24th International Congress on Modelling and Simulation, Sydney, NSW, Australia, 5 to 10 December 2021
mssanz.org.au/modsim2021

An instance space analysis of combat simulations to understand the impact of force and information advantage on survival ratios

K. Smith-Miles [a], **S. Bista** [b] **and L. Jiang** [b]

[a] School of Mathematics and Statistics, University of Melbourne, [b] Defence Science and Technology Group, Department of Defence, Canberra
Email: smith-miles@unimelb.edu.au

Abstract: Instance Space Analysis (ISA) is a new methodology to rigorously "stress-test" algorithms to gain visual insights into their strengths and weaknesses. A diverse and comprehensive set of test instances is used to construct a 2D visualisation of the entire space of possible test instances, within which the performance of algorithms and the sufficiency of the available test instances can be scrutinised. In particular, test instances can be scrutinised for their diversity, unbiasedness, discrimination power, and real-world-likeness. Regions in the instance space where an algorithm has statistical evidence of good performance are generalised to form an algorithm "footprint", where machine learning methods are used to predict expected good performance on untested instances. The properties of algorithm footprints, including their area, density and purity, provide objective measures of comparative algorithm performance, rather than the traditional approach of on-average reporting of a performance metric over a test suite. ISA essentially unlocks the test suite to expose algorithm strengths and weaknesses, explaining how test instance characteristics affect algorithm performance.

The aim of this study is to explore for the first time how ISA can provide insights into combat simulations to understand the impact of defence force design on combat outcomes. Specifically, the study has focused on exploring how force advantage and information advantage, in the form of additional joint force assets and extended technological capabilities, can affect improved survival ratios of force assets at the end of the combat. Employing ISA to explore such questions has required a novel reinterpretation of the terms "algorithm" and "test instance" in order to map the ISA methodology onto the combat context. The study analyses data based on simulation runs from the JFOrCE agent simulation tool (Au et al., 2018), which simulates a fictitious combat between blue and red teams. Two data sets have been generated using the JFOrCE simulation comprising: i) 1854 force scenarios where the red and blue teams have identical initial assets with varying technological capabilities; and ii) 57 force scenarios where the red and blue teams have varying initial force assets with identical technological capabilities. An instance space is created using these datasets where "algorithm" success is defined as the Blue team (Experimental Force) having a better survival ratio of all assets compared to the Red team (Opposing Force). Analysis of the instance space reveals how the simulation parameters that define red and blue force attributes determine the outcomes of a simulated battle, with particular focus on those attributes that represent a significant force size or information advantage through technological capability.

The results show that identical initial force data is unbiased and quite comprehensive, with clear indications of how key force capability attributes determine the probability of success. It is clear from the visualisations obtained that loss is inevitable if the opposing force has an advantage in the form of superior jet sensor range, when loss is defined as poorer survival ratio of assets. Likewise, a win is guaranteed if the team has a superior jet sensor range. There are other more nuanced conditions involving combinations of force attributes, that make probability of win or loss high. These include likely wins if submarine sensor ranges are higher, and if shared communications offer an information advantage. Exploring the varied initial force scenarios, the study was also able to confirm that an advantage in the number of jets can overcome disadvantage in terms of jet capabilities. These findings roughly support those of previous studies with related but different simulation datasets (Au et al., 2018). Since the datasets are fictitious, these conclusions are inconsequential. However, the aim of this paper is to demonstrate the utility of the ISA methodology, by reassuring visualisations of these sensible relationships, and to show the potential for greater insights with additional simulation data. Scrutinising the diversity of the entire instance space, we show that there are simulation scenarios that could be explored where the combat outcomes are currently less predictable. There is also an opportunity to explore simulation outcomes for other parameters that define combat rules and strategies.

Keywords: *Instance space analysis, combat simulation, information advantage, force advantage*

1. INTRODUCTION

In this paper we describe a proof-of-concept study that aimed to explore the potential of Instance Space Analysis (ISA) to provide insights into how joint force attributes – such as force composition, size and technological capabilities – affect the outcomes of combat in an agent-based simulation. ISA was developed by Smith-Miles and co-authors in order to understand how problem instance characteristics affect algorithm performance, originally in the field of optimisation (Smith-Miles et al., 2014) but later generalised to many other fields including machine learning (Munoz et al., 2018) and forecasting (Kang et al., 2017). ISA involves running an <*algorithm*> on multiple <*test instances*> within a <*problem domain*> to generate <performance metrics> defining success; e.g. running a <forecasting method> on multiple <time series> within the <forecasting domain>, with <small errors> defining success (Kang et al., 2017). ISA reveals how properties of test instances (i.e. features of time series) determine success of algorithms (i.e. forecasting methods).

