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

# Simulation and Data Analytics for a Defence Workforce Transition

K. Mortimer [a], V. Nguyen [a], T. Caelli [b] and B. Hill [b]

[a] *Defence Science and Technology Group, 506 Lorimer St, Fishermans Bend, VIC, 3207, Australia*
[b] *Deakin University, Waurn Ponds, Geelong, Vic 3218, Australia*
*Email: katie.mortimer-murphy@dst.defence.gov.au*

**Abstract:** The Australian Army Aviation (AAvn) is transitioning their Armed Reconnaissance Helicopter (ARH) capability from the Tiger ARH to the Apache Guardian helicopter. This transition requires retraining of all current Tiger-trained personnel in a specific time period from 2025 to 2029. While these personnel are being transitioned, the ARH capability in AAvn must be maintained. The complexity of the Defence workforce, with its hierarchical nature and highly interconnected structure, means that this transition has many workforce risks that need to be addressed through advanced planning. Further, there are many unknown variables such as workforce and resource availability, training requirements and workforce attrition. These challenges are magnified when it comes to pilots, where resource constraints and the requirement for a highly-trained workforce makes planning more difficult. Workforce analysts must have a good understanding of their workforce, and how a transition affects it, to create a robust workforce transition plan and ensure that workforce risk is minimised.

In this workforce transition analysis, a combination of simulation, Design of Experiments (DoE) and visual analysis was used to provide workforce planners with an understanding of how the AAvn pilot workforce is affected by the transition. A discrete time simulation (DTS) that models individuals, as they progress through their career and complete postings, was used to simulate the pilot workforce. The simulation engine was able to take into account personnel eligibility for particular postings and promotions.

Due to the large number of unknown variables in planning for a workforce transition, a DoE approach was used to ensure the entire problem space was effectively explored. The focus of this analysis was to determine how variations in transition courses affect the ability of AAvn to maintain its critical ARH capability, as well the entire AAvn workforce. Variations in the transition courses include transition course length, numbers transitioned per course and the delay between transition courses. This gave 42 experimental factors, so the experimental matrix was determined using a nearly orthogonal and balanced data farming design from the Naval Postgraduate School SEED (Simulation Experiments & Efficient Designs) Center for Data Farming [SEED, 2021].

Visual analytics were used to visualise the the large amounts of data produced. Complex data analytics techniques were then used to explore the AAvn workforce and find vulnerabilities, and understand why they were occurring through the analysis of relationships between inputs and outputs and between different outputs. Highly interactive visualisations of these analysis techniques were built to provide workforce analysts with the ability to further explore and understand various scenarios and outcomes.

Time series analysis was used to analyse the vulnerabilities of AAvn units throughout the workforce transition. Correlation analysis between simulation inputs and unit performance was used to determine how relationships between inputs and outputs affect these vulnerabilities. This analysis was displayed to the user via a correlation heatmap, with the user being able to explore these relationships further by clicking on the node to display the input-output relationship of that node.

A Bayesian network was fitted to the data using the PC algorithm [Spirtes et al., 1993] and parameter estimation was completed using the Expectation-Maximisation method. Effective visualisation and interactivity of the Bayesian network showed the complexity and interconnectedness of the AAvn pilots workforce, as well as allowing the user to explore in more detail specific relationships and variables. The use of Bayesian network inference allowed users to perform *what-if* analysis, without the need for further simulations, giving them a clearer understanding of the direct effect of the relationships between variables.

***Keywords:*** *Data analytics, visual analytics, simulation, data farming*

## 1 Introduction

The Australian Army Aviation (AAvn) corps provide a critical capability to the Australian Army and the wider Defence Force. Pilots are a crucial part of this capability but can be one of the most difficult to train and sustain in the workforce. This is because training a pilot is a time-consuming process that requires many resources. Aircrew training schools within the Australian Defence Force (ADF) are shared by AAvn, the Royal Australian Navy (RAN) and the Royal Australian Air Force (RAAF). This means resources must be shared and the needs of each service must be balanced against each other. Once AAvn pilots are flight-trained they also receive further platform-specific training. Even within AAvn there are competing needs as the trained personnel must be split between multiple platforms. These competing needs, as well as unforeseen circumstances such as particularly high failure rates, workforce separation or low recruitment numbers, can mean that pilot inflow into the system can vary from year to year. It is necessary for workforce scheduling to be resilient enough to withstand these variations.

