Statistics and Its Interface Volume 12 (2019) 573–583

A new rank sensitivity metric for decision support

Anil Dolgun*, Haydar Demirhan, Andrew Gill, Dion Grieger, Stella Stylianou, and Stelios Georgiou

Assessing the change in the relative performance of competing systems across a factor space generated by a combination of input variables is a common problem in decision making. We propose a new metric to assess the sensitivity of the performance rankings of a set of options when input variables are changed. The proposed metric is useful in foreseeing the impact of changing values of input variables on an output metric in complex systems through computer simulation experiments. Numerical characteristics of the proposed metric are illustrated and discussed and an application is provided to illustrate use of our metric in decision support.

AMS 2000 subject classifications: Primary 90B50, 62K99; secondary 90C31.
Keywords and phrases: Decision support systems, Design of experiments, Weighted Spearman footrule distance, Sensitivity, Performance ranking.

1. INTRODUCTION

Computer simulation experiments involve modeling of a process or a system in a way that the model mimics the response of the actual (physical) experiments which are too costly or even impracticable. Computer experiments are required in many fields as a decision support tool for predicting behavior of systems under different input settings and quantifying and ranking the effects of input variables on the response (Fang et al., 2005). Moreover, when there are several systems to compare, assessing the relative performance given the input variables is also of interest. Some examples include the comparison of different dynamic optimization algorithms with several parameters (del Amo and Pelta, 2013), assessment of effect of different factors on the relative performance of different classifiers used in machine learning (Villacorta and Sáez, 2015), and ranking of the operational effectiveness of alternative military options in a combat simulation (Chau et al., 2017).

In addition to ranking the performance, decision makers often need to know how the performance rankings of competing systems change as the input variables change. For example, consider a combat simulation that seeks to compare three military options A, B, and C over a factor space composed of many different combinations that might affect the performance metric (e.g., lethality). For a given factor combination (or scenario), the performance of the options could be ranked as A>B>C where as for a different scenario, performance rankings could be B>C>A. In such cases, decision makers would be interested in identifying how sensitive the rankings of the options are across the entire factor space and across a range of metrics. This can be seen as a sensitivity analysis of performance rankings of the options as the input variables change.

In this paper, our purpose is to propose a metric to measure the sensitivity of the relative performance of competing systems across the entire input/factor space. For this purpose, we measure the relative performance of the competing systems using a statistical ranking of the systems based on the all pairwise comparisons as given by del Amo and Pelta (2013) and Villacorta and Sáez (2015). Then, we propose a distance based sensitivity metric to assess the change in the performance rankings of the competing systems across the factor space. This approach utilizes the statistical ranking of the competing systems (del Amo and Pelta, 2013; Villacorta and Sáez, 2015) along with the properly defined weighted Spearman Footrule distance among the factor space. Using the normalization of this distance metric, a new similarity metric called "weighted rank sensitivity" is introduced and evaluated numerically. The proposed metric is shown to be sensitive to the changes in the ranking vectors as inputs change and has a calibrated interpretation. Without loss of generality, we specifically focus on the sensitivity of the effect of a set of options on the output metric across a set of scenarios defined by the input variables to illustrate use of our sensitivity metric.

This paper is organized as follows. In section 2, the motivation of the paper is provided via a military context. In section 3, the proposed sensitivity metric is introduced and illustrated over a toy example. An algorithm used to evaluate the proposed metric is explained and the interpretations of the proposed metric is given in section 4. The sensitivity metric is then illustrated on a combat team attack mission in section 5 before general conclusions are drawn in section 6.

2. MOTIVATION

Combat simulations are used to estimate and compare the operational effectiveness of alternative military options using an experimental format. These simulation experiments are distinct from traditional experiments in a way that they

*Corresponding author.

are stochastic, have a very large parameter space and often compare a large number factors (e.g., default speed of a vehicle, ambient light level, etc.) across multiple performance metrics (e.g., lethality, survivability, mobility, etc.). In the evidence-based decision making of Defence projects, comparison of the operational effectiveness of alternative military options and identifying how sensitive the relative performance of options are across the entire factor space are often of interest.

Consider a combat simulation setting, where a set of military options, O_j $(j = 1, 2, \ldots, M)$ are evaluated against a set of independent performance metrics, M_t $(t = 1, 2, \ldots, m)$, over an experimental design composed of k factors $(f_1, f_2, \ldots, f_k)$ having N combinations, each of which we refer to as a scenario S_i, $i = 1, 2, \ldots, N$. In the experimental setting, we replicate all options for each scenario n times. This problem framework is illustrated in Figure 1.

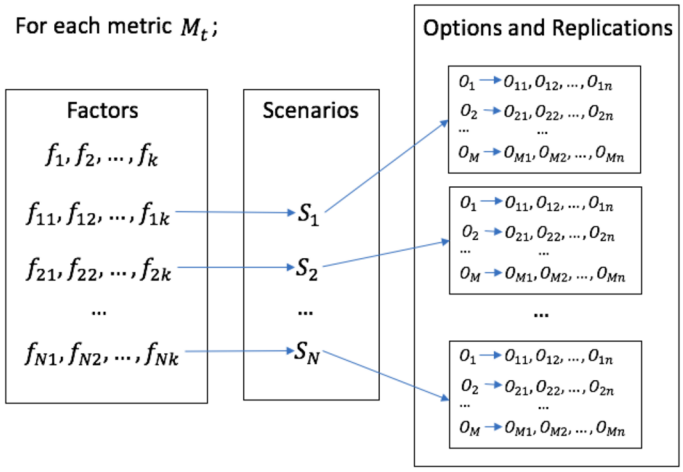


Figure 1. Summary of problem framework.

In the general sense, our focus is to propose a metric that measures the similarity between two scenarios based on the given factor space over the replications of a set of available options. Each design of experiment (DOE) methodology applied here provides a different factor space, and correspondingly, we get a different bunch of scenarios out of each factor space. This factor space can be created using full factorial, fractional factorial, orthogonal arrays, supersaturated designs, etc. to decrease its size in an appropriate way. The sensitivity metric introduce in this study is generically applicable under any DOE methodology.

The desired main features of a useful similarity metric are that i) it should distinguish the scenarios based on their sensitivity to the measured dependent variable – military metrics in the example; ii) it should take into account the magnitude of the change in the metric across the scenarios; iii) it should be a calibrated metric in terms of its interpretation.

The problem reduces to ranking of option means under each scenario and then comparing the rank vectors corresponding to each scenario with each other to give a conclusion about the similarity between the scenarios. Therefore, we have two sub-problems here: i) how the options are ranked under each scenario? ii) how the rank vectors are combined together to construct a sensitivity metric?

We will use a toy example to explain the development process of our metric. Suppose we have three options O_j, $j = 1, 2, 3$ and seven scenarios S_i, $i = 1, \ldots, 7$, which come out of a DOE methodology. For a given scenario S_i, we have the artificially created mean performance values ($\bar{O}_1$, $\bar{O}_2$, and $\bar{O}_3$) given in Table 1.

