Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

COMBAT SIMULATION ANALYTICS: REGRESSION ANALYSIS, MULTIPLE COMPARISONS AND RANKING SENSITIVITY

Andrew Gill
Dion Grieger
Martin Wong
William Chau

Defence Science and Technology Group
23 Laboratories, Third Avenue
SA, 5111, AUSTRALIA

ABSTRACT

The design and analysis of simulation experiments and the analysis of simulated alternatives are critically important tasks when employing combat simulations in support of Army modernization. Effective sensitivity and ranking analyses enable insight to be gained on the marginal contributions of sub-systems to overall operational effectiveness, as well as comparatively assessing competing system alternatives. This paper makes some unifying discoveries related to sensitivity analysis via linear multiple regression, and discusses some empirical findings concerning ranking analysis via score-based, partition-based and consensus-based methods. Finally, an investigation into the combined ranking sensitivity problem is also attempted.

1 INTRODUCTION

Combat simulations often form part of a multi-method approach to compare the operational effectiveness of alternative military systems or tactics. These simulations complement other methods such as war-gaming, field trials and historical analysis while allowing for the exploration of a larger parameter space and the ability for a scenario to be replicated many times in order to conduct statistical comparisons. Data from events related to movement, acquisition, fires, communications and resource consumption can be captured and analyzed to investigate the cause-and-effect relationships in military environments.

Warfare is extremely complex which makes the parameterization of the detailed physical and behavioral aspects of war-fighting a challenging task for the combat simulation community. As a result there is a large multi-dimensional space of uncertain parameters associated with each simulated military scenario (Sanchez et al. 2012). Examples of these uncertain parameters could include the detection algorithm chosen (or the value of the constants within those algorithms), the fidelity with which the terrain or environment is modeled, or the likelihood that an entity will choose a particular course of action. As these parameters are usually internalized settings of the combat simulation, each combination represents a plausible model of a real world situation and it follows that the insights generated from the performance rankings of the alternative systems should consider the entire parameter space in order to reduce the risk of presenting an outlier result. However, it is unrealistic to enumerate, simulate and collate results from every single possible model.

The Design and Analysis of Simulation Experiments (DASE) is the process of deciding which input parameter combinations should be used to interrogate the simulation for an output response and how the corresponding input/output transformation should be analyzed (Kleijnen 2015; Fang et al. 2005; Santner et al. 2003). In practice, the parsimony principle often applies, where only relatively few parameters of combat simulations have a substantial influence on the simulated response, and for which screening designs are critical (see Shen and Wan (2009)). While uncertainty analysis and robust optimization are

also common activities performed within DASE (see Kleijnen (2015)), our particular interest is with the sensitivity analysis of these influential parameters. Quantifying the form and magnitude of the effects by fitting mathematical or statistical meta-models to the input/output transformation allows ranking of each parameter's importance and answering various 'what-if' questions, which may be used to assess the marginal contributions to the overall operational effectiveness of proposed technology enhancements to military systems.

The Analysis of Simulated Alternatives (AoSA) is the process of estimating the (relative) effectiveness of a discrete number of alternative systems by analyzing the ensemble of corresponding simulation responses (Goldsman and Nelson 1994; Hochberg and Tamhane 1987). Our particular interest is with ranking analysis, whereby a complete ordering of the set of alternatives from best to worst is sought, which may be useful when a full tender evaluation of competing alternatives is required.

We present new DASE research findings and empirical observations made during several AoSA case studies performed in support of the Australian Army in modernizing its combat vehicle systems. Section 2 reexamines classical linear regression within sensitivity analysis and the various modifications for departures from the standard normal, independent and identically distributed assumptions. Section 3 compares and contrasts several approaches for ranking analysis while Section 4 explores the problem of conducting sensitivity analysis on a ranking analysis. We conclude by providing some summary remarks in Section 5.

2 REGRESSION ANALYSIS RESEARCH

2.1 Normal, Independent and Identically Distributed Responses

Kleijnen (2015) is one of the seminal texts on DASE. Linear regression over q input parameters approximates the simulation response at each design point $\mathbf{x}_i$ (combination of input parameter level settings) by a first-order polynomial meta-model $\hat{y}_i = \sum_{j=0}^{q} x_{ij}\hat{\beta}_j, \quad i = 1,\ldots,n$ where $x_{i0} = 1\,\forall i$. When we have m_i replications at each design point, this approximation can be fitted using Ordinary Least Squares (OLS), leading to the linear equations:

$$\sum_{i=1}^{n} x_{ik} m_i \sum_{j=0}^{q} x_{ij}\hat{\beta}_j^{OLS} = \sum_{i=1}^{n} x_{ik} m_i \overline{w}_i, \quad k = 0,\ldots,q$$

which in matrix form is $X'MX\hat{\beta}^{OLS} = X'M\overline{\mathbf{w}}$ where M is an $n \times n$ diagonal matrix with entries $M_{ii} = m_i$; X is the $n \times (q+1)$ design matrix; $\overline{\mathbf{w}}$ is the $n \times 1$ vector of simulation output averages; and prime denotes matrix transpose. So the OLS estimator for $\hat{\beta}$ is a linear function $\hat{\beta}^{OLS} = L\overline{\mathbf{w}}$ where $L = (X'MX)^{-1}X'M$ is a $(q+1) \times n$ matrix and in the case where the number of replications of each design point is constant, $M = m\mathbf{I}$ and L reduces to the usual $(X'X)^{-1}X'$. To compute confidence intervals for individual regression parameters $\hat{\beta}_j^{OLS} = \sum_{i=1}^{n} L_{ji}\overline{w}_i$ we treat the sample averages $\overline{w}_i$ as random variables, so that