The resilience of the workforce is especially important throughout workforce transitions such as the one currently occurring within AAvn as they transition their Armed Reconnaissance Helicopter (ARH) capability from the Tiger ARH to the Apache Guardian helicopter. This transition is occurring over the period 2025 to 2029 and requires retraining of the current Tiger-trained personnel to the new Apache helicopter while, at the same time, requiring the Australia's ARH capability to be maintained. To ensure that the workforce transition occurs in this time-frame, with as little disruption to the capability as possible, a robust workforce transition plan needed to be created. This transition plan had to include planning for many of the risks associated with a workforce transition, such as resource and training delays.

Simulation plays an important role in workforce planning and supply analysis and has proven useful in Defence-specific workforce applications such as this [Davenport et al., 2007; Nguyen et al., 2017]. However, in this case there are a large number of unknown variables in the AAvn workforce transition, such as attrition, future personnel inflow, and workforce availability. Additionally, workforce planners need to plan for retraining current personnel, training of new personnel, resource availability, and changes in personnel capability requirements.

Addressing these challenges requires a range of approaches to be used in conjunction with simulation. Firstly, simulation sessions need to sparsely sample the complete range of policies and variable values. To produce such data, a Design of Experiments (DoE) approach should be used to avoid the factorial explosion of all the possible variable combinations. Data Farming is a DoE approach for designing large simulation experiments to maximise the useful information gained and has been studied extensively by the Naval Postgraduate School SEED (Simulation Experiments & Efficient Designs) Center for Data Farming [Kleijnen et al., 2005]. It has already been used in various Defence applications, including homeland security [Lucas et al., 2007], NATO decision support [Huber et al., 2019], aircraft fleet management [Marlow et al., 2015] and in the analysis of Defence workforce [Hill et al., 2019].

Due to the complexities in these types of simulations, a combination of visualisation and data analytics is necessary for the end user to understand resultant simulator input-output relationships - visual analytics [Thomas and Cook, 2005; Lavigne et al., 2011] was used in this study. Visual analytics encompasses a wide range of visualisation and analytical techniques, including machine learning, statistical analysis, descriptive analysis and graph analysis. While visual analytics has been used in Defence applications such as maritime domain awareness [Lavigne et al., 2011] and cyber security [Ionita and Patriciu, 2015], its use in the analysis of workforce simulation data has so far been limited.

In this study, the focus will be on exploring workforce risk, relationships between inputs and outputs and between different outputs, in order to answer questions such as: where are vulnerabilities in the workforce and why are particular vulnerabilities occurring; how does the workforce respond to changes caused by the workforce transition; and how do workforce vulnerabilities at different points in time affect other parts of the workforce.

## 2 Defence Workforce Simulation and Experiment Design

Given the challenges present in planning for the AAvn ARH workforce transition, an analysis into how variations in the workforce transition training affected the ability of the operational units in AAvn to meet their required personnel numbers, was performed. The simulation engine is explained in subsection 2.1.

### 2.1 Simulation Engine

Modelling of the AAvn pilot workforce was implemented in a discrete time simulation (DTS) that modelled individual personnel as they progress through their career and complete postings. Pilots in ranks Lieutenant (LT), Captain (CAPT), Major, (MAJ) and Lieutenant Colonel (LTCOL) were modelled and move through the workforce as in Figure 1. Units were defined as groupings of positions and included operational units or regiments, training schools and headquarters. Postings refer to a particular combination of rank and unit, and have a target requirement, which the simulator will attempt to fill with eligible personnel, a priority value that defined the order in which it is filled, and may also have a platform endorsement requirement. Posting priority, target and eligibility requirements are defined by AAvn and cannot be altered. Personnel attributes, such as current posting, platform endorsement, rank and current time in posting, determined if they were eligible to be promoted to the next rank or fill a particular posting. Personnel may be promoted to the higher rank if they have reached the minimum time in rank constraint and there is an available posting at the next rank.