Table 1. Option mean – Scenario combinations for the toy example

S_i	$\bar{O}_1$	$\bar{O}_2$	$\bar{O}_3$
1	40	30	20
2	80	50	20
3	40	20	30
4	30	40	20
5	20	30	40
6	20	40	30
7	30	20	40

We further assume for each scenario that every pairwise comparison of the options returned a statistically significant result. In the second scenario, only the magnitude of option means increase compared to the first scenario. In the rest of the scenarios, the magnitude of the change is determined by the changing ranks of options.

3. THE WEIGHTED RANK SENSITIVITY METRIC

In this section, we first discuss the ranking of the option space, and then, we introduce our weighted sensitivity metric for the comparison of two scenarios in terms of their sensitivity.

3.1 Statistical ranking of the option space

For a given scenario S_i, assume that options can be ranked according to their performance metric. The ranking straightforwardly can take place between options depending on the absolute values of the performance metrics y_{ijr} measured in the scenario i, option j, and replication r. However, one of the important requirements is to quantify the statistical rankings of the options according to a performance metric. A ranking scheme given by Villacorta and Sáez (2015) that depends on all pairwise multiple comparisons of options in a given scenario fulfills this requirement. For each pairwise comparison (of options), this approach assigns a score from $\{-1, 0, 1\}$ to the options according to the statistical significance criteria and the final rankings of the options (x_{ij} – the rank of option j for scenario i) are calculated as the sum of all scores received from all possible pairwise comparisons.

Therefore, x_{ij}'s store the result of how many options are better, equal or worse than the option being tested, where the decision on each pairwise comparison is made by a pairwise nonparametric multiple comparison test. In this perspective, this algorithm analyses the relative performance of options rather than the absolute performance (Villacorta and Sáez, 2015). The implementation of this ranking approach is done by the R package SRCS (Villacorta, 2015). From now on, we will refer the ranking approach of Villacorta and Sáez (2015) as the VS approach.

To illustrate the VS ranking approach, we refer to the example of Section 2. The rankings of the options according to the VS approach are given in Table 2. Note that in this table, the first and second scenarios will have the same rankings for the options, while other scenarios will have different rankings.

Table 2. Option mean for each scenario and corresponding rank vector assigned by the VS approach

	Option means			VS rankings		
S_i	$\bar{O}_1$	$\bar{O}_2$	$\bar{O}_3$	x_{i1}	x_{i2}	x_{i3}
1	40	30	20	2	0	-2
2	80	50	20	2	0	-2
3	40	20	30	2	-2	0
4	30	40	20	0	2	-2
5	20	30	40	-2	0	2
6	20	40	30	-2	2	0
7	30	20	40	0	-2	2

The VS ranking scheme is plausible for our problem in the sense that it uses a statistical decision criteria while assigning ranks to the options. We can assess whether one option is statistically better or worse than another using this approach. The strength of this approach is in its computational efficiency as it uses a simple ranking algorithm. However, it should be noted that, this ranking scheme does not take into account the absolute differences between options; hence, we cannot figure out how much O_j is better than O_l. Moreover, the selection of pairwise comparison test would have a substantial effect on the option rankings. Therefore, the most powerful pairwise comparison tests should be used while assessing significance (Demirhan et al., 2010; Dolgun and Demirhan, 2017).

3.2 Sensitivity of scenarios

For a given scenario S_i, we first generate the rank vector $\boldsymbol{x}_i = (x_{ij})$ corresponding to M options employing the VS approach of section 3.1. When we have M options to be compared, there will be $\binom{M}{2}$ possible pairwise comparisons. After we assign the scores to the options, the maximum score attainable by an option will be $M-1$, which means it outperforms the rest, and the minimum is $-(M-1)$, meaning it is outperformed by the rest. As this score stores the aggregated information of how many times one option outperformed another option or is outperformed by another one, they provide a suitable basis for a weighted rank similarity metric.

For each scenario, possible values of ranks assigned by the VS approach are $-(M-1), \ldots, -1, 0, 1, \ldots, M-1$. When we consider the scenarios as pairs, a strong similarity between two scenarios implies a strong agreement between the rank vector assigned to the scenarios while a weak similarity implies a weak agreement between two rank vectors. Thus, we can consider the ranks as the ordinal ratings and approach the problem from the perspective of measuring the agreement of two raters (see Gwet (2014) for the details of inter-rater agreement studies). Note that the existing statistics for measuring associations between ranks, such as Kendall's τ coefficient and Spearman's rank correlation coefficient are not suitable to capture the similarity in our context as they would only take into account the ranking orders of two rank vectors.

The common approach in inter-rater agreement studies is to attach weights to the difference between ordinal ratings in order to capture the severity of the disagreements. In our context, these ordinal ratings correspond to the possible values of ranks assigned by the VS approach. The magnitudes of linear, quadratic, and radical weights defined in inter-rater agreement studies rely on the scores attached to the ordinal ratings (Gwet, 2014). We propose use of the ranks assigned by the VS approach in place of scores in these weighting approaches. Then, a weighted metric using the linear, quadratic, and radical weights will capture the effect of the magnitude of difference between ranks on the similarity of the scenarios S_i and S_l when they are used to combine rank vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$ over a distance measure.

For each scenario pair (S_i, S_l), maximum distance between scores of scenarios will be $(M-1)-(-(M-1)) = 2(M-1)$ and the minimum distance between scores of scenarios will be 0. Then, for a given option j, the distance between the scores of scenarios relative to the maximum achievable distance gives us the vector of linear weights $\boldsymbol{\omega'_{il}} = (w'_{ilj})$ for the scenario pair (S_i, S_l):

$$w'_{ilj} = \frac{|x_{ij} - x_{lj}|}{2(M-1)}, \tag{1}$$

where $0 \leq w'_{ilj} \leq 1$ and $j = 1, \ldots, M$. Note that in Eq. (1), x_{ij}'s refer to the rank scores (i.e., [-2, 0, 2]) instead of ranks (i.e, [1,2,3]). Linear weights are based on the magnitude of the absolute difference of two scores relative to the range of all possible values of scores.

In a similar way, quadratic weights are defined as follows:

$$w'_{ilj} = \left(\frac{x_{ij} - x_{lj}}{2(M-1)}\right)^2. \tag{2}$$

The quadratic weights are generally smaller than the linear weights in value and they follow a quadratic pattern for increasing values of absolute difference (Cohen, 1968).

The radical weights are defined as follows:

$$w'_{ilj} = \sqrt{\frac{|x_{ij} - x_{lj}|}{2(M-1)}}. \tag{3}$$

Lastly, we multiply the preliminary weights defined in Eq. (1)–(3) by an additional weight that accounts for the change in the rank order of options between scenarios:

$$w_{ilj} = w'_{ilj} \cdot (u_{ilj})^{1/M}, \tag{4}$$

where

$$\boldsymbol{u_{il}} = \frac{1}{2}(\boldsymbol{r}_i + \boldsymbol{r}_l), \tag{5}$$

$\boldsymbol{r}. = \ll \boldsymbol{x}. \gg$ and $\ll \cdot \gg$ returns a vector including the ranks of elements of inner vector such as $\ll [-2, 0, 2] \gg = [1, 2, 3]$. The additional weight u_{ilj} increases the effect of each weight according to the order of swapping options in the compared scenarios. The exponent in the additional weight smooths the effect of it to account for the increasing amount of information when we have a large number of options. Then, the additional weight is based on the average rank index corresponding to each option. This definition ensures the symmetry in the comparison of scenarios in terms of the proposed metric.