$$var(\hat{\beta}_j^{OLS}) \quad = \quad \sum_{i=1}^{n}\sum_{i'=1}^{n} L_{ji}L_{ji'}cov(\overline{w}_i, \overline{w}_{i'}) = \sum_{i=1}^{n}\sum_{i'=1}^{n} \frac{L_{ji}L_{ji'}\sigma_{ii'}}{\max(m_i, m_{i'})} \tag{1}$$

using the bi-linearity property of the (co)variance function; the fact that w_{ir} and $w_{i'r'}$ are independent except for when $r' = r$; and where $\sigma_{ii'} = cov(w_i, w_{i'})$. Note that these results generalize and simplify those in Kleijnen (2015) (pages 36,103) who treats the cases of constant/non-constant number of replications separately.

To estimate the covariance matrix Σ_w, if the w_i are independent and identically distributed (i.i.d), then $\Sigma_w = \sigma^2(w)\mathbf{I}$ and this constant variance $\sigma^2(w)$ can be estimated by pooling the n (classic) sample variance estimators $s^2(w_i)$ using their degrees of freedom as weights

$$\sigma^2(w) \approx s^2(w) = \frac{\sum_{i=1}^{n}(m_i - 1)s^2(w_i)}{\sum_{i=1}^{n}(m_i - 1)}$$

so that (1) becomes (where $N = \sum_{i=1}^{n} m_i$)

$$var(\hat{\beta}_j^{OLS}) \approx \hat{var}(\hat{\beta}_j^{OLS}) = \sum_{i=1}^{n}\sum_{i'=1}^{n} L_{ji}^2 \sum_{r=1}^{m_i} \frac{(w_{i'r} - \overline{w}_{i'})^2}{m_i(N-n)}, \quad j = 0, \ldots, q.$$

Kleijnen (2015) also proposed in this i.i.d case an alternative variance estimator for $\sigma^2(w)$, supposedly based on the mean squared residual (*MSR*) of the OLS regression

$$\sigma^2(w) \approx \frac{\sum_{i=1}^{n} m_i(\hat{y}_i - \overline{w}_i)^2}{\sum_{i=1}^{n} m_i - (q+1)}. \tag{2}$$

However, (2) can be rearranged to the following form

$$\sigma^2(w) \approx \frac{\sum_{i=1}^{n} m_i \left[\sum_{r=1}^{m_i}(\hat{y}_i - w_{ir})/m_i\right]^2}{\sum_{i=1}^{n} m_i - (q+1)}$$

where the numerator represents the sum of weighted squared average residuals. This is qualitatively different to the usual sum of squared residuals, and as such, risks some residuals canceling each other out in the calculation of the average residual at each design point and therefore underestimating the *MSR*. Furthermore, it is easy to verify that the Kleijnen (2015) formula ((2.26)) does not correctly reduce to the equation for the case of a constant number of replications ((2.25)) (while the numerators are the same, the denominators, $n-(q+1)/m$ vs. $n-(q+1)$ differ). The correct calculation of the *MSR* is in fact

$$\sigma^2(w) \approx MSR = \sum_{i=1}^{n}\sum_{r=1}^{m_i}(\hat{e}_{ir})^2/(\sum_{i=1}^{n} m_i - (q+1)) = \sum_{i=1}^{n}\sum_{r=1}^{m_i}(\hat{y}_i - w_{ir})^2/(N-(q+1)).$$

Consequently, the Kleijnen (2015) lack-of-fit F-statistics ((2.30) and (2.31)), which is the ratio of the *MSR* to the classic variance estimator $s^2(w)$, is incorrect and risks suggesting an adequate regression when it may not be so. The correct, and general, expression is in fact

$$F_{N-q,N-n} = \frac{MSR}{s^2(w)} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{m_i}(w_{ir} - \hat{y}_i)^2/(N-(q+1))}{\sum_{i=1}^{n}\sum_{r=1}^{m_i}(w_{ir} - \overline{w}_i)^2/(N-n)}.$$

2.2 Equivalence of Methods for Departures from NIID

In practice it is rare for combat simulation response data to follow the i.i.d assumptions. For example, if common random numbers (CRNs) are used (popular as a variance reduction technique), then Σ_w is no longer diagonal, but can be approximated by the sample covariances (here a constant number of replications for each design point must be used ($m_i = m$)) so

$$var(\hat{\beta}_j^{OLS}) \approx \sum_{i=1}^{n}\sum_{i'=1}^{n} L_{ji}L_{ji'} \sum_{r=1}^{m} \frac{(w_{ir} - \overline{w}_i)(w_{i'r} - \overline{w}_{i'})}{m(m-1)}.$$

Kleijnen (2015) also proposed an alternative method that (supposedly) does not require this estimation of Σ_w and was inspired by one of the seminal texts on simulation modeling (Law 2007). Kleijnen (2015) defines $\hat{\beta}_{j;r}^{LAW} = \sum_{i=1}^{n} L_{ji} w_{i;r}$ as the r-th point estimate for the j-th regression coefficient, from which the sample mean and sample variance are used to construct the associated confidence intervals. However

