Statistical validation of a new Python-based military workforce simulation model

Stephen Okazawa, Patricia Moorhead, Abe Jesion, Stan Isbrandt
Defence Research and Development Canada, Ottawa, Canada
stephen.okazawa@drdc-rddc.gc.ca
patricia.moorhead@drdc-rddc.gc.ca
abe.jesion@forces.gc.ca
stan.isbrandt@drdc-rddc.gc.ca

Abstract: The Canadian Department of National Defence uses military workforce modelling and simulation to inform the decision-making process with regard to the management of military personnel. Defence Research and Development Canada (DRDC) has a suite of workforce models that have been used in a variety of studies over many years to answer questions, generate forecasts and analyse various scenarios under consideration. Over time, these simulation models have been refined, errors corrected, and results validated against historical data. Consequently, the level of confidence in these models is very high. As computer technology advances, the need arises periodically to update these models to take advantage of newer technology and to provide more advanced capabilities. DRDC is updating its workforce modelling and simulation technology through the development of a Python-based discrete event simulation environment that is intended to replace various commercial simulation software products in which existing DRDC workforce models have been built. DRDC has begun the process of rebuilding select workforce models in this new environment. Rebuilding a simulation model using new technology usually results in some loss of confidence in the model even if the features of the new technology are impressive. This is due to the possibility of reintroducing errors, inexperience with the new technology, and a lack of validation against hard data. To establish confidence in the new model implementation, it is necessary to validate its equivalence to the older trusted version. This can be done by subjecting both implementations to identical input scenarios and comparing simulation output. However, because workforce models frequently make use of random effects modelled using probability distributions (e.g. age at recruitment, release events, and course failures), variability in simulation output from the same input is expected. Therefore, the challenge is to determine whether this variability is A) due to the known random effects that are part of the simulation model, or B) an indication that the behaviour of the new model implementation is not equivalent to the old. In this paper, we demonstrate the use of non-parametric and time series statistical techniques to test the hypothesis that the results obtained from Arena-based and Python-based implementations of the same military rank structure model are equivalent. Important measures of system performance that were tested include population size, time in rank, promotions, releases and course qualifications. This methodology will be of broad use to analytics practitioners who face the challenge of testing whether two independently-developed simulation models of the same system are in fact statistically equivalent.

Keywords: workforce model; simulation; Python; non-parametric statistics; time series; statistical validation

1. Introduction

1.1 Background

The management of military human resources (HR) in the Canadian Department of National Defence (DND) is a complex area in which decision-making affecting the careers of military personnel often produces unforeseen and long-lasting consequences. Without careful analysis, a decision to change personnel force structure, training programs, or operational commitments can unknowingly produce shortages and/or excesses of personnel in various occupations at various times and ranks and qualification levels. Ultimately, these unintended consequences may result in a reduction of or interruption in operational readiness.

In order to support the decision-making process in this domain, Defence Research and Development Canada (DRDC) develops models of military HR systems and conducts simulations in order to generate forecasts of the system effects of potential courses of action. Examples of HR-related information produced by this work include forecasts of: training demand, training output and occupation health (Corbett 2013; Latchman and Hunter 2002; Straver, Okazawa, and Wind 2009; Séguin 2011); manning levels at all ranks in various occupations (Zegers and Isbrandt 2010); and readiness for deployed operations (Scales, Okazawa, and Ormrod 2011; Moorhead, Wind and Halbrohr 2008).

Military HR planning occurs on an ongoing basis and is very dynamic due to changing policies, priorities, budgets, and political and economic events. Operating in this context, military HR models are often used repeatedly and modified frequently over a long lifespan. Over this lifespan, simulation objectives typically evolve to become more ambitious, and computer hardware and software capabilities consistently advance at a rapid pace. Therefore, it is inevitable that older model implementations must eventually undergo significant upgrades to exploit technological developments in order to continue to meet the evolving requirements of simulation exercises.