In order to adopt ISA for understanding the impact of force attributes on combat outcomes we propose a novel mapping of the defence application that treats simulation parameters defining force attributes as a test instance, and the "algorithms" are combat simulations of Experimental (Blue) and Opposing (Red) forces whose performance is measured as the survival ratio of assets remaining at the end of the simulated combat. In other words, ISA involves running a <combat simulation> on multiple <force structure scenarios> within a <combat setting> to generate survival ratios defining win or loss.

The study utilises data from the Joint Future Operating Concept Explorer (JFOrCE): an agent-based simulation model (Au et al., 2018) that enables studies of engagements between competing forces with variations in force structure. The force structure is defined as a set of generic military assets such as fighter jets, submarines, air warfare destroyers, etc. Simulating the model leads to force engagements between the Blue and Red team assets, with interactions occurring via technological capabilities comprised of sensing platforms, weapons systems, and information sharing mechanisms. The extent of these capabilities for each competing force is provided by simulation parameters, along with parameters controlling the number of weapons available, and the probability of success (kill) when deploying a weapon. Given these parameters defining a simulation scenario, the model follows a set of defined engagement rules (e.g. shooting at a target when in range, or retreating from a threat if no weapon is available) that determine the actions of agents, with dynamic movement tracked via geospatial positioning, until the end of the simulation which records each force's remaining assets.

The purpose of the JFOrCE model is to test trade-offs in capability under the widest range of potential scenarios to assist with investment decision making. Two main types of capability are of interest: individual force assets with technological specifications (e.g. jets with certain sensor ranges), and collective sharing of information between force assets, e.g. via an Airborne Early Warning and Control (AEW&C) system, which is capable of sharing its own sensor information with other air defence capabilities, including Ground Based Air Defenders (GBADs) and Air Warfare Destroyers (AWDs). An information advantage is created by passing locations of enemy assets within the sensor range of the AEW&C onto all Blue GBADs and AWDs for target execution.

In a previous study by Au et al. (2018), various scenarios were explored where Blue has a force advantage with more initial assets and/or an information advantage provided by the presence of AEW&C assets to share information. In particular, the question of whether Blue information advantage helps overcome Red force advantage is explored through simulations. The conclusions provided include:

- Blue information advantage amplifies the tendency to win for similar force sizes;
- Blue force advantage of more AWDs is sufficient to magnify Blue's winning probability (more so than jets or GBADs) in the absence of information advantage;
- Blue information advantage helps counterbalance Red's force advantage with more jets or GBADs only;
- Blue information advantage does not overcome Red's force advantage with more AWDs.

Based on the simulation scenarios considered, it is clear that more AWDs provides either side with a force advantage; information sharing via AEW&C provides either side with an information advantage; but the information advantage cannot overcome an opposing force advantage if the opposing side has more AWDs. The data employed by Au et al. (2018) generated scenarios that explored the number of such critical assets (AWDs and AEW&C) required for Blue success. In contrast, the data provided for our current study largely assumes each side has the same number of assets but has varied the technological capabilities in terms of sensor ranges and speeds. We can still consider force and information advantage, however, in terms of difference in sensor speed and sensor range capabilities of the assets. Furthermore, a small set of additional simulations are conducted to vary the number of initial assets while assuming each side's assets have similar technological capabilities. While the data generated by JFOrCE for this study is therefore different from Au et al. (2018), we aim to explore whether ISA reveals similar conclusions using its visualisation capabilities, and how ISA can guide the design of additional simulation scenarios that would be useful to generate additional insights.

