Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

AN INFORMATION THEORETIC APPROACH TO COMPUTER SIMULATION SENSITIVITY ANALYSIS

John W. Dalle Molle
Douglas J. Morrice

Department of Management Science and Information Systems
Graduate School of Business, The University of Texas at Austin
Austin, Texas 78712-1175

ABSTRACT

In this paper, statistical information theory-based procedures are applied to sensitivity analysis in computer simulation. Information theory, through use of the conditional entropy functional, provides a nonparametric approach to qualitatively assessing the sensitivity of the distributional relationships of the input and output processes of a simulation model. Since the conditional entropy functional quantifies the amount of uncertainty in the distribution of a set of random variables, it can be used as the basis for a methodology to assess the relative strengths of the statistical dependencies among the input/output processes. The application of information theory in this paper focuses on assessing the uncertainty in the simulation output processes attributable to the simulation input processes. This approach to sensitivity analysis is illustrated by an example.

1 INTRODUCTION

In computer simulation, sensitivity analysis usually concerns the sensitivity of output performance measures (such as the expected value) to changes in specified deterministic input factors (or parameters). Specifically, this type of sensitivity analysis is concerned with the quantification of changes in simulation performance measures to changes in deterministic input factors and is therefore a parametric approach. Procedures such as *factor screening* (see, for example, Kleijnen (1987)) and *gradient estimation* (see, for example, Glasserman (1991)) are representative forms of this type of sensitivity analysis.

This paper considers a different type of sensitivity analysis using the concepts of *entropy* and *information theory.* The approach examines the joint statistical dependencies in the distribution of the simulation input and output processes. Since the entropy functional operates on the joint probability of occurrences of a set of random variables, it is nonparametric in nature. In contrast to the typical deterministic parametric approaches to sensitivity analysis, this nonparametric entropic approach is applicable if the simulation input processes are generated by parametric probability distributions, empirical distributions, or trace driven data. The objective of this type of procedure is to identify input processes which account for a significant portion of the uncertainty in the simulation output process. Additionally, this type of information can be used to determine the allocation of resources for improving the quality (i.e., reducing the uncertainty) of simulation input data.

The remainder of the paper is organized in the following way. Section 2 contains a description of entropy and techniques in information theory that are used in the sensitivity analysis procedure presented in this paper. Section 3 describes an approach to simulation sensitivity analysis using conditional entropy and related information theoretic measures. A queueing example is provided to illustrate this procedure. Section 4 contains some concluding remarks and future research directions.

2 INFORMATION THEORY AND ENTROPY

Entropy is a qualitative measure of the degree of dependence between a set of random variables. It is an accepted nonparametric measure of the statistical dependencies between the input and output processes in a stochastic system. Entropy-based information theory can be used to identify those input processes which account for a statistically significant amount of uncertainty in the distribution of the output processes. The statistical entropy function is defined as the logarithmic decomposition of the joint probability distribution (see equation (1) below), and requires only a few assumptions for its use.

The *unconditional entropy function* is the basic measure used in calculating the measures of uncertainty in the input/output distribution that are the basis for the sensitivity analysis procedure. To define the entropy function, let $\mathbf{X} = (X_1, \ldots, X_n)$ be a set of random variables where each $X_j, j = 1, \ldots, n$ has a finite number outcomes that are not necessarily integer-valued. The $m - th$ order unconditional entropy $H_u(X_1, \ldots, X_m)$ where $m \leq n < \infty$ can be expressed as:

$$H_u(X_1, \ldots, X_m) = -\sum_{i_1=1}^{k_1} \cdots \sum_{i_m=1}^{k_m} p_{i_1,\ldots,i_m} \ln(p_{i_1,\ldots,i_m}) \tag{1}$$

The indices $k_1, \ldots, k_m$ are defined as the number of possible outcomes in each of the respective sample spaces of the random variables $(X_1, \ldots, X_m)$. The quantity $p_{i_1,\ldots,i_m}$ is the multivariate joint probability of the set random variables $(X_1, \ldots, X_m)$ taking on the set outcomes $(x_{i_1}, \ldots, x_{i_m})$.