$\hat{\beta}_j^{LAW} = \sum_{r=1}^{m} \hat{\beta}_{j;r}^{LAW}/m = \sum_{r=1}^{m}\sum_{i=1}^{n} L_{ji} w_{i;r}/m = \sum_{i=1}^{n} L_{ji}\overline{w}_i$ which is the same expression for $\hat{\beta}_j^{OLS}$, and

$$\begin{aligned} var(\hat{\beta}_j^{LAW}) &\approx \sum_{r=1}^{m} (\hat{\beta}_{j;r} - \hat{\beta}_j^{LAW})^2 / [m(m-1)] \\ &= \sum_{r=1}^{m} \left[\sum_{i=1}^{n} L_{ji}(w_{ir} - \overline{w}_i) \right]^2 / [m(m-1)] \\ &= \sum_{r=1}^{m} \sum_{i=1}^{n} L_{ji}(w_{ir} - \overline{w}_i) \sum_{i'=1}^{n} L_{ji'}(w_{i'r} - \overline{w}_{i'}) / [m(m-1)] \\ &= \sum_{i=1}^{n} \sum_{i'=1}^{n} L_{ji} L_{ji'} \sum_{r=1}^{m} (w_{ir} - \overline{w}_{i'})(w_{i'r} - \overline{w}_{i'}) / [m(m-1)] \end{aligned}$$

which is the same expression approximating $var(\hat{\beta}_j^{OLS})$. Thus this alternative method by Law (2007) is identical to the OLS estimation discussed by Kleijnen (2015).

The other assumption often challenged by combat simulation response data is that of output normality. When this is the case, an approach suggested by Kleijnen (2015) is to use jackknifing, where the r-th jackknifed pseuduovalue $J_{j;r} = m\hat{\beta}_j^{OLS} - (m-1)\hat{\beta}_{j;-r}^{OLS}$ is a weighted difference of the OLS estimators based on the simulation response averaged over all m replications and the simulation response averaged over all replications excluding the $r-$th. However

$$J_{j;r} = m\sum_{i=1}^{n} L_{ji}\overline{w}_i - (m-1)\sum_{i=1}^{n} L_{ji}\overline{w}_{i;-r} = \sum_{i=1}^{n} L_{ji}\left(\sum_{r'=1}^{m} w_{ir'} - \sum_{r'=1;r'\neq r}^{m} w_{ir'}\right) = \sum_{i=1}^{n} L_{ji} w_{ir}$$

which is the expression for $\hat{\beta}_{j;r}^{LAW}$. Thus jackknifing discussed by Kleijnen (2015) is identical to the alternative method by Law (2007). This sampled-based method by Law (2007) is therefore equivalent to both approaches recommended by Kleijnen (2015) and is thus a robust method to use if concerned about departures from the classical assumptions.

3 RANKING ANALYSIS RESEARCH

3.1 Comparison of Multiple Nonparametric Groups

As mentioned above, the use of CRNs and the presence of non-normally distributed simulation responses are typical situations with combat simulation analysis. It follows then that nonparametric statistical tests for paired, or dependent, samples should be employed in the AoSA. The Friedman test (Friedman 1937) can be used for discrete or continuous data to jointly compare all alternatives and determine if there is a statistical difference between any of the alternatives (or the Cochran test (Cochran 1950) in cases where the data are dichotomous). The Nemenyi post-hoc test (Nemenyi 1963) can then be used to determine which pairs of alternatives show a statistical difference. A limitation of these omnibus approaches is that the results may change if an alternative is added or removed from analysis. It is also possible that the omnibus null hypothesis is rejected but the post-hoc procedure fails to find any individual statistical differences.

An alternative approach is to conduct multiple pairwise comparisons using the Wilcoxon Signed Rank test (Wilcoxon 1945) or the McNemar test (McNemar 1947) in the case of dichotomous data. While avoiding the issues of the omnibus test, they introduce another potential inconsistency related to the transitivity of the results. An example is if alternative A is assessed as being statistically better than B and B better than C, but without A being better than C. The use of multiple comparisons also requires the use of family-wise error adjustment techniques (see Miller (1981)).

Fey and Clarke (2012) suggest there is no single nonparametric approach that completely addresses all of these issues. It follows that the challenge is to present the information to decision makers in a manner that is easy to understand but also acknowledges any limitations.

3.1.1 Score-Based Methods

One approach to overcome the intransitivity issues is to calculate a cumulative score for each alternative based on all pairwise comparisons. Villacorta and Sáez (2015) describe such a process. Starting from a score of zero, one point is added (subtracted) for each alternative it is statistically better (worse) than. If both alternatives are statistically equivalent then no change to the score is made. If there are K alternatives each begins with a score of zero and can achieve a maximum score of $K-1$ and a minimum of $-(K-1)$.

