A MULTIVARIATE RANK SUM TEST FOR NETWORK SIMULATION VALIDATION

Ann E. M. Brodeen and Malcolm S. Taylor
(410) 278-8947/6638

U.S. Army Research Laboratory
Advanced Computational and Information Sciences Directorate
Aberdeen Proving Ground, MD 21005-5067

ABSTRACT

Simulation is a widely accepted means of analyzing systems that are too complex to model analytically. Most communications systems fall into this category. But simulation credibility suffers when a continuing verification and validation program is not undertaken, thereby diluting the value of analyses that simulations support.

A primary goal of any verification and validation process should be to enhance both the correctness of a simulation and the confidence placed in its results. One challenge is to develop a process that is at the same time feasible and compatible with an organization’s needs, and can be applied to both existing simulations as well as new ones.

Multivariate methods can be used to test the hypothesis of agreement between simulated predictions and empirical observations. This paper describes a statistical test useful for the validation of simulations of battlefield communications networks. The method employs a multivariate nonparametric rank sum test with the aid of a computer-intensive randomization procedure to assess the significance of the defined test statistic. For illustrative purposes, the validation procedure is applied to a simulation that was developed to duplicate a configuration tested in FY91, in which "messages" were passed over a communications network using the combination of the Tactical Fire Direction System (TACFIRE) protocol and Single-Channel Ground and Airborne Radio System (SINCGARS) Combat Net Radios (CNR).

The purpose of this research is to strengthen the link between experimentation and simulation, both of which should be utilized in evaluating communications systems’ measures of performance (MOP).

INTRODUCTION

Nonparametric multivariate statistical methods address a broad class of problems in data analysis in which the assumption of normality is not feasible, and where the data occur naturally as n-tuples (vectors) rather than scalar values. This is the data structure that is most common in the engineering sciences and, coincidentally, the least tractable. A computer-intensive approach to the analysis of these data, usually referred to as randomization or permutation procedures, will be the specific focus of this work. Tests based on permutations of observations are nonparametric tests. This study considers a multivariate extension of the well-known Kruskal-Wallis rank sum test as a method for hypothesis testing, a technique commonly employed for simulation validation. The test statistic investigated is a nonparametric analogue of the classical Hotelling $T^2$ used for the normal theory model.

This undertaking is part of a broader-based Army research program, the goal of which is to improve the ability of communications networks to deliver critical information on the battlefield when and where it is needed despite a rapidly changing and often hostile environment. It will also support the ongoing effort to formalize the validation process for network simulation that, in turn, provides the groundwork for exploring alternatives and testing hypotheses throughout the research program. This formalization of the validation process can be readily transmitted to other organizations that rely on network simulations for their analyses.
LIMITED BANDWIDTH TACTICAL NETWORKS

The purpose of a network is to serve as a carrier of information from one point to another. On a limited bandwidth tactical network, the number of nodes and the amount of information to pass can be large, especially during peak battle periods. The effective distribution of information can enhance the decision process on the battlefield, while the impact of making decisions from old information can be catastrophic (Ref. 1).

To measure a network's effectiveness, one must determine whether the messages the network services arrive at their destination correctly and in time to be useful. The amount of correctly passed information is referred to as "network throughput," and the amount of time required to pass that information as "network delay." There are a number of parameters that can impact throughput and delay, for example, the number of messages to be transmitted, the size of the messages, the number of nodes on the network, the communications protocol, and the communications hardware. If the interaction of these network parameters is understood, the network's effectiveness can be optimized.

One way to examine the interaction of network parameters is through simulation. But communications protocols are often too complex to model precisely. The simulations often take required input, such as the probability two or more messages will collide, the expected delay in message transmission, or the arrival rate of messages at a given node, and extrapolate those estimates to a large scenario of multiple nodes. These drastic assumptions, usually made to simplify the simulation, may actually result in an unrealistic representation of the protocol. Controlled experimentation with the actual communications protocol on the intended hardware offers much insight into the behavior of the protocol under various conditions, facilitating the modeling and simulation efforts (Ref. 1).

Experimentation with a simulation is only a surrogate for actually being able to experiment with an existing or proposed system. A reasonable goal of validation is to ensure that a simulation is developed that can actually be used by a decision-maker to make the same decision that would have been made if it were feasible and cost-effective to experiment with the actual system. Validation should enhance both the correctness of a simulation and the confidence placed in its results. The challenge is to develop a validation process that is at the same time effective to experiment with the protocol under various conditions, facilitating the modeling and simulation efforts (Ref. 1).

