history.

Metric – the most improvement in a regression fit to the time series if one point is removed.

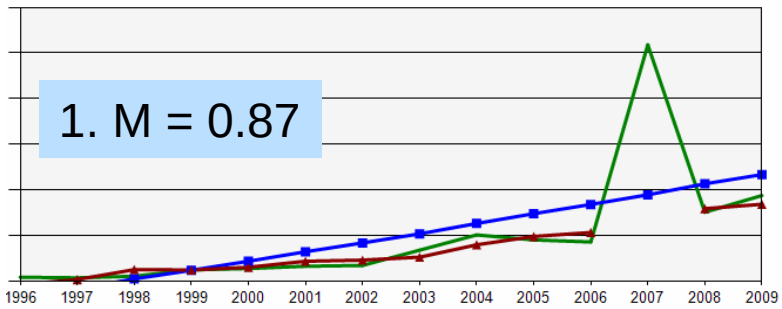
$$M = 1 - \sqrt{\min_i Ri / R}$$

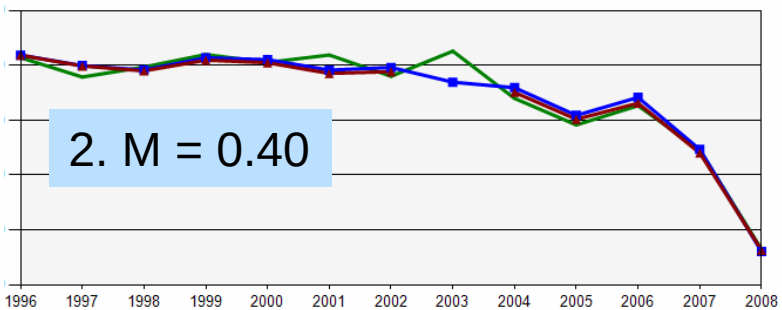
Where R is the sum of the squares of the residuals and R_i is the sum of the residuals when the i^{th} point is removed from the sequence.

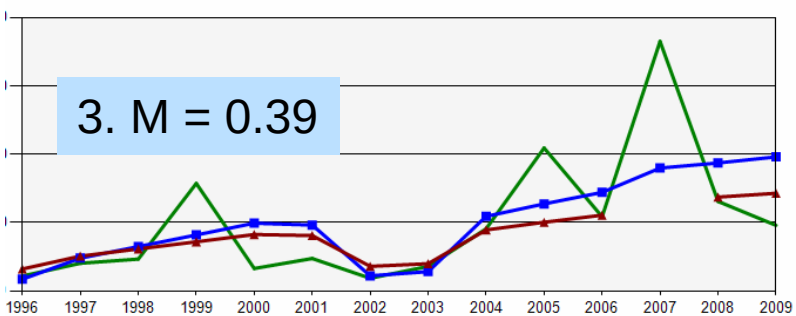
Display – a sorted list of metric values with a "drill down" to details of the regressions that lead to that value of the metric

One Bad Point Examples

NORTHROP GRUMMAN







In these plots, the green line shows the costs, after AFTOC processing, for one aircraft in one cost category. The blue line is the regression to the full set of data. The red line is the regression with one point removed.

1.High metric - removing the costs for 2007 dramatically improves the fit. This is the type of time series the test is designed to find.

2.Low metric – The fit to the full interval is quite good removing one point does not help dramatically.