Dominance hierarchy approaches have been widely used to identify an individual's ordinal rank. The approach takes into account a scoring mechanism for dominance of one alternative over another. Kendall (1955) showed an example of a dominance hierarchy procedure through a chess tournament represented by preference matrices. The representation of the pairwise procedure allocated 1 for a win, 0.5 for a draw and 0 for a loss. Kendall (1955) described that continuous matrix multiplication eventually converges to resultant rankings. Reynolds (1976) described in the application of behavioral sciences the power of an individual as the row-sum of the matrix $D+D^2$ where D is the dominance matrix of results. In social dominance studies, Boyd and Silk (1983) described multiple dominance interactions can be recorded into a matrix containing the frequencies of wins and losses. Jensen (1986) illustrated a similar approach where a dominance matrix was used to find the eigenvalues and eigenvector scaling to achieve ordinal rankings to the pick-the-winner preference votings.

3.1.2 A Partition-Based Method

To address the issue of intransitivity Emond and Turnbull (2006) proposed the Breakpoint Analysis with Nonparametric Data Option (BRANDO) post hoc procedure, which draws on the knowledge that any rejection of the omnibus null hypothesis implies that there is at least two subgroups within the group of alternatives. BRANDO initially orders the alternatives according to their sample statistics and then combines the p-values from the two omnibus tests to determine the most likely split of those alternatives into two subsets. This concept is applied recursively, while maintaining the initial ordering until no further statistically significant subsets can be identified. A limitation of this approach for the analysis of combat simulations that utilize CRNs is that BRANDO assumes the alternatives are independent. We have developed an improved implementation of BRANDO that employs paired sample nonparametric tests is as follows:

1. Conduct the Friedman omnibus test to determine if any of the K alternatives differs significantly from the others. If the null hypothesis is accepted, then there is no need to proceed further and all alternatives can be ranked equally. Note that the original BRANDO algorithm uses the Kruskal-Wallis test whenever multiple datasets are compared. The improved algorithm proposed here replaces any use of the Kruskal-Wallis test with the Friedman test to account for the paired samples.
2. Order the K alternatives by increasing value of the average $rank_k = \sum_{r=1}^{m} rank_{kr}/m$ where $rank_{kr}$ is the rank of alternative k in replication r and m is the total number of replications (same for all alternatives). Note that the original BRANDO algorithm calculates the average rank when all observations from all alternatives are considered as one group which is not appropriate in the case of paired samples.
3. Consider the $K-1$ ways of breaking the K alternatives into two subgroups while maintaining the initial ordering. For each possible breakpoint conduct the Friedman test on each of the two resulting groups and take the product of the two observed p-values as the relative likelihood of the given breakpoint. The breakpoint for which this product is the greatest is selected.
4. Continue by testing the null hypothesis for each of the subgroups and, upon rejection, repeat the procedure in step 3 to determine further breakpoints within the subgroups. Continue until the null hypothesis is accepted in all subgroups, i.e. no more breakpoints can be found and noting that when there is only one alternative in a group, the p-value will be equal to one.

The BRANDO approach addresses the transitivity issue and also ensures that at least one breakpoint is found following rejection of the initial omnibus test. Emond and Turnbull (2006) also discuss family-wise error as it relates to BRANDO and provide an example of how the technique preserves the desired false alarm rate. However, an issue still exists with BRANDO in that the relative comparison between two alternatives may change if some other alternative is added or removed from analysis. It is also not possible to draw a statistical conclusion between all possible pairs of alternatives as not all combinations are tested. It is possible that two alternatives separated by a breakpoint actually fail to reject the null hypothesis when specifically compared.

3.2 Consensus Ranking

When alternatives are ranked individually across a number of different scenarios that are representative of the uncertain parameter space, it may be desirable to determine an overall, or consensus, ranking for those alternatives. The consensus ranking literature is vast and is summarized succinctly by Cook (2006). Of particular interest to the AoSA is the Emond and Mason (2002) approach which uses a rank correlation coefficient to account for ties between alternatives - a common occurrence in combat simulations.

Another approach is to extend the dominance matrix approach where the results can be assigned to indices from multiple scenarios. Boyd and Silk (1983) assigned multiple observations of dominance into a single matrix of dominance interactions. From this, a single ranking of the alternatives can be achieved from the aggregated dominance matrix by completing the power ranking procedure.

If each scenarios is assumed to be an equally plausible representations of the real world, then a third alternative is to treat the results from the entire set of scenarios as a single dependent sample. For instance if K alternatives were simulated in n equally plausible scenarios and replicated m times, then a single sample of size nm can be obtained for each alternative. The pairwise or partition based statistical approaches described above can then be applied to these K dependent samples to obtain a single ranking.

3.3 Ranking Analysis Case Study

The multiple comparison and consensus ranking approaches were applied to a case study typical of the Australian AoSA context. Consider $K = 8$ military systems (Alt1,...,Alt8) being evaluated for potential acquisition (e.g. a sensor, a weapon, or an entire platform). Each alternative was examined in $n = 6$ scenarios, $(S_1, \ldots, S_6)$ representing equally plausible combinations of uncertain parameters (e.g. weather conditions, opposing force behaviors or route options)and $m = 200$ replications were conducted for each scenario. The alternatives were assessed via eight discrete or continuous metrics $(M_1, \ldots, M_8)$ that capture the simulation outcomes of interest such casualties, detections, or time required to complete specific tasks. Initially, alternatives were ranked using three approaches - the score based approach of Villacorta and Sáez (2015) (referred to as VS approach), the dominance matrix approach and the partitioned based BRANDO method. Pairwise comparisons were made using the Wilcoxon Signed Rank Test and the Bonferroni method was chosen to adjust for family-wise error as the desire was to minimize the likelihood of Type 1 errors.