Simulation and modeling are widely accepted means of analyzing real-world systems that are too complex to model analytically. Most communications networks fall into this category. But model credibility suffers when a continuing verification and validation program is not undertaken, thereby diluting the value of analyses the models support. It is not uncommon within a military organization to find several groups each developing a network simulation that performs essentially the same tasks; the differences usually lie in the model assumptions and/or definitions of simulation responses. An independent evaluator is called upon to assess the performance of several simulations against limited empirical data. The product of this research will be the formalization of a multivariate multisample rank sum test that will enhance long-term efforts to standardize the process of building, verifying, and validating command, control, and communications simulations for flexibly addressing issues related to low-level information distribution on the battlefield and serve to strengthen the link between experimentation and simulation, both of which should be utilized in evaluating communications protocols' MOP.

CONDITIONAL NONPARAMETRIC HYPOTHESIS TESTS

We consider the construction of nonparametric (distribution-free) hypothesis tests whose critical regions are determined from information gained from observed data. The critical region is thus conditional, since it can be created only after the data have been observed. Nonetheless, the test procedure has overall significance level $\alpha$ because the critical region is constructed to assure the conditional probability of rejecting a valid null hypothesis $H_0$ remains $\alpha$. Conditional hypothesis tests are discussed at several levels of theoretical intensity, ranging from Conover (Ref. 2), Noreen (Ref. 3), through Randles and Wolfe (Ref. 4), Edgington (Ref. 5), to Puri and Sen (Ref. 6). Our ultimate interest lies in hypothesis testing in a multivariate multisample framework.

The construction of conditional tests finds application in fundamental considerations of multivariate analysis where counting and ranking techniques do not lend themselves effectively to small sample situations. Puri and Sen (Ref. 6) provide a rigorous treatment of the use of conditional tests in dealing with problems in multivariate data analysis. The approach in the following sections corresponds in the main to their development.

Let

$$X^k_j = \left( x_{1j}, \ldots, x_{pj} \right)^\prime,$$

$$j = 1, \ldots, n_k,$$

$$k = 1, \ldots, c$$

(1)
be independent vector-valued random variables from continuous distributions with the cumulative distribution functions (c.d.f.) of $X^k$ denoted by $F_k(x)$, $k = 1, \ldots, c$. The data structure is that of a multivariate multisample ($2 \leq c$) location problem, i.e.,

$$F_k(x) = F(x - \delta_k), \quad k = 1, \ldots, c \quad (2)$$

and the interest is in testing $H_0: \delta_1 = \cdots = \delta_c$ against the alternative $\delta_r \neq \delta_s$ for some $r \neq s$.

The combined sample of these data is naturally represented as a matrix of observations of the form

$$X = \begin{bmatrix}
X^1_{11} & \cdots & X^1_{1n_1} & \cdots & X^c_{1n_c} \\
\vdots & & \vdots & & \vdots \\
X^1_{p1} & \cdots & X^1_{pn_1} & \cdots & X^c_{pn_c}
\end{bmatrix} \quad (3)$$

where $X$ is a $p \times N (= \Sigma n_k)$ matrix in which the columns are the vector-valued observations, i.e.,

$$X = \begin{bmatrix} X_1^1, \ldots, X_{n_1}^1, X_1^2, \ldots, X_{n_c}^c \end{bmatrix} \quad (4)$$

The development of the data analysis proceeds under the influence of the following consideration. In the construction of distribution-free procedures, the data are often replaced by their ranks. This transformation may be attractive for a number of reasons, a principal one being the distribution of a test statistic need be established only once. Otherwise, a customized test must be constructed for every set of data. The customized test in which the data themselves serve as scores was an idea originally advanced by Fisher (Ref. 7)—rank tests and tests based on rank scores are descendants. The development to follow, consistent with Puri and Sen (Ref. 6), will use the rank representation, but with the understanding that different score functions remain a viable alternative approach.

If the $N$ observations on the $j$th variate $X_j^k$, $j = 1, \ldots, n_k$, $k = 1, \ldots, c$, are arranged in ascending order, and $R^k_{ij}$ denotes the rank of $X_j^k$, the observation matrix $X$ gives rise to a corresponding matrix of ranks

$$R = \begin{bmatrix}
R^1_{11} & \cdots & R^1_{1n_1} & \cdots & R^c_{1n_c} \\
\vdots & & \vdots & & \vdots \\
R^1_{p1} & \cdots & R^1_{pn_1} & \cdots & R^c_{pn_c}
\end{bmatrix} \quad (5)$$

Each row of this matrix is a random permutation of the integers $1, \ldots, N$, and thus $R$ is a random matrix that can have $(N!)^p$ possible realizations.

