DATA, DEFINITION, DEDUCTION:

AN EMPIRICAL VIEW OF OPERATIONAL ANALYSIS

Springer Cox

Digital Equipment Corporation
146 Main Street
Maynard, MA 01754

The theoretical aspects of operational analysis have been considered more extensively than matters of its application in practical situations. Since its relationships differ in their applicability, they must be considered separately when they are applied. In order to do this, the foundations of three such relationships are examined from an empirical point of view.

To further demonstrate the intimate connection between data, definitions, and performance models, the problem of measurement artifact is considered.

1.0 Introduction

A continuing discussion of operational analysis has appeared in the literature, most recently in [2]. For the most part, the theoretical structure of the methodology has received more attention than empirical matters. In the developments which follow, we consider the relationship of operational analysis to its applied environment.

Our discussion weaves around three performance relationships that have been included in operational analysis as defined in [1]. These relationships have been named the interactive response time formula, the utilization law, and and the class of homogeneity assumptions. Together they provide a large part of the power offered by operational analysis. We then proceed to a brief discussion of the relative testability of operational and stochastic models. We conclude with a consideration of some of the consequences of artifact, the contamination of data by its measurement.

2.0 Operational Analysis

Operational analysis consists of relationships of different strengths and applicability. For this reason it is useful to consider them separately in practical applications. In the present treatment we will restrict ourselves to the interactive response time formula, the utilization law, and the class of homogeneity assumptions. The reader is assumed to be familiar with these.

2.1 The Interactive Response Time Formula

The interactive response time formula is an identity which may be applied to any collection of sets of real numbers. For our

purposes, we consider a finite collection of finite sets. When we construct an interpretation of these data, each element represents a value of time which forms both the beginning of one period (response or think) and the end of another. Each set can be associated with a class or source of work units--usually a terminal or initiator. In a sense, each timestamp represents a penetration of a theoretical system boundary[3].

For each set we add and subtract all the timestamps from the elapsed time. We then can identify individual responses and think times by pairing the differences. By adding all such equations and carefully defining the mean think time and the mean response time, we arrive at the interactive response time formula.

This development of the interactive response time formula is superior to direct application of Little's result when considering typeahead with negative think times. In addition, it illustrates the looseness of the connection to the measured system.

Accuracy and precision are not issues at this point. In the derivation of the interactive response time formula we need only be concerned with the numbers themselves, not what they represent. By using different measurements of the same system at the same time, a new collection of real numbers may result which still follows the interactive formula. These new measurements could represent the reality of different measurement instrumentation. They could represent intrinsic variability or erroneous measurements. The interactive response time formula still holds because we have based the definitions of all external performance metrics on this single collection of sets of real numbers. There is no requirement for any connection to a specific system for the indentity to hold.

In practice, the interactive response time formula is extremely useful. Its most immediate application is to check the internal consistency of a set of performance data. There have been many such data collection errors identified in the past. In performance prediction it can be used to rule out many impossible situations. But it is weak when used alone.

It is a sign of both versatility and weakness that data collected from any measurement instance must always satisfy this relationship. The predictive weakness follows because although any future measurement instance must also fall into line, the set of possible external metrics is not constrained in a very useful way. It is true, however, that sometimes inconsistent results can be ruled out. For example, we may have information on some of the variables in the formula. We then can make deductions with respect to the others. But in general, details regarding the structure of the system must be added in order to predict system behavior.

2.2 The Utilization Law

The utilization law differs from the interactive response time formula in both foundation and applicability. When supplied with an auxiliary assumption, it is probably the most widely applied performance prediction technique. The usual auxiliary assumption is that "no resource can be busier than all the time." Although this

may seem axiomatic, when dealing with measured data one sometimes finds counter-examples arising from approximations and errors. Generally, though, a strong inequality is imposed on each measured utilization. These constraints are then connected to the external performance metrics by assuming that the mean service time required per work unit can be predicted. This connection can often be made easily and reliably. We then have a strong statement about feasible behavior of the external performance metrics. In fact, these constraints force performance models which follow the conservation of work to fall into the same ballpark of accuracy. This is not to say that model accuracy is always sufficient for the objectives at hand. But it does explain somewhat, how the simplest models can deliver surprising accuracy.

2.3 Homogeneity and Testability

Assumptions of homogeneity provide another pathway to predictive power. When we use them we are led to the same predictions we find with the class of queuing network models with product form solutions. Unfortunately, it is most unlikely that such assumptions hold precisely in a given measurement instance.

Also, it is known that this class of models is excellent in some applications, and inadequate in others. For example, experience from queuing network models indicates that homogeneity assumptions may be inappropriate when the physical memory of a system is overcommitted or when requests are blocked in an I/O subsystem. Therefore our deductions are weaker than those based on the relationships discussed above. The question then is what inaccuracy are we led to by assumptions of homogeneity. In a given situation, we may choose not to make such assertions.

Let us turn briefly, now, to stochastic models. Surely, no measured computer system is an exact instance of a continuous time stochastic model. Finite precision and several other reasons argue otherwise. The question then, is how accurately can future measurements be predicted by the model.

