THE USE OF CLUSTER ANALYSIS TECHNIQUES IN SPACEFLIGHT PROJECT COST RISK ESTIMATION

G. Fox, D. Ebbeler and E. Jorgensen
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, California

ABSTRACT

Project cost risk is the uncertainty in final project cost, contingent on initial budget, requirements and schedule. For a proposed mission, a dynamic simulation model relying for some of its input on a simple risk elicitation is used to identify and quantify systemic cost risk. In addition, we display in a simple tree the relationship of the new proposed mission to historical missions. The tree is constructed by cluster analysis, a data mining technique for identifying similar objects based on a simple measure of the similarity between two missions. The cost history of missions identified by cluster analysis as closest to the proposed mission are used to modify the systemic cost risk. A case study demonstrates this approach to the early evaluation of total cost risk for proposed spaceflight projects.

INTRODUCTION

Cluster analysis is a data mining technique for identifying similar objects, represented as records in a database of common attributes. We classify and display, in a simple tree, the relationship of a new proposed mission to historical missions in a database of system attributes such as spacecraft mass, development duration and cost and mission operations period. Missions are grouped by defining, then comparing, the separation between sets of missions, based on a simple measure of the similarity, or distance, between missions. This grouping is used to identify missions similar to the proposed mission. For the proposed mission, simple cost risk elicitation is used to identify and quantify any systemic cost risk, i.e. potential cost changes that arise from project tasks, not from external causes. The historical missions identified by cluster analysis as closest to the proposed mission are used to estimate the ratio of total cost risk to internal cost risk and this factor is used to modify the internal cost risk simulation of the proposed mission. The following sections present our cost risk and cluster analysis methodology and a simple case study illustrating their use.

ESTIMATING COST RISK

The traditional NASA cost estimation process used for early assessment of cost risk for proposed spaceflight projects is based on historical data and uses simple regression models with system level characteristics, such as mass and peak power supplied, used as explanatory variables to estimate final costs and model uncertainty. This approach implicitly assumes a static world, where the only source of data is typically at mission completion. Costs are adjusted for average resource inflation to correct for chronology, but the dynamic evolution to final cost is not modeled. Both the proposed mission and missions used in the historical database are assumed to be homogeneous with respect to explanatory variables and are assumed to have the characteristics of a random sample of missions selected from a hypothetical universe of feasible missions. None of these implicit assumptions have real world validity. Furthermore, cost risk is taken to be the uncertainty in model fit. In fact, that measure captures observational error, confounding effects among mission characteristics and model misspecification. It cannot be expected to be a credible predictor of cost risk for a proposed mission.

Historical cost and performance data for space missions and spacecraft are typically extracted, analyzed and revised after a program is completed. Data collection takes place during actual development and operation, but it is rare that a sufficiently complete cost history is maintained and retained. This lack of necessary information prevents explicit explanatory modeling of dynamic project cost. Another source of information loss with historical data is that a project WBS is often based on performing organization structure rather than the subsystems, components, tasks and their interfaces. This hides the true realization of component and task cost and the relationships among them. Information about elements common to more than one subsystem is typically recorded arbitrarily in any one of the common subsystems,

so correct assignment of cost is further obscured. This can be a significant source of error when using technical parameters at the subsystem level to explain system cost. Due to the unique nature of spaceflight missions and the evolving technology employed, the cost and performance tradeoffs among subsystems in a new mission are rarely duplicated from historical missions. It is precisely these potentially synergistic trades which reduce total system cost. Many spaceflight projects evolve over decades, from initial proposed concept to completion of operations; inflation is not uniform in its effects on resource costs over time, especially for these long periods. The varied pace of technical progress is another confounding factor, e.g. Moore's Law, that invalidates static models. Even if these confounding effects, typical of historical cost data, were to be absent, there is a fundamental problem with using such data for predicting spaceflight project cost risk. At best, historical cost observations, taken to be final cost values, capture the historical development path dependence of cost for a project. Except under the uninteresting hypothetical circumstance where (1) No missions in the cost database experienced unusual historical development anomalies with respect to cost, (2) The database is free of confounding effects and (3) Systemic effects, after observed costs are corrected for externalities, are homogenous, *historical cost model CERs (with model uncertainty) alone cannot be used to predict cost risk for a proposed spaceflight project.*