Across the 48 different combinations of metric and scenario the rankings of the three approaches were consistent in 34 cases (Table 1). The majority of the inconsistencies were in cases where the VS approach produced more groupings (splits) than the other approaches. This is an intuitive result given that the VS approach scores ties and losses differently. The results also show that it is possible for the BRANDO approach to occasionally produce more splits than the pairwise comparison approach. Presumably these cases are an occurrence of the example described earlier were the BRANDO test will always produce a split among groups whenever the omnibus test rejects the null hypothesis.

Table 1: Comparison of partition-based and score-based approaches across six scenarios and eight metrics. Three different outcomes are described. The first outcome (blank cell) describes different order of rankings among the approaches. The second outcome (in italics) shows similar order of rankings occurring with the approaches. The final outcome (in bold) is that the same rankings have occurred.

	Scenario 1	**Scenario 2**	**Scenario 3**	**Scenario 4**	**Scenario 5**	**Scenario 6**
M_1						
M_2	**Fewer splits for BRANDO**	*Fewer splits for Dominance approach*	*Fewer splits for Dominance approach*	**More splits for VS**	**More splits for VS**	**More splits for VS**
M_3	**More splits for VS**					
M_4				*More splits for BRANDO*		
M_5						
M_6		*More splits for BRANDO*			**More splits for VS**	**More splits for VS**
M_7	*More splits for BRANDO*					
M_8					**More splits for VS**	**More splits for VS**

A number of different consensus ranking approaches were then employed to offer further insight for metric M_2. The approaches used were:

1. Ave_Rank: The average ranking of each option across all 1200 replications and is calculated in the same way that average rank is calculated within the BRANDO algorithm. This was done in order to generate and ordering with no ties with which comparisons with other methods could be made.
2. SS_BR: Single sample implementation of the BRANDO approach.
3. SS_PW: Single sample implementation of the Villacorta and Sáez (2015) pairwise scored-based approach.
4. EM_1200: Takes all 1200 sets of rankings and determines the consensus rank using the Emond and Mason (2002) approach.
5. EM_BR6: Uses the Emond and Mason (2002) approach to determine the consensus rank of the BRANDO rankings corresponding to the six different scenarios.
6. EM_PW6: Uses the Emond and Mason (2002) approach to determine the consensus rank of the six pairwise scored-based rankings calculated using the Villacorta and Sáez (2015) approach.
7. EM_DM6: Uses the Emond and Mason (2002) approach to determine the consensus rank of the six pairwise scored-based rankings calculated using the dominance matrix approach.
8. AggDM6: Uses the aggregated dominance matrix approach and the results from the six sets of pairwise Wilcoxon Signed Rank tests to determine the consensus rank of the alternatives.

The consensus rank for the eight alternatives for each of these approaches for metric M_2 are shown in Figure 1. All produce different rankings with the exception of the single sample score-based approach (SS_PW) and the Emond and Mason consensus of the six BRANDO rankings (EM_BR6). There does appear to be some groupings between certain pairs of alternatives in many of the rankings. For example, Atlt1 and Alt3 are always ranked equally as are Alt5 and Alt7. However, there are some inconsistencies with the number of total groups within each set of rankings.

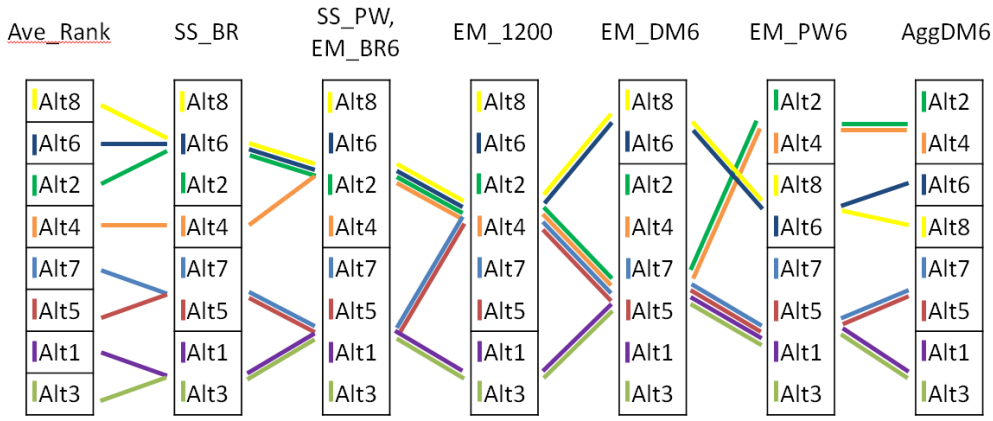

Figure 1: Comparison of consensus ranking approaches.

Examining all 1200 replications and calculating for each alternative how many other alternatives performed better in that particular replication (Figure 2) offers a richer insight into the comparison of the various consensus ranking methods and also provides some indication as to the deltas between the various alternatives. The results reinforce the pairs of alternatives that appear to be similarly ranked and also highlight that any method that splits Alt2 and Alt4 (SS_BR) or Alt8 and Alt6 (AggDM6) appears to be counter-intuitive. The reasons for the conflicting results in Figure 1 around the rankings of alternatives 2, 4, 6 and 8 can be better understood by noting that despite Alt8 and Alt6 having more instances where they are not outperformed by other alternatives that Alt2 and Alt4 there are also more cases where they are outperformed by five or more other alternatives.

