Proceedings of the 1994 Winter Simulation Conference
ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila

VECTOR-AUTOREGRESSIVE INFERENCE FOR EQUALLY SPACED, TIME-AVERAGED, MULTIPLE QUEUE LENGTH PROCESSES

John M. Charnes

The University of Kansas
School of Business
Summerfield Hall
University of Kansas 66045–2003 USA

Evelyn I. Chen

Stanford University
Department of Operations Research
Terman Engineering Center
Stanford, California 94305-4022 USA

ABSTRACT

This paper investigates the performance of the vector-autoregressive method of analyzing multivariate output data (numbers in subsystem) from queueing network models *vis-a-vis* three other methods of multivariate analysis—Bonferroni batch means, multivariate batch means, and spectral analysis. Differences in performance for all methods are found when time averages of numbers in subsystem are used rather than discretized observations taken at equally spaced points in simulated time. Further investigation is made into the effect of varying the spacing of averaging times for the methods. The results show that the analysis of time averages rather than discretized observations leads to slightly improved performance for all methods considered but that there is little difference in the relative performance of the methods considered.

1 INTRODUCTION

The VAR (vector-autoregressive) method of making statistical inferences on the mean vector of simulation output was studied by Jow (1982), and Charnes and Kelton (1993). In the latter, both open- and closed-system multiple queueing networks were studied by analyzing the vectors of "snapshots" (discretized vector observations) of the numbers in subsystem at equally spaced moments in simulated time. It was found that the VAR method worked quite well relative to the other output analytic methods with which it was compared.

However, the data resulting from taking snapshots of numbers in subsystem are integer-valued, and the VAR method uses the continuous-space vector-autoregressive model for making inferences. An obvious issue to investigate is whether the VAR method might work better for continuous output than it does for discrete. This paper reports such an investigation.

The specific questions considered here are: (*i*) Is the coverage of VAR improved when time averages of numbers in subsystem are analyzed instead of snapshots of numbers in subsystem, and (*ii*) How is the coverage of VAR confidence regions affected by varying the spacing of the times at which the averages are taken? The next section of this paper describes briefly the VAR method of output analysis. Following that, the experiment is described and the results given. The concluding section gives implications and directions for future research.

2 VAR OUTPUT ANALYSIS

The basic notion underlying the VAR method of output analysis is to model the simulation's steady-state data-generation process as a vector-autoregressive process, estimate the VAR parameters from the simulation output, and then use the estimated parameters to construct confidence regions on the mean vector of the steady-state simulation output process.

The VAR model is

$$\mathbf{X}_t - \underline{\mu} + \mathbf{A}_1(\mathbf{X}_{t-1} - \underline{\mu}) + \cdots + \mathbf{A}_p(\mathbf{X}_{t-p} - \underline{\mu}) = \underline{\epsilon}_t$$

where $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{dt})'$ is the $d \times 1$ vector of observations at time t, $E[\mathbf{X}_t] = \underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)'$ is the parameter on which inference is to be made, and the $\mathbf{A}_k = [-a_{ij}^k]$ are $d \times d$ matrices of autoregression coefficients. The vector of random errors at time t, $\underline{\epsilon}_t = (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{dt})'$, is multivariate white noise (not necessarily Gaussian) with $d \times d$ variance-covariance matrix $\mathbf{\Sigma}$, i.e.,

$$E[\underline{\epsilon}_t] = 0_{d\times 1} \text{ and } E[\underline{\epsilon}_{t+h}\,\underline{\epsilon}_t'] = \begin{cases} \mathbf{\Sigma} & \text{if } h = 0 \\ \mathbf{0}_{d\times d} & \text{otherwise} \end{cases}$$

(0 denotes an all-zero matrix of the indicated dimensions).