2. INSTANCE SPACE ANALYSIS

In many fields that rely on algorithms, such as operations research and machine learning, there has been long-standing criticism of how algorithms are typically tested to establish their trust and reliability (Hooker 1995; McGeoch, 1996). Standard academic practice in these fields typically produces a suite of test problems (from real-world scenarios, or randomly generated problems), and reports the performance of an algorithm on average across this test suite. If the performance is acceptable on average, and certainly if it is better than alternative approaches, a new algorithm is deemed successful since its reliability has been tested on some random, presumably unbiased, data. The problem with this approach is that it offers no scrutiny over whether the chosen test instances are truly unbiased. Furthermore, reporting performance on-average can mask risk of failure, offering no insight into how to avoid deployment disasters that may only be encountered for specific scenarios, but are lost on average if such scenarios are infrequent in the test data. If we are to trust the performance evaluation, we must challenge the choice of test problems, and have the means to scrutinise their properties: are they unbiased, diverse, challenging, discriminating, and representative of real-world scenarios? It is unusual for randomly generated test problems to possess these important properties. We must also move away from on average reporting and analyse at the per-instance level in the test set to gain insights into the conditions under which a model or algorithm is expected to fail or succeed. It has long been recognised as an open challenge to develop a more rigorous methodology to establish trust that test instances are fit for purpose, and to support a more "empirical science of algorithms" (Hooker, 1994).

An alternative "stress-testing" methodology – known as *Instance Space Analysis (ISA)* – offers visualisations and analytics to support reliable decision-making and trust in algorithms. The key steps ISA are summarised in Figure 1, and we refer the interested reader to several key papers (Smith-Miles et al., 2014; Munoz et al., 2018) and the online tool[1] for more details. By creating a 2D instance space of test problems, including a mathematically defined boundary beyond which no instances can theoretically exist, ISA creates a 2D map which essentially "cracks open" a test suite, offering insights that are otherwise hidden by on-average analysis. Most critically, ISA offers:

(i) scrutiny of the suitability of test instances – with metrics quantifying diversity, bias, discrimination, and real-world-likeness;
(ii) visual insights into how test instance properties affect algorithm performance, and why;
(iii) objective assessment of strengths and weaknesses of algorithms by measuring the area of their "footprint", based on empirical evidence-based machine learning predictions; and
(iv) guidance on the generation of useful additional test instances with controllable properties to fill within the mathematically defined boundary of the instance space.

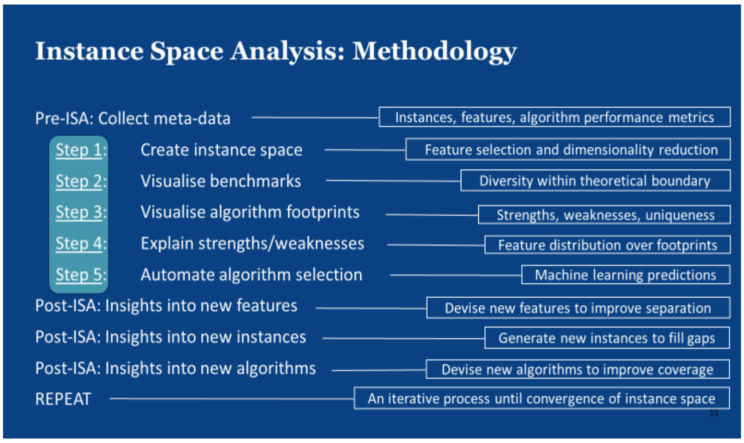
The ISA methodology has been tested within numerous fields of algorithmic science with well-established benchmark test instance repositories to support the above-mentioned claims of its advantages. In this paper however, we are more interested in how the ISA methodology can be effectively mapped to a completely different context as a first step towards gaining similar insights in the field of combat analysis.

3. SIMULATION DATA

Two datasets were provided for this study comprising simulation runs from JFOrCE, showing the remaining assets for Red and Blue teams, under various scenarios related to asset parameters such as sensor and weapon ranges, and speeds, and probability of kill. For dataset 1, the number of initial assets was identical for both teams, but the technological capabilities were varied to provide a range of scenarios where each team has force and/or information sharing advantage. For dataset 2 the capabilities were fixed but the initial force size was varied asymmetrically. The list of available assets is provided in the Appendix (Table 1), along with the parameters defining a simulation scenario (Table 2). For dataset 1, a set of 92 different scenarios were considered that systematically vary the parameters in Table 2 within ranges. For a given parameter range

[1] https://matilda.unimelb.edu.au

defining a scenario, around 20 parameter samples[2] are generated. For each sample, 200 iterations of the simulations are performed, due to the stochastic nature of the simulation model, and the average number of assets remaining for each team over these 200 trial runs are recorded. Consequently, there are average outcomes recorded for 1854 scenarios in dataset 1. For dataset 2, 57 scenarios were considered by varying the parameters in Table 2. The result is a combined set of 1922 scenarios with average assets remaining. We define the performance metric as the survival ratio of Blue team assets (i.e. remaining total assets per initial total assets).