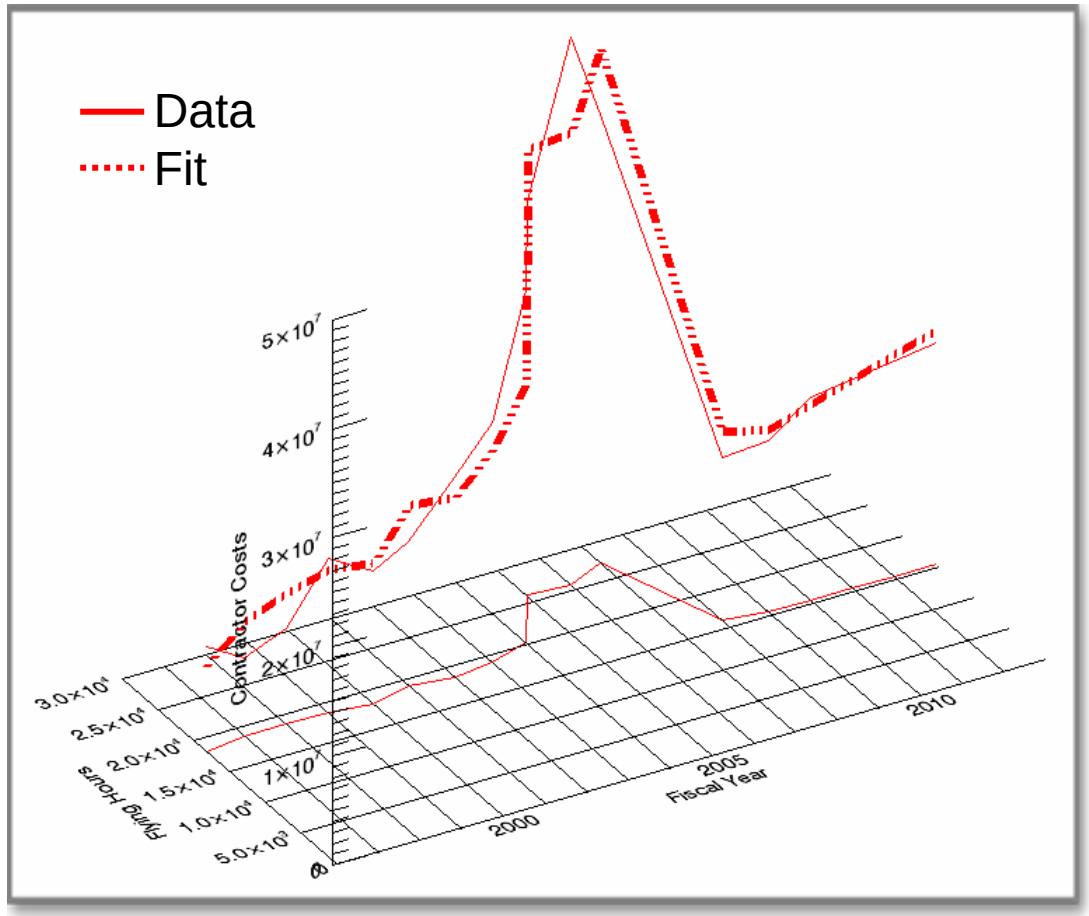
3.Low metric – removing any single point still leaves an irregular time series.

A systematic approach allows the analyst to focus on the most interesting cases.

Why not just look at the graphs?

NORTHROP GRUMMAN

- The option of trying to understand a thousand of these graphs each quarter is challenging
- 3-D graphs need to be viewed from at least two different perspectives
- What is needed is a metric to find the worst cases.



At first glance, this looks like a big anomaly, but there is a large drop in flying hours, so the fit is quite good.

Application to Other Warehouses

NORTHROP GRUMMAN

- There are many competing goals for warehouse resources such as: performance, reliability, new features, and data quality
- The patterns (models) in the data and the metrics used will be unique to the data. For example
  - "Seasonality" of data may be important for a retail data warehouse
  - Long term trends may not be relevant (does data from 1920's help show today's stock market was reported accurately?)
- There is a trade-off between getting the data right and getting it published in a timely manner.
  - This tradeoff will be different for each warehouse
  - In some cases, it may be necessary to update the warehouse before all the quality issues are analyzed.
- Everyone who operates a data warehouse should be able to address the question "How accurate and complete is the data?" Although the answer is never simple, the techniques presented here provide a partial answer.

Recommendations for Other Warehouses

NORTHROP GRUMMAN

- A data quality program needs three approaches
  - Field and record validation of all inputs
  - Process controls over the extract, translate and load procedures
  - A structured approach to look for anomalies in the data
- Data quality needs to become part of the culture of operating a data warehouse

Am I running a data warehouse full of precious data that is easy to find, use and understand?

- Or -

Am I running a data "attic" with a jumble of old and useless data?

Summary

- Data quality in data warehouses is a significant problem and an interesting challenge
- Improving data quality requires
  - Dedication to quality
  - a structured approach to focus on the largest problems
  - Collaboration to analyze and fix problems
- There is a payoff for effort required
- The AFTOC team is proud of quality of its data