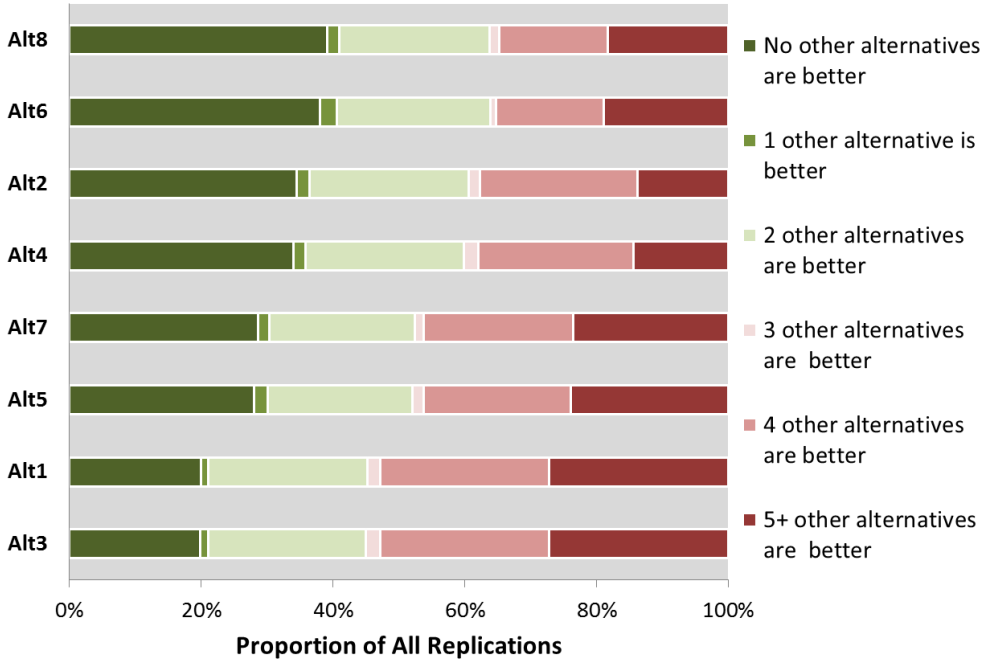

Figure 2: Summary of each alternative for all 1200 replications.

This case study clearly does not provide evidence for the selection of a single optimal method of either ranking alternatives within a single scenario, or determining a consensus ranking across multiple scenarios. However, it does highlight two important considerations for combat simulation data analysis. The first is to carefully consider the military problem in terms of the importance of penalizing a losing alternative in any pairwise comparison. The treatment of ties and losses is different among the ranking methods explored and thus the ranking method chosen should be aligned to the problem statement accordingly. The second is that there are risks associated with deriving insights from rankings alone. The underlying data should be examined in order to confirm the validity of the rankings and to offer further insights into the deltas between tiers within a ranked set of alternatives; using an approach similar to that shown in Figure 2.

4 RANKING SENSITIVITY ANALYSIS RESEARCH

4.1 Ranking Sensitivity Meta-Model

In the AoSA context sensitivity analysis approaches are required to understand the uncertainty in the final rank order of a set of K alternatives evaluated via a set of end of run metrics collected using a stochastic combat simulation. Due to the initial choice of model parameters affecting the metrics final value and thereby affecting the final rankings, the final rank order is sensitive to the choice of initial parameters. The aim is to identify (and rank) the set of model parameters that have a high probability of changing the result of our rankings. This shall be referred to this as the ranking sensitivity problem. In this section three approaches to the ranking sensitivity problem are described. The prevailing idea here is that confidence in the reported parameter rankings should increase if all three approaches present similar results. The approaches involve using a logistic regression model; a variance based method; and using a decision tree.

4.1.1 Logistic Regression

The ranking sensitivity problem in relation to a particular alternative can be phrased in terms of the probability of the ranking of that alternative changing as a result of changes in parameter $\mathbf{x}$. Sensitive parameters would then be those that quickly shift the probability of a rank change from low to high for relatively small input change sizes. The problem is addressed with a logistic regression model $P(Z = 1||\Delta\mathbf{x}|) = [1+\exp(-\beta^T|\Delta\mathbf{x}|)]^{-1}$ that takes as inputs changes in absolute magnitudes of the input parameters with the corresponding binary outcome variable Z of whether the rank changed or not. This involved generating samples of recorded differences between the input parameter collected via simulation runs as well as observed rank value changes for two different parameter settings.

Noting that the probability of the rank changing approaches one as the exponential term $\exp(-\beta^T|\Delta\mathbf{x}|) \to 0$, the relative effect that each parameter has on the probability of a rank change can be determined by comparing the relative magnitudes of the β coefficients as larger positive coefficient values contribute more towards increasing $P(Z = 1)$. This would assume that the inputs scales of the parameters have similar magnitude values. Hence the first sensitivity ranking index is to directly rank the magnitudes of the β coefficients obtained through a fitted logistic regression model.

