Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

INSIGHTS INTO THE HEALTH OF DEFENCE SIMULATED WORKFORCE SYSTEMS USING DATA FARMING AND ANALYTICS

Brendan Hill

School of Information Technology
Deakin University
Tower 2, 727 Collins St
Docklands VIC, 3008, AUSTRALIA

Damjan Vukcevic

School of Mathematics and Statistics
The University of Melbourne
Parkville VIC, 3010, AUSTRALIA

Terrence Caelli

School of Engineering
University of South Australia
101 Currie St
Adelaide SA, 5001, AUSTRALIA

Ana Novak

Joint and Operations Analysis Division
Defence Science and Technology Group
506 Lorimer St
Fishermans Bend VIC, 3207, AUSTRALIA

ABSTRACT

This work is motivated by the need for the Australian Defence Force to produce the right number of trained aircrew in the right place at the right time. It is critical to understand the impact of structural/resourcing policies on the ability to maintain squadron capability, both for individual squadrons and jointly across the Force. By combining techniques in experimental design, simulation, data analysis and optimization, we have created an automated system to efficiently identify significant relationships between simulation parameters and squadron capability, as well as propose optimal parameter ranges for each individual squadron. The interplay of competing demands between squadrons was then analysed in the context of the entire Force and the stability of these extrapolations over sampling noise was also evaluated. Finally, we present compact summaries and visualisations of the most important insights about the most significant contributors to the "health" of the whole system.

1 INTRODUCTION

A workforce supply strategy refers to the recruitment and training of new staff/students along with the scheduling of their training to deliver required qualifications within a specified time-frame. When these skill-sets are highly specialized, the cost and length of training is high and the infrastructure required is limited. The task therefore becomes quite complex.

Optimal recruitment and workforce supply strategies have been studied for a number of decades using various approaches including Linear Programming models (Azimi et al. 2013), Integer Linear Programming (ILP) (Akinyele 2007) and Markov Decision Processes (Udom 2013). Previous research by the authors utilized Stochastic search (Pike et al. 2018), ILP, and heuristic methods (Hill et al. 2018), modelling the objective as satisfying a risk tolerance while minimizing recruitment. The common challenge to all these solutions is the large number of variables and design points to consider.

In this work we consider the training of Australian Defence Force (ADF) pilots, observers and aircrew. This problem domain has a number of unique features, one of which being the requirement that some graduated students be posted as instructors, creating a feedback loop in the supply network. While external

instructors are sometimes available, they are costly, and their exclusive use has other drawbacks. In general the courses have limited places, student numbers are relatively small, and failure rates are high with much variability. These features necessitate the development of new ways of understanding how to optimize the training continuum.

Previous work by the authors focused on optimization strategies within a fixed configuration of the training program, providing insights and solutions into recruitment numbers and risk incurred by each squadron. However, the lack of variation in the resourcing parameters have meant that these insights are of limited value. There is a broader need to understand the effects of different resourcing policies on squadron risk by considering variations of such policies, and determine alternative resourcing parameters which may help to alleviate or mitigate risk.

In the field of Operations Research (OR), simulation has traditionally been perceived as a "method of last resort", and regarded as intrinsically inferior to analytic modelling methods (Lucas et al. 2015). However, simulation optimisation (SO), being a "black box" approach to simulation (that is, a paradigm involving only the observations of inputs and outputs), allows for arbitrarily complex models without the need to sacrifice model fidelity for the sake of analytic tractability. What makes this approach increasingly accessible is the rapid increase in cheap, fast and parallelizable computing resources.

Hence, in recent years, the integration of techniques from SO and the statistical design of experiments (DoE) has lead to the emergence of data farming (DF). In general, SO alone is lacking in techniques for experimenting with large numbers of factors with complex effects on response variables, addressed by invoking a statistically rigorous experimental design. Conversely, the traditional application of DoE has been to create design matrices for experiments which may be expensive to run in terms of time and cost. However, in simulation, it is typically straightforward to run a large number of simulations for arbitrary design points with sufficient parallelization. Data farming is then considered the application of designed experiments to simulation and the subsequent analysis of the individual and joint effects of factors on one or more response variables of interest within the simulation. A terminology mapping between the two fields is as follows, and hereafter the terminology from DoE is adopted for consistency with existing literature:

Simulation		**Design of Experiments**
Simulation Parameter	$\leftrightarrow$	Factor
Scenario/Vignette	$\leftrightarrow$	Design point
Set of scenarios/vignettes	$\leftrightarrow$	Design matrix / experiment
Simulation run	$\leftrightarrow$	Observation/sample
Set of simulation runs	$\leftrightarrow$	Resultant dataset
Simulation variable	$\leftrightarrow$	Response variable

Data farming has been successfully used in defence applications such as homeland security (Nannini and Wan 2011), NATO support (Horne and Seichter 2014), aircraft fleet management policies (Marlow, D.O. and Sanchez, S.M. and Sanchez, P.J. 2015), risk readiness (Kang et al. 2006) and other areas. In the pursuit of rigorous experimental designs for simulation, much work has been done by the SEED Center for Data Farming at the US Naval Postgraduate School in the development of near optimal design matrices (Kleijnen et al. 2005). When a large number of factors are involved, factorial designs are intractable. Factor screening techniques such as Sequential Bifurcation (Cheng 1997), and derivative methods (Bettonvil and Kleijnen 1997; Sanchez et al. 2005; Shen et al. 2010) can help identify significant factors for a single response variable, but lack natural extensions for dealing with multiple response variables. A Latin hypercube design can be applied to a large number of factors and can facilitate analysis of effects on multiple response variables. A desirable property of a design matrix is that the factors are orthogonal, and work has been done to produce and publish spreadsheets with nearly orthogonal Latin hypercube designs (NOLH) (Cioppa and Lucas 2007; Hernandez et al. 2012).