DRDC is currently in the process of updating its workforce modelling and simulation technology through the in-house development of a novel discrete event simulation (DES) environment called the Right Person, Right Qualifications, Right Place, Right Time, Human Resources (R4 HR) simulation environment. This Python-based application is intended to replace various commercial simulation software products in which existing DRDC workforce models have been built. Several such models were originally built in the Arena DES software. While commercial applications like Arena are powerful, they generally cater to a wide audience in industry and are therefore not specialized for the particular needs of military HR simulation. Some of these specialized needs are the ability to easily code custom model logic, the ability to use relational databases and database querying during simulation execution, and the ability to interconnect separately developed models in order to conduct integrated simulation exercises (Okazawa 2013). These are the gaps that are addressed by the R4 HR simulation software.

DRDC has begun the process of rebuilding select workforce models in this new environment. The models chosen to be upgraded have an established record of informing military decisions over many years. The new versions of these models benefit from the specialized features of the R4 HR platform which facilitate the modelling process and provide more room for improvement as simulation requirements evolve. However, before the new versions can be used to make recommendations in real decision-making scenarios, they must undergo a validation process. The validation must demonstrate empirically that, subject to identical input scenarios, the new model implementation produces equivalent results to the older version in which confidence has already been established.

1.2 Aim

The problem of demonstrating the equivalence of two different implementations of the same stochastic simulation model is the subject of this paper. As an example, we use a DRDC military rank structure model that was previously implemented in Arena and saw extensive use over many years, informing real-world military decision making in DND. This simulation model has now been re-implemented in the R4 HR software. The two model implementations, while logically equivalent, are not direct “line-by-line” translations of each other: they use different data structures, different programming techniques, and different random number generators for simulating stochastic processes. As a result of the latter, variation in simulation output is expected. However, it is necessary to determine whether these variations are just different random outcomes of the same underlying stochastic processes, or an indication of non-equivalence of the two model implementations. In order to test the hypothesis of equivalence, non-

parametric and time series statistical techniques are used. Failure to reject this hypothesis provides validation of the equivalence of the two model implementations.

The paper is organized as follows. First, we briefly introduce the R4 HR simulation environment, identifying the features it provides that are suited to conducting military HR simulations, and indicating how the underlying concepts fundamentally differ from those of conventional DES applications. We then describe the military rank structure model, and highlight the differences in implementation between the Arena and R4 HR versions of the model that result from the different DES approaches. The methodology for statistically testing equivalence of the model implementations is then presented. Finally we discuss the test results which indicate, to a high degree of confidence, that the outputs of the two model implementations are statistically equivalent.

2. Python-based simulation environment

The military HR models developed by DRDC to inform DND's decision-making processes have typically been built in general purpose simulation applications, particularly Arena. While powerful, certain aspects of these applications were found to be not well-suited to the unique modelling requirements of military HR problems. The objective of the R4 HR software is to address these issues in order to make it easier to build military HR models and to reuse and modify them over their lifespan. Specifically, the main goals of R4 HR software are to provide:

- the flexibility of a programming language for defining model logic but without the steep learning curve normally associated with writing custom model code;
- the scalability of a relational database for storing simulation data and the ability to query the data using structured query language (SQL) during simulation execution; and
- a modular architecture such that model logic is easily reusable and interchangeable and allows individual models to be interconnected to create larger, integrated models.

As of the date of this publication, DRDC has developed a functioning prototype of the R4 HR simulation environment, and is investing in the development of a commercial grade version for use within DRDC and distribution to sister organizations in allied nations. This new software provides a number of unique features designed to meet the goals outlined above.

In R4 HR, simulation events are defined by writing compact routines in Python. These routines, called "code parts", are linked to each other and to other data elements to build the model. R4 HR does not provide the standard set of DES building blocks such as assign and decide blocks; instead, it allows the user to write equivalent Python code. The rationale for this is that, in practice, the user-defined instructions that must be entered into conventional DES building blocks quickly become more complicated and difficult to manage than plain code. Thus, we opt for implementing model logic in code from the start. This requires only a basic level of Python programming knowledge for simple models, but also allows the logic to seamlessly scale to increasingly complex behaviour by adding to the code. Python was chosen because it is a general purpose, interpreted programming language that is used extensively in scientific research. It is powerful enough for large-scale scientific and web applications, while also having a straightforward syntax that is accessible to non-programmers. Furthermore, it is supported by an impressive variety of scientific libraries including statistical modeling, machine learning, large-scale data processing, and scientific plotting.