4.1.2 Variance-Based Method

Variance based methods (Saltelli et al. 2008) approach sensitivity analysis from the viewpoint of attributing the output variance of the simulation response to each of the individual input parameters. The sensitivity of the inputs is then ranked in terms of the how much output variance each input parameter was responsible for. If it is assumed that x_1 is an important and sensitive parameter with respect to the simulation response W, then it should be possible to detect noticeable changes in the mean values of W across different fixed values of x_1. Therefore the variance of the means of W conditioned on x_1, $V[E[W|x_1]]$, would provide some indicator of the importance of x_1. Normalising the variance of means with respect to the total variance $V(W)$ gives the first order sensitivity index $S_i = \frac{V[E[W|x_i]]}{V(W)}$ as described in Saltelli et al. (2008).

The first order sensitivity index was calculated across two different data encodings. The first used the original dataset encoding where each input column was a categorical type and the output column was an integer signifying the rank achieved by the alternative. This gave an overall sensitivity rating for each input parameter. The second was performed over a one hot encoding model of the data. This gave an overall sensitivity rating for each categorical value across all parameters.

4.1.3 Decision Tree

A decision tree is used to predict the likely outcome of an input based on a set of features present in that input. The nodes on the tree are constructed such that each level represent the best splits in the dataset according to the input variables of the data (in this case the model parameters). The metric used to determine the best splits is known as the Gini impurity index and it looks at the probability of misclassification of a sample according to the criteria I where J represents the number of classes and$p(i)$ is the probability of an observation being classified into class i $I = 1 \sum_{i=1}^{J} p(i)^2$. The terminal leaf nodes in the tree represent the possible classes in the target variable, (in this case the determined ranks of the alternative). Each branch sequence from the root to a leaf node can then be read as a set of features used to predict the target class represented by the leaf. For trees constructed utilizing the Gini impurity criteria, the gini importance (Breiman 2001) is used as the measure of variable importance.The gini importance is measuring the (normalized) decrease in the Gini impurity of the nodes involving a parameter mentioned across all nodes in the tree weighted by the proportion of samples reaching those nodes. Hence, the importance of a parameter is dictated by not only how high its node is on the tree but also on how many of its nodes in total are present in the tree and how well those nodes manage to split the incoming samples.

4.1.4 Application of Sensitivity Indices

The dataset analyzed was made up of categorical values that represented input parameter settings as well as details about the alternative and the rank that it attained under those settings. This encoding was used by all sensitivity indices to calculate the parameter rank. The parameter rank is the ranking of all parameters from most sensitive to least sensitive. The parameter setting rank is the sensitivity ranking of all available categorical settings across all parameters. To determine the parameter setting rank, the original encoding is transformed into a set of binary valued columns via a one-hot encoding schema, which replaces the original categorical columns with binary columns that encode the presence or non-presence of each categorical value. All three sensitivity indices were calculated using a one-hot encoding representation to find the parameter setting rank.

To determine the level of consensus among the three models, a comparison of the top three (out of eight) ranked parameters over all combinations of metrics and alternatives was made. For parameter settings, the top seven (out of twenty one) settings were compared. A score of three was given if all three parameters were reported by all models (disregarding rank position), two if two out of three were present, one if only one was common and zero if there were no common parameters. The mean score determines a consensus rating. A consensus rating of three would indicate that all three models reported the same three parameters as the most sensitive. Table 2 presents the results of the consensus ratings across five examplar metrics. Metrics that have a high number of common parameter and/or parameter settings as reported by the consensus ratings could be used as a starting point for conducting further sensitivity analysis. For example, Table 2 would suggest looking at the common parameters in metric 4 and 5 and the common parameter settings in metric 1 due to their relative high consensus rating averages.

5 SUMMARY

The DASE and the AoSA are important tasks when employing combat simulations in support of Army modernization. Effective sensitivity and ranking analyses enable insight on the marginal contributions of sub-systems to overall operational effectiveness, as well as comparatively assessing competing system

Table 2: Average parameter consensus ratings across all alternatives.

Metric	Average parameter consensus	Average parameter setting consensus
1	1	2.167
2	1	1.833
3	0.667	1.5
4	1.333	2
5	1.333	1.5

alternatives. This paper presents unifying discoveries related to sensitivity analysis via linear regression, and some empirical findings concerning ranking analysis via score-based, partition-based and consensus-based methods. An investigation into the combined ranking sensitivity problem was also attempted.

Combat simulations often depart from the classical normal, independent and identically distributed residuals assumptions of linear regression and Kleijnen (2015) offers separate remedies involving OLS, jackknifing and an alternative proposed by Law (2007). This paper has shown analytically the equivalence of these techniques, thus simplifying their presentation. A small flaw in the lack-of-fit F-statistic in the white-noise case in Kleijnen (2015) was also discovered and remedied. The selection of a ranking analysis method for an AoSA study for which a complete ordering of the set of alternatives is sought needs to consider whether a complete set of pairwise comparisons is required. The importance and implication of treating ties differently to losing alternatives also needs to be clearly defined during the problem definition stage. More importantly, the underlying data should also be presented in conjunction with any ranking analysis in order to provide further evidence and insights pertaining to the deltas between different groupings within the ranked set of alternatives. Finally, the efficient search for unique results in the AoSA requires knowledge about the sensitivity of the system to its input parameters. As there are multiple means of calculating a parameter's sensitivity index, a consensus based method between multiple models was utilized to increase confidence in the results reported. Possible future directions for the work include ranking across heterogeneous parameter types, using an ensemble based methodology for multiple sensitivity models and methods for validating ranking sensitivity results.