The next stage is to attach the weights to a distance measure to complete the sensitivity metric with desired properties. We used the Spearman Footrule distance as the distance metric because it measures the total distance between two rank vectors (Deza and Deza, 2009) instead of minimum number of swaps to achieve a complete match between vectors (i.e., Kendall's and Cayley's distance metrics). When we incorporate the weights in the formulation of Spearman Footrule distance, we get the following weighted Spearman Footrule distance between the scenarios S_i and S_l:

$$\Delta_{il}^{F_\omega} = \sum_{j=1}^{M} w_{ilj}|x_{ij} - x_{lj}|. \tag{6}$$

Note that the unweighted version of the Spearman Footrule distance is obtained by using identity weights $w_{ilj} = 1$ in Eq. (6). A drawback of the measure obtained in Eq. (6) is that it is not normalized. So, we normalize the weighted Spearman Footrule distance using the maximum value of that distance to produce a sensitivity measure that has a range of $[-1, 1]$ as follows:

$$\rho_{il}^{F_\omega} = 1 - 2\frac{\sum_{j=1}^{M} w_{ilj}|x_{ij} - x_{lj}|}{\max(\boldsymbol{\Delta}^{\boldsymbol{F_\omega}})}, \tag{7}$$

where $\boldsymbol{\Delta}^{\boldsymbol{F_\omega}} = (\Delta_{il}^{F_\omega})$.

Table 3 presents the sensitivity metrics with identity weights (Spearman Footrule distance with $\omega_{ilj} = 1$) for the data given in Table 2.

Table 3. Sensitivity matrix using identity weights for the scenarios in Table 2

$\rho_{il}^{F_L}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	1						
S_2	1	1					
S_3	0	0	1				
S_4	0	0	-1	1			
S_5	-1	-1	-1	-1	1		
S_6	-1	-1	-1	0	0	1	
S_7	-1	-1	0	-1	0	-1	1

It is clear from Table 3 that the unweighted version does not capture the changes in the values of rank scores and gives a vague inference since it has only the values of $-1, 0$, and 1. This situation causes a high rate of false positive decisions in favour of high sensitivity of scenarios when they are not actually.

The linearly weighted Spearman Footrule sensitivity metrics calculated using Eq. (7) with the weights in Eq. (1) are calculated for the same scenarios and given in Table 4.

Table 4. Sensitivity matrix with linear weights for the scenarios in Table 2

$\rho_{il}^{F_L}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	1.00	1.00	0.55	0.46	-1.00	-0.50	-0.50
S_2	1.00	1.00	0.55	0.46	-1.00	-0.50	-0.50
S_3	0.55	0.55	1.00	-0.50	-0.50	-1.00	0.46
S_4	0.46	0.46	-0.50	1.00	-0.50	0.55	-1.00
S_5	-1.00	-1.00	-0.50	-0.50	1.00	0.46	0.55
S_6	-0.50	-0.50	-1.00	0.55	0.46	1.00	-0.50
S_7	-0.50	-0.50	0.46	-1.00	0.55	-0.50	1.00

We get $\rho_{12}^{F_L} = 1$ for the exact same rankings for (S_1, S_2) pair, and we have $\rho_{36}^{F_L} = -1$ for (S_3, S_6) pair which has the maximum change in the values of ranks. So, the proposed $\rho_{il}^{F_L}$ consistently captures the similarity in relation to the magnitude of change in the rank vectors corresponding to the scenarios. The values of $\rho_{il}^{F_L}$ close to -1 imply a strong sensitivity and those tend to 1 show a strong insensitivity between the scenarios. It is possible to compare sensitivity of two pairs of scenarios by comparing the corresponding values of $\rho_{il}^{F_L}$ and $\rho_{jk}^{F_L}$ since the proposed metric is a calibrated measure; hence, smaller values imply more sensitivity between the considered scenarios. For example, when we compare $\rho_{13}^{F_L} = 0.55$ to $\rho_{34}^{F_L} = -0.50$, we can conclude that the degree of sensitivity for passing S_4 from S_3 is more than the sensitivity of passing from S_1 to S_3. Also, consistently to this inference, it shows a moderate amount of sensitivity with $\rho_{13}^{F_L} = 0.55$.

Besides the change in the rankings of the options, the proposed weighted sensitivity metric (in terms of distance and similarity) incorporates weights and has a reasonable interpretation within a normalized range (i.e., $[-1, 1]$ for the weighted similarity measure.

Table 5. Sensitivity matrix with quadratic weights for the scenarios in Table 2

$\rho_{il}^{F_Q}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	1.00	1.00	0.77	0.73	-1.00	-0.25	-0.25
S_2	1.00	1.00	0.77	0.73	-1.00	-0.25	-0.25
S_3	0.77	0.77	1.00	-0.25	-0.25	-1.00	0.73
S_4	0.73	0.73	-0.25	1.00	-0.25	0.77	-1.00
S_5	-1.00	-1.00	-0.25	-0.25	1.00	0.73	0.77
S_6	-0.25	-0.25	-1.00	0.77	0.73	1.00	-0.25
S_7	-0.25	-0.25	0.73	-1.00	0.77	-0.25	1.00

Table 6. Sensitivity matrix with radical weights for the scenarios in Table 2

$\rho_{il}^{F_R}$	S_1	S_2	S_3	S_4	S_5	S_6	S_7
S_1	1.00	1.00	0.36	0.24	-1.00	-0.70	-0.70
S_2	1.00	1.00	0.36	0.24	-1.00	-0.70	-0.70
S_3	0.36	0.36	1.00	-0.70	-0.70	-1.00	0.24
S_4	0.24	0.24	-0.70	1.00	-0.70	0.36	-1.00
S_5	-1.00	-1.00	-0.70	-0.70	1.00	0.24	0.36
S_6	-0.70	-0.70	-1.00	0.36	0.24	1.00	-0.70
S_7	-0.70	-0.70	0.24	-1.00	0.36	-0.70	1.00

The sensitivity matrices obtained by using the quadratic and radical weights through the proposed measure are given in Table 5 and 6.