Under $H_0$, all the observations, $X_j^k$, $j = 1, \ldots, n_k$, $k = 1, \ldots, c$, have the same distribution. Consequently, for each variate $i$, the mean of the ranks assigned to the $k$th sample

$$T^k_i = \frac{1}{n_k} \sum_{j=1}^{n_k} R^k_{ij} \quad (6)$$

should be close in value to the overall mean $\bar{E}_i$, where

$$\bar{E}_i = \frac{1}{N} \sum_{k=1}^{c} n_k T^k_i \quad (7)$$

[The expression for $\bar{E}_i$ is unnecessarily cumbersome for this application, since $\bar{E}_i$ is simply the mean of the integers $1, \ldots, N$, therefore, $\bar{E}_i = \frac{N + 1}{2}$, $i = 1, \ldots, p$. It is written in this form to allow for subsequent inclusion of scores other than ranks.]

A test for $H_0$ based on the contrasts between the mean scores $T^k_i$ is intuitively appealing. The set of $p(c - 1)$
contrasts $T_i^k - \overline{E}_i$, $i = 1, \ldots, p$, $k = 1, \ldots, c$, should, under $H_0$, be numerically small stochastically. For a global assessment of $H_0$, a test based on a function of the contrasts that would be sensitive to the numerical largeness of any contrast seems appropriate. A positive definite quadratic form in $T_i - \overline{E}_i$ will accommodate this. Puri and Sen (Ref. 6) advance as a test statistic

$$L_N = \sum_{k=1}^{c} n_k \left[ \left( T^k - \overline{E} \right) V^{-1}(R) \left( T^k - \overline{E} \right) \right]$$

where $T^k = \left( T^k_1, \ldots, T^k_p \right)$ and $\overline{E} = \left( \overline{E}_1, \ldots, \overline{E}_p \right)$. $L_N$ is a weighted sum of $c$ quadratic forms in $\left( T^k - \overline{E} \right)$, $k = 1, \ldots, c$, with a common discriminant $V^{-1}(R)$.

The quadratic form $L_N$ is attractive in that the correlation structure between the variates $i = 1, \ldots, p$, is taken into account through the covariance matrix $V(R)$. Scaling of the variates was simultaneously accomplished by assignment of ranks.

**TACTICAL NETWORK EXPERIMENT AND SIMULATION DEVELOPMENT**

A controlled laboratory experiment was conducted at the U.S. Army Research Laboratory's (ARL) Command, Control, Communications, and Computers (C4) Research Facility during the summer of 1991 to evaluate the performance of a tactical communications protocol over CNR (Ref. 8). The approach was to quantify the effects of message arrival rate and message length on the throughput and delay of a small combat radio net. The results provided statistically sound baseline information to be used as input for network simulations, partial guidelines for designing network architectures and communications protocols, and for future experiments on combat radio nets. Eventually a small communications network simulation was developed utilizing the OPNET tool, duplicating the configuration of the aforementioned experiment.

The two factors tested in the experiment were message arrival rate and message length. Four levels of message arrival rate were tested with each of four levels of message length (i.e., a full-factorial design), yielding sixteen test combinations. The levels for message arrival rate were 100, 250, 350, and 500 messages per node. The levels for message length were 48, 144, 256, and 352 characters.

It was decided the shortest reasonable time to test any of the 16 combinations was 1 hr. Since the testing of all 16 combinations required a minimum of 16 hr for a single replication, which realistically could not be completed in 1 day, a randomized incomplete block design was constructed in order that day-to-day variability would not influence the results. The 16 combinations were divided into blocks of size 4, and the 4 blocks were run over a 4-day period. The assignment of the combinations into blocks was based on a confounding scheme. This scheme, in which a different set of three of the nine degrees of freedom for the interaction term were completely confounded within each replication, assured the effects of message arrival rate and message length, as well as their interaction, on network throughput and network delay could be measured. Three replications of the design matrix were made to insure the incomplete block design was balanced, thereby facilitating the analysis, although part of the precision of the estimate of the interaction effect was sacrificed (i.e., the relative information available for the interaction term was two thirds).