ACKNOWLEDGMENTS

The authors thank Professor Jack P.C. Kleijnen from Tilburg University for fruitful discussions surrounding regression analysis and Dr. Adrienne Turnbull from Defence Research and Development Canada for information pertaining to the BRANDO method.

REFERENCES

Boyd, R., and J. B. Silk. 1983. "A Method for Assigning Cardinal Dominance Ranks". *Animal Behaviour* 31(1):45–58.

Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1):5–32.

Cochran, W. G. 1950. "The Comparison of Percentages in Matched Samples". *Biometrika* 37(3/4):256–266.

Cook, W. D. 2006. "Distance-Based and ad hoc Consensus Models in Ordinal Preference Ranking". *European Journal of Operational Research* 172(2):369–385.

Emond, E., and A. Turnbull. 2006. "BRANDO: Breakpoint Analysis with Nonparametric Data Option". Technical Report 40, DRDC CORA, Canada. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc61/p527685.pdf.

Emond, E. J., and D. W. Mason. 2002. "A new Rank Correlation Coefficient with Application to the Consensus Ranking Problem". *Journal of Multi-Criteria Decision Analysis* 11(1):17–28.

Fang, K.-T., R. Li, and A. Sudjianto. 2005. *Design and Modeling for Computer Experiments*. Chapman and Hall/CRC Computer Science and Data Analysis Series. New York, USA: Chapman & Hall.

Fey, M., and K. A. Clarke. 2012. "Consistency of Choice in Nonparametric Multiple Comparisons". *Journal of Nonparametric Statistics* 24(2):531–541.

Friedman, M. 1937. "The use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance". *Journal of the American Statistical Association* 32(200):675–701.

Goldsman, D., and B. L. Nelson. 1994. "Ranking, Selection and Multiple Comparisons in Computer Simulations". In *Proceedings of the 26th Conference on Winter Simulation*, edited by M. S. Manivannan and J. D. Tew, WSC '94, 192–199. San Diego, USA: Society for Computer Simulation International.

Hochberg, Y., and A. C. Tamhane. 1987. *Multiple Comparison Procedures*. New York, USA: John Wiley & Sons.

Kendall, M. G. 1955. "Further Contributions to the Theory of Paired Comparisons". *Biometrics* 11(1):43–62.

Kleijnen, J. 2015. *Design and Analysis of Simulation Experiments*. 2nd ed. New York, USA: Springer.

Law, A. 2007. *Simulation Modeling and Analysis*. 4th ed. Boston, USA: McGraw-Hill.

McNemar, Q. 1947. "Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages". *Psychometrika* 12(2):153–157.

Miller, R. 1981. *Simultaneous Statistical Inference*. New York, USA: Springer-Verlag.

Nemenyi, P. 1963. *Distribution-Free Multiple Comparisons*. Ph.D. thesis, Princeton University.

Reynolds, T. J. 1976. "The Analysis of Dominance Matrices : Extraction of Unidimensional Orders within a Multidimensional Context". Technical Report No. 3, USA. http://www.dtic.mil/dtic/tr/fulltext/u2/a029450.pdf.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola. 2008. "Variance-Based Methods". In *Global Sensitivity Analysis. The Primer*, 155–182. West Sussex, England: John Wiley and Sons.

Sanchez, S. M., T. W. Lucas, P. J. Sanchez, C. J. Nannini, and H. Wan. 2012. "Designs for Large-Scale Simulation Experiments, with Applications to Defense and Homeland Security". In *Design and Analysis of Experiments*, edited by D. J. Balding et al., Chapter 12, 413–441. Wiley-Blackwell.

Santner, T. J., W. B., and N. W.. 2003. *The Design and Analysis of Computer Experiments*. New York, USA: Springer-Verlag.

Shen, H., and H. Wan. 2009. "Controlled Sequential Factorial Design for Simulation Factor Screening". *European Journal of Operational Research* 198(2):511–519.

Villacorta, P. J., and J. A. Sáez. 2015. "SRCS: Statistical Ranking Color Scheme for Visualizing Parameterized Multiple Pairwise Comparisons with R". *R JOURNAL* 7(2):89–104.

Wilcoxon, F. 1945. "Individual Comparisons by Ranking Methods". *Biometrics Bulletin* 1(6):80–83.

AUTHOR BIOGRAPHIES

ANDREW GILL is the data analytics discipline leader in Land Capability Branch of the Joint and Operations Analysis Division at the Defence Science and Technology Group. His research focus is on the design and analysis of simulation experiments and his e-mail address is andrew.gill@dst.defence.gov.au.

DION GRIEGER is an operations research specialist in Land Capability Branch of the Joint and Operations Analysis Division at the Defence Science and Technology Group. His research focus is on the analysis and visualization of simulation experiments and his e-mail address is dion.grieger@dst.defence.gov.au.

MARTIN WONG is a data analyst in Land Capability Branch of the Joint and Operations Analysis Division at the Defence Science and Technology Group. His research focuses around the analysis of simulation results and machine learning for computational agents. His email address is martin.wong@dst.defence.gov.au.

WILLIAM CHAU is a data analyst in Land Capability Branch of the Joint and Operations Analysis Division at the Defence Science and Technology Group. His research focus is on the design and analysis of simulation experiments and his e-mail address is william.chau.dst.defence.gov.au.