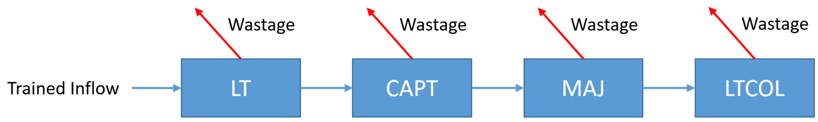


**Figure 1**. Flow of pilots through their career, with the application of inflow and wastage shown.

The model had initial states and posting requirements that changed over time as the Tiger ARH capability was reduced and the Apache ARH capability increased. There were also a number of adjustable input parameters of the model, such as yearly inflow, wastage rates and transition course numbers, length and delays. These parameters were varied throughout the experiment as discussed in subsection 2.2.

### 2.2 Design of Experiment

In this analysis, the focus was on how variations in the transition courses affect the units that were being transitioned, as well as the entire workforce. The factors investigated were the transition course length, transition course numbers and transition course delay. For each transition course session and each rank within a session, giving a total of 42 factors, a reasonable expected range was given. The output variables of interest were the numbers of personnel in each unit over ten years. The inflow and wastage rates were kept constant.

Data Farming was used to determine the experimental matrix. Based on the number of factors in this experiment, a nearly orthogonal and balanced (NOB) design was chosen and implemented using the ”NOB mixed design worksheet” available from the Naval Postgraduate School SEED Center for Data Farming [SEED, 2021]. This provided a NOB design of experiment with 512 design points [Vieira et al., 2011]. These designs could be *stacked* to provide a larger design matrix that was able to fill the space better and, as such, two NOB design matrices were stacked in each experiment so that 1024 experimental runs were completed. Each design point was also replicated 20 times and averaged to ensure stability of the simulation results.

## 3 Data Analytics and Interactive Visualisation

Visual analytics [Thomas and Cook, 2005] was then used to analyse the large amount of data gathered. Techniques should be chosen depending on the types of data and the types of questions to be answered. In this case, workforce analysts were particularly interested in the questions described in section 1. As such, three interactive visualisations of data analytics techniques were employed to answer these questions. These were:

1. Visualisation of time varying data - a visual analysis of the scheduler dynamics for AAvn units over time and throughout the simulation runs to discover which units were consistently being under or over supplied.

2. Correlation Heatmap - a correlation analysis between the simulation inputs and unit performance to find relationships between inputs and outputs.

3. Network Analysis - a Bayesian Network analysis of the simulation data to further analyse input to output relationships, as well as to analyse output to output relationships. Bayesian Network querying and inference methods supported further analyses of the effect of these relationships on the performance of the AAvn units throughout the transition.

### 3.1 Visualising the scheduler input and output dynamics

The visual analysis of the workforce transition began with finding points in the workforce that were particularly vulnerable: consistently being under or over-supplied. The simplest way to do this was to aggregate the results over the simulation runs and display this to the analyst via a time series plot. This had the advantage of being able to display the $>$1000 simulation runs in an intuitive and familiar way, allowing for quick identification of units that were particularly vulnerable. The aggregation used was the maximum and minimum values of the simulation runs over time, as this was an easily understood concept and clearly showed the impact of the varying inputs on the workforce.

The time series display for the experiment analysing the effect of transition variables on the workforce can be seen in Figure 2. A table on the right-hand side displays a further aggregated ranking of the units, shown as a percentage of the bar graph how often the different units achieved their required capability, or were under or over supplied. The analyst could then use this to click through the time series plots of the data and determine which units or ranks are particularly risky. Using this display, it can be seen that some units, such as PILOT School Unit 1, PILOT Other, and PILOT School Unit 2 performed well throughout the simulations. However PILOT Unit 1 can be seen to be consistently under-supplied, while PILOT Unit 2 is alternatively under-supplied, well-supplied and over-supplied. Analysis of the time series plots reveals points in time that are particularly vulnerable. However, the user is still unclear on why these vulnerabilities are occurring. Correlation heatmaps were used to explore this further.