Instance Space Analysis: Methodology

Pre-ISA: Collect meta-data — Instances, features, algorithm performance metrics

Step 1: Create instance space — Feature selection and dimensionality reduction

Step 2: Visualise benchmarks — Diversity within theoretical boundary

Step 3: Visualise algorithm footprints — Strengths, weaknesses, uniqueness

Step 4: Explain strengths/weaknesses — Feature distribution over footprints

Step 5: Automate algorithm selection — Machine learning predictions

Post-ISA: Insights into new features — Devise new features to improve separation

Post-ISA: Insights into new instances — Generate new instances to fill gaps

Post-ISA: Insights into new algorithms — Devise new algorithms to improve coverage

REPEAT — An iterative process until convergence of instance space

Figure 1. Summary of ISA methodology

Within the ISA framework, the "algorithm" (simulated combat outcome) is deemed good if Blue's survival ratio is no less than Red's, and bad otherwise. The test scenarios are based on the values of parameters in Table 2, and pre-processed to generate features defining the advantage of Blue over Red in each asset or capabilities (e.g. adv_jets = initial blue jets – initial red jets). The goal of ISA in this combat analysis study is to visually explore how the force scenarios influence performance of red/blue survival ratios and combat outcomes, and to assess the adequacy of the test instances.

4. INSTANCE SPACE ANALYSIS FOR COMBAT SIMULATION DATA

The available meta-data comprises 1911 instances (scenarios), described by a set of 21 features that capture blue advantage in terms of assets and capabilities, with good "algorithm" performance defined by having a superior (or not inferior) Blue survival ratio compared Red.

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 0.6878 & -0.3699 \\ -0.1344 & -0.0749 \\ -0.8006 & -0.7114 \\ -0.1662 & -0.1382 \\ -0.1139 & -0.1629 \\ -0.5767 & 0.3462 \\ -0.6245 & 0.2626 \end{bmatrix}^T \begin{bmatrix} adv-jets \\ adv-jet-speed \\ adv-jet-sensor-range \\ adv-info-range \\ adv-AWD-sensor-range \\ AIM9-PK \\ AGM88-PK \end{bmatrix} \qquad (1)$$

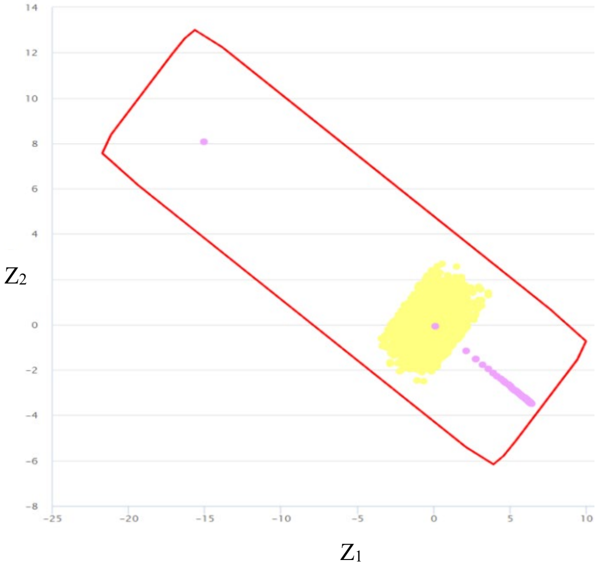


Figure 2. Location of 1854 instances (scenarios) with equal assets (dataset 1 in yellow), and 57 instances with varied assets (dataset 2 in pink).