The *conditional entropy function* quantifies the amount of information in the joint distribution of the *contingent variables* at different known levels of the *explanatory variables*. In the framework of a simulation experiment, the contingent variables are the output processes and the explanatory variables are the input processes. The conditional entropy can also be interpreted as the amount of uncertainty remaining in the distribution of the output after the input processes are realized. From logic and algebraic techniques (see Guiasu (1977) or Yaglom and Yaglom (1983)), the conditional entropy can be expressed in terms of the unconditional entropy in the following manner:

$$H_c(\mathbf{X}|\mathbf{Y}) = H_u(\mathbf{X}, \mathbf{Y}) - H_u(\mathbf{Y}) \tag{2}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$ are the n contingent variables and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ are the m explanatory variables.

The next measure concerns the *amount of shared or redundant information* that a set of explanatory variables supplies about a set of contingent variables. Using the measures of unconditional and conditional entropy, the amount of shared or redundant information is defined as:

$$\begin{aligned} H_{si}(\mathbf{X}, \mathbf{Y}) &= H_u(\mathbf{X}) - H_c(\mathbf{X}|\mathbf{Y}) \qquad (3) \\ &= H_u(\mathbf{X}) + H_u(\mathbf{Y}) - H_u(\mathbf{X}, \mathbf{Y}) \end{aligned}$$

where the last step follows by substituting equation (2) into equation (3). From the definition of unconditional and conditional entropy, equation (3) can be interpreted as the difference between the amount of uncertainty in the contingent variables before and after having observed the explanatory variables. In other words, it quantifies the reduction in uncertainty in the contingent variables due to the information furnished by the explanatory variables. In the simulation framework, H_{si} is a measure of the uncertainty in the conditional relationship between the input and the output processes and can be used in a qualitative procedure to assess which input processes account for significant amounts of the uncertainty in the distribution of the output process.

The quantity H_{si} provides a means for statistical testing of the strength of a relationship between random processes over the continuum from complete independence to complete determinacy. If the set of contingent variables $\mathbf{X}$ is independent of the set of explanatory variables $\mathbf{Y}$, then H_{si} calculated from a sample of size N, converges to zero at a rate of $(1/N)$. It can be shown that under these conditions $2{*}N{*}H_{si}$ is asymptotically distributed as a chi-square distribution with $(k_{X_1} \cdots k_{X_n} - 1)(k_{Y_1} \cdots k_{Y_m} - 1)$ degrees of freedom where $k_{X_i}, i = 1, \ldots, n$ and $k_{Y_j}, j = 1, \ldots, m$ are the number of possible discrete responses for the contingent and explanatory variables, respectively (for a discussion on the convergence rate and the asymptotic distributional properties, see Gokhale and Kullback (1978)).

The final measure is the *percentage of shared information*. It is defined as:

$$\begin{aligned} H_{ps}(\mathbf{X}, \mathbf{Y}) &= \frac{100(H_{si}(\mathbf{X}, \mathbf{Y}))}{H_u(\mathbf{X})} \qquad (4) \\ &= \frac{100(H_u(\mathbf{X}) - H_c(\mathbf{X}|\mathbf{Y}))}{H_u(\mathbf{X})} \end{aligned}$$

where the last step follows by substituting equation (3) into equation (4). The percentage of shared information is similar to the regression-derived explained variance measure R^2. Both can be viewed as attempts to assess the input/output dependencies as a function of the reduction in the amount of uncertainty inherent in a set of contingent variables from the measurement or observation of a set of explanatory variables. However, unlike the R^2 criterion, the H_{ps} does not require that the relationship between the contingent and explanatory variables to be linear and is a nonparametric measure that can be readily calculated (in theory) for any bounded order of dependency of the distribution (see Golden et al. (1990), Appendix 1) .

The quantities H_{si} and H_{ps} are the primary measures used in the next section for simulation sensitivity analysis.

3 INFORMATION THEORETIC SENSITIVITY ANALYSIS

In order to mimic stochastic systems in business and engineering, computer simulation models are driven by stochastic input processes and therefore generate stochastic output processes. The information theoretic measures based on entropy described in section 2 can be used to quantify uncertainty in the input/output relationships in a computer simulation model. More specifically, these measures can be used to quantify the uncertainty in the distribution of a stochastic output process which can be attributed to the uncertainty in the distribution of a stochastic input process. This type of information is very important because it can be utilized to assess the quality of simulation input. For example, if a large amount of uncertainty in a simulation output process is attributable to a particular input process, then resources should be allocated to improve the quality of (i.e., reduce the uncertainty in) the data associated with this input process.