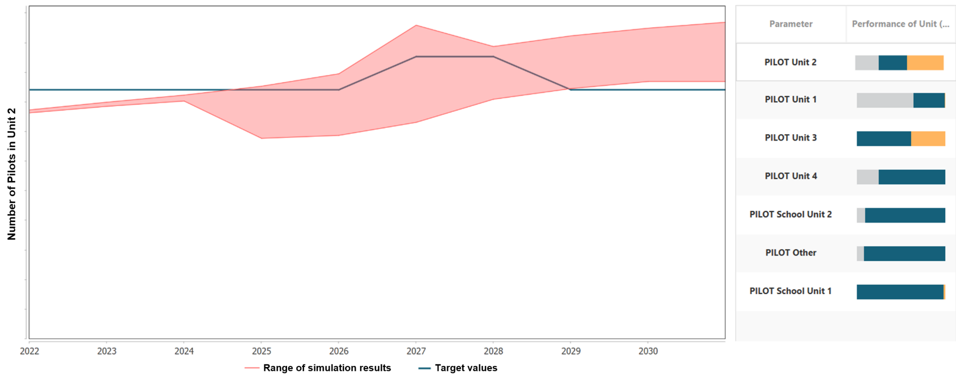


**Figure 2**. Time Series Visualisation of the various units in AAvn. This time series visualisation shows minimum and maximum simulation results as a shaded pink area, as well as target values in blue. Bar graphs on a table (right hand side window) allow the user to see how different units perform throughout the simulations, with the amount of time a unit spends under-supplied represented by blue, time spent over-supplied represented by yellow and time on target represented by grey. The users can interact with the visualisation by clicking on the units to see the corresponding time series analysis.

### 3.2 Correlation Heatmaps

Correlation analysis was used to find significant relationships between simulation inputs and outputs to determine how vulnerable portions of the workforce were affected by changes in workforce inputs. It could be reasonably assumed that these relationships were monotonic due to the nature of the workforce. For instance, a larger delay between two transition courses or a smaller number of personnel transitioned at a course would reduce the number of personnel available for the new capability. As such, Spearman's (rank order) correlation coefficient [Pirie, 2006] was used to find the correlation between the input variables and the outputs at each time step. This created an $n \times m$ matrix, where $n$ is the number of input variables and $m$ is the number of output variables, of correlation coefficients between -1 and 1.

These correlations were displayed to the analyst via a correlation heatmap using colour maps, as in Figure 3. This allowed the analyst to see where relationships were particularly strong and how vulnerable parts of the workforce might be influenced by various inputs. Further exploration of the heatmap occurred when the user double clicks on a square in the heatmap, displaying a scatter-plot of simulation results for the input and output.

The correlation heatmap exposed the users to relationships between inputs and outputs that may explain under and over-supplies in the results. For instance, PILOT Unit 1 was found to be consistently under-supplied