To construct the instance space of all possible scenarios we used the online tool MATILDA, with the default parameter setting except that a maximum of 8 features per team were permitted in the feature selection stage, instead of the default limit of 3. MATILDA has selected 7 key features from the 21 candidate features that best describe the performance difference between Blue and Red having the better survival ratio. The projection from the selected 7D feature space to a 2D instance space is given by the following linear transformation shown in Equation (1), with the coefficient matrix based on an optimisation algorithm (see Munoz et al., 2018) that aims to achieve visualisation of near-linear trends across the instance space in terms of algorithm performance metric and instance features. The location of each of the 1911 scenarios within the instance space is given in Equation (1) by their unique coordinates (Z_1, Z_2), and are shown in Figure 2 for the two datasets. The mathematical boundary (shown in red in Figure 2) is defined by projecting

[2] Some scenarios have 19 samples, others have 35, but most have 20 samples per parameter scenario.

the hypercube vertices defined by the upper and lower bounds of each of the 7 features into the 2D space using the same projection matrix and connecting the vertices. The two datasets generated for this study do not fill the entire space of possible test scenarios, although many parts of this theoretical instance space would be unlikely in real-world context. Nevertheless, it is useful to assess the diversity of the chosen test scenarios, and to consider where additional scenarios may be useful. Clearly the equal asset dataset 1 is denser and quite limited in its variation, while the varied asset dataset 2 is spread sparsely across the negative diagonal of the instance space, with many more variations possible but not currently explored. The two datasets have different characteristics, but how do these characteristics affect the outcomes of the battle?

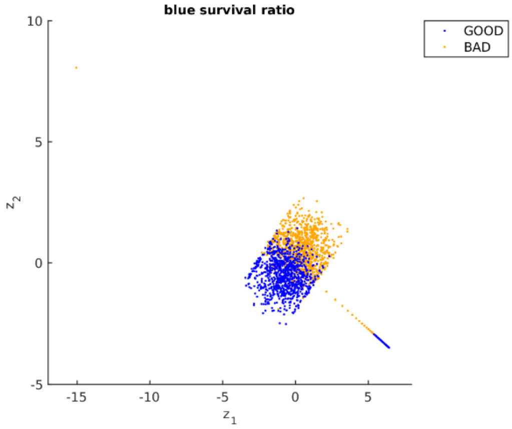


Figure 3. Observed good performance of "Blue algorithm" with better survival ratio shown in blue, and worse survival ratio shown in orange

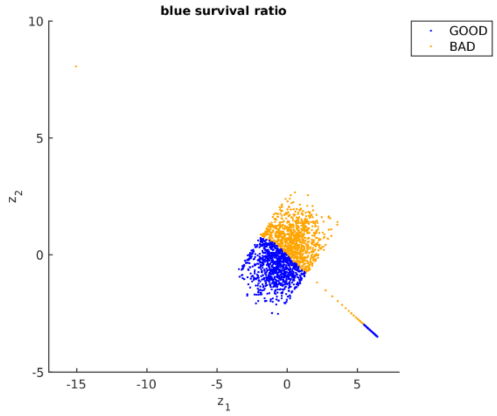


Figure 4. Support Vector Machine prediction of good performance of "Blue algorithm" shown in blue, and worse survival ratio shown in orange

Figure 3 shows the scenarios in which Blue has better survival ratio performance than Red. We see in Figure 3 that scenarios falling in the bottom of dataset 1, and those at the far-right tail of the dataset 2 ensure that Blue wins, while the converse results in a Red win. Clearly the location of a scenario in the instance space tells us the likely outcome, and the location is based only on the feature values of the scenario. In order to establish if some features are predictive of whether Blue will have a better survival ratio or not for a given force scenario, we can use machine learning methods to predict the outcome for untested scenarios. Figure 4 shows the results of a support vector machine (SVM) that has been trained using 10-fold cross-validation to identify the predicted regions of win (good performance) and loss (bad performance) for the Blue team. A similar SVM has also been trained for predicting Red team wins. Combining these two SVMs, we have an SVM recommendation that achieves an 86.3% accuracy in predicting which team will have a better survival ratio given their initial assets and capabilities. By way of comparison, if we assume Blue wins for all scenarios, the accuracy of this naïve model is 47.7% for the scenarios tested in both datasets.

Beyond machine learning prediction of combat outcomes for a given scenario, the instance space provides the opportunity to understand why the location of a scenario, completely determined by its feature vector, affects the combat outcomes. Figure 5 shows the distribution of two of the key features across the instance space, with blue colour indicating minimal values of the feature, and yellow indicating maximal values. Inspecting such feature distributions across the instance space reveals the gradients that determine win or loss are primarily correlated with the feature adv_jet_sensor range (Blue wins for large values in scenarios where both teams have identical initial assets), and the feature adv_jets (Blue wins for scenarios (in yellow) where the Blue team has 13 more initial jets compared to the red team). Blue having an information advantage in the form of AEW&C sensor range, and AWD sensor range, also supports a Blue win. For all other features, their variations are not predictive of outcomes, and so they have been omitted from Figure 5.

