Proceedings of the 1984
Winter Simulation Conference
S. Sheppard, U. Pooch, D. Pegden (eds.)

USING OPERATIONAL ANALYSIS ASSUMPTION ERRORS

Neal M. Bengtson
Department of Computer Science
North Carolina State University
Raleigh, North Carolina 27695-8206

ABSTRACT

Assumptions that are the basis for operational analysis models of devices have the characteristic that they can be proved to hold by observing the data. Error measures are defined for the main operational analysis assumptions. A method for deriving correction terms is described. These terms are functions of the error measures and can be used to get exact results of behavior sequence performance measures of interest, such as, mean number of jobs at a device. These performance measures will be exact no matter how badly the operational assumptions are met by the data. Formulas for performance measures that were developed assuming homogeneous arrivals and services were found to give exact results under less restrictive conditions. Since the performance measure correction terms can only be calculated exactly with an amount of data that would be required to obtain direct performance measure results, ways to estimate the correction terms with reduced data collection are suggested.

INTRODUCTION

Operational analysis (OA) is a term defined by Jeffrey Buzen [1] for a type of analysis of observed (i.e. operational) data provided by a system. This type of analysis is used to calculate performance measures (PMs) such as mean number of jobs at a device, throughput, and response times for a particular time series. What characterized the foundation of OA was the use of

1. Testable assumptions,
2. Measurable variables,
3. Finite observation periods.

By testable assumptions is meant assumptions that can be proved to hold by examining the data from the system. For example, a common assumption is job flow balance: the number of jobs that enter a device equals the number that leave. This can be determined by simply keeping track of the two variables, arrival and completion times, and comparing them. These variables are measurable. No variable that we can't measure or derive from the data is used in OA. Working with a specific set of output data implies a finite period of observation in which the data was collected. OA assumptions are about the behavior of system data. OA says nothing about the underlying nature of the system which generated the data. Because of this characteristic OA seems to be less useful as a tool for prediction than stochastic analysis.

DEFINITIONS

Some of the fundamental definitions are given below. These are from [2]. For simplicity only a single device is considered.

n = The state of a device. The number of jobs at the device.

N = Largest state of the device during the period of observation.

$A(n)$ = Number of arrivals when $n(t) = n$.

$C(n)$ = Number of completions when $n(t) = n$.

$T(n)$ = Total time n jobs are at the device.

A = Total number of arrivals at the device.

$= A(0)+A(1)+A(2)+...+A(N-1)$

C = Total number of completions from the device.

$= C(1)+C(2)+...+C(N-1)+C(N)$

T = Total time period of observation.

$= T(0)+T(1)+T(2)+...+T(N-1)+T(N)$

$p(n)$ = The proportion of time n jobs are at the device.

$= T(n)/T$

$p_A(n)$ = The fraction of arrivals that find n jobs at the device.

$= A(n)/A$

$p_C(n)$ = The fraction of completions that leave n jobs of the device.

$= c(n+1)/C$

As an example to illustrate these terms consider the sequence of arrivals and completions given in Figure 1.

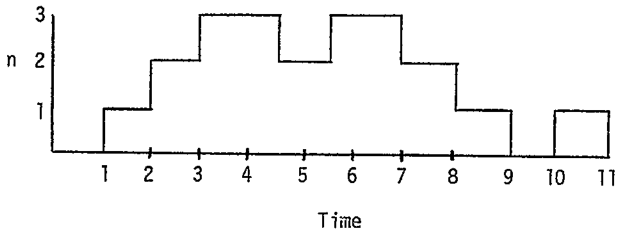

Figure 1. Sequence of arrivals and completions at a device.