When we use the quadratic weights, we get similar inferences with the case of linear weights. The difference in the sensitivity results with the linear and quadratic weights is that we have an increased resolution in the sensitivity matrix. The magnitude of sensitivity/insensitivity is the same across the pairs, for example, (S_1, S_3) $(\rho_{13}^{F_L} = 0.55)$ and (S_1, S_6) $(\rho_{16}^{F_L} = -0.50)$ when the linear weights are used. However, the magnitude of sensitivity $(\rho_{16}^{F_Q} = -0.25)$ is less than that of insensitivity $(\rho_{13}^{F_Q} = 0.77)$ for the same pairs of scenarios with quadratic weights. The radical weights also give different magnitudes for sensitivity/insensitivity for the considered pairs of scenarios. However, when compared to the quadratic weights, they tend to behave in favour of sensitivity as radical weights produce a smaller value for the (S_1, S_3) pair $(\rho_{13}^{F_R} = 0.36)$ and a larger (absolute) value for the (S_1, S_6) pair $(\rho_{16}^{F_R} = -0.70)$. When compared to the linear weights, radical weights gives the magnitude of sensitivity closer to the end points of the interval $[-1, 1]$.

If we focus on the pairs (S_1, S_5) and (S_1, S_6) for which the total absolute difference between the ranks are the same with different rank orders, we see that the proposed metric gives different magnitudes of sensitivity; hence, it simultaneously captures the change in the values of ranks and the changing order of the ranks with any of the linear, quadratic, or radical weights. However, quadratic weights assign a smaller (absolute) value to the (S_1, S_6) pair than the other weights.

In the next section, we present a numerical analysis of the behavior of the proposed weighted metric.

4. A NUMERICAL EVALUATION AND INTERPRETATION

In this section, we will discuss numerical features of the proposed metric to measure sensitivity of the scenarios across two cases with four and five options. The VS ranking scheme is based on the the number of the options to be considered. We have the following constraints for the ranks assigned by the VS method:

- $\sum_{j=1}^{M} r_{ij} = 0 \ \forall i$,
- $\big(||r_{ij} = (M-1)|| \wedge ||r_{ij} = -(M-1)||\big) \leq 1$,

where $i = 1, \ldots, N$, $j = 1, \ldots, M$, and $||\cdot||$ is the number of cases with inner condition is satisfied. Under these conditions, it is possible to generate all possible rank vectors and all possible pairs of rank vectors.

We apply the following algorithm to generate the population of all possible comparisons.

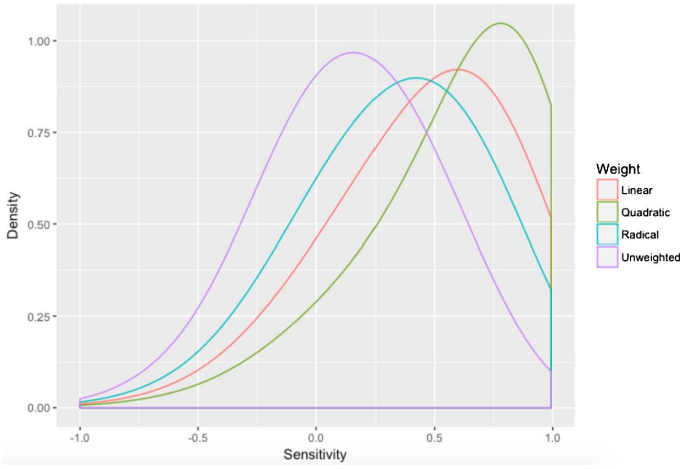
Algorithm 1.

1. Generate all possible combinations of ranks that can be assigned to M options under the constraints given above. Let the number of all possible combinations generated be $\widetilde{N}$.
2. Set $\widetilde{N}$ to the number of scenarios and assign a scenario to each combination of ranks.
3. Create all pairs of $\widetilde{N}$ scenarios.

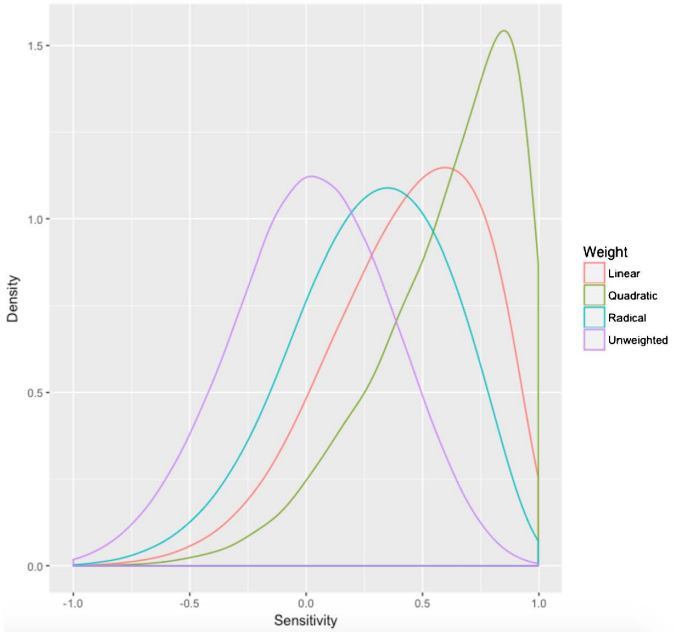
In practice, any set of scenarios is a sample taken with replacement from the set of scenarios created by the Algorithm 1. Note that Algorithm 1 does not allow repeating rank vectors. Values of any metric calculated over the pairs of scenarios generated by Algorithm 1 correspond to the population of the metric (all possible values of that metric) for a specific number of options. Therefore, any inference drawn over the scenarios resulting from Algorithm 1 can be generalized for the considered number of options. We use this approach to figure out the characteristics of the proposed sensitivity metric.

To work out the effect of our weighting approach, we calculate normalised Spearman Footrule distance in Eq. (7) over the pairs generated by Algorithm 1 for the cases of $M = 4$ and $M = 5$. Corresponding kernel density estimates of probability distribution function (pdf) of unweighted and weighted normalised Spearman Footrule distances are given in Figure 2 for the cases of $M = 4$ and $M = 5$. Implementation of Algorithm 1 for $M > 5$ is computationally troublesome due to the memory limitations.

The pdf of unweighted normalised Spearman Footrule distance is slightly skewed to left for both of $M = 4$ and $M = 5$ cases. So, proportions of very similar or dissimilar ranking schemes are very small and most of the rankings are moderately different from each other. Near symmetry in the population pdf of unweighted normalised Spearman Footrule distance confirms that it treats each unit change in ranks equally in terms of sensitivity. However, in our