R4 HR includes an integrated relational database allowing simulation data to be stored in database tables and queried using SQL during simulation execution. For military HR models, the ability to store and

process large volumes of data is essential. In many cases, relational databases and SQL are the ideal tools to handle personnel data during simulations.

R4 HR provides a modular model architecture through the implementation of a graph-alias naming system (GANS) (Okazawa 2013). In GANS, interactions between the elements of a model (i.e. the code, data, and other simulation objects) are handled explicitly using one-to-one connections rather than implicitly using scopes. Scopes are used in existing simulation software to define the set of model elements that can be accessed by a given element of model logic or line of code. For example, in Arena, data arrays, queues and the simulation clock are part of the global scope and are therefore accessible anywhere in the model. The disadvantage of scopes is that all names in a scope must be unique. If more than one element uses the same name, the name becomes ambiguous (known as a naming collision). Thus, it is generally not safe to select a group of model elements and duplicate it within the same model or move it into a new model because of the likelihood of creating naming collisions. In GANS, any model logic (up to and including the entire model) can be duplicated as-is and reused and reconnected with any other model because the connections between model elements are explicitly defined and are preserved with the model logic. Thus, duplicate names, if they occur, do not produce a naming collision.

R4 HR also facilitates modularity and model reuse by eliminating the concept of a global simulation clock and instead allowing the model developer to create multiple local clocks. The developer can independently set each clock's time and tick rate. As the simulation proceeds, all clocks advance together in accordance with their relative tick rates. When a model element makes use of time (e.g. to create a delay), it must be connected to a local clock, and it then specifies points in time as measured by that clock. In this way, when a developer builds a large model by interconnecting several sub-models, each sub-model brings its own clock which may operate in different ways. For example, one sub-model may use a clock that ticks once a year and where the zero-time is the year 2000, while another sub-model may use a clock as a timer that ticks once a day and where the time is periodically reset to zero by the model logic.

Finally, R4 HR generalizes the concept of the conventional DES entity. Conventionally, an entity contains information that describes one instance of an object moving through a process (e.g. a parcel moving through the mail system), and entities are the primary information carriers in the model. However, in larger models, there are often many types of information that must be communicated, some of which do not fit the entity concept well. For example, one sub-model in a military HR model may request information from another sub-model regarding the availability and qualifications of a list of military personnel. This type of communication does not readily fit the entity concept. To address this, R4 HR provides a single mechanism that allows any information to be transmitted within the model as needed including simple values (numbers and strings), compound types (lists and dictionaries), custom objects, and traditional DES entities.

These fundamental features of R4 HR allow model developers to easily build models that are flexible, scalable and modular without requiring advanced knowledge in programming or modelling techniques.

3. Military rank structure model

The military rank structure model focusses on career progression in a military occupation, or group of occupations. The model is rank-oriented and rank-driven, and simulates members in a military occupation for up to 20 years forward. The projected number of promotions and the number of people leaving various ranks typically provide the drivers from which other aspects follow, such as the expected number of recruits needed, and the projected demand for training courses.

A schematic of the simulation model logic is provided in Figure 1. As a simulation proceeds from year to year, there is a sequence of events that changes the status of members in an occupation according to rules which are defined by career progression policies and the occupation's organizational structure. Once each run year, attrition rates are sampled from user defined probability distributions, and applied to the current population to determine releases at each rank level. After attrition has been applied, any required training is assigned to the remaining members of the population, taking into account factors such as prerequisite courses, course priorities, and course capacities. Training is applied at the lowest rank first, and then proceeds up through the ranks. The next event is the application of terms of service conditions; at pre-defined career points, members are offered a new service contract or are released. The next step is to apply promotions. Promotions into the senior ranks are based upon time served and qualifications held, and are subject to there being a vacancy at the next rank. Promotion into the junior ranks is based upon time served and qualifications held; no vacancy at the next rank is required. New recruits are introduced at the lowest rank level. Simulation information and outputs are recorded at various points during, and at the end of, each model run year.