True, *a priori,* project cost risk is the uncertainty in final project cost, contingent on initial budget, requirements and schedule. This pressure on cost results from the dynamic interaction of schedule and technical performance risk and the evolving funding negotiation and commitment environment, i.e. the institutional constraints under which project development and eventual mission operation unfold. The current computational power of desktop computing opens the door to the use of inexpensive but realistic simulations of this dynamic project cost generation process. We have developed just such a methodological framework with two goals in mind: (1) The data requirements should be the minimum necessary to support realistic cost and cost risk estimates, (2) The models should be transparent to the user, i.e. effects should be traceable to causes. We have developed a cost risk model, consistent with those goals, that relies on parsimonious input information, yet captures the dynamic aspects essential to characterizing systemic spaceflight project cost risk. The modeling information required is (1) a high level Work Breakdown Structure (WBS) with budgets assigned to each task and, (2) for each task with significant risk, a pessimistic cost estimate that is elicited from knowledgeable technical experts. The experts for each task are asked to provide a 99th percentile cost estimate for the task, called the pessimistic cost, conditional on the initial budget, schedule and required technical performance. All tasks are assumed to finish on schedule, all receivables and deliverables are supplied and the required technical performance is achieved. This information is input to a simple dynamic model of cost commitment evolution that is evaluated by Monte Carlo simulation. Thus, the information developed in the initial budget estimation and plan development is used to propagate potential cost changes to a final cost probability distribution function (S-curve). The initial budget is relevant only to the extent that it is a starting point for evolving project cost commitments. Alternative initial budgets can lead to different final cost S-curves. It is useful to perform this budget sensitivity exercise in order to gauge the impact of the initial budget on cost risk. The basic dynamic model has been extended to include (1) the impact on cost risk of initial budget misspecification, where available funding is significantly lower than that necessary for successful project completion, and (2) project schedule slips resulting from intertask schedule precedence relationships and uncertain task completion dates. That extended model is presented in another paper in this volume, "Dynamic Cost Risk Estimation and Budget Misspecification", by D. Ebbeler, G. Fox and H. Habib-agahi.

Pessimistic task cost estimation by knowledgeable technical experts drives the systemic cost risk estimate. In this paper we will show how cluster analysis of the attributes of a historical mission database can be used to conjecture a total cost risk S-curve for a proposed mission, given the S-curve characterizing systemic risk for the proposed mission.

CLUSTERING A DATABASE OF HISTORICAL MISSIONS

For our cluster analysis inputs we used a data base of 20 historical and ongoing JPL missions with 16 mission attributes including programmatic variables such as development phase (A, B, C/D, E) durations, costs and JPL labor work-years, plus the technical variables dry mass, launched mass and cruise duration. These 16 were winnowed down from a much larger set by a process of eliminating fields with missing values and fields deemed highly correlated with the 16 remaining variables. No subsystem information was included in the resulting database. Table 1 displays the final mission attributes selected.

Selected Spaceflight Mission Attributes

Phase A Duration (Months)	Phase A/B Cost ($M-FY01)	Phase A/B JPL Labor (wk-years)	Cruise Duration (Years)
Phase B Duration (Months)	Phase C/D Cost ($M-FY01)	Phase C/D JPL Labor (wk-years)	Dry Mass (kg)
Phase C/D Duration (Months)	Phase E Cost ($M-FY01)	Phase E JPL Labor (wk-years)	Total Launched Mass (kg)
Phase E Duration (Months)	Phase (A-E) Cost ($M-FY01)	Total JPL Labor (wk-years)	Total Cost incl. Launch ($M-FY01)