To our knowledge our research represents the first application of these methods to workforce planning and capability risk analysis. Unique challenges in this domain involve the need to consider factor effects on the short, medium and long term, and the need to balance the needs of multiple squadrons, which require recruits of different types from different paths with shared resources. We have therefore developed methods to identify individual and joint effects on response variables, and techniques to highlight potential conflicts in the joint solution. The application of these techniques is within an automated data analytics module intended to be presented to a non-technical audience. Therefore, we describe automated techniques to produce insights in textual and graphical form. Also, we favor statistical techniques which produce parsimonious models to highlight the most significant relationships and extract results with practical significance.

2 AIRCREW TRAINING PIPELINE SIMULATION

The ADF provides training for various aviation-related personnel including aircrew, engineers and technicians. Recruits enter the training program through agencies such as Australia Defence Force Academy (ADFA), the Royal Naval College (RNC) or via direct entry. Pilots undergo initial flight screening to assess aptitude, and successful applicants officially enrol in the pilot training stream. Modelled at a high level, each individual recruit passes through a sequence of schools and courses from their initial intake point, and if successful, they reach eventual deployment in a squadron.

The ADF training pipelines are susceptible to a great deal of variability in student intakes, changing policies and aircraft types. Within the pipeline, the pass rates for individual courses can be highly variable, and attrition rates within the squadrons can be unpredictable, and influenced by external factors such as industry recruitment. Taking into account the dependencies between the various stages of training and critical resourcing issues, planning over an extended horizon becomes quite a challenging process requiring a robust formulation capable of adapting to such constraints. Consequently, we model the ADF training pipeline as a directed acyclic graph comprised of intake points, courses C, schools K and squadrons Q. The arcs define the prerequisites and possible paths for individual recruit types R through the program (see Figure 1). We assume that an unlimited number of recruits are available at the intake points at any time. Each course $c \in C$ is repeated on a fixed set of sessions E_c per year, and each school $k \in K$ requires certain instructor types allocated to specific recruit types r, $I_{k,r} \subseteq I$, the set of instructor types. Each squadron $q \in Q$ has a set of recruit types $r \in R_q \subseteq R$.

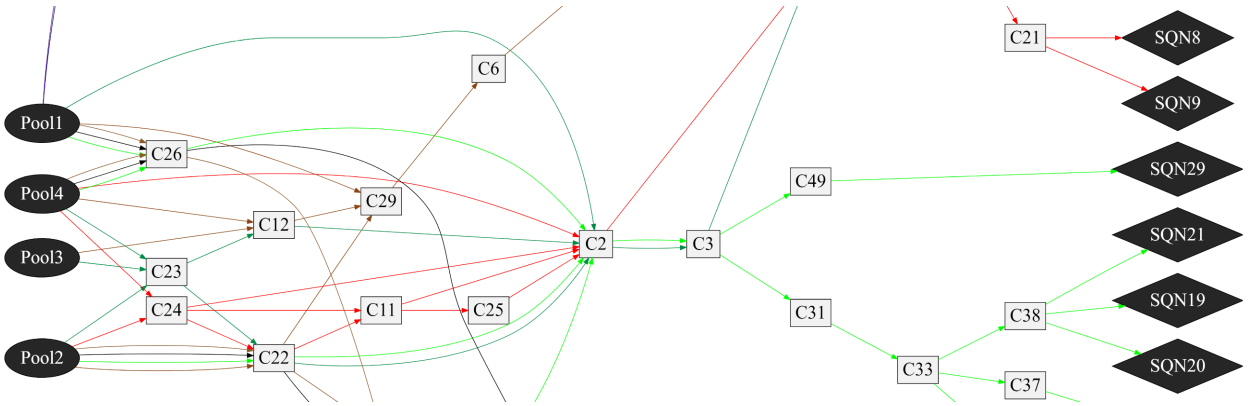

Figure 1: A subset of the full training pipeline with intake points (Pool*), courses (C*) and squadrons (SQN*). Arrows indicate possible training progressions with a different color for each trainee type.

Each recruit type $r \in R_q$ in squadron $q \in Q$ has a predefined target number $\mathbf{O}_{q,r}$. Staying at or above this target defines operational capability. These targets are fixed, and not varied in our analysis. Resource policies are represented in the model as resource constraints and targets. In particular, each course session $e \in E_c$ has a maximum capacity $\mathbf{M}_e$, as well as individual capacities per recruit type $\mathbf{M}_{e,r}$. In addition, each

school k maintains a target number of instructors of each instructor type $i \in I_k$ for each recruit type $r \in R_k$ denoted $\mathbf{I}_{k,i,r}$, which the simulation attempts to satisfy by drawing sufficiently experienced recruits from squadrons. These instructors are removed from the squadrons and placed in schools for extended periods, so although increasing instructor targets may facilitate a large number of recruits in the long term, it may affect squadron capability in the short term due to the workforce shortfall.