It is assumed that the process is stationary, i.e., that the roots of

$$\left| \mathbf{I}_d + \mathbf{A}_1 z + \mathbf{A}_2 z^2 + \cdots + \mathbf{A}_p z^p \right| = 0$$

are outside the unit circle in the complex plane ($\mathbf{I}_d$ is the $d \times d$ identity matrix), so that neither $\underset{\sim}{\mu}$ nor the $\mathbf{A}_k$'s depend on time. For such a process is defined the lag-h autocovariance matrix

$$\mathbf{\Gamma}(h) = E\left[(\mathbf{X}_{t+h} - \underset{\sim}{\mu})(\mathbf{X}_t - \underset{\sim}{\mu})'\right].$$

From an observed output sequence $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ on a stationary VAR(p) model the estimator of $\underset{\sim}{\mu}$ used is

$$\widehat{\underset{\sim}{\mu}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{X}_t.$$

The autocovariance matrices $\mathbf{\Gamma}(h)$ can then be estimated by

$$\widehat{\mathbf{\Gamma}}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(\mathbf{X}_{t+h} - \widehat{\underset{\sim}{\mu}})(\mathbf{X}_t - \widehat{\underset{\sim}{\mu}})',$$

for $h = 0, 1, \ldots, n-1$.

The algorithm reported in Charnes and Kelton (1993) proceeds as follows. Use the BIC criterion (Lütkepohl 1985, 1993) to select the order p of the VAR model, and the Durbin-Levinson algorithm (Durbin 1960, Levinson 1946, Whittle 1983) to solve recursively the Yule-Walker equations with the estimated autocovariance matrices,

$$\begin{aligned}\widehat{\mathbf{\Gamma}}(0) + \mathbf{A}_1\widehat{\mathbf{\Gamma}}(-1) + \cdots + \mathbf{A}_p\widehat{\mathbf{\Gamma}}(-p) &= \mathbf{\Sigma}\\ \widehat{\mathbf{\Gamma}}(h) + \mathbf{A}_1\widehat{\mathbf{\Gamma}}(h-1) + \cdots + \mathbf{A}_p\widehat{\mathbf{\Gamma}}(h-p) &= \mathbf{0},\end{aligned}$$

for $h = 1, \ldots, p$, to obtain the estimators $\widehat{\mathbf{A}}_k$ of $\mathbf{A}_k$ for $k = 0, 1, \ldots, p$ (by definition, $\mathbf{A}_0 = \mathbf{I}_d = \widehat{\mathbf{A}}_0$) and $\widehat{\mathbf{\Sigma}}$ of $\mathbf{\Sigma}$. Letting

$$\widehat{\mathbf{\Phi}} = \sum_{k=0}^{p}\widehat{\mathbf{A}}_k,$$

the VAR confidence region is the d-dimensional ellipsoid

$$\left\{\underset{\sim}{\theta} : n(\widehat{\underset{\sim}{\mu}} - \underset{\sim}{\theta})'\widehat{\mathbf{\Phi}}'\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Phi}}(\widehat{\underset{\sim}{\mu}} - \underset{\sim}{\theta}) \le \chi^2_{1-\alpha}(d)\right\}$$

where $\chi^2_{1-\alpha}(d)$ is the $100(1-\alpha)\%$ percentile of the chi-square distribution with d d.f. Because the VAR model approximates the relationships among output vectors observed at different points in simulated time, it accounts for both autocorrelation and cross-correlation in the simulation data-generation process.

The methods to which VAR is compared here, BBM (Bonferroni batch means), MBM (multivariate batch means), and SPA (spectral analytic), account for autocorrelation and cross-correlation in different ways. BBM is a method for combining the results of two or more univariate analyses; MBM is a generalization of the univariate batch means method; and SPA is a frequency-domain time-series technique (as opposed to time-domain, as is VAR). See Charnes and Kelton (1993) for further discussion of these methods.

Design Point	Factor Levels System Type		
		d	n
1	open	2	4,096
2	open	2	8,192
3	open	4	4,096
4	open	4	8,192
5	closed	2	4,096
6	closed	2	8,192
7	closed	4	4,096
8	closed	4	8,192

Table 1: Factor Combinations for Experiments

3 DESCRIPTION OF EXPERIMENTS

The experiments reported here used both open- and closed-system queueing networks having exponentially distributed service times and Poisson arrivals to the open system. See Charnes and Kelton (1993) for a fuller description of these networks. The same systems are used here because the interest is in comparing the previous results (obtained with snapshots of numbers in subsystem) to the results obtained here (with time averages of numbers in subsytem).

We consider only exponential distributions for service times here both in the interest of conserving space in the *Proceedings* and because a common belief among researchers in simulation output analysis is that the highly skewed exponential distributions yield the "hardest" data to analyze, i.e., if an output analytic method works well for data from queueing models using exponential distributions, then it is likely that it will also work well for models using less-skewed distributions.