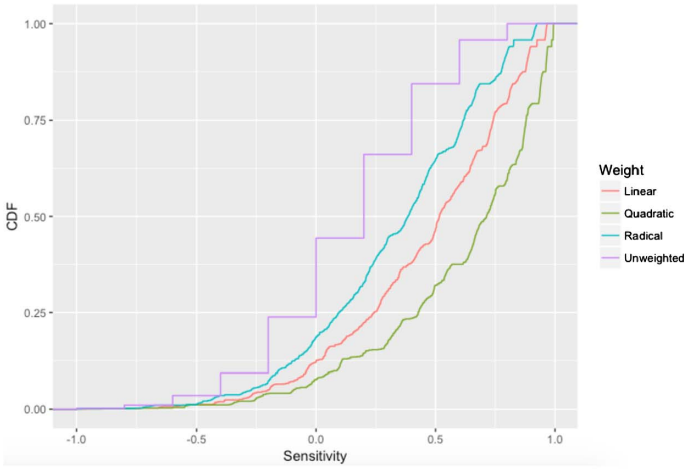


(a) $M = 4$

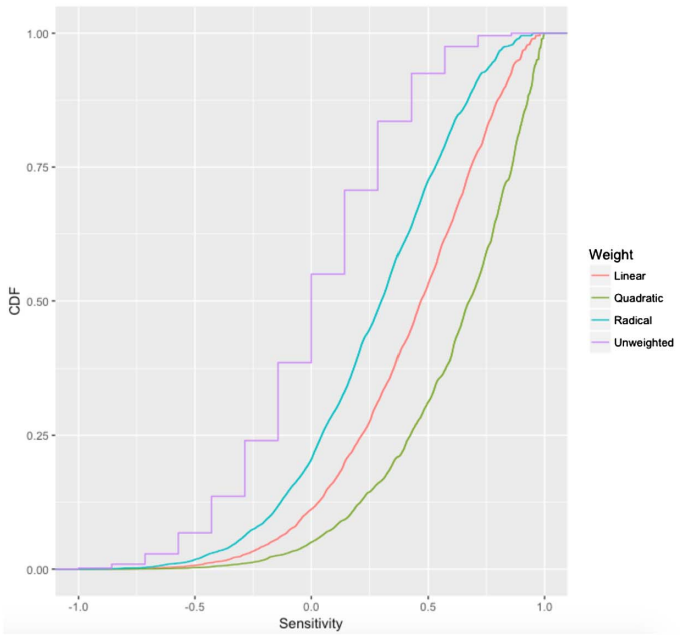


(b) $M = 5$

Figure 2. Population distributions of normalised Spearman Footrule distance.



(a) $M = 4$



(b) $M = 5$

Figure 3. Population cumulative distribution functions of normalised Spearman Footrule distance.

combat simulation setting, the effect of one, two, three,... unit changes in ranks are not equal across the possible values of ranks. Another drawback of the unweighted metric is that due to its low resolution, it qualifies a considerable amount scenario pairs as highly sensitive. Therefore, the pdf of unweighted metric has a thicker right tail implying that false positive rate of it is significantly high. Therefore, it is not suitable to use the unweighted normalised Spearman Footrule distance directly.

For both of $M = 4$ and $M = 5$ cases, all the population pdf's of the proposed metric are left-skewed for all weights. In accordance with the weighting schemes, the metric with radical weights has a thicker left tail than those with linear and quadratic weights; hence, this measure is the most liberal measure in terms of detecting sensitivity. The metric with the quadratic weights is the most conservative one in this sense. It qualifies more than half of the possible comparisons as either mildly sensitive or insensitive. The linear weights are in between radical and quadratic weights.

Because the distribution of the proposed metric is not symmetric, a perfectly symmetric interpretation over the $[-1, 1]$ range is not appropriate. To provide a clear and suitable interpretation for the proposed metric with each weighting scheme, we use cumulative distribution functions (cdf's) of sensitivity metric, which are given for unweighted and weighted metrics in Figure 3.

Table 7. Interpretation of the proposed metric

Interpretation	Linear	Quadratic	Radical
Sensitive	$[-1, 0.23)$	$[-1, 0.43)$	$[-1, 0.11)$
Neutral	$[0.23, 0.73)$	$[0.43, 0.87)$	$[0.11, 0.60)$
Insensitive	$[0.73, 1]$	$[0.87, 1)$	$[0.60, 1]$

We divide the probability range $[0, 1]$ into three intervals such that $[0, 1] = [0, 0.25) \cup [0.25, 0.75] \cup (0.75, 1]$. Each interval corresponds to the regions where the pairs of scenarios are qualified as "Sensitive", "Neither sensitive nor insensitive (Neutral)", and "Insensitive". Alternatively, these intervals can be divided into subintervals to increase the resolution of interpretation. Based on this approach, Table 7 gives the inferences and corresponding limits of the metric for each weight. The limits are found using the inverse cdf for each weights at the boundary points of the intervals $[0, 0.25)$, $[0.25, 0.75]$, and $(0.75, 1]$.

5. SENSITIVITY OF COMBAT TEAM ATTACK SCENARIOS

In this application, we focus on making decisions based on the sensitivity of scenarios in a combat team attack mission where the set of military options being compared are represented by four different vehicles: A, B, C and D; hence, in our setting, O_1 =A, O_2 =B, O_3 =C, and O_4 =D. The following three ($m = 3$) primary metrics are used to compare the four options:

- "MissionSuccess" (M_1): A binary indicator as to whether Blue was able to successfully complete the mission.
- "RedVehiclesDamaged" (M_2): A discrete value indicating how many Red vehicles were damaged.
- "RedInfantryDefeated" (M_3): A discrete value indicating how many Red infantry were defeated.

A full factorial design was conducted to explore the impact of uncertainty surrounding the following three ($k = 3$) modelling factors on the rankings of the options:

- "Plan" (f_1): Two different sets of tactics (making a direct or indirect advance to the area of interest).
- "Range" (f_2): Withdrawal range (in metres) of vehicle entities after an engagement (only the upper and lower range of values were modelled: 75m and 25m).
- "Protection" (f_3): Fidelity / complexity of the algorithms which model vehicle survivability (High, Medium, Low).

All factors and combinations of factors were considered to be equally plausible. In total, $n = 200$ replications were run for each of the $N = 12$ different plausible scenarios given in Table 8.

The problem we focus on in this decision problem is to compare sensitivity of each combination of levels of Plan, Range, and Protection factors to the selection of vehicles A, B, C, and D. For each metric, we compute our sensitivity metric with linear, quadratic and radical weights over Eq. (1)–(7), as well as using identity weights (i.e., $w_{ilj} = 1$).

Table 8. Scenarios in the combat team attack problem

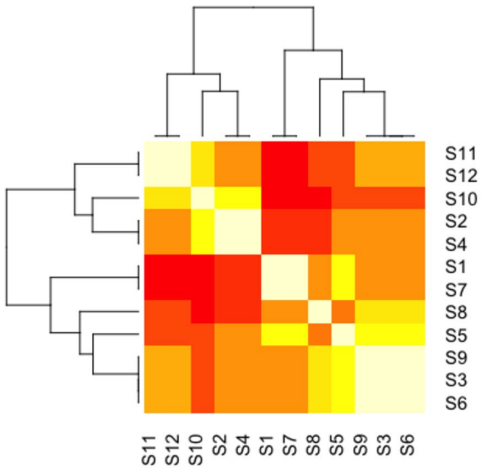
Scenario	Plan	Range	Protection
S_1	Direct	75m	High
S_2	Indirect	75m	High
S_3	Direct	25m	High
S_4	Indirect	25m	High
S_5	Direct	75m	Medium
S_6	Indirect	75m	Medium
S_7	Direct	25m	Medium
S_8	Indirect	25m	Medium
S_9	Direct	75m	Low
S_{10}	Indirect	75m	Low
S_{11}	Direct	25m	Low
S_{12}	Indirect	25m	Low

We summarize the results using heatmap plots with dendograms to be able to draw inferences for decision support. Heatmaps for the sensitivity matrices corresponding to MissionSuccess, RedVehiclesDamaged, and RedInfantryDefeated metrics are given in panels (a), (b), and (c) of Figures 4–7, respectively.