The experiment consisted of four nodes, each of which was a SUN workstation, communicating over a tactical network. Each contained a message driver, providing communications loading, and data collection software to log the sending and receipt of messages and acknowledgments as well as information on queues, as depicted in Figure 1. The nodes were connected to modems to enable communications via radios that could communicate in single-channel (SC) or in frequency-hopping (FH) mode. It was decided to simulate only the SC capability. The modems allowed communication using a specified tactical net-sensing algorithm and communications protocol. To minimize error rates, the radios were placed no more than 3 ft apart and were, therefore, set to low power. Resistor loads were used in place of antennas to avoid interference. The analogous four-node simulation configuration utilizing the OPNET tool is represented in Figure 2. Figure 3 depicts the structure of an individual tactical node. Each node has three processor modules, a queue module that performs the bulk of the channel-access processing, and a pair of radio receiver and transmitter modules.

The four message arrival rates emulated the rate of actual user generated messages and specific nodes' ability to respond to incoming messages. For the experiment, the arrival rate, $\lambda$, represented the number of messages generated during a 1-hr test cell and queued for transmission on the net, not the number of messages actually transmitted during the hour. A message was assumed to enter network service when it reached the
A scenario generator was written to create "messages" of character strings of a specified length and arrival rate over a 1-hr period. The simulation, then, had to accommodate varying message lengths and arrival rates. Once the message was generated, the communications protocol added several layers of information to insure the message arrived at its destination. This included five error correction/detection bits for each seven-bit character, four synchronization characters, and a preamble to bring the transmitter to full power before the message was sent. Acknowledgments, though shorter in length, were wrapped with similar overhead bits. In the experiment, the numbers of messages generated for transmission each hour by each node were assumed to be mutually independent Poisson-distributed random variables with parameter, \( \lambda_1 \). The messages were equally distributed among the four nodes. For example, if the arrival rate was 2,000 messages/hour, the scenario generator created a file of 500 messages for each node. Each of these details was represented in the simulation. Thus, the simulation represented an actual system using a real and less-than-trivial protocol.

**APPLICATION TO NETWORK SIMULATION VALIDATION**

We consider the special case of comparing two systems (i.e., "real-world" and simulated) on the basis of several carefully selected performance measures. We effect this comparison by determining whether \( F_1(x) \) and \( F_2(x) \) differ in location. The data consist of two independent vector-valued random samples. The \( k^{th} \) random sample is of size \( n_k \), where \( k = 1, 2 \). Denote the empirical observations as \( X_j^1, j = 1, 2, 3 \); denote the simulated observations as \( X_j^2, j = 1, \ldots, 7 \). The total number of observations is \( N = n_1 + n_2 = 3 + 7 = 10 \). There are no missing observations or tied values to consider.

Although data for a number of MOP were collected during the experiment, comparisons between empirical and simulation results were limited to network throughput, network delay, and utilization. These were the only measures that could be defined by continuous random variables.

Network throughput is the average number of information bits that were successfully transmitted and acknowledged over a one-hour test cell. Throughput does not include such overhead as the acknowledgments themselves or, in the event of collisions, message retransmissions. It does, however, include error
Network delay is the average time that passes between a message's arrival at a host's modem until the acknowledgment returns to the host. Messages that were never completely serviced during the running of a test cell were not considered in computing network delay.

Network utilization for a particular time interval is the amount of time spent actually transmitting messages, message retransmissions, or acknowledgments during that interval, divided by the amount of time in the interval. Messages, retransmissions, and acknowledgments include the preamble and other protocol overhead in addition to actual transmission bits.

While 16 combinations of message arrival rate and message length were tested in the 1991 experiment, only 8 were chosen for the validation study. The eight combinations were not chosen in a purely random fashion as it was desirable to ensure the simulation would be evaluated at the two extremes of both parameter ranges (i.e., arrival rate of 400 messages and message length of 48 characters; arrival rate of 2,000 messages and message length of 352 characters). One component of the highest-order interaction was confounded such that the 16 combinations were divided into 2 blocks of 8 units each. The principal block was selected as it contained the two extreme conditions mentioned previously. Given the data were not all collected under the same conditions, each combination was treated as a homogeneous grouping and, therefore, each served as an independent case to test the null hypothesis.

The appropriate empirical observations were taken from each of the three replications performed for the 1991 experiment. The simulation was not run with the scenarios generated for the experimental test cells to ensure the independence of the sample observations. The capability to utilize actual message scenarios as simulation input does, however, allow the developer a useful tool for verification.