The resulting quantities for arrivals are

$$\begin{array}{l} A(0) = 2 \\ A(1) = 1 \\ \underline{A(2) = 2} \\ A \quad = 5 \end{array}$$

and for completions are

$$\begin{array}{l} C(1) = 2 \\ C(2) = 1 \\ \underline{C(3) = 2} \\ C \quad = 5\ . \end{array}$$

Notice there are no arrivals in state N or completions in state 0. The time for each state is

$$\begin{array}{l} T(0) = 2 \\ T(1) = 3 \\ T(2) = 3 \\ \underline{T(3) = 3} \\ T \quad = 11\ . \end{array}$$

The resulting proportions are

$$\begin{array}{lll} p_A(0) = 2/5 & p_C(0) = 2/5 & p(0) = 2/11 \\ p_A(1) = 1/5 & p_C(1) = 1/5 & p(1) = 3/11 \\ p_A(2) = 2/5 & p_C(2) = 2/5 & p(2) = 3/11 \\ & & p(3) = 3/11\ . \end{array}$$

The individual fractions sum to 1 in each case.

Figure 2 gives some examples of quantities which can be calculated from the A(n), C(n) and T(n) values. Notice that mean time between completion, S, is not the same as mean service time because there can be periods where no job is at the device. These operational quantities may be manipulated to derive relationships among them which are called operational laws. Some of the possibilities are given in Figure 3.

ASSUMPTIONS

Using the general characteristics of operational analysis as a guide the analyst is free to make any assumptions about the behavior of data that are convenient. Some of the simpler and more natural assumptions are;

1. One-step behavior - No more than one arrival or completion may occur at a time,
2. Job-flow balance - The overall arrival rate is equal to the output rate,
3. Homogeneous arrivals - The arrival rate is constant for all states in which arrivals occur,
4. Homogeneous services - The mean time between end of services is constant for all states in which completions occur.

If there is a network of devices then we might add

5. Routing homogeneity - The job flow rate between any pair of devices depends only on the state at the source of the flow.

$S(n)$ = Mean time between completions when $n(t)=n$
= $T(n)/C(n)$

S = Overall mean time between completions
= B/C

B = Total busy time
= $T-T(0)$

U = Utilization
= B/T

X = Output rate
= C/T

W = Job seconds of accumulated waiting time
= $\sum_{n=1}^{N} nT(n)$

Q = Mean number of jobs at the device
= W/T

R = Mean response time per completed job
= W/C

$Y(n)$ = Arrival rate when $n(t)=n$
= $A(n)/T(n)$

Y_0 = Overall arrival rate
= A/T

Y = Restricted arrival rate
= $A/(T-T(N))$

Figure 2. Example operational quantities [2].

$$p_A(n) = p(n)(Y(n)/Y_0)$$

$$Y/Y_0 = 1/(1-p(n))$$

$$Y_0 = \sum_{n=1}^{N-1} p(n)Y(n)$$

$$S = \sum_{n=1}^{N} p_C(n-1)S(n)$$

$$X = \sum_{n=1}^{N} p(n)/S(n)$$

$$U = SX = 1-p(0) \qquad \text{(utilization law)}$$

$$R = Q/X \qquad \text{(Little's law)}$$

Figure 3. Example operational laws [2].

These assumptions are the ones used for this work. Other authors have chosen other assumptions as needed. For example, another version of the homogeneous arrival assumption is

> "The mean queue length seen by arriving customers...is equal to the mean queue length seen by an outside observer" [3].

Assumptions can become quite elaborate as the homogeneity of residuals assumption shows.

> "The total of the forward service period residuals seen by arrivals equals the total of the backwards residuals" [3].

Here for arrival j the "forward residual is either the time remaining in the service period during which j arrives or zero if arrival j begins a service period... Similarly backward residual is either the time since the beginning of the service period during which j arrives or zero if arrival j begins a service period."

The use of assumptions allows us to calculate PM's from the collection of some basic data. From Figure 2

$$S_i(n) = T_i(n)/C_i(n)$$

assumes one-step behavior at device i. If job flow balance is assumed we may say the completion distribution is the same as the arrival distribution, then each $p_A(n) = p_C(n)$. By assuming both homogeneous arrivals and service the average number of jobs at device i is

$$\overline{n}_i = \frac{U_i}{1-U_i-p_i(n)} (1-(N+1)p_i(n)) .$$

ASSUMPTION ERROR MEASURES

One of the objectives of this work is to contribute to the understanding of the behavior of OA assumption errors in observed data and to find ways to improve OA estimates of performance measures at devices and for networks of devices. Since the OA assumptions are defined such that they can be verified by observed data it is possible to define terms that measure to what degree the assumptions are not met by that set of data. These assumption error measures in themselves can indicate a degree of confidence we should have in any PM's calculated with formulas derived using the assumptions whose errors are measured. Using the assumption error measures it is possible to develop correction terms which can be applied to PM formulas to give exact results. That is, the results that would have been obtained by the formulas had the assumptions been met.