There are two stochastic elements in the simulation. Firstly, the random pass rates for each course/session are sampled from Beta-Binomial distributions. This distribution allows us to parameterize the mean and variance, which was done in alignment with historical estimates. Secondly, the random attrition rates are sampled from Poisson distributions. This allows us to model random attrition at each time-step in accordance with a parameterized yearly average, which was similarly aligned with historical estimates.

The model was implemented in an agent based discrete event simulation (DES) where resourcing policies were parameterized. The DES followed a heuristic optimization strategy which computes expected demand on each node of the graph as projected over time, with the objective of minimizing recruitment while satisfying a capability risk target for each squadron. This strategy determined the main DES control actions including initial recruitment numbers at each time step and movements of recruits through the pipeline. The details of this strategy are provided in (Hill et al. 2018) and are outside the scope of this paper. Instead, this strategy was simply taken to be an element of the dynamics of the simulation. Importantly, within a given simulation run, the resourcing policies were not varied or used as control actions. For the purposes of this analysis we simulated with a 15 year horizon with monthly time-steps for a total of 180 discrete time-steps, denoted $T = \{1, \ldots, 180\}$.

3 EXPERIMENTAL FRAMEWORK AND DESIGN

The challenge in DoE is to develop efficient methods for selecting design points such that the insights resulting from the subsequent data analysis can be maximized. Our aim is to identify the most significant practical relationships between policy parameters and squadron capability, as well as form a suggestion for alternative parameter values which improve results over the horizon. We have approached this from the perspective of statistical experimental design by varying the parameters in a principled manner, then analyzing their collective effects on squadron capability, which forms the set of response variables.

3.1 Experimental Factors and Design Matrix

As an exploratory tool, we do not require on the prior definition of particular factors and ranges. Instead, we take all resourcing parameters related to course capacities and instructor targets as the set of design factors resulting in a total of 164 integer valued factors, denoted by:

$$F = \{F_1, \ldots, F_{164}\} \equiv \{\mathbf{M}_e, \ldots, \mathbf{M}_{e,r}, \ldots, \mathbf{I}_{k,i,r}, \ldots\}$$

For each factor $f \in F$ in the design matrix, it is necessary to define a plausible range of variation. In this model, student numbers and course capacities are typically in the range of 0—50. While in general these ranges can be specified manually, for exploratory purposes we have defined ranges of variation automatically relative to the baseline value in simulation. The lower bound LB_f is simply the default value in configuration. Then, the upper bound is double this value, or this value increased by at least some minimum constant δ:

$$UB_f = \max\{2 \cdot LB_f, LB_f + \delta\}$$

This upper bound is determined on the assumption that reasonable increases to resourcing parameters will be up to double the current value, thereby ensuring that a reasonable range is explored in the experimental design relative to the default value. However in practice these ranges could be specified manually based on domain knowledge. A value of $\delta = 5$ was used in our work so that larger ranges were explored for parameters with small default values, i.e. where $LB_f \leq 4$. The primary purpose of this work is to identify

which resource constraints need to be relaxed, so we only consider upwards variation of these factors. However, with the alternative goal of identifying which resource targets could be reduced without adversely effecting the operationality of the ADF, downwards variation could be accommodated as well.

It is desirable, in general, for the design factors of the design matrix to be uncorrelated (Hernandez et al. 2012) in the system's output so that inadvertent correlation between factors do not unduly bias the results. In space filling terminology, if two design factors are highly correlated, then the off-diagonal "corners" of their parameter subspace are underrepresented by the design points, so results around those regions may be less accurately reflected in the output dataset. For example, high correlation in values of two course capacity parameters across the design points may stifle the discovery that only one of them needs relaxing in order to resolve a training bottleneck.

Consequently we adapted a design matrix from the "NOB mixed design worksheet" as published on the SEED Data Farming website (SEED 2019), which provides a NOLH with 512 design points D and up to 300 factors of varying ranges. Although this worksheet is presented for manual construction of design matrices, we automated this process by selecting columns corresponding to the ranges defined for each design factor and shifting the values accordingly. Hence, the end users of this exploratory tool need not be exposed to the technical details of the design matrix. The value of the factor f in the d'th design point is denoted $X_f^{(d)}$ such that:

$$LB_f \leq X_f^{(d)} \leq UB_f \qquad d \in D, f \in F$$

The choice of replication number N per design point is important for ensuring the stability of the results. For this analysis we consider an experimental design with $N = 10$ replications of each design point, and the stability is evaluated and discussed in section 4.3.

3.2 Definition of Response Variables

The simulation environment models the number of people at each stage of the training program at each time-step, including recruits in training, waiting, in squadrons and in a larger sustainment pool. After running a number of simulations, the resultant dataset forms a multi-sample, multivariate time series of non-negative integer valued variables. The ultimate goal is to maintain capability of each squadron over the horizon and as such it is necessary to introduce additional variables which reflect the state of being at or above capability. Therefore, in post-processing we computed binary variables $C_{q,r,t}^{(d,i)}$ for each squadron q, recruit type r, and time step t, indicating whether the squadron was at or above operational target in the i'th replication of design point d:

$$C_{q,r,t}^{(d,i)} = \begin{cases} 1 & \text{if } Q_{q,r,t}^{(d,i)} \geq \mathbf{O}_{q,r} \\ 0 & \text{otherwise} \end{cases} \qquad d \in D, 1 \leq i \leq R, q \in Q, r \in R_q, t \in T$$

where $Q_{q,r,t}^{(d,i)}$ is the corresponding number of recruits observed for the given indices.