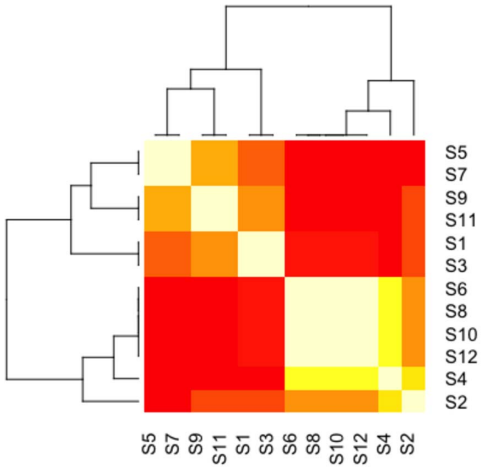
Figures 4–7 present the heatmap plots of the sensitivity matrices computed using identity, linear, quadratic and radical weights, respectively. As a general rule $w^I_{ilj} \geq w^R_{ilj} \geq w^L_{ilj} \geq w^Q_{ilj}$, where w^I_{ilj}, w^R_{ilj}, w^L_{ilj}, w^Q_{ilj} represent the identity, radical, linear and quadratic weights assigned. When we compare the sensitivity matrices obtained by different weighting schemes, we observe that using identity weights always produces highest sensitivity values (i.e., sensitivity values close to -1, see Figure 4) because they assign more contribution to unit changes than the other weighting schemes. The second and the third highest sensitivities are observed in sensitivity matrices obtained using radical and linear weights (see Figures 7 and 5). On the other hand, the lowest sensitivities (i.e. sensitivity values close to 1) are observed in the quadratic ones (see Figure 6) as quadratic weights assign less contribution to the same amount of change.

Throughout this section we only interpret the sensitivity matrices using linear weights (i.e., see Fig. 5) as they provide a weighting scheme in between identity and quadratic weights. One can also use identity, radical or quadratic weights depending on the amount of contribution to be assigned to a unit change.

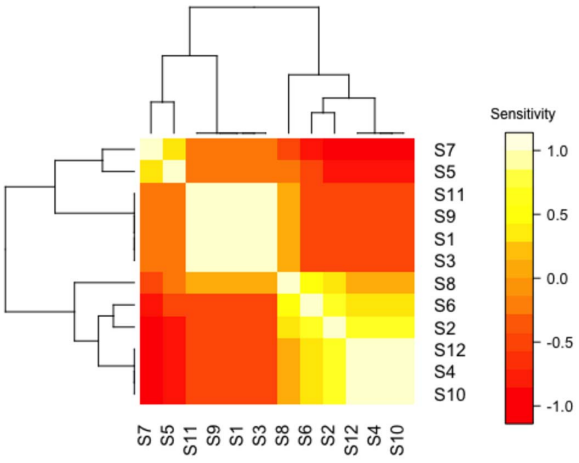
For the success of mission (MissionSuccess metric), the sensitivity of choosing different vehicles is maximized when we change the combination of factors from S_1 to S_{10} or from S_7 to S_{10}. In practice, the performance rankings of the four vehicles is highly sensitive with respect to the success of the mission when the combat team makes an indirect advance



(a) MissionSuccess

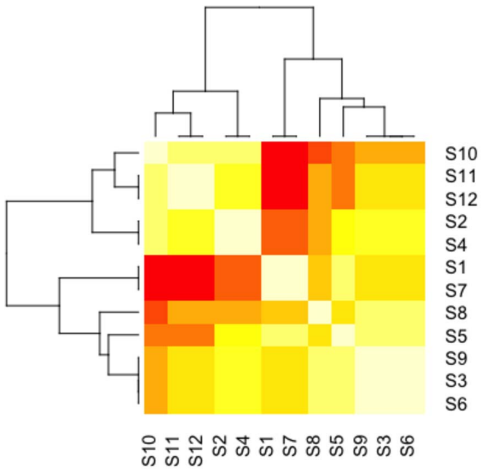


(b) RedVehiclesDamaged

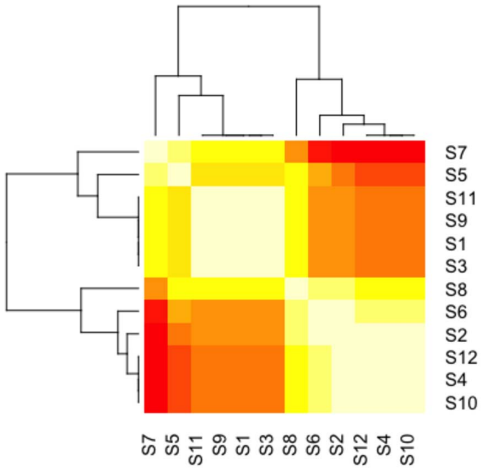


(c) RedInfantryDefeated

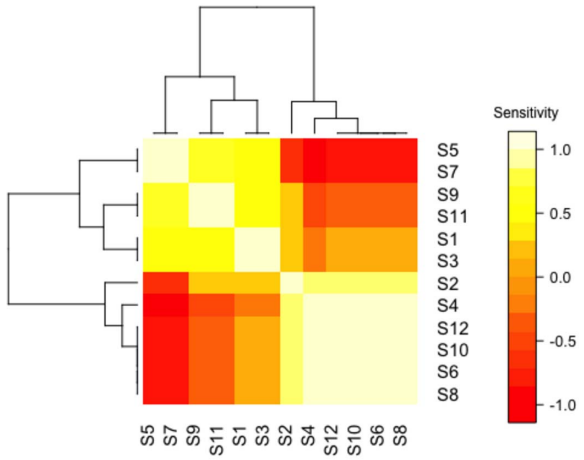
Figure 4. Heatmap plots of the sensitivity matrices that are computed using identity weights (unweighted) for the considered metrics.



(a) MissionSuccess



(b) RedVehiclesDamaged



(c) RedInfantryDefeated

Figure 5. Heatmap plots of the sensitivity matrices that are computed using linear weights for the considered metrics.

to the area of interest with a low fidelity protection algorithm, instead of a direct advance to the area of interest with a high fidelity protection algorithm (Panel (a) of Figure 5). Also, for example, the performance rankings of the four vehicles is totally insensitive with respect to the success of the mission when the withdrawal range is changed from 75m to 25m due to close to zero sensitivity values of pairs (S_1, S_3) and (S_2, S_4) for high protection; (S_5, S_7) and (S_6, S_8) for medium protection; and (S_9, S_{11}) and (S_{10}, S_{12}) for high protection.