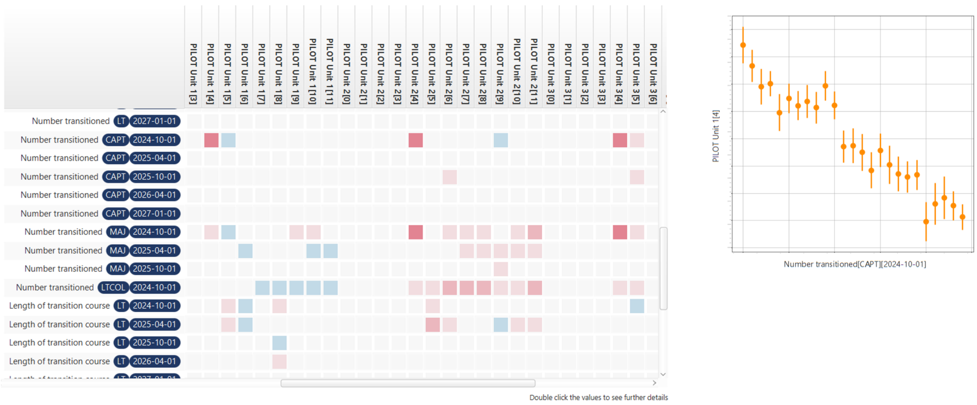


**Figure 3**. Part of the correlation heatmap visualisation. Rows represent inputs and columns represent outputs. The colourmap defined large negative correlations by darker reds, large positive correlations by darker blues and low correlations by grey. Double clicking on a coloured square shows the relationship of the input-output simulation data (right hand side window).

throughout the simulations. It can be seen in Figure 3 that PILOT Unit 1 has particularly strong positive correlations with the variables related to the transition course held in 2024, and in particular the numbers of CAPTs and MAJs transitioned in this course, and the length of this transition course. It follows that the under-supply may be caused and, furthermore, solved by focusing on the numbers of pilots transitioned during the transition course in 2024. These inferences are then more explicitly and syntactically examined via Bayesian network inference techniques.

### 3.3 Bayesian Network Analysis

The next step in the analysis was to explore how the relationships between variables affect the workforce in a way that allows for an explicit, syntactic understanding of relationships and dependencies between key input and output variables. Analysts can then understand how, for example, vulnerabilities in one node might affect another part of the workforce.

Bayesian Networks are probabilistic graphical models that represent the conditional dependencies between a set of nodes using a directed acyclic graph (DAG). This graph structure is able to display large, complex and highly interconnected data, such as the pilot data, in an intuitive manner. Bayesian networks are also able to estimate parameter conditional probability distributions. This allows for querying of the model, where a probability distribution of a node's value is estimated based on some "evidence". These functions of the Bayesian Network are useful in the analysis of the workforce transition. That is, from the simulation data, the relationships can be further explored through the network enabling immediate analysis of the effect of variations in both workforce inputs and outputs using a network querying capability.

To model the Bayesian Network structure and estimate parameter distributions the *Bayes Server* platform *BayesFusion* SMILE (Structural Modeling, Inference and Learning Engine) was used [BayesFusion, 2021]. It uses a range of standard parameter estimation and structure learning algorithms. For the model parameter estimation the standard Expectation-Maximisation (E-M) option was used where different models are assessed via their log-likelihood scores [Lauritzen, 1995]. As for the more difficult problem of structure learning, an initial structure was created using the correlation results from subsection 3.2, where links were added if the correlation was above a threshold value of 0.7. This value was chosen to avoid a network that was overly populated and uninterpretable. The initial structure then seeded the BayesFusion model estimator using the PC algorithm [Spirtes et al., 1993] to learn a final network structure. Constraints were placed on the links such that the direction of links in the network could not go backwards in time, nor could they go backwards from an output variable to an input variable.

Visualisation of the network needs to be intuitive allowing for interactive exploration of the network. As such a grid layout was used to display the nodes, with each row in the grid representing a particular unit and the columns representing a yearly simulation time step. Links between nodes were then drawn, showing a highly connected Bayesian network, as in Figure 4. To reduce the number of links shown, links between the same units at $t$ and $t + 1$ were not shown as these links were present, but uninformative in the analysis. Interacting with the network by hovering over or clicking on particular nodes reduces the number of links further by showing only links that are part of the node's Markov blanket - the parent, children and sibling variables of the node. Clicking on a node also brings up further details, as in Figure 5. A Gaussian plot shows the probability distribution of that node, as estimated by the Bayesian network, with colours on the Gaussian representing when the node is under-supplied, acceptable or over-supplied. The nodes inside the Markov blanket are also given in a table, where new query values for these nodes can be set. Once the query is run, the new probability distribution of the node based on the new evidence will be displayed on the Gaussian distribution.