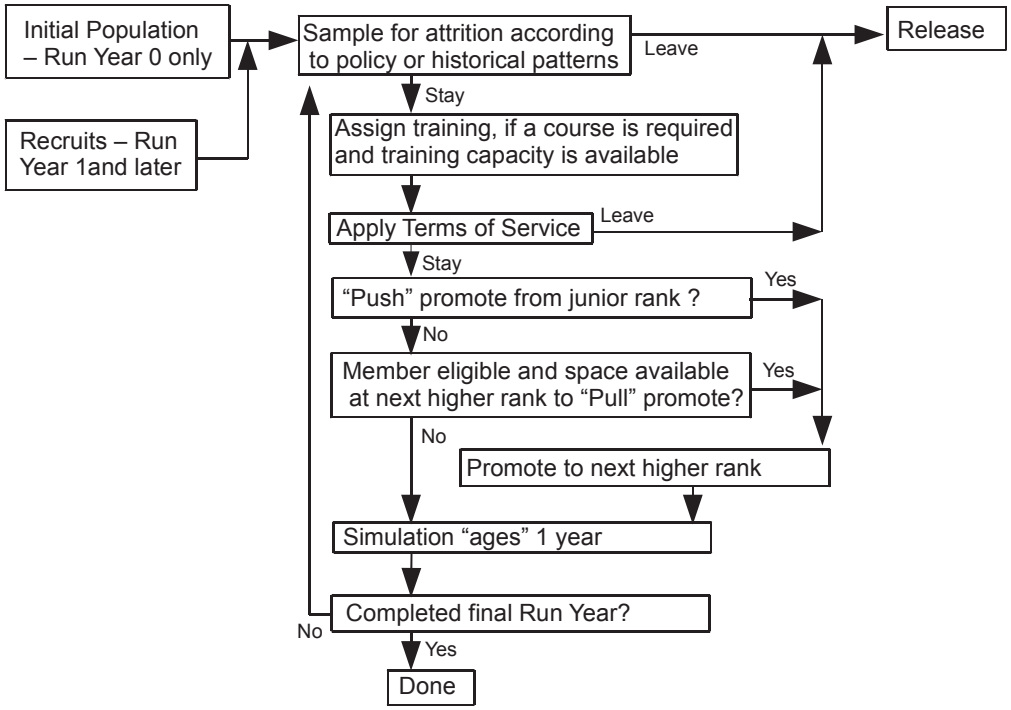

Figure 1: Simulation logic for the military rank structure model

From the recorded simulation output, various measures of occupation health are tallied by model run year. The projection of occupation health over time provides expected trends in a number of areas (e.g. promotion and release rates over time, average time in rank over the years, etc.) which can assist career managers with assessment of alternative management policies, as well as providing leading indications of potential future problems with overall occupational flows.

4. Differences in model implementations

The same model logic sequencing shown in Figure 1 is used in both the R4 HR and Arena implementations of the military rank structure model. The two model implementations employ the same assumptions (e.g. identical probability distributions for random events), and the mechanisms for the

various simulation processes (such as recruitment, training and promotion) are logically equivalent. However, as the two modelling environments differ greatly, the R4 HR version cannot be a “line by line” translation of the Arena version. Some significant examples of the differences are as follows:

- In the Arena version, the model logic is largely implemented using standard DES building blocks. However, certain complex logic makes use of Arena’s embedded Visual Basic for Applications (VBA) coding to invoke custom behaviour. While the R4 HR version retains the same high level structure and sequence of events shown in Figure 1, it makes use of the Python code part and SQL database querying to implement the bulk of the lower level model logic.

- In the Arena version, members are represented by distinct entities which flow through defined paths in the model representing their advancement through the ranks and eventual release. In the R4 HR version, members being modelled correspond to rows in a database table, with their attributes stored in the table’s fields. This allows different portions of the model to easily access a member, and to apply and track changes (e.g. setting a new rank due to promotion).

- In the Arena version, while entities are used to carry information about individual military members, Arena’s signals (event triggers that broadcast to the whole model) are required to coordinate the flow of individuals moving through the occupation. For example, signals are used to ensure that all releases are processed before initiating training. In the R4 HR version, information about the status of individuals, and information used to coordinate their movement through the occupation are both transmitted by a single means: R4 HR’s generalized information transmission mechanism.