The simulation is populated with a realistic initial state representing current enrolments, numbers in squadrons and instructor levels. This initial state is shared among all simulation runs regardless of the factor values defined in each design point. Early analysis of the simulation suggested that the system has an initial "correction" period with respect to its initial state and takes a number of years initially to reach some steady state. Initial exploratory work confirmed that some factors have different effects on some responses in different periods. Therefore it is informative to break up the 15 year horizon into three periods: years 1-5, years 6-10, and years 11-15. So additional variables for each squadron q, recruit type r and period $p \in \{0,1,2\}$ were computed by averaging the binary capability variables within each period across all replications of each design point d:

$$Y^{(d)}_{q,r,p} = \frac{1}{N}\frac{1}{\tau}\sum_{i=1}^{R}\sum_{t=\tau p+1}^{\tau(p+1)} C^{(d,i)}_{q,r,t},$$

where $\tau = 60 = 5 \times 12$ months is the number of discrete timesteps per period.

These aggregate "health" variables then represent the expected proportion of time the squadron is at or above capability during each period, which are taken to be the response variables. The analysis task is then defined as identifying relationships between the set of factors $X_{...}$ and the set of responses $Y_{...}$, across the design points $d \in D$:

$$\left\{X^{(d)}_i : 1 \leq i \leq 164\right\} \sim \left\{Y^{(d)}_{q,r,p} : q \in Q, r \in R_q, p \in \{0,1,2\}\right\}$$

4 DATA ANALYSIS AND VISUALIZATION

Having established the experimental design and performed the required simulations, the next step is to extract the most relevant insights from the dataset and present them efficiently in an interpretable manner. The trade-off between the granularity of this information and the stability of the information over sampling error must be considered and carefully controlled to establish confidence in the results. We consider the analysis, the visualization and the evaluation of stability as follows.

4.1 Data Analysis

4.1.1 Discovering Pairwise Monotonic Relationships

A reasonable expectation in this exploratory analysis is that significant relationships between factors and responses will be monotonic. For example, squadrons which are frequently below capability due to a training bottleneck earlier in the pipeline will typically resolve once that resource constraint is relaxed above a certain level. Conversely, squadrons which are typically maintaining capability may be adversely affected if instructor targets are raised too high, since more recruits will be withdrawn from the squadrons to meet those targets. To identify such relationships, we adopted Spearman's rank correlation coefficient (Pirie 1988), which computes the Pearson correlation coefficient of the ranks of the observations. This was computed pairwise between all factors and responses yielding a matrix of point estimates of the degree of monotonicity present in each factor/response relationship. This can be interpreted as a quantification of factor significance with regard to statistical significance, and visualizations are explored in section 4.2.1.

4.1.2 Discovering Relationships of Practical Significance with Regression

In addition to the pairwise monotonicity measure, it is possible to identify relationships which are of practical significance with the fitting of a multiple regression model. We modelled each response variable separately using multiple regression, in each case including all factors as the predictor variables. Then, the magnitude of model coefficients can be interpreted as quantifications of factor significance with respect to each response. (In the fitting of these regression models, we note a natural advantage of the NOLH approach which guarantees the absence of collinearity among the predictors.)

To compare the relative significance of different factors, as well as apply a common significance threshold, the resulting set of model coefficients must be comparable. Hence it is essential to scale the factors to a common measure. Our experimental design involved sampling factor values between their default baseline values and some upper bound. Hence, for regression purposes, we transformed the predictor variables into the common measure of amount incremented above baseline value (e.g. see Figure 6 horizontal axis).

The selection of an appropriate regression model is key to the success of this method. The response variables are proportions, bounded on the unit interval, and expected to have monotonic relationships with certain factors. In particular, as resourcing parameters reach critical values, we may expect bottlenecks in

the pipeline to be cleared and proportion at capability within a period to increase significantly (an example is given in Figure 2). Therefore we adopted a Fractional Logistic Regression Model (FRM) (Ramalho et al. 2011), the adaptation of the Logistic Regression model for fractional responses.

With the application of L_1 regularisation, the factors with small effect are omitted, leaving a smaller number of factors that have the most impact on the response variable. Once the model is estimated for a given response variable Y_i, the estimated coefficients $\hat{\beta}_{Y_i}$ are interpreted as quantifications of factor significance with respect to response Y_i, yielding both magnitude of significance, and direction of effect (positive/negative). Visualizations of these results are developed in section 4.2.1.

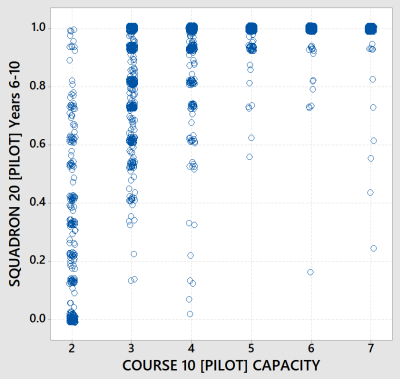

Figure 2: Example of response variable vs. factor with sigmoidal relationship.