Even without assuming an overall stochastic mechanism to explain system behavior, it is useful to account for measurement variability by treating observations as random variables. We then have at our disposal powerful statistical methods such as Analysis of Variance and Regression Analysis [4].

When we test a stochastic model we determine if the observed results are predicted to represent a rare event. If so, the model is discarded. If not, we say that the data do not contradict the assumptions used in building the model. We do not prove that the model exactly represents the real system. The model can achieve credibility only as it is used. We must quantify both accuracy and domain of usability.

The testability advantage of the homogeneity assumptions is a moot point. In principle, we can test the assumptions by measurement, but in practice massive amounts of data must be collected--thereby perturbing the system.

We must keep in mind that we are trying to predict performance. When applied in performance prediction, we can test, in principle,

assumptions such as homogeneity if we have access to a real system. But even if we can confirm them in an existing environment, we still must reassert them in the untested environment.

Just as the existence of a stochastic mechanism is never proved empirically, the confirmation of homogeneity in one measurement instance does not prove that it will occur in any other, including a replication of the first. Even in operational analysis these generalizations are inductive. Once we have made or extended an assumption of invariance or homogeneity we may proceed deductively in the predicted environment.

As far as proving the existence of a specific mechanism in the real system, neither modelling approach is testable. Credibility is established only by successful application.

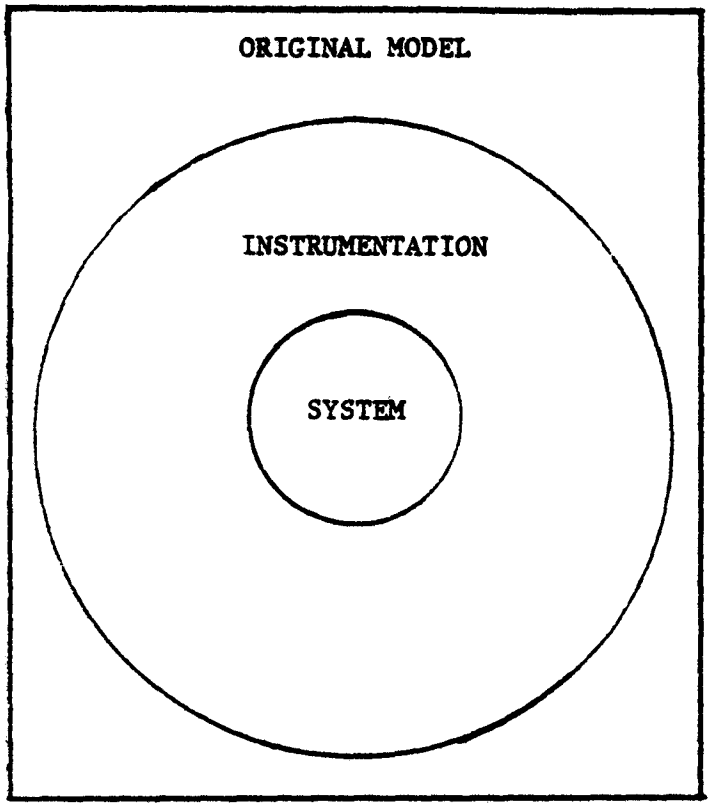
3.0 Artifact

There is a surprising degree of lattitude that exists when measurement instrumentation is implemented. What was thought to be a well-defined theoretical entity often leads to a choice of several implementations -- none of which is exactly what is desired.

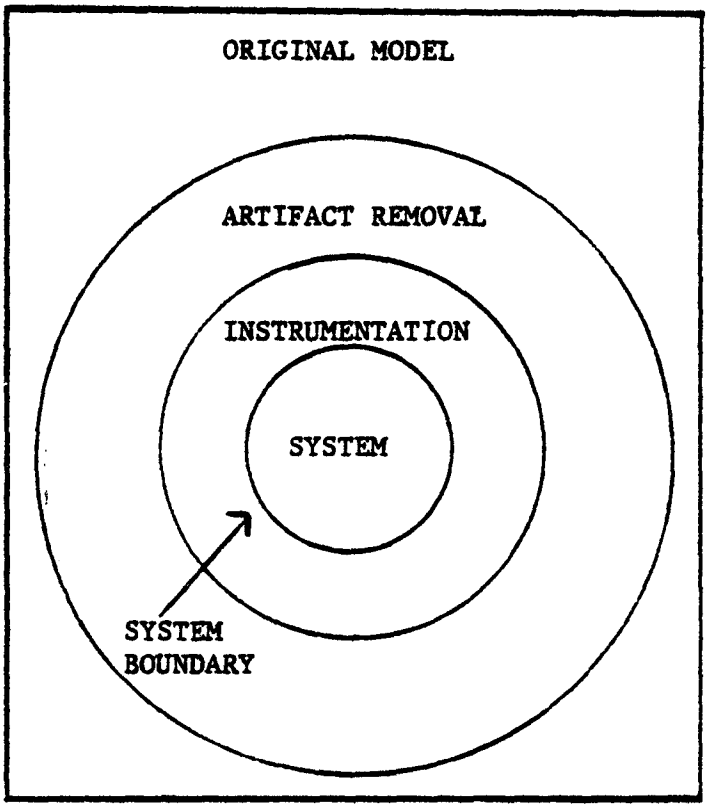
This situation arises both from the realities of data collection and from a mismatch between the theoretical model and the real system. For example, finite buffering, approximations, measurement overhead, output device limitations, and design errors, are properties of data collection which may modify the data itself. Time sequences and the system's resistance to measurement, perhaps due to interrupt disablement, can force the instrumentation to compromise the data being collected. One example is that many accounting measurement routines cannot charge the interrupt handling overhead of the system to the proper account because of the intolerable overhead that would be required. The result is that without applying a crude correction factor, the resource time accounted-for falls far short of that actually used. Clearly, the data actually delivered to the model are products of both system and instrumentation.