Table 1. Spaceflight Mission Attributes used in Clustering

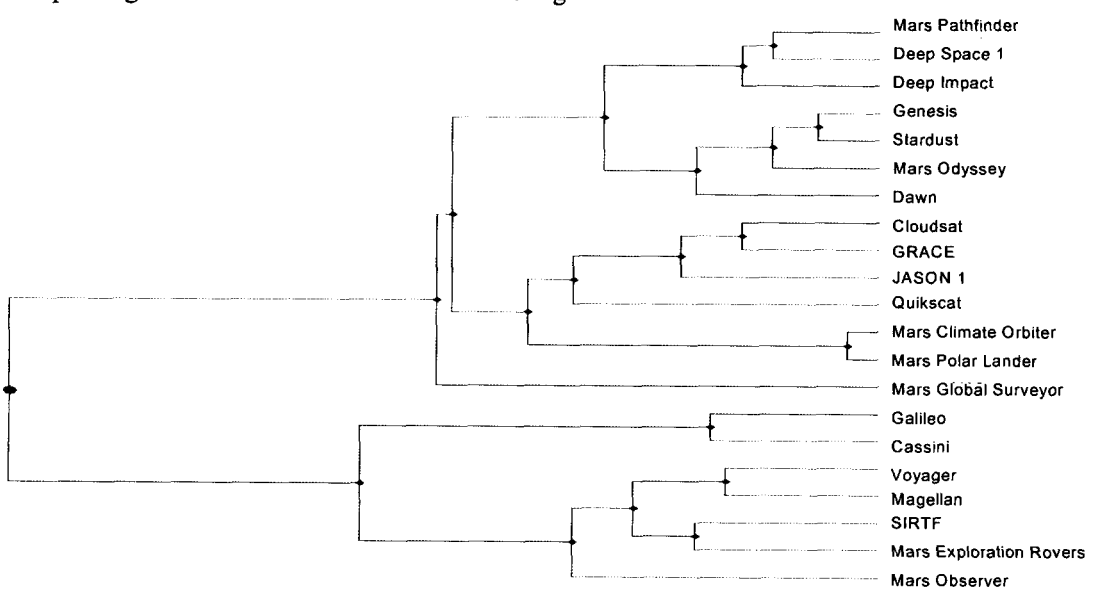

Figure 1. Phylogram Format for Spaceflight Mission Clustering

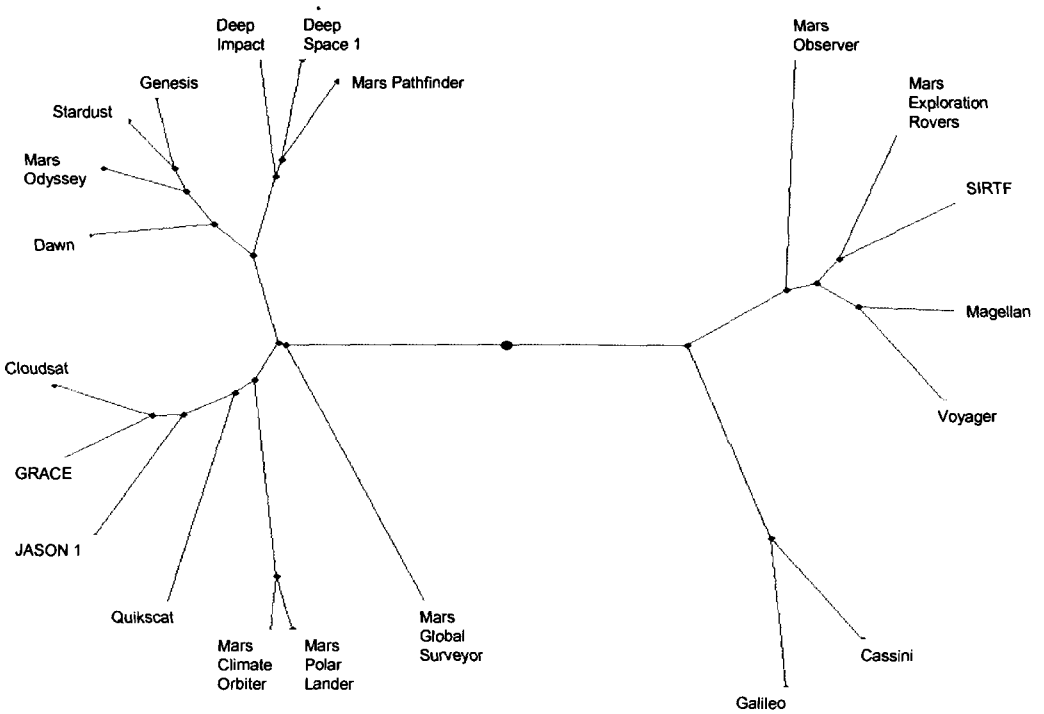

Figure 2: Unrooted Tree Format for the Same Spaceflight Mission Clustering

Clustering requires a similarity metric to be defined between objects; we use a dissimilarity measure to the same effect. We selected a Euclidean metric in the 16-dimensional ***log*** space (real numbers) of the attribute values. The metric is also referred to as the distance between missions.