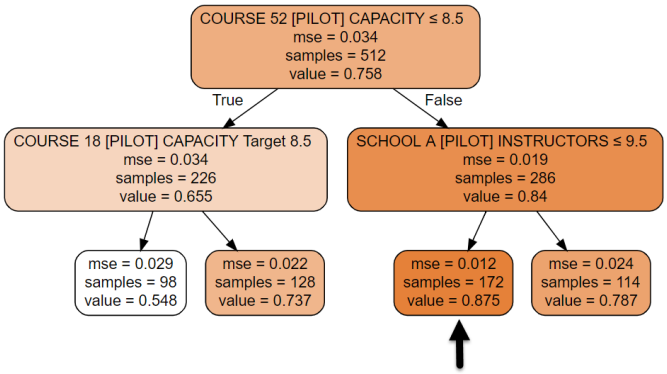

Figure 3: Example of extraction of optimal factor ranges from best leaf in decision tree.

4.1.3 Optimizing Factor Ranges for Individual Responses with Decision Trees

The analysis so far identifies significant relationships between factors and responses. In addition to this, we wish to identify critical ranges for factors with respect to each response. To achieve this we introduces an application of Decision Trees (DT) which efficiently identifies such ranges. The process is simply to grow the decision tree using all factors (with some complexity constraints), identify the "best" leaf, then extract the factors and ranges on the path from that leaf to the node as the best identified solution for a particular response variable, e.g. see Figure 3.

The most common application of decision trees is to produce predictive models, typically in the context of an ensemble method such as random forests. In that context, reducing the complexity of the tree is important only insofar as it improves the generalizability of the model, so typically highly complex trees are produced.

For parsimony and stability, therefore, it is essential to reduce the complexity of the tree such that only significant factors and critical values are identified in the best branch. Common methods of constraining tree complexity are to set a minimum numbers per split, set a maximum depth, or a minimum impurity decrease per split. The choice of parameters and values must be determined manually by inspection of the resulting trees, depending on the context and type of response values. For overly complex trees, the best leaf will be highly variable and include factor ranges based on sampling noise, whereas overly simplistic trees will miss important relationships, so tuning these complexity constraint parameters is critical, and dependent on the dataset. We found that the minimum impurity decrease parameter had a strong effect on the tree complexity, and a maximum depth was unnecessary. Performing a sweep on the minimum impurity decrease parameter on the range [0, 0.1] found a sharp decrease in node count at around 0.001, and this value was selected and fixed for subsequent analysis.

Concise and interpretable visualizations of the findings extracted from decision trees for responses both individually and globally are presented in sections 4.2.2, 4.2.3.

4.2 Visualization and Presentation of Insights

Key to the success of this method is the ability to present the most significant insights compactly and comprehensibly to a non-technical audience. We introduce two visualizations of factor/response relationships, a textual summary of the results for each response, and a visualization of the optimal factor ranges as discovered across all responses.

4.2.1 Direct Relationships as Heat-maps and Bipartite Graphs

Both the results of Spearman rank correlation, and regression model coefficients can be taken as a matrix with factors as rows and responses as columns. An efficient way to identify relationships of statistical or practical significance then is as as a heat-map with the neutral color centered at value 0, as shown in Figure 4. In this view, significant relationships are clearly highlighted along with their direction and strength.

Equivalently, the same information can be rendered as a bipartite graph, which depicts at a glance the complexity of interdependencies between the factors and responses. The strength of relationships determines the width of each edge, and the sign determines the color. An appropriate significant threshold value must be selected to determine which arcs deserve inclusion in the graph.

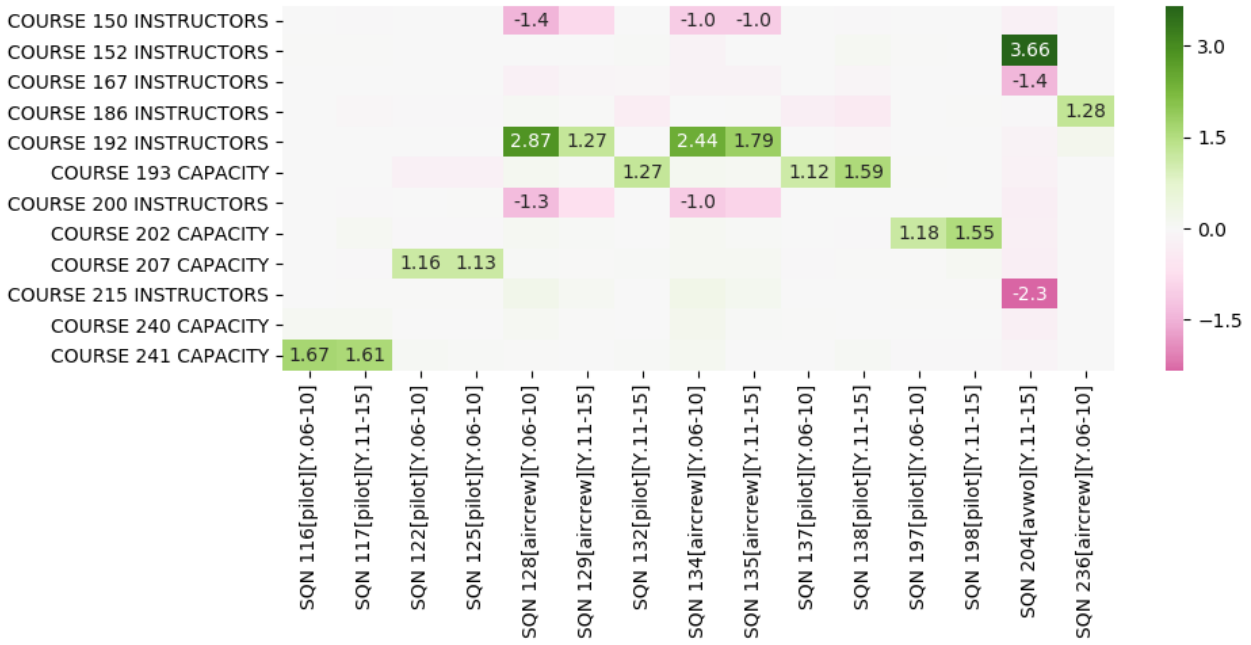

Figure 4: Heat-map of FRM coefficients between factors and responses.