There are two different remedies to the corruption of data by the instrumentation. We can remove the artifact and leave the model unchanged, or we can redefine our performance metrics and modify the model to account for the artifact. These alternatives are represented in figure 1. In figure 1A we have the uncorrected situation. The instrumentation gets in the way so that we cannot see the system as it really exists. Figure 1B shows our first alternative. We apply artifact corrections to the data and deliver the transformed data to the original model. This approach may be preferable when internal metrics like utilizations are measured. The objective is to return to a characterization of the original system boundary as closely as possible. In practice, known synthetic workloads and independent measurements are useful in the quantification of artifact.

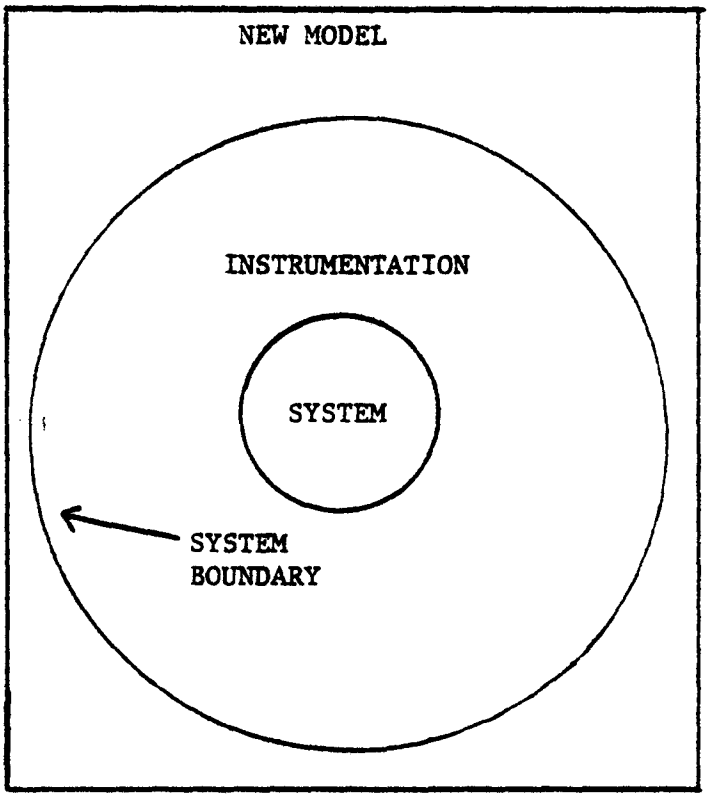
The second approach is represented in figure 1C. Here we create a new model with redefined performance metrics. In this case, the realities of measurement are included in the performance model. We have expanded the theoretical extent of the system leading to a different definition of performance metrics. As discussed above, this



1A. NO TREATMENT



1B. ARTIFACT REMOVAL



1C. REDEFINITION OF METRICS

Figure 1. MEASUREMENT ARTIFACT

presents no problem with respect to the applicability of operational relationships. Although the new metrics are different from the old, they preserve their internal consistency. This redefinition of extent may be preferable to artifact removal when external performance metrics like response times and think times are measured.

4.0 Conclusions

Operational analysis is not a completely unified structure. The major relationships within it have varying strengths and applicability. Because of this we must consider them separately when we apply them.

From an empirical point of view we cannot separate our understanding of system behavior from the measurements by which we achieve that understanding. Regardless of the modelling approach, we do not study the system independently of the measured data. In fact, it is often desirable to adjust our definitions and models to reflect the realities of measurement.

5.0 References

[1] Denning, P. and Buzen, J.,"The Operational Analysis of Queuing Network Models", ACM Computing Surveys, Vol 10, No. 3, 1978, pp. 225-261.

[2] Buzen, J., "A Note on Operational Assumptions", Performance Evaluation Review, Vol. 10, No. 2, 1981, pp. 76-79.

[3] Cox, S.W., "Interpretive Analysis of Computer System Performance" , Proc. Sigmetrics, 1974, pp. 225-275.

[4] Cox, S. W., "Performance Constraints from Regression Analysis", Proc CMG XI, 1980, pp. 75-85.