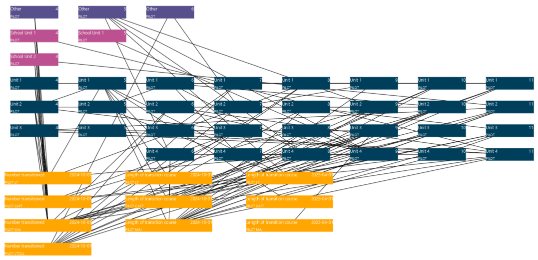

**Figure 4**. Bayesian network visualisation of the transition experiment. The nodes are sorted so that the y-axis represents specific units with input variables at the bottom, while the x-axis represents timesteps between those units. Only node with dependencies are shown. Colour is used to group nodes as input variables, operational units, school units, and other units.

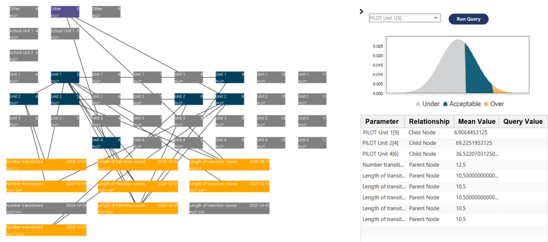

**Figure 5**. Bayesian network visualisation of the transition experiment, when a node has been clicked on. All nodes that are not in the Markov blanket have been greyed out, with corresponding links removed. The Gaussian plot displays the parameter estimation done by the Bayesian network model. Child, sibling and parent variables are shown in the table. Queries can be run from inside this table.

The Bayesian network allows users to take their analysis of the transition a step further, with the ability to discover and explore relationships and dependencies between inputs and outputs, and between different outputs through the use of effective visualisation and *what-if* analysis. In this instance, time series visualisation showed that PILOT Unit 1 should be a focus of this analysis. Correlation analysis further analysed the relationships between inputs and outputs and showed that the number of personnel transitioned in the course held in 2024 was a particularly important factor. These relationships were seen in the Bayesian network, however it could also be seen that PILOT Unit 1 has dependencies with PILOT Unit 2, PILOT Unit 4 and the headquarters unit, further showing how interconnected the system is. Inference of PILOT Unit 1 at a particular point in time showed the large effect that these parameters, and possible solutions to the under-supply problem were investigated, such as reducing the number of personnel transitioned in 2024 and reducing the course length.

## 4 Discussion and Conclusion

Data analysis and the corresponding visualisation of results are an important part of exploring simulation data. In the case of an ADF workforce transition with many unknown or uncertain variables, various analytics and visualisation techniques have been used to explore workforce risks and vulnerabilities. In an interactive time series plot, the user was able to explore the performance of units or ranks throughout the experiment and find points that were particularly vulnerable. This was followed by an interactive correlation heatmap that provided the user with an understanding of how simulation inputs and outputs were related to each other. The final step in the visual analysis process was an interactive Bayesian network. This allowed the user to investigate further the relationships found in the correlation heatmap, and the relationships between different outputs. The user was able to explore the network through mouse hovers and clicks, as well as by utilising the ability to make queries of the Bayesian network and see the direct effect of the relationships between parameters.

There are a number of advantages of using techniques such as these to analyse simulation data. Firstly, discussions with subject matter experts (SMEs) in the military have shown that these visualisations are intuitive and understandable. Involving SMEs in this manner is a critical part of being able to draw real-world conclusions from the simulation data and then use the insights gained as part of the workforce planning process. The intuitive nature of the visualisations means that SMEs that lack experience in simulation and data analytics are able to use and take advantage of these techniques in their workforce planning.

Another advantage of these techniques is that the results from this analysis allow for further, more detailed analysis of the workforce. For instance, the Bayesian Network of the transition experiment showed that only seven factors out of 42 had a large effect on the units during the transition. The process of Data Farming and consequent analysis could be repeated with a focus on these parameters. Optimisation could also be used on this smaller subset of critical parameters to find an optimal workforce transition plan. In this case, the results from this analysis were used to create scenarios that were run in a higher fidelity simulation engine, that modelled resource requirements on courses, the training portion of the career before becoming a LT and career courses attended while part of the workforce.