The particular error measures derived are as follows [4]:

Job-flow Balance - $e_F(n) = (N+1)(p_A(n)-p_C(n))$

$$e_F = \sum_{n=0}^{N-1} e_F(n)$$

Homogeneous Arrivals - $e_A(n) = p_A(n) \frac{T-T(N)}{T(n)} - 1$

$$e^*_A = \sum_{n=1}^{N-1} \frac{nT(n)}{T-T(N)} e_A(n)$$

Homogeneous Services - $e_S(n) = \frac{1}{p_C(n-1)} \frac{T(n)}{T-T(0)} - 1$

$$e^*_S = \sum_{n=1}^{N} np_C(n-1)e_S(n)$$

Homogeneous Routing - $e_{Ri}(n,N) = p_{Ai}(n,N)-p_i(n,N-1)$

$$e_{Ri}(N) = \sum_{n=1}^{N} ne_{Ri}(n,N)$$

Notice that in each case a state dependent error measure and an overall error measure are given. The chief characteristic of each error measure should be that it equals zero if and only if the assumption it measures holds. In the case of job flow balance if any $e_F(n)\neq 0$ then $e_F\neq 0$. This is because if the error at one state is positive the error at any other state will also be positive. This condition does not hold for homogeneous arrival and service errors. From the example given in Figure 1 we can calculate the following for service error.

$$\begin{array}{ll}
S(1) = 3/2 & e_S(1) = -1/6 \\
S(2) = 3/1 & e_S(2) = 2/3 \\
S(3) = 3/2 & e_S(3) = -1/6 \\
S = 9/5 &
\end{array}$$

$$e^*_S = \sum_{n=1}^{N} np_C(n-1)e_S(n) = 0.$$

This illustrates that we can have homogeneous errors in each of the states and still get a zero value for the particular measure defined by e^*_S. Does this mean that the formula for e^*_S is inadequate? It turns out that even though e^*_S is a weak homogeneity assumption a value of $e^*_S=0$ is all that is necessary to use PM formulas derived with the homogeneity assumption. The same weak condition applies for homogeneous arrivals as well.

At a single device the values of the job flow balance error measures will go to zero as the length of the period of data observation increases. The same can't be said for the homogeneity error measures. For example, for a device in a closed system the

$$\lim_{t\to\infty} E[e^*_A] = Q_A-L_N$$

where: Q_A = Mean number at the device seen by an arriver,

L_N = Mean number at the device excluding the state N,

and

$$\lim_{t\to\infty} E[e^*_{S}] = L_0-Q_C$$

where: L_0 = Mean number at the device excluding the state 0,

Q_C = Mean number at the device seen by a completer.

Once assumption error measures have been found we can develop correction terms. When the correction times are added to PMs calculated using formulas derived under various assumptions the result will be exact values for particular set data. In general, for the PM θ we want to find a correction term

$$C = \theta-\theta^{OA}$$

$$= F(\text{error measures, simple quantities})$$

where θ^{OA} is the PM of interest calculated using operational analysis formulas based on specified assumptions.

Example 1. If we have homogeneous arrivals then

$$p^A_A(n) = p(n)/(1-p(n)) \quad n=0,1,\dots,N-1$$

where the superscript indicates homogeneous arrivals is assumed. We want a correction term $C(p^A_A(n))$ such that

$$p_A(n) = p^A_A(n)+C(p^A_A(n)) \quad n=0,1,\dots,N-1$$

after some algebraic manipulation the correction term is found to be

$$C(p^A_A(n)) = e_A(n)\,\frac{T(n)}{T-T(N)} \quad n=0,1,\dots,N-1.$$

Example 2. If services are homogeneous then the mean number at the device seen by an arriver is