$d(x,y) = (\Sigma_A\, (\boldsymbol{log}(A_x/A_y))^2)^{1/2}$
where x and y are missions, A_x and A_y are attribute A values for missions x and y (***log*** is the natural logarithm)

Like the Euclidean metric in the 2-D plane or the 3-D space we inhabit, the distance is always non-negative. The ***log*** transformation acts approximately like taking the percent difference between attributes and using that as the distance between them, e.g. a mission with a 5% higher cost than a reference mission is closer to the reference mission than is a mission with a 10% higher cost than the reference mission. For small deviations the effect is linear when using the ***log*** function. Between attributes, the Euclidean metric enforces an egalitarianism between dimensions, i.e. a 10% distance in one attribute dimension has the same impact as a 10% distance in another.

There are a number of different clustering methods, each with its own special effect on the distance matrix that translates into different cluster characteristics. In our case we chose a method called *Complete Clustering* to define the distance between clusters, i.e. the distance between two clusters is the maximum distance found from all possible pairs of elements, one selected from each cluster.

$d(S,T) = \max d(x,y)$
where $x \in S$ and $y \in T$, S and T are clusters.

Complete Clustering results in clumps of bushy clusters rather than the long strings of missions obtained in *Single Link Clustering*.

There are many commercial and public domain tools available to do the clustering, using as input the distances between missions. In addition, many statistical packages can execute the *Simple* and *Complete Clustering* algorithms. We used **Data Desk 6.1** for the clustering and **PhyloDraw** (Department of Computer Science, Pusan National University, Pusan, Korea) for the display of cluster trees.

A DYNAMIC COST RISK MODEL

Our approach to modeling cost risk is to begin with the functional decomposition of the mission development and operations into the activities used in the construction of the initial budget. An associated cost probability distribution is generated for each activity in that decomposition. Cost risk is

characterized by the cost probability distribution formed by combining cost probability distributions for individual activities.

The knowledge and experience of experts is used to ascertain the activities for which cost volatility needs to be considered and those that can be approximated as fixed. For each of those activities that contribute to cost risk, we will construct a model of systemic risk constrained by the NASA principal-agent management structure, contingent on a correctly specified initial budget. When total cost risk is to be estimated, the impact of external shocks is treated as a multiplier of the systemic cost realization of each relevant WBS element calculated during Monte Carlo simulation of the model.

For each activity cost risk element, for a fixed schedule and performance requirements, systemic cost uncertainty is the consequence of a sequence of random events whose cost impacts can be described by a time dependent stochastic (Wiener) process; that is, incremental departures from a correctly specified initial activity budget are normally distributed. Under the NASA principal-agent management structure, the rational behavior of project management reacting to such events implies that systemic costs will never fall below the initial activity budget. The Wiener model with this constraint produces a systemic cost evolution equivalent to a random walk with a non-absorbing boundary constraint. It can be described by the stochastic difference equation

$dC(t) = a \cdot C(t)dt + s \cdot C(t)dw$ with constraint $dC(t) \geq a \cdot C(t)dt$ for $t \in [0,T]$

Choice of the units for t is application dependent. Assuming t is measured in years,

C(t) = predicted cost at time t in year t dollars
C(0) = initial activity budget estimate in base year dollars
a = inflation rate (%/year)
s = volatility parameter (%/year)
T = activity duration (years)
dw = a random variable distributed N(0,dt), i.e. dw is normally distributed with mean zero and variance dt

In order to determine the volatility parameter for the Wiener process corresponding to an activity cost evolution, it is mathematically sufficient to elicit an estimate of any percentile of the potential cost overrun for that activity.