4.2.2 Textual Summary of Insights

For efficient summarization of a single response variable, a textual description of the findings is useful. We developed the following transformation of the numeric findings into a natural language equivalent.

As with the bipartite graph, significant factors can be identified by thresholding the significant values (either Spearman correlation or FRM regression coefficients). The best leaf solution from the decision tree yields a set of factors and ranges. A baseline value for each the response variable is estimated by running 100 simulations of the base scenario with no modifications to any parameters. These results are summarized in text form given in Figure 5.

```
"SQUADRON 143 [Y.06-10]" is significantly affected by "COURSE 198 CAPACITY" and
"COURSE 231 INSTRUCTORS". The following factor ranges are suggested to bring
the average proportion at capability from 47% to 87% during this period:
  - "COURSE 231 INSTRUCTORS" at least 9
  - "COURSE 198 CAPACITY" at least 8

"SQUADRON 122 [Y.11-15]" is not significantly affected by any factors.
No changes to factor ranges are suggested.

"SQUADRON 180 [Y.06-10]" is significantly affected by "COURSE 131 CAPACITY".
The following factor ranges are suggested to bring the average proportion at capability
from 51% to 90% during this period:
  - "COURSE 150 INSTRUCTORS" between 15 and 16
  - "COURSE 131 CAPACITY" at least 4
```

Figure 5: Text based summary of insights.

4.2.3 Global Requirements and Potential Conflicts

Across the inferred optimal factor ranges for each squadron, there is the possibility that conflicting constraints for a given factor will emerge. These reflect competing demands on resourcing requirements—for example, the optimal range for one squadron may define a lower bound for the number of instructors at given a school, but since this draws trained members from other squadrons, their optimal ranges may define upper bounds. If this set of ranges has no mutually overlapping region, then there is a conflicting set of demands.

It is therefore informative to view all constraints placed on the factors across all response variables simultaneously. For this purpose, we introduce the heat-map based visualization shown in Figure 6. The vertical axis defines the bounded factors, and the horizontal axis represent shifts from the default value in configuration. Each arrow represents a bound placed by a particular response variable. For each factor, for each incremental value, we compute the proportion of constraints at each value rendered from red (0.0) to green (1.0). Hence green means all constraints are satisfied by the given value, while red means no constraints are satisfied, whereas orange indicates some intermediate proportion of constraints are satisfied. For example, if one squadron places +2 lower bound on a factor and another squadron places +5 upper bound on the same factor, then 100% of the constraints are satisfied within the range +2 to +5, and 50% are satisfied at all other values.

This depicts at a glance the amount by which each factor must be increased to satisfy all bounds specified by the individual responses. Furthermore, where no green region exists for a given factor, there are conflicting constraints. This is an important discovery, potentially indicating a systemic problem in the structure of the pipeline deserving special attention.

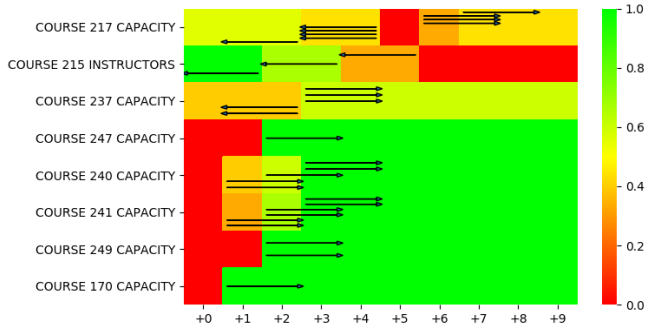

Figure 6: Optimal ranges for influential factors. Color indicates proportion of constraints satisfied for each increment.

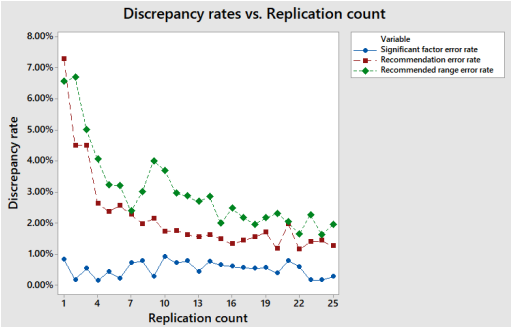

Figure 7: Average discrepancy rates as a function of replication count.

4.3 Evaluation of Stability of Insights over Sampling Error

The extraction of this discrete set of discoveries from the dataset is sensitive to sampling error, choice of significance thresholds and tuning of decision trees. As these discoveries are presented as final results it is important to understand their stability over different samples and select an appropriate number of replications of each design point. To evaluate this we performed a sweep on replication number and present discrepancy estimates for different elements of the results.