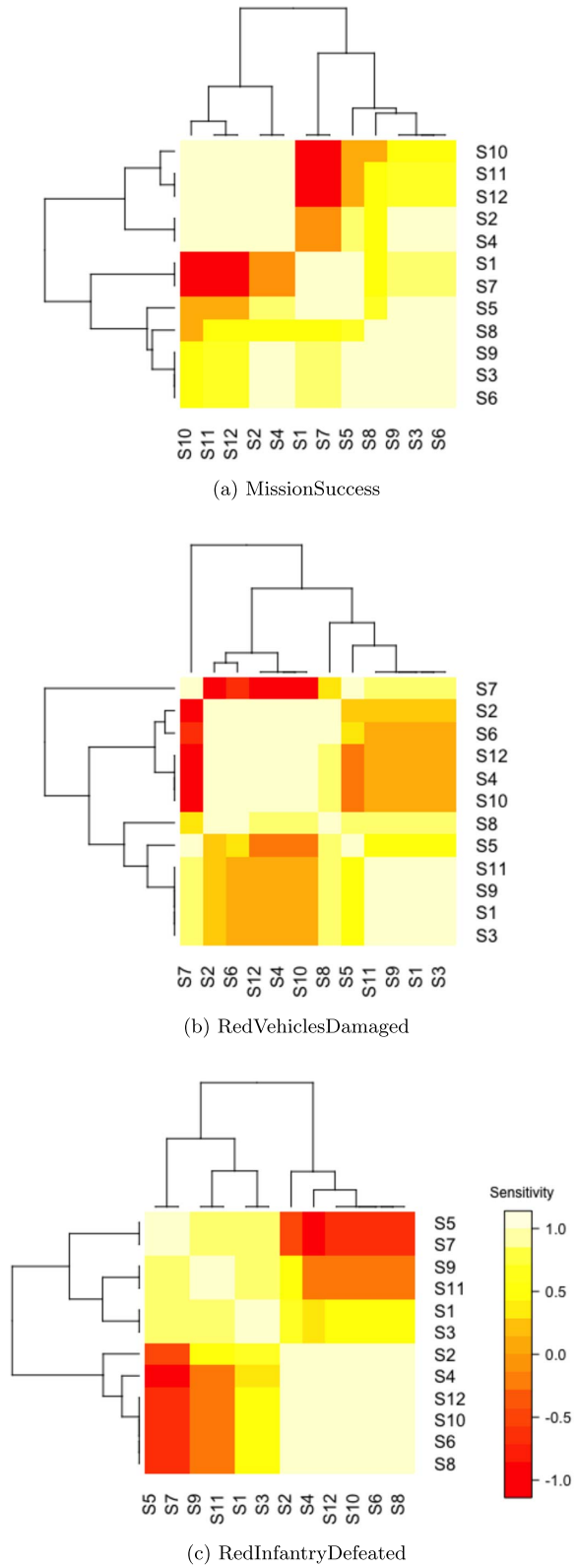
In terms of the number of enemy vehicles destroyed, the sensitivity of the performance rankings of the four vehicles is maximized when we change the combination of factors from S_2 to S_7, from S_7 to S_{10}, and from S_7 to S_{12}. The scenario seven is highly or moderately sensitive when paired with $S_2, S_4, S_6, S_8, S_{10}$, and S_{12} in terms of the effect of the performance rankings of the four vehicles with respect to the the number of enemy vehicles destroyed. The deviances from the strategy that the combat team makes a direct advance to the operation area with a withdrawal range of 25m under a medium fidelity protection algorithm creates high sensitivity of the performance rankings of the four vehicles.

For the number of enemy infantry defeated, the sensitivity of the performance rankings of the four vehicles is maximized when we change the combination of factors from S_4 to S_5 and from S_4 to S_7. When we change the advance plan to the area of operation from indirect to direct along with changing the protection algorithm from high to medium, the effect of the performance rankings of the four vehicles is one of the highest levels of sensitivity.

Overall, the scenario S_7 is in the pairs which constitute highly sensitive cases in terms of the sensitivity of the effect of vehicle choice on all three success metrics. Use of heatmaps provide this information in a clearer way showing the highly sensitive and insensitive scenarios in clusters. In terms of number of damaged enemy vehicles, the performance rankings of the four vehicles will be very sensitive upon changing the approach plan to direct from indirect (Figure 5) since scenario pairs including the direct approach plan $(S_1, S_3, \ldots, S_{11})$ and indirect approach plan $(S_2, S_4, \ldots, S_{12})$ constitute highly sensitive pairs; and hence, included in the same (right) arm of the hierarchical classification provided on top the heatmap.

For the mission success metric, changing the protection algorithm fidelity from medium to low creates high sensitivity in terms of the performance rankings of the four vehicles unless the team advances to the area directly in a 25m withdrawal range with low protection (S_7 and S_{11}) since the pairs $(S_5, S_9), (S_6, S_{10})$, and (S_8, S_{12}) are all either insensitive or mildly sensitive.

In terms of the number of destroyed enemy vehicles, variations from the seventh scenario are highly important on the effect of the performance rankings of the four vehicles when the plan is changed to indirect from direct since S_7 constitutes highly sensitive pairs with the scenarios including



(a) MissionSuccess

(b) RedVehiclesDamaged

(c) RedInfantryDefeated

Figure 6. Heatmap plots for the sensitivity matrices for the considered quadratic metrics.

indirect approach plan. Also, when the approach strategy to the operation area is changed from direct to indirect, the effects of the protection algorithm fidelity and withdrawal range are important on the performance rankings of the four vehicles since nearly all the scenarios including direct approach constitute highly or moderately sensitive pairs with scenarios including direct approach.

For the number of enemy infantry defeated, we have a similar sensitivity scheme for the effect of the performance rankings of the four vehicles on this metric, with the change of plan for approaching the operation area being important since the scenarios $(S_2, S_4, S_6, S_8, S_{10}, S_{12})$ and $(S_1, S_3, S_5, S_7, S_9, S_{11})$ are separated into two subsets at the top level of hierarchical clustering.