As indicated in section 1 and illustrated in section 2, entropy-based measures provide a nonparametric approach to sensitivity analysis. These measures examine the distributional relationship via the associated probabilities of the possible realizations of the stochastic processes driving the simulation and the stochastic processes generated by the simulation. The input processes need not be generated by parametric probability distributions. They may, for example, be generated by empirical distributions or trace driven historical data. Even if the data is generated by parametric distributions, the analysis remains nonparametric because the parameter values remain fixed.

It is important to note that the objective of the information theoretic approach to sensitivity analysis differs from the objective of a procedure like factor screening. The objective of factor screening is to distinguish between the input factors that have a significant impact on an output performance and those input factors that do not. Factors which have little or no impact on the performance measure are eliminated from consideration in any subsequent sensitivity analysis (for example, gradient estimation). The objective of the information theoretic to sensitivity analysis is not to eliminate input factors or even input processes. It is to identify the processes which contribute the most to the uncertainty in the distribution of the stochastic output processes. Those input processes which contribute very little to the uncertainty in the distribution of the output are not necessarily unimportant, but instead, may have very little uncertainty associated with their distribution. An extreme example of this is a constant input process whose value directly impacts an output process. Since this input process remains unchanged during the analysis, no uncertainty in the data series of the output process can be attributed to this factor.

The following example is an illustration of the implementation and analysis of the entropic approach to sensitivity analysis.

Example: Consider an $M/M/1$ queue with first-in, first-out service discipline with infinite queue capacity. The model has two input processes: an exponential interarrival process and an exponential service process. The output data series of interest is the system waiting time. The objective of this analysis is to measure the amount of uncertainty in the system waiting time attributable to the interarrival and service processes. Using the notation of section 2, let X be a random variable representing the system waiting time (the contingent variable) and let Y_1 and Y_2 be random variables representing the interarrival and service times respectively (the explanatory variables).

The information theoretic measures used in this analysis are those found in equations (1), (2), (3) and (4). It is important to note that since the underlying processes have continuous probability distributions, it is the relative values (not the actual values) of the entropies which provide meaningful information (see Papoulis (1984), page 525). By definition, this type of relative information is provided by H_{si} and H_{ps}. In the following analysis relative comparisons will also be made across different run scenarios (i.e., for different traffic intensities and different sample sizes).

Table 1 contains estimates of H_u, H_c, H_{si}, and H_{ps} along with the observed significance level (p-values) for the chi-square test statistic described in section 2 for H_{si}. For each traffic intensity, the data were produced by making 1,000 independently seeded replications (or simulation runs) of 2,000 customer service completions per run (i.e., a total of either 2,000,000 service completions). Each replication was initialized by sampling waiting times from the steady state distribution. On each replication, the sample mean for each process was computed and used as a single data point in the estimation of H_u, H_c, H_{si}, and H_{ps}. Therefore, the data in Table 1 were actually derived from 1,000 independent sample means from each of the input and output processes. The experiment was designed in this way to circumvent the difficulties associated with estimating the entropy functional for serial correlated data in the waiting time process. It turned out that using sample means as a surrogate measure for each of the processes provided very useful information.

Table 1: Measures of Shared Information for $M/M/1$ Queue with 1,000 Replications 2,000 Observations Initialized in Steady State

Traffic Intensity	$H_u(X)$	Between Y_1 and X				Between Y_2 and X			
		H_c	H_{si}	H_{ps}	P-value	H_c	H_{si}	H_{ps}	P-value
0.3	1.87	1.87	0.00	0.00	1.00	1.55	0.31	16.9	0.00
0.5	1.85	1.83	0.02	1.10	1.00	1.64	0.20	11.0	0.00
0.7	1.60	1.54	0.06	3.76	0.00	1.46	0.14	8.71	0.00
0.9	1.46	1.39	0.08	5.26	0.00	1.35	0.11	7.59	0.00
0.95	1.48	1.41	0.08	5.10	0.00	1.39	0.09	6.09	0.00

Table 2: Measures of Shared Information for $M/M/1$ Queue with 1,000 Replications 1,000 Observations Initialized in Steady State