4.3.1 Stability Evaluation Method and Results

The stability evaluation was performed by producing a dataset with 100 replications per design point, and considering replication counts N ranging from 1 to 25. For each N, this full set of 100 replications was randomly partitioned into subsets of N replications each, and 4 of these partitions were selected at random. The analysis was repeated in full for each replication subset and the insights were extracted independently. The insights yielded by each independent partition were then compared. We consider the stability of three different classes of information: 1) the identification of "significant" factors per response; 2) the choice of factors to form recommendations on per response; and 3) the specification of lower and upper bounds for the recommended factor ranges. For each response variable and each information element, the discrepancy rate is taken to be the proportion of partitions in which that element of information differed from the mode. The average discrepancy rate across all responses, for each discrepancy type, is shown in Figure 7.

The discrepancy rate in the identification of significant factors is low ($\leq 1\%$) and does appear to improve with greater replication counts. The information extracted from the decision trees becomes significantly more stable as the replication count increases, with discrepancy rates reducing from $\sim 8\%$ to around $\sim 2\%$ by around 15+ replications, demonstrating that stability can be improved with sufficient replications. Further tuning of the complexity of the decision tree, more replications or alternative techniques may be required to reduce the discrepancy rates further.

4.3.2 Discussion

The usage of decision trees to extract optimal factor ranges is efficient and the results are highly interpretable, but achieving stability of these results is difficult, requiring a large number of replications, or a significant reduction in tree complexity resulting in a loss of detail. Furthermore, manual inspection of the decision trees reveals that the "best leaf" solution very often has an average value similar to some of the next best leaf solutions. These alternative solutions indicate alternative ways to achieve desirable response values and may be of comparable interest or value to the analyst; in particular their incorporation may assist in the construction of a consistent global solution. Further work on the incorporation of decision trees is required to control their stability without sacrificing the usefulness of the information.

We note that none of the optimal factor values identified were at the upper bounds of the parameter space explored, hence the choice of factor upper bounds described in Section 3.1 did not result in any binding constraints. However since many optimal factor values equalled the default values, it is possible that the lower bounds defined are binding constraints. Downward variation of the factor values as described earlier is necessary to determine whether this is the case or not.

5 SUMMARY

We have presented a design for a fully automated exploratory data analysis module for a simulation of an aircrew training system, translating the raw simulation output into interpretable, non-technical insights regarding the influences of each factor on the multiple responses of concern. These outputs range from highlighting significant pairwise relations and suggesting alterations to parameters, to a global view of the combined constraints and conflicts in the requirements of different squadrons on the resourcing parameters

of the system. As an exploratory tool, this has the potential to rapidly speed up diagnostic analysis of the aircrew training program, identify training bottlenecks and propose candidate solutions.

Although this project has focused on aircrew training, the techniques can in theory be applied to any simulated environment, with appropriate selection of models and significance thresholds. The data farming paradigm in general permits claims of causality to be made since the definition of each design point represents an intervention at $t = 0$. Therefore the insights derived are powerful and extend beyond mere correlation. However, an important qualification is that the acuity of these insights is limited by the accuracy and fidelity of the underlying simulation.

The pipeline presented here performs a single iteration of the experimental design $\rightarrow$ simulate $\rightarrow$ analyse $\rightarrow$ insight process. A future research direction will be the development of an iterative control method using insights from previous iterations to determine the experimental design for the next (a.k.a. adaptive design, or active learning). In this manner we may be able to converge on globally optimal sets of parameters ranges. Regarding the information content in the insights, there is a significant trade-off between the level of detail and the stability of the insights. While both can be improved by increasing the replication count, the trade-off can be controlled by the selection of significance thresholds and decision tree complexity parameters.

In conclusion, the framework presented is a collection of practical, scalable techniques for exploring the dynamics and reactions of a simulated environment with a large number of parameters and response variables of interest, automating the process of generating insights for a non-technical audience.

REFERENCES

Akinyele, S. T. 2007. "Determination of the Optimal Manpower Size Using Linear Programming Model". *Research Journal of Business Management* 1(1):30–36.

Azimi, M., R. R. Beheshti, M. Imanzadeh, and Z. Nazari. 2013. "Optimal Allocation of Human Resources by Using Linear Programming in the Beverage Company". *Universal Journal of Management and Social Sciences* 3(5):48–54.

Bettonvil, B., and J. P. Kleijnen. 1997. "Searching for important factors in simulation models with many factors: Sequential bifurcation". *European Journal of Operational Research* 96(1):180–194.

Cheng, R. C. 1997. "Searching for important factors: Sequential bifurcation under uncertainty". In *Proceedings of the 1997 Winter Simulation Conference*, edited by S. Andradttir, K. J. Healy, D. H. Withers, and B. L. Nelson, 275–280. Atlanta, Georgia: Institute of Electrical and Electronics Engineers, Inc.

Cioppa, T. M., and T. W. Lucas. 2007. "Efficient nearly orthogonal and space-filling Latin hypercubes". *Technometrics* 49(1):45–55.

Hernandez, A. S., T. W. Lucas, and M. Carlyle. 2012. "Constructing nearly orthogonal Latin hypercubes for any nonsaturated run-variable combination". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 22(4):20.

Hernandez, A. S., T. W. Lucas, and P. J. Sanchez. 2012. "Selecting random Latin hypercube dimensions and designs through estimation of maximum absolute pairwise correlation". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher, 1–12. Berlin, Germany: Institute of Electrical and Electronics Engineers, Inc.

Hill, B., D. Kirszenblat, B. Moran, and A. Novak. 2018. "Optimizing recruitment to achieve operational capability conditional on appetite for risk". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3801–3812. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Horne, G., and S. Seichter. 2014. "Data Farming in support of NATO operations - methodology and proof-of-concept". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 2355–2363. Savanah, Georgia: Institute of Electrical and Electronics Engineers, Inc.