For the experiments, eight design points were used as a result of considering three factors at two levels each. The factors are: (1) type of queueing network, open and closed; (2) the dimension of the vectors of data to be analyzed, $d = 2$ and $d = 4$; and (3) the number of vector observations to be analyzed, $n = 4096$ and $n = 8192$. Table 1 gives the assignment of design point number to each combination of factors.

For each design point the steady-state distribution of the vectors of numbers in subsystem are known. The corresponding steady-state distribution was used to generate the initial numbers in subsytem for each replication. Thus there was no initial transient bias in the observations generated for the experiments reported here.

In Charnes and Kelton (1993) the observations that were analyzed consisted of d-dimensional vector snapshots of numbers in subsystem observed at points in time that were spaced equally every $w = 3.0$ units of simulated time. Question (i) in §1 asks if the coverage of VAR is improved if time averages are analyzed instead of snapshots. To answer this, the experiments run previously were run again exactly as they were when the snapshots were obtained; however, in the reruns time averages were calculated during each period of time $w = 3.0$ instead of using the discrete numbers in subsystem at each observation time. Question (ii) in §1 asks how the coverage of VAR confidence regions is affected by varying the spacing of the times at which the averages were taken. To answer this, the interobservation time, w, was varied between 1.0 and 12.0.

4 RESULTS OF EXPERIMENTS

4.1 Snapshots vs. Time Averages

Denote the vector of numbers in subsystem at time, t as $\mathbf{Q}(t) = (Q_1(t), Q_2(t), \ldots, Q_d(t))^T$, where $Q_j(t)$ is the number of customers in queue j at time t, plus one if the jth server is busy. The observations analyzed when snapshots are used are then $\mathbf{Q}(t_1), \mathbf{Q}(t_2), \ldots, \mathbf{Q}(t_n)$ where $t_i = iw$. The observations analyzed when time averages are used are then $\overline{\mathbf{Q}}(t_1), \overline{\mathbf{Q}}(t_2), \ldots, \overline{\mathbf{Q}}(t_n)$ where $t_i = iw$ and $\overline{\mathbf{Q}}(t_i) = \int_{t_{i-1}}^{t_i} \mathbf{Q}(t)dt/w$.

For the queueing models simulated here, the true mean vector is known. Thus, the coverage of each method can be estimated empirically by making several independent runs of the model, calculating the confidence regions from the observations generated, and then checking whether the true mean vector fell within the d-dimensional confidence region. The experiments reported here consisted of 100 independent runs at each design point. Thus the empirical coverage reported for each method is the proportion of the 100 runs for which the indicated confidence region contained the true mean. This empirical coverage is denoted in the graphs of the next subsection as $\hat{\gamma}$.

Table 2 compares the coverage obtained with 90% confidence regions calculated from snapshots (in the column labeled "Snap") and time averages (in the column labeled "TAvg") for the BBM, MBM, SPA, and VAR methods at each of the eight design points. While the coverage is improved at many design points, it is not improved at every point. In particular, the coverage of every method is much lower with time averages than with snapshots at Design Point 5; in addition, the VAR coverage is much lower with time averages than with snapshots at Design Point 8.

Design	BBM		MBM	
Point	Snap	TAvg	Snap	TAvg
1	0.85	0.87	0.84	0.89
2	0.86	0.86	0.82	0.87
3	0.87	0.92	0.83	0.87
4	0.89	0.90	0.89	0.88
5	0.93	0.87	0.86	0.80
6	0.97	0.98	0.86	0.88
7	0.95	0.89	0.88	0.86
8	0.86	0.86	0.84	0.83
Avg	0.90	0.89	0.85	0.86

Design	SPA		VAR	
Point	Snap	TAvg	Snap	TAvg
1	0.86	0.89	0.79	0.86
2	0.83	0.86	0.78	0.82
3	0.84	0.90	0.80	0.89
4	0.89	0.90	0.86	0.88
5	0.89	0.81	0.86	0.79
6	0.89	0.91	0.84	0.89
7	0.91	0.87	0.85	0.85
8	0.84	0.84	0.84	0.79
Avg	0.87	0.87	0.83	0.85

Table 2: Comparison of Coverage for Snapshot Observations vs. Time Average Observations

4.2 Spacing of Averaging Times

To investigate the effects on coverage of varying the spacing of the times at which the time averages are observed, the queueing models described above were used to generate the same number of observations, n, as specified in Table 1. Thus the simulations were run for a total of nw units of simulated time.