Traffic Intensity	$H_u(X)$	Between Y_1 and X				Between Y_2 and X			
		H_c	H_{si}	H_{ps}	P-value	H_c	H_{si}	H_{ps}	P-value
0.3	1.85	1.85	0.00	0.00	1.00	1.54	0.31	16.5	0.00
0.5	1.60	1.59	0.01	0.63	1.00	1.42	0.19	11.6	0.00
0.7	1.33	1.30	0.03	1.89	1.00	1.21	0.12	9.01	0.00
0.9	1.36	1.31	0.06	4.14	0.01	1.26	0.11	7.71	0.00
0.95	1.40	1.35	0.05	3.52	0.09	1.32	0.07	5.20	0.00

Calculation of the statistics in Table 1 was performed by a FORTRAN computer program developed by Dalle Molle (1989). A complete description of the details of the calculations is given in this reference. For each of the processes (i.e., interarrivals, services, and waiting times), the sample mean data was categorized into ten intervals in order to estimate the entropies. Based on this categorization, the statistics $2*1,000*H_{si}(X,Y_1)$ and $2*1,000*H_{si}(X,Y_2)$ have approximate chi-square distribution with 81 degrees of freedom.

Each row of Table 1 contains results for a different system traffic intensity. In all cases the mean interarrival was held at one and the mean service time was changed to control the traffic intensity. A clear and explainable pattern emerges for the relationship between the interarrival and waiting processes. For lower traffic intensities (i.e., 0.3 and 0.5), the p-values are very close to one indicating that the uncertainty in the waiting time process has very little to do with uncertainty in the arrival process. These results can be understood by noting that at low traffic intensities, the waiting times are almost completely determined by the service times and the interarrival times contribute very little to the uncertainty in the waiting times. As the traffic intensities increase to 0.7, 0.9, and 0.95, on average, more customers spend time waiting in queues. The portion of the time a customer spends waiting in a queue is directly impacted by interarrival times. Hence, at higher traffic intensities, the uncertainty in the interarrival time process has a much greater impact on the uncertainty in the waiting time process (note the smaller p-values).

At all traffic intensities, a significant amount of the uncertainty found in the waiting time process was attributed to the uncertainty in the service time process (p-values are close to zero in all cases). This is to be expected since waiting times always contain a service time component. Therefore, the uncertainty in the latter contributes to the uncertainty in the former at all traffic intensities. Note that the magnitude of $H_{ps}(X|Y_2)$ decreases as the traffic intensity increases because of an increase in the variability of the waiting time process attributable to the interarrival process.

To test the effects of sample size on the results, Tables 2 and 3 contain the same information as Table 1 for $1,000$ replications of $1,000$ and 100 observations, respectively. It appears from these results and other empirical evidence collected by the authors that more observations per replication provide more definitive results. For example, the p-values for the shared information between interarrivals and waiting times decrease to zero as the traffic intensity increases for $2,000$ observations per replication. For $1,000$ observations per replication the same general pattern occurs, but the drop off of the p-values is not quite

Table 3: Measures of Shared Information for $M/M/1$ Queue with 1,000 Replications 100 Observations Initialized in Steady State

Traffic	$H_u(X)$	Between Y_1 and X				Between Y_2 and X			
Intensity		H_c	H_{si}	H_{ps}	P-value	H_c	H_{si}	H_{ps}	P-value
0.3	1.69	1.68	0.01	0.72	1.00	1.33	0.36	21.4	0.00
0.5	1.58	1.55	0.03	2.16	0.84	1.32	0.26	16.4	0.00
0.7	1.50	1.45	0.06	3.87	0.01	1.31	0.19	12.6	0.00
0.9	1.57	1.53	0.04	2.72	0.35	1.48	0.09	5.55	0.00
0.95	1.62	1.60	0.02	1.37	1.00	1.58	0.04	2.23	0.75

Table 4: Measures of Shared Information for $M/M/1$ Queue with 1,000 Replications 1,000 Observations Initialized Empty and Idle