5. DISCUSSION AND CONCLUSIONS

This paper has described a proof-of-concept study for how a recently developed methodology known as Instance Space Analysis can be mapped to a combat analysis context to provide visualisations of simulation scenarios exploring how force design impacts combat outcomes. The power of ISA lies in its ability to tease apart a set of test instances, in this case simulation scenarios, to understand how comprehensively they fill the theoretical space of all possible instances, and to assess any bias in testing. We have seen from this study that the available simulation scenarios, generated from two datasets that independently varied technological capabilities and initial force assets (but didn't vary both types of attributes in tandem due to computational limitations) had limited success in filling the entire instance space. Certainly dataset 1, with initial force assets and varying technological capabilities is very comprehensive and densely fills the central region around the origin of the instance space. However, the limited available simulations in dataset 2, varying initial force assets with identical technological capabilities, has created a sparse line of additional instances with much unexplored territory between the tested scenarios. Arguably though, the tested instances represent more realistic scenarios, and there may be little practical value in generating more scenarios merely to comprehensively fill the instance space. The available datasets revealed clear relationships between how the measurable features of the scenarios, based in chosen simulation parameters, affect the combat outcomes. Machine learning methods were able to predict simulated combat outcomes with 86.3% accuracy, and the crucial role of jet sensor range advantage, more than any other advantage, has been revealed.

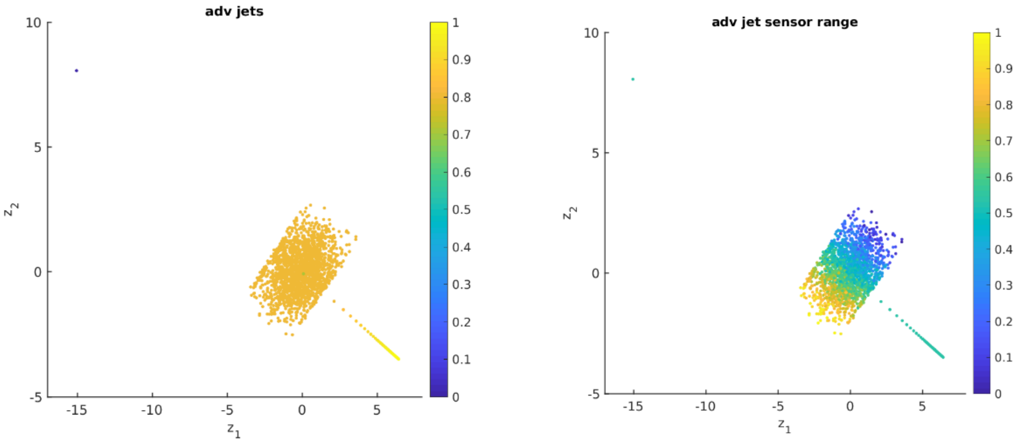


Figure 5. Distribution of four key features across the instance space with minimal feature values shown in blue, and maximal feature values in yellow.

We recommend extending the present study to examine the impact of various force structures, strategies, and assumptions about Opposing Force strategies when exploring the simulation capabilities offered by more recent tools such as Command PE and AFSIM. An ISA for simulations generated by such tools would enable a commander to test various courses of action, with confidence that the tested instances fill practical regions of the instance space, to support decision making based on "algorithms" that have been rigorously "stress-tested".

ACKNOWLEDGMENTS

This project was funded by the Modelling Complex Warfighting Strategic Research Investment of Defence Science and Technology Group. The MATILDA toolkit was developed with funding from the Australian Research Council through the Australian Laureate Fellowship project FL140100012. The authors would like to thank DSTG colleagues, Dr Andrew Gill, Dr Steven Wark, Dr Ed Lo, and Dr Peter Hoek for useful discussions on dataset 1 and making JFOrCE available for this research.

REFERENCES

Au, T. A., Hoek, P. J., Lo, E. H. S., 2018. Combat Analysis of Joint Force Options using Agent-Based Simulation, 2018 Military Communications and Information Systems Conference (MilCIS) pp. 1-7.