Figures ??-?? are plots of $\hat{\gamma}$ (coverage) vs. w (interobservation time) for each design point. The plotted data were obtained by specifying w at the integers 1, 2, ..., 12. The nominal coverage of the regions (0.90) is indicated on each plot by a horizontal arrow.

In general, the plots show low coverage for all methods when w is small (1 or 2), with improved coverage as w increases. The variability of the coverage for $w \geq 3$ is what is to be expected as the standard error for 90% confidence intervals with 100 independent replications is $\sqrt{(.90(.10))/100} = .03$ and the coverages for larger w mostly fall within the two-standard-error range of .84 to .96.

An interesting characteristic of the plots is the tendency of the coverages to follow each other. That is, in general no one method vastly outperforms any other method, and the coverages of all the methods seem to be relatively close to each other. This effect is more pronounced in the open networks (Figures 1–4) than in the closed networks (Figures 5–8).

5 CONCLUSION

This paper is part of an investigation of the behavior of the VAR method of output analysis. The two questions posed in the Introduction are answered here: (*i*) The coverage of VAR *is* improved when time averages are taken instead of snapshots (but not in every case). Apparently the continuous-space VAR model does a fairly adequate job of modeling even when the process being modeled has discrete output, but the time-averaging may provide central limit effects that help improve the performance modestly. (*ii*) The coverage of VAR is affected by the spacing of the time averages, but the effect is only obvious for low values of w. After w is about 3 or 4, the plots shown here indicate that the coverage fluctuates as would be expected due to sampling error. This implies that the spacing w need not be more than about 3 or 4 (this should be compared to the interarrival time of 1.0 for open queueing systems and service times of .8, .7, .6, .5, and .4 for both open and closed queueing systems).

In comparing the methods BBM, MBM, SPA, and VAR to each other, it appears that there is little difference in the coverage of the methods. This implies that the choice of which method to use should be made on some other criterion than coverage. Previous work has shown that the VAR method yields confidence regions that are somewhat smaller (more precise) on average than the other methods. However, the final choice of a method should probably be based upon what an analyst is comfortable with. The use of a method that an analyst does not understand could very well lead to erroneous conclusions.

Future work might consider how the comparisons reported here fare with queueing models having higher traffic intensities. With the arrival rate set at 1.0 in the open sytems, and the service rates in both systems at .8, .7, .6, .5, and .4, the congestion level of the queueing networks might not be as high as would be experienced in practice. Higher levels of congestion in the systems could give different results.

REFERENCES

Charnes, J.M. and W.D. Kelton, "Multivariate Autoregressive Techniques for Constructing Confidence Regions on the Mean Vector," *Management Science*, 39 (1993), 1112–1129.

Durbin, J., "The Fitting of Time Series Models," *International Statistical Institute Review*, 28 (1960), 233–244.

Jow, Y.-L. L., "An Autoregressive Method for Simulation Output Analysis," Ph.D. Thesis, Department of Operations Research, Stanford University, December, 1982.

Levinson, N., "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction," *Journal of Mathematics and Physics*, 25 (1946), 261–278.

Lütkepohl, H., "Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process," *Journal of Time Series Analysis*, 6 (1985), 35–52.

Lütkepohl, H., *Introduction to Multiple Time Series Analysis* second edition, Springer-Verlag, Berlin, 1993.

Whittle, P., *Prediction and Regulation by Linear Least-Square Methods*, (Second Ed., Rev.), University of Minnesota Press, Minneapolis, 1983.

ACKNOWLEDGEMENT

The first author gratefully acknowledges support for this work from the University of Minnesota Supercomputer Institute.

AUTHOR BIOGRAPHIES

JOHN M. CHARNES is Associate Professor of Statistics and Quality Management in the School of Business at The University of Kansas. His research interests are in the area of statistical analysis of multivariate output. He is a member of **ASA**, **ASQC**, **TIMS**, and is Editor of the *TIMS College on Simulation Newsletter*.

EVELYN I. CHEN was graduated in 1994 from the University of Miami in Coral Gables, Florida, with a B.S. Systems Analysis degree earned in the School of Business, Department of Management Science. She was recently admitted to the masters degree program in the Department of Operations Research at Stanford University.