Without these technologies and relying only on the DTS places significantly more demands on the analyst and clients. However, further development of these techniques may be able to provide improved results and visualisations. In the case of Bayesian Networks, their usefulness has been proven however they are limited by how many links and nodes can be displayed. In more complex problems, with more workgroups or factors, the network could explode to an incomprehensible amount. It may be useful to allow the user to explore an aggregated version of the network, with more related nodes and links shown on request. The correlation analysis used in the correlation heatmap, as well as to provide the initial structure for the Bayesian network, could also be further investigated. Possible areas of research for these applications include machine learning techniques, principal component analysis and regression techniques.

## ACKNOWLEDGMENTS

We wish to thank AAvn for their support.

## REFERENCES

BayesFusion [2021], 'Smile: Structural modeling, inference, and learning engine'.
**URL:** *https://www.bayesfusion.com/smile/*

Davenport, J., Neu, C., Smith, W. and Heath, S. [2007], Using discrete event simulation to examine marine training at the marine corps communication-electronics school, *in* '2007 Winter Simulation Conference', pp. 1387–1394.

Hill, B., Vukcevic, D., Caelli, T. and Novak, A. [2019], Insights into the health of defence simulated workforce systems using data farming and analytics, *in* '2019 Winter Simulation Conference (WSC)', pp. 2491–2502.

Huber, D., Horne, G., Kallfass, D., Hodicky, J. and Reus, N. d. [2019], Data farming services: Micro-services for facilitating data farming in nato, *in* '2019 Winter Simulation Conference (WSC)', pp. 2455–2466.

Ionita, M.-G. and Patriciu, V.-V. [2015], Cyber incident response aided by neural networks and visual analytics, *in* '2015 20th International Conference on Control Systems and Computer Science', pp. 229–233.

Kleijnen, J. P. C., Sanchez, S. M., Lucas, T. W. and Cioppa, T. M. [2005], 'State-of-the-art review: A user's guide to the brave new world of designing simulation experiments', *INFORMS Journal on Computing* **17**(3), 263–289.

Lauritzen, S. L. [1995], 'The em algorithm for graphical association models with missing data', *Computational Statistics & Data Analysis* **19**(2), 191–201.
**URL:** *https://www.sciencedirect.com/science/article/pii/0167947393E0056A*

Lavigne, V., Gouin, D. and Davenport, M. [2011], Visual analytics for maritime domain awareness, *in* '2011 IEEE International Conference on Technologies for Homeland Security (HST)', pp. 49–54.

Lucas, T. W., Sanchez, S. M., Sickinger, L. R., Martinez, F. and Roginski, J. W. [2007], Defense and homeland security applications of multi-agent simulations, *in* '2007 Winter Simulation Conference', pp. 138–149.

Marlow, D., Sanchez, S. and Sanchez, P. [2015], 'Testing aircraft fleet management policies using designed simulation experiments'.
**URL:** *https://calhoun.nps.edu/handle/10945/57856*

Nguyen, V., Novak, A., Shokr, M. and Pash, K. [2017], Aircrew manpower supply modelling under change: an agent-based discrete event simulation approach, *in* 'WinterSim 2017'.

Pirie, W. [2006], *Spearman Rank Correlation Coefficient*, Vol. 8.

SEED [2021], 'Seed center for data farming software downloads'.
**URL:** *https://nps.edu/web/seed/software-downloads*

Spirtes, P., Glymour, C. and Scheines, R. [1993], *Causation, Prediction, and Search*, Vol. 81 of *Lecture Notes in Statistics*, 1 edn, Springer-Verlag New York, The address.

Thomas, J. J. and Cook, K. A. [2005], Illuminating the path: The research and development agenda for visual analytics.

Vieira, H., Sanchez, S., Kienitz, K. H. and Belderrain, M. C. N. [2011], 'Generating and improving orthogonal designs by using mixed integer programming', *European Journal of Operational Research* **215**(3), 629–638.