For these data, the number of permutations possible is \(N!\) and the number of distinct values of the \(L_N\) statistic possible is \(\left( n_1! n_2! \cdots n_c! \right) \). Since this validation study deals with small values of \(N\) (\(N = 10\)) and \(p = 3\), the \(L_N\) statistic may be calculated for all permutations of the observations that lead to distinct values of the statistic and, thus, an exact application of the permutation test is possible. We have elected to replace the ranks \(R_{ij}\) by a rank score function of the form \(\frac{R_{ij}^k}{(N+1)}\), thereby reducing \(T_1^k\) to \((N+1)^{-1}\) times the average rank of the \(k^{th}\) sample, \(i^{th}\) variate observations among the combined sample \(i^{th}\) variate observations. [The motivation for this choice was the univariate case \((p = 1)\) for which the statistic \(L_N\) reduces to the Kruskal-Wallis test.] This test was applied to an OPNET communications network simulation in order to validate simulated output with respect to empirical observations. The hypothesis (Section 3) was tested for the MOP outlined earlier in this section. The significance or \(P\)-value is the proportion of the \(10!3!7! = 120\) data permutations providing an equivalent or larger test statistic than that obtained for the reference, or observed, set. Assuming an a priori significance level of 0.05, the null hypothesis was rejected in five of the eight test combinations. The observed test statistic values and the resultant \(P\)-levels are summarized in Table 1.

We note that the testing procedure and, hence, the conclusion reached, depends on the measure of location shift that is employed. Alternative measures could be employed; however, use of a different estimator for \(\delta\) will likely produce a different conditional rejection region. Since we are free to choose the test statistic, an alternate statistic suggested by Chung and Fraser (Ref. 9) was considered.

The theoretical contributions of Chung and Fraser, while substantial, have generally gone unnoticed. They proposed several randomization tests for the multivariate two-sample problem that were initially developed for the normal-theory two-sample problem for which the Hotelling T² test does not exist; however, the tests are also valid in a more general context as nonparametric tests. The approach of Chung and Fraser is to select a statistic suitable for the univariate case, apply it to each of the \(p\) variates, and add the resulting expressions. This approach does not take into account covariances as is required with the nonparametric counterpart of the Hotelling's T² statistic. For measuring shift in location alternatives, a rank test may be obtained by recording ranks as done previously and using the absolute value of the difference in sample means as a test statistic. One of the forms of the Chung and Fraser statistic, and the one utilized in this study, is \(\sum_{i=1}^{p} \left| \bar{x}_i - \bar{y}_i \right|\), appealing in its simplicity (Ref. 9).

The observed test statistic values and the resultant \(P\)-levels for Chung and Fraser's rank test are summarized in Table 2. The null hypothesis was rejected for six of the eight test conditions. The conditional rejection region...
REFERENCES


differs from the one established previously. Indeed, the same conclusion regarding the simulation's validity was established for only three of the eight test conditions. However, exploratory data analysis suggested a correlation structure among the three variates. This was expected from theoretical considerations of the communications network. This structure is not incorporated in the Chung and Fraser statistic and may account for the observed differences.

<table>
<thead>
<tr>
<th>Input Condition (messages, characters)</th>
<th>H0: ( \delta_1 = \cdots = \delta_k = 0 )</th>
<th>Observed Statistic</th>
<th>P-Value</th>
<th>Reject/Fail to Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td>400, 48</td>
<td>7.184996</td>
<td>0.04167</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>400, 256</td>
<td>9.035234</td>
<td>0.00833</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>1,000, 144</td>
<td>6.970858</td>
<td>0.05833</td>
<td>Fail to Reject</td>
<td></td>
</tr>
<tr>
<td>1,000, 352</td>
<td>9.651814</td>
<td>0.00833</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>1,400, 48</td>
<td>7.82677</td>
<td>0.02500</td>
<td>Reject</td>
<td></td>
</tr>
<tr>
<td>1,400, 256</td>
<td>6.581197</td>
<td>0.10000</td>
<td>Fail to Reject</td>
<td></td>
</tr>
<tr>
<td>2,000, 144</td>
<td>6.754517</td>
<td>0.07500</td>
<td>Fail to Reject</td>
<td></td>
</tr>
<tr>
<td>2,000, 352</td>
<td>9.210527</td>
<td>0.00833</td>
<td>Reject</td>
<td></td>
</tr>
</tbody>
</table>

SUMMARY

As reliance upon computer simulations to model processes that resist analytical description increases, so does the need to validate the simulations themselves. An impartial approach to simulation validation is through statistical hypothesis testing. In this paper, an application of a nonparametric multivariate procedure to assess the validity of a communications network simulation model, whose intent is to emulate a limited bandwidth combat radio net, is detailed. The procedure, sometimes described as a permutation or randomization test, offers considerable flexibility to the analyst charged with maintaining the fidelity of the modeling effort.