$$\bar{n}_A = \bar{n}^s_A+C(\bar{n}^s_A)$$

where the superscript s indicates the homogeneous service assumption. In this case

$$c(n^s_A) = e_F-e^*_s.$$

Example 2 brings up a couple of interesting points. First, it can be shown that if e^*_s is zero then e_F will also be zero. That is, if we have weak homogeneity of services job flow balance is guaranteed to exist. The converse is not true. This will be true not just for this calculation but for all cases. Second, even though the model for $\bar{n}^s_A$ is determined assuming homogeneous services $\bar{n}^s_A$ will give an exact value for $\bar{n}_A$ under the less restrictive condition of weak homogeneity.

USING THE CORRECTION TERMS

Two points should be emphasized about the use of correction terms in operational analysis. One is that the assumption error measures derived for use in PM correction terms are for a particular set of data. A new run with new data may produce different error measures, particularly for a short behavior sequence. Second, the amount of information needed to get the error measures is the same as is needed to measure the PM's directly.

In practice what will be needed are good estimates of the error measures and using them, correction terms. Several approaches for getting these can be investigated. Three are

1. Find bounds on the error measures.
2. Estimate correction term values in "short" runs and use them for estimating PM values in longer runs.
3. Determine theoretical correction term values and use them with simplified OA models to estimate the PM's.

As an example of the error bound approach let

$$|e_s(n)| \le \delta \quad n=1,2,\dots N$$

Then

$$e^*_s = \sum_{n=1}^{N} np_c(n-1)e_s(n)$$

$$\le \sum_{n=1}^{N} np_c(n-1)|e_s(n)|$$

and

$$e^*_s \le \delta\bar{n}_c$$

So, given a known upper bound, δ, for each $e_s(n)$ the value of e^*_s must be less than or equal to δ times $\bar{n}_c$. But we know the $e_s(n)$'s are not independent so tighter bounds can be found by solving

$$\max e^*_s = \sum_{n=1}^{N} np_c(n-1)e_s(n)$$

$$\text{s.t.} \quad \sum_{N=1}^{N} p_c(n-1)e_s(n)=0$$

$$|e_s(n)| \le \delta \quad n=1,2,\dots,N.$$

Using Kunn-Tucker it can be shown that the solution is

$$e^*_s \le \delta(\bar{n}_c-\bar{n}_{ct})$$

where $\bar{n}_{ct}$ is the average number of jobs at the device seen by a completer truncated at the median state. This can greatly reduce the maximum value on e^*_s.

In a time series of observed data if the correction term is calculated periodically as new data is generated the value of the correction term may stabilize before the PM value itself. In that case we can fix a correction value and from that point in the series use formulas which simplify the PM calculations based on OA assumptions to speed up data collection. At the end of the time series (e.g. a simulation run) we can then just add the correction term to improve the PM estimate. This should lead to a reduction in the length of the time series necessary to get some desired confidence.

For specific OA models it is possible to determine expected values for correction terms. If we are dealing with similar models then we can use the OA models derived under the correction term assumptions to simplify data collection and apply the theoretical correction terms as estimates to improve the PM estimates derived under the OA assumptions.

These techniques are areas for future research.

SUMMARY

The error measures defined by this work in themselves can indicate the reliability of OA assumptions about observed data. The OA assumptions are used to develop relationships which can simplify the process of estimating performance measures of simulation output. Correction terms for these PM estimates can be found which will give PM values closer to the actual PM's for a set of observed data. These correction terms can be used to reduce data collection and performance measure calculations.

REFERENCES

[1] Buzen, J. P., "Fundamental Operational Laws of Computer Systems Performance." *Acta Informatica* , 7, pp. 183-195, 1976.

[2] Buzen, J. P. and P. J. Denning, "Measuring and Calculating Queue Length Distributions," *Computer* , Vol. 13, No. 4, pp. 33-44, April 1980.

[3] Brumfield, J. A., *Operational Analysis of Queueing Phenomena,* Ph.D. Thesis, Purdue University, December 1982.

[4] Bengtson, N. M., *Development and Use of Operational Analysis Model Error Measures*, Ph.D. Thesis, Purdue University, May 1983.