Kang, K., K. H. Doerr, and S. M. Sanchez. 2006. "A design of experiments approach to readiness risk analysis". In *Proceedings of the 2006 Winter Simulation Conference*, edited by L. F. Perrone, F. P. Wieland, J. Liu, B. G. Lawson, D. M. Nicol, and R. M. Fujimoto, 1332–1339. Monterey, California: Institute of Electrical and Electronics Engineers, Inc.

Kleijnen, J. P., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa. 2005. "State-of-the-art review: a users guide to the brave new world of designing simulation experiments". *INFORMS Journal on Computing* 17(3):263–289.

Lucas, T. W., W. D. Kelton, P. J. Sanchez, S. M. Sanchez, and B. L. Anderson. 2015. "Changing the paradigm: Simulation, now a method of first resort". *Naval Research Logistics (NRL)* 62(4):293–303.

Marlow, D.O. and Sanchez, S.M. and Sanchez, P.J. 2015. "Testing Aircraft Fleet Management Policies Using Designed Simulation Experiments".

Nannini, C. J., and H. Wan. 2011. "Designs for large-scale simulation experiments, with applications to defense and homeland security". *Design and Analysis of Experiments, Volume 3: Special Designs and Applications* 810:413.

Pike, C., B. Moran, D. Kirszenblat, B. Hill, and A. Novak. 2018. "A Stochastic Programming Approach to Optimal Recruitment in Australian Naval Aviation Training". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3753–3764. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Pirie, W. 1988. "Spearman rank correlation coefficient". *Encyclopedia of statistical sciences*.

Ramalho, E. A., J. J. Ramalho, and J. M. Murteira. 2011. "Alternative estimating and testing empirical strategies for fractional regression models". *Journal of Economic Surveys* 25(1):19–68.

Sanchez, S. M., H. Wan, and T. W. Lucas. 2005. "A two-phase screening procedure for simulation experiments". In *Proceedings of the 2005 conference on Winter simulation*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 223–230. Orlando, Florida: Institute of Electrical and Electronics Engineers, Inc.

SEED 2019. "SEED Center for Data Farming, Software Downloads". https://my.nps.edu/web/seed/software-downloads, accessed 21st July.

Shen, H., H. Wan, and S. M. Sanchez. 2010. "A hybrid method for simulation factor screening". *Naval Research Logistics (NRL)* 57(1):45–57.

Udom, A. U. 2013. "A Markov Decision Process Approach to Optimal Control of a Multi-level Hierarchical Manpower System". *CBN Journal of Applied Statistics* 4(2):31–50.

AUTHOR BIOGRAPHIES

BRENDAN HILL is a Researcher at Deakin University specialising in Operations Research. Brendan holds a B.Sc (Mathematics and Statistics) from the University of Melbourne. He has accumulated 19 years experience in commercial software development across many industries, in roles ranging from technical lead to management. Research interests include optimization, simulation, machine learning and artificial intelligence. His website is http://brendanhill.com.au and his email address is brendan.hill@gmail.com.

DAMJAN VUKCEVIC is a statistical data scientist, specialising in statistical genetics. He completed a Bachelor of Science (Honours) degree at the University of Melbourne in 2004, including an Honours project in bioinformatics with Professor Terry Speed. He was then awarded a Commonwealth Scholarship to study at the University of Oxford, where he completed a DPhil in statistical genetics with Professor Peter Donnelly. Dr Vukcevic has contributed to a number of important genetic studies, including the landmark multi-disease genome-wide association study by the Wellcome Trust Case Control Consortium, published in *Nature* in 2007. This received a number of awards, including Research Leader of the Year from *Scientific American*. Dr Vukcevic has experience working in both an academic and non-academic environment, with a strong grounding in statistical theory, computation and practical data analysis. His website is http://damjan.vukcevic.net/ and his email address is damjan.vukcevic@unimelb.edu.au.

TERRENCE CAELLI is a Research Professor in Division of Engineering at the University of South Australia funded by the Defence Science and Technology Group at Port Melbourne, Victoria. Previous to this he has held a number of senior positions with National ICT Australias (NICTA) including Laboratory Director and Director of NICTA Health Program. His interests lie in Signal Processing, Human and Machine Perception, Cognitive Engineering, Machine Learning and their applications in Health, Environment and Defence. He is a Fellow of the International Association for Pattern Recognition (FIAPR) and a Fellow of the Institute for Electronic and Electrical Engineers (FIEEE). He is also a Convocation Medalist from the University of Newcastle. He has spent 15 years in North American universities and research institutes (Bell Laboratories and NASA Commercial Space Centre), has been a DFG Professor, Germany, Killam Professor of Science, the University of Alberta, Canada. He has served on the editorial boards of many international journals including IEEE: PAMI, Pattern Recognition and numerous international conference committees. His email address is terry.caelli@gmail.com.

ANA NOVAK joined DST Group in 2010 as an Operations Research Scientist after four years as a specialist consultant in a boutique demand, inventory and supply chain optimization firm. After immigrating to Australia in 2000, she received her B.Eng. (Honours) in Information Technology and Telecommunications from University of Adelaide in 2003, followed by a Ph.D. degree in Mathematics from The University of Melbourne in 2006 in Queueing Theory. Her research interests include planning and scheduling systems, simulation and modeling, optimization and modeling uncertainty, and her email address is Ana.Novak@dst.defence.gov.au.