Traffic	$H_u(X)$	Between Y_1 and X				Between Y_2 and X			
Intensity		H_c	H_{si}	H_{ps}	P-value	H_c	H_{si}	H_{ps}	P-value
0.3	1.85	1.85	0.00	0.00	1.00	1.54	0.31	16.5	0.00
0.5	1.60	1.59	0.01	0.64	1.00	1.42	0.19	11.7	0.00
0.7	1.32	1.29	0.03	1.95	1.00	1.19	0.12	9.16	0.00
0.9	1.24	1.18	0.06	4.85	0.00	1.14	0.10	8.28	0.00
0.95	1.46	1.37	0.09	6.30	0.00	1.34	0.13	8.83	0.00

as abrupt and the results seem to be less stable for higher traffic intensities. For 100 observations per replication, the results appear to be very unstable at higher traffic intensities. The problem of instability of the results at higher traffic intensities is caused by high variability in the data. As the number of observations increases per replication, the sample means exhibit less variability making the entropy statistics easier to estimate. By comparing H_u and H_c for 1,000 and 2,000 observations, the results appear to have almost reached a stable level by 1,000 observation per replication (i.e., the H_u's and the H_c's are very similar across the scenarios of 1,000 and 2,000 observations per replication). Since the corresponding results are quite different for 100 observations per replication, this sample size does not appear to be large enough to properly characterize the uncertainty in the distributions of the input and output processes.

To conclude this example, the data in Table 4 illustrates the effect of initialization bias on the information theoretic statistics. In each case the simulation run was initialized with an empty queue and an idle server. Even with initialization bias, this procedure is able to detect the same general pattern of sensitivity noted from the results in Table 1.

4 CONCLUDING REMARKS

This paper has illustrated an information theoretic approach to computer simulation sensitivity analysis. The approach, which is based on the conditional entropy function, is nonparametric and requires very few assumptions for its use. This procedure provides information on the uncertainty in the distribution of the simulation output processes attributable to the simulation input processes. In turn, this information can be used as an aid in decisions regarding resource allocation for the improvement of the quality of simulation input data.

Future research includes applying another information theoretic measure to simulation sensitivity analysis. The *marginal incremental information* measures the conditional relationship between a set of contingent variables and a set of explanatory variables after accounting for the conditional information in the set of contingent values with respect to a second disjoint set of explanatory variables. In the simulation context, this measure allows for quantification of the relative importance of a specific subset of the input processes in addition to the information already supplied by another subset of input processes.

In the queueing example, the sample means of the processes were used in the analysis rather than the raw data. Although this approach provided good results, it is conceivable that some information was lost due to the smoothing effect associated with averaging the data. Therefore, a future research direction is to develop a procedure which makes direct use of the uncertainty in the raw data. Finally, it is important to note that input/input and output/output distributional relationships could also be examined using the sensitivity analysis procedure developed in this paper.

REFERENCES

Dalle Molle, J. W. 1989. Program Entropy. Working Paper.

Glasserman, P. 1991. *Gradient Estimation Via Perturbation Analysis.* Boston: Kluwer Academic Publishers.

Gokhale, D. V. and S. Kullback. 1978. *The Information in Contingency Tables.* New York: Marcel Dekker.

Golden, L. L., P. L. Brockett, and M. R. Zimmer. 1990. An Information Theoretic Approach for Identifying Shared Information and Asymmetric Relationships Among Variables. *Multivariate Behavioral Research* 25:479-502.

Guiasu, S. 1977. *Information Theory with Applications.* New York: McGraw-Hill International Book Company.

Kleijnen, J. P. C. 1987. *Statistical Tools for Simulation Practitioners*, New York: Marcel Dekker.

Papoulis, A. 1984. *Probability, Random Variables, and Stochastic Processes.* Second Edition. New York: McGraw-Hill, Inc.

Yaglom, A. M. and I. M. Yaglom. 1983. *Probability and Information.* Dordrecht, Holland: D. Reidel Publishing Company.

AUTHOR BIOGRAPHIES

JOHN W. DALLE MOLLE is an academic visitor at Imperial College in London, England. He holds Masters degrees in both Mathematics and Petroleum Engineering and a Ph.D. in Management Science from The University of Texas at Austin. His research interests include the statistical analysis of simulated data and higher order spectral analysis of time series data.

DOUGLAS J. MORRICE is an assistant professor in the Department of Management Science and Information Systems at The University of Texas at Austin. He received his undergraduate degree in Operations Research at Carleton University in Ottawa, Canada. He holds a M.S. and Ph.D. in Operations Research and Industrial Engineering from Cornell University. His research interests are in the statistical design and analysis of large scale simulation experiments and the statistical aspects of quality control. He is a member of the The Institute of Management Science, the Operations Research Society of America, and the American Statistical Association.