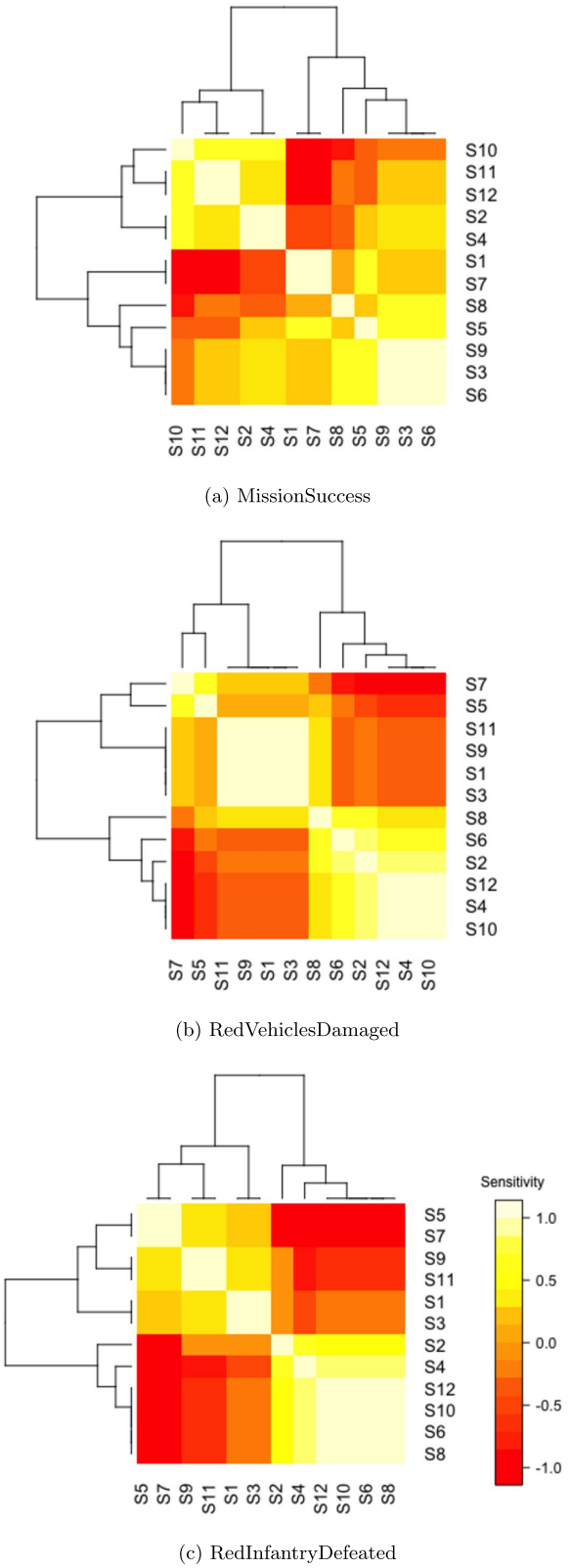
6. DISCUSSION

In this article, we focus on evaluating the change in the relative performance of competing systems across a factor space generated by the design of experiment methodology. It is different from the usual problems arising from design of experiments in the sense that we have an additional set of options to be considered and each run in the factor space is considered as a different scenario. The set of options is another factor that is considered along with the factor space and we need to compare the sensitivity of runs over the set of options. When we consider each run as a different strategy in a decision making problem, our focus turns into finding a sensitivity metric to assess the sensitivity of different strategies under different options to provide decision support.

In this paper, we propose a weighted sensitivity metric based on the rankings of options according to a dependent variable. The proposed metric is useful in predicting the impact of changing values of input variables on an output metric in complex systems through computer simulation experiments. Since we can figure out the effect of switching between alternative options on the target variable under the presence of related factors, decision makers can assess the cost of each decision more accurately by using our measure and end up with choosing the optimal option for each case.

The effect of identity, linear, quadratic, and radical weighting strategies are discussed. In terms of detecting sensitivity, the radical weights are found to be the most liberal weights and linear weights are in between radical and quadratic weights in this sense. We explore distributional properties of the proposed metric for the cases with four and five options. Based on distributional properties, we calibrated the interpretation of the proposed metric.

An application of the proposed sensitivity metric is given for a combat team attack operation in a military setting. Based on the results of the sensitivity analysis of scenarios in the combat team attack operation, we show how performance rankings of the vehicles are affected by changes in the scenarios composed of approach plan, withdrawal range, and protection algorithm fidelity.



(a) MissionSuccess

(b) RedVehiclesDamaged

(c) RedInfantryDefeated

Figure 7. Heatmap plots for the sensitivity matrices for the radical weight considered metrics.

As future research, we are planning to link the sensitivity matrices to the factor space to be able to apply screening procedure and reduce the size of the factor space based on the sensitivity of runs.

ACKNOWLEDGEMENT

We thank the anonymous reviewers for their constructive comments and suggestions which helped us to improve the quality of our paper.

Received 8 June 2018

REFERENCES

Chau, W., Gill, A., Grieger, D., 2017. Using combat simulation and sensitivity analysis to support evaluation of land combat vehicle configurations. In: 22nd International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand. MODSIM2017. URL https://www.mssanz.org.au/modsim2017/D1/chau.pdf.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological bulletin 70 (4), 213.

del Amo, I. G., Pelta, D. A., 2013. SRCS: a technique for comparing multiple algorithms under several factors in dynamic optimization problems. In: Metaheuristics for Dynamic Optimization. Springer, pp. 61–77.

Demirhan, H., Dolgun, N. A., Demirhan, Y. P., Dolgun, M. O., 2010. Performance of some multiple comparison tests under heteroscedasticity and dependency. Journal of Statistical Computation and Simulation 80 (10), 1083–1100. MR2759906

Deza, M. M., Deza, E., 2009. Encyclopedia of distances. In: Encyclopedia of Distances. Springer, pp. 1–583. MR2538177

Dolgun, A., Demirhan, H., 2017. Performance of nonparametric multiple comparison tests under heteroscedasticity, dependency, and skewed error distribution. Communications in Statistics-Simulation and Computation 46 (7), 5166–5183. MR3698518

Fang, K.-T., Li, R., Sudjianto, A., 2005. Design and modeling for computer experiments. CRC Press. MR2223960

Gwet, K. L., 2014. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters. Advanced Analytics, LLC.

Villacorta, P. J., 2015. SRCS: Statistical Ranking Color Scheme for Multiple Pairwise Comparisons. R package version 1.1. URL https://CRAN.R-project.org/package=SRCS.

Villacorta, P. J., Sáez, J. A., 2015. SRCS: Statistical ranking color scheme for visualizing parameterized multiple pairwise comparisons with R. The R Journal 7 (2), 89–104.

Anil Dolgun
School of Science, Mathematical Sciences
Royal Melbourne Institute of Technology
Melbourne, VIC 3001
Australia
E-mail address: anil.dolgun@rmit.edu.au

Haydar Demirhan
School of Science, Mathematical Sciences
Royal Melbourne Institute of Technology
Melbourne, VIC 3001
Australia
E-mail address: haydar.demirhan@rmit.edu.au

Andrew Gill
Defence Science and Technology Group
PO Box 1500, Third Avenue
Edinburgh, SA 5111
Australia
E-mail address: andrew.gill@dst.defence.gov.au

Dion Grieger
Defence Science and Technology Group
PO Box 1500, Third Avenue
Edinburgh, SA 5111
Australia
E-mail address: dion.grieger@dst.defence.gov.au

Stella Stylianou
School of Science, Mathematical Sciences
Royal Melbourne Institute of Technology
Melbourne, VIC 3001
Australia
E-mail address: stella.stylinaou@rmit.edu.au

Stelios Georgiou
School of Science, Mathematical Sciences
Royal Melbourne Institute of Technology
Melbourne, VIC 3001
Australia
E-mail address: stelios.georgiou@rmit.edu.au