Hooker, J., 1994. Needed: An empirical science of algorithms. Operations Research 201-212.

Hooker, J., 1995. Testing heuristics: We have it all wrong. Journal of Heuristics 1(1), 33-42.

Kang, Y., Hyndman, R., Smith-Miles, K., 2017. Visualising Forecasting Algorithm Performance using Time Series Instance Spaces. International Journal of Forecasting 33(2), 345-358.

McGeoch, C., 1996. Toward an experimental method for algorithm simulation. INFORMS Journal on Computing 8(1), 1-15.

Muñoz, M. A., Villanova, L., Baatar, D., Smith-Miles, K. A., 2018. Instance Spaces for Machine Learning Classification. Machine Learning 107(1), 109-147.

Smith-Miles, K. A., Baatar, D., Wreford, B., Lewis, R., 2014. Towards Objective Measures of Algorithm Performance across Instance Space. Computers and Operations Research 45, 12-24.

APPENDIX

Table 1. Assets and their initial quantities [ranges] for red and blue teams for each simulation run

Asset name and description	Dataset 1	Dataset 2
AIM9: Short-range air-to-air missile, used by jets to shoot at other jets, using optical (IR) guidance system	1 per jet	1 per jet
AIM120: Long-range air-to-air missile, used by jets to shoot at other jets beyond visual range, using radar guidance system	1 per jet	1 per jet
AGM88: Air-to-Ground missile, used by jets to shoot at GBAD and tanks	1 per jet	1 per jet
AEW&C: Airborne Early Warning & Control aircraft, with long-range radars, used to guide jets to targets (all remain)	3	3
AWD: Air Warfare Destroyer (ship), equipped with surface-to-arm missiles to shoot jets	5	Blue 2; Red 1
GBAD: Ground Based Air Defence, equipped with surface-to-arm missiles to shoot jets	all remain	Blue 1; Red 4
Jets: fast jets that shoot missiles at jets, AWD, GBAD, tanks	32	Blue [11,68] Red [10,25]
Subs: Submarines that shoot at subs and AWD	3	0
Tanks: Tanks that shoot at enemy tanks	10	Blue 10; Red 25

Table 2. Asset parameters [ranges] defining a simulation scenario

Capability	Description	Dataset 1	Dataset 2
jet_speed	speed of friendly/hostile jets used in a simulation	[1200,2800]	1200
jet_sensor_range	sensor range of friendly/hostile jets used in a simulation	[25,180]	150
AEW&C_speed	speed of friendly/hostile AEW&C aircraft used in a simulation	[100,1600]	Blue 760 Red 700
AEW&C_sensor_range	range of sensor range of friendly/hostile AEW&C aircraft used in a simulation	[150,650]	700
GBAD_sensor_range	sensor range of friendly/hostile GBAD platforms used in a simulation	[45,200]	100
GBAD_weapon_range	weapon range of friendly/hostile GBAD platforms used in a simulation	[70,225]	100
AWD_sensor_range	sensor range of friendly/hostile AWD ships used in a simulation	[50,300]	Blue 95 Red 100
AWD_weapon_range	weapon range of friendly/hostile AWD ships used in a simulation	[95,250]	100
subs_speed	speed of friendly/hostile subs used in a simulation	[24,54]	Blue 280 Red 290
subs_sensor_range	sensor range of friendly/hostile subs used in a simulation	[65,265]	Blue 280 Red 290
subs_weapons_range	weapon range of friendly/hostile subs used in a simulation	[69,270]	Blue 280 Red 290
AIM120-PK	probability of kill of AIM120 missiles used by both friendly/hostile jets	[0,1]	0
AIM120-Range	range of AIM120 missiles used by both friendly/hostile jets.	[76,138]	0
AIM9-PK	probability of kill of AIM9 missiles used by both friendly/hostile jets	[0,1]	0.17
AIM9-Range	range of AIM9 missiles used by both friendly/hostile jets	[19,36]	95
AGM88-PK	probability of kill of AGM88 missiles used by both friendly/hostile jets	[0.5,1]	0.45
AGM88-Range	range of AGM88 missiles used by both friendly/hostile jets	[67,141]	85
AWD-PK	probability of kill of missiles used by both friendly/hostile AWD ships	[0.5,1]	0.45
GBAD-PK	probability of kill of missiles used by both friendly/hostile GBAD systems	[0.5,1]	0.4