For each activity with volatile cost, it is proposed that an estimate of the 99th percentile cost overrun for that element, conditional on a schedule and nominal budget, be elicited. The estimates are elicited from cognizant individuals for each such activity. The 99th percentile means that the cognizant individuals judge, given a schedule and nominal budget that do not change, that there is only a 1% chance of exceeding the elicited value. Elicitation is preferred because our focus is on capturing past experience together with knowledge about the specific task and any conditions that may make the anticipated bounds differ from the past. 1% was selected because it is sufficiently large that the corresponding cost overrun is still of plausible likelihood of occurrence. The Wiener process volatility parameters are determined empirically by calibrating the Monte Carlo simulations so that each simulated cost 99% value is approximated by the corresponding elicited 99% value.

ILLUSTRATION FOR A PROPOSED SPACEFLIGHT MISSION

Two proposed missions were analyzed for total cost risk using cluster analysis to determine potential analogy missions using the following steps. A development plan, including task budgets and schedules, was provided to us by each proposal team. This is a standard product developed by all proposal teams. Next, after determining the appropriate WBS level to estimate, pessimistic costs for each significant risk task were elicited from the proposal team using explicit instructions we developed and provided in an Excel worksheet.

The proposal team developed values for the 16 programmatic variables in Table 1 above for each proposed mission. Most, if not all, of these values will have been identified when developing the program plan. Clustering is then performed on the expanded database. Figure 3 shows the results of clustering the two proposed missions we wished to evaluate. It's useful to compare the old cluster tree without the new missions to the new tree to see if any clusters have changed. Major changes may indicate an anomaly in the new or existing missions, which should be investigated further.

Figure 3 seems to show that our two proposed missions are probably minor variations on a theme.

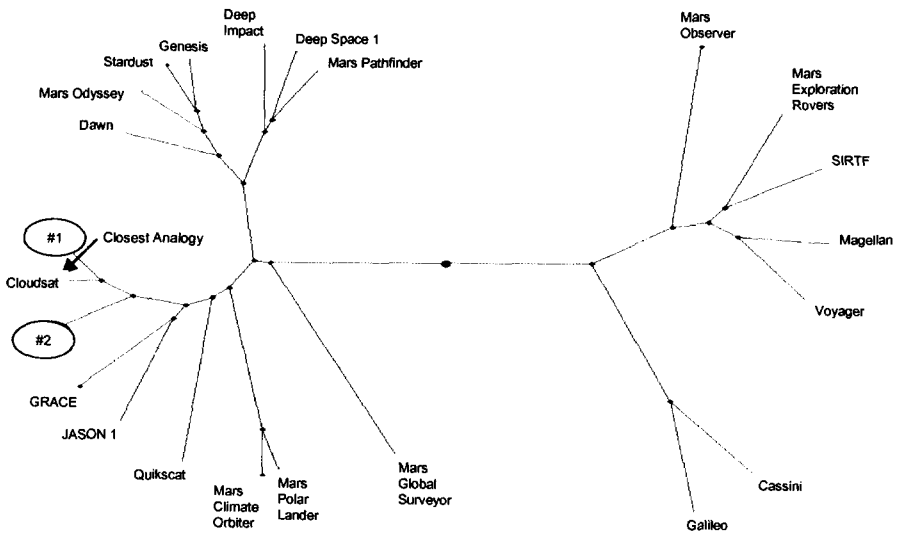

Figure 3. Cluster Tree of Spaceflight Mission Database and New Proposed Missions

Cluster analysis identified a mission, labeled in Figure 3 as the "Closest Analogy", that is near the two proposed missions. The quantitative analysis charts in Figure 4 show the normalized values of the 16 attributes for the analogy mission and the 5 additional missions in the large cluster in the lower left quadrant in Figure 3.

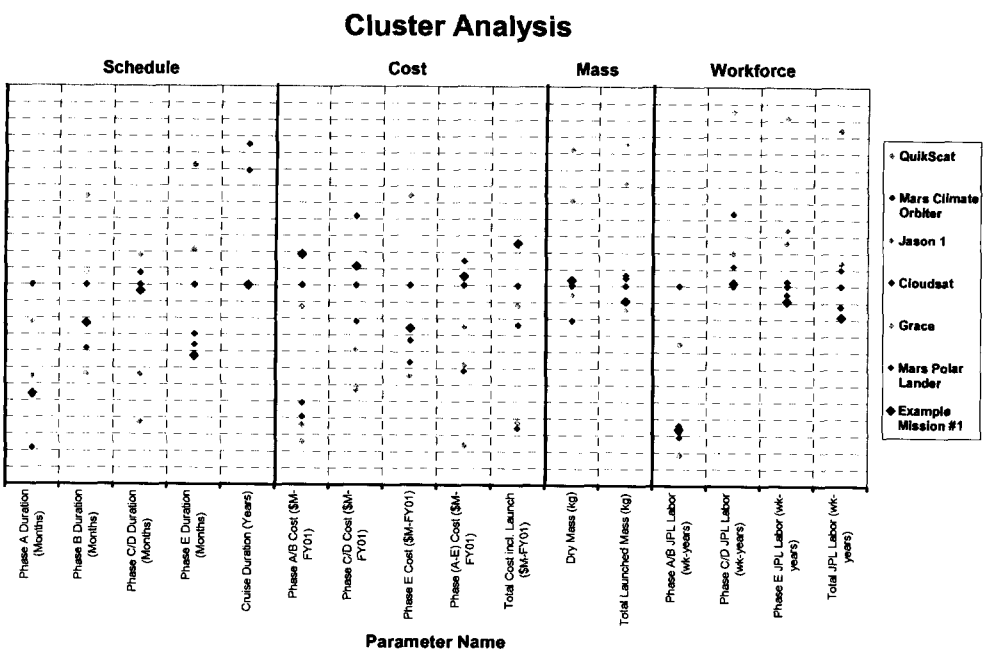

Figure 4. Quantitative Comparison of Attributes for Example Mission and Analogous Missions

In the case of the closest analogy mission, the total cost overrun was attributed to a single WBS element and the external overrun was 60% of that total cost overrun. We scaled the cost risk distribution for the relevant WBS element in example mission #1 to calculate total cost risk and reran the analysis, the results of which are illustrated in Figure 5. A similar figure was also generated for example mission #2, but is not included in this paper.

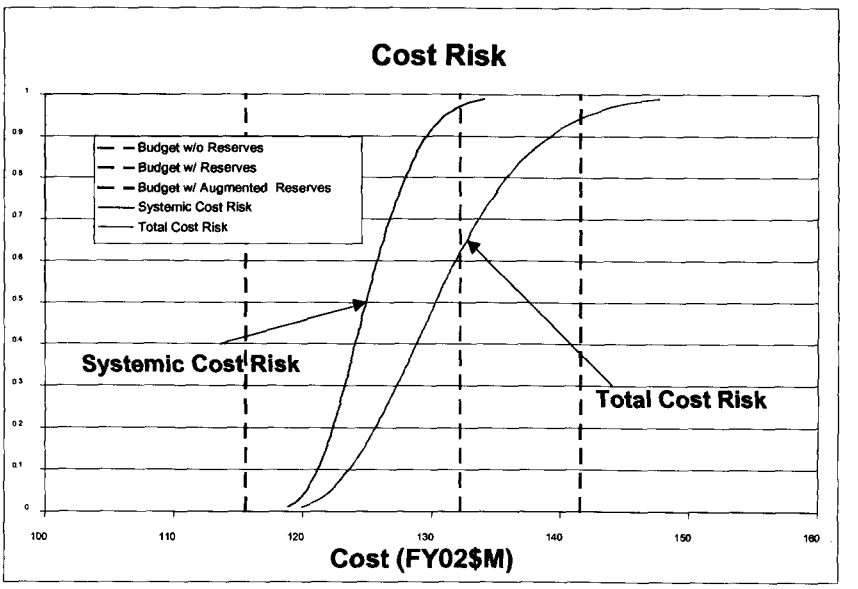

Figure 5. Systemic and Total Cost Risk for Example Mission #1.

CONCLUSIONS

Our example relates systemic cost risk to total cost risk, which includes the effects of potential external cost risk. Projects should not, however, be held accountable for risks beyond their control. The information provided by the total cost risk S-curve is an early warning to the funding agency of potential cost overruns and could be useful in its higher level planning which includes other projects and programs competing for scarce funds.

As illustrated by our example, cluster analysis can be a useful addition to the cost analyst's toolbox, used to identify and estimate spaceflight project cost risk based on a historical database of past missions.

ACKNOWLEDGMENT

The research described in this publication was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration Cost Analysis Division.