In addition to the above points, the two model implementations use different random number streams, so the same simulation processes will result in different specific outcomes for each run year, within each replication.

5. Statistical validation methodology

To validate the equivalence of the R4 HR and Arena implementations of the rank structure model, each version was used to analyse the same scenario, and the simulation outputs were compared. Given the stochastic nature of many aspects of the model (e.g. age at recruitment, release events and course failures), variability in the simulation output from the same input is expected. In order to determine whether the observed variability is A) due to different random outcomes of the same underlying stochastic processes, or B) an indication that the behaviours of the two model implementations are not equivalent, statistical non-parametric and time series techniques were used to compare the simulation outputs.

5.1 Test scenario

The military scenario utilized is fictional, and focusses on the management of a Maritime Operations occupation during a period of downsizing and capability transition. A demographic profile, incorporating attributes such as age, rank, time in rank, years of service, and qualifications held for each person in the starting population was created along with assumptions governing personnel policies, operational requirements and training capacity. The same input data and assumptions, which included both deterministic and stochastic elements, were used for the Arena and R4 HR implementations of the model.

Each implementation simulated 20 years of promotions, releases, recruitment, and training within the Maritime Operations occupation, and consisted of 50 model replications. Several personnel system performance measures were recorded, of which five that are key determinants of military occupation health were selected for the validation analysis:

- Population strength (the size of the Maritime Operations occupation at each rank level);
- Average time in rank for each rank level;
- Promotions to each rank level;
- Releases at each rank level; and
- Qualification strength (the number of people with particular qualifications at each rank level).

5.2 Non-parametric statistical tests

For each of the performance measures selected, the simulation output consisted of two data sets (one from each model implementation), each containing 50 replications of a 20-year simulation period. As expected, the data sets generated did not follow normal distributions, negating the use of many parametric two-sample statistical tests to compare the data samples. To avoid making distribution assumptions that may not be valid, non-parametric techniques were employed to test equivalence of the outputs from the two model implementations. While very conservative, the Kolmogorov-Smirnov test was selected, as the only assumptions required are that the two samples are mutually independent, and come from continuous populations (Hollander and Wolfe 1973), assumptions which are satisfied. More powerful tests, such as the Wilcoxon rank sum test for location and Ansari-Bradley test for dispersion, require assumptions regarding the dispersion and location, respectively, of the continuous populations (Hollander and Wolfe 1973), assumptions that may or may not be valid. The null hypothesis for the Kolmogorov-Smirnov test is that the two samples come from the same continuous population, with the alternative hypothesis being they do not.

The Kolmogorov-Smirnov testing was conducted on each performance measure independently. For a given performance measure, the test was conducted 20 times—once for each model run year—with each test comparing the sample of 50 observations generated by R4 HR with the corresponding sample of 50 observations from Arena. Given that multiple hypothesis tests were being conducted, the Holm-Bonferroni method (Holm 1979) was used to control the family-wise Type I (false positive) error rate at level $\alpha = 0.05$.

5.3 Time series analysis

The non-parametric tests ignored the fact that the simulation output data were time series. For each performance measure, there were 50 independently generated time series covering a 20-year time period from each of the two model implementations. As the focus of the analysis was on pattern comparison rather than model fitting, formal statistical tests which require the fitting of time series models (e.g. autoregressive and/or moving average) to the data were not conducted.

Assuming the two model implementations are equivalent, the time series generated by the two implementations would follow the same patterns. This includes reproducing the same pattern of autocorrelations between model run years for a given performance measure. For each individual time series generated by the simulations, the sample autocorrelation coefficients at lag k, $k = 1,\ldots,20$, were calculated. Some of the observed time series contained trends, however these were not removed prior to calculating the sample autocorrelation coefficients, as the intent was to compare existing patterns rather than fitting time series models to the data in order to predict future patterns. For each set of 50 time series, the 5^{th}, 50^{th}, and 95^{th} percentiles of the sample autocorrelation coefficients were calculated, and plotted. Any visually notable deviations between the percentile patterns for corresponding data samples were interpreted as indications of non-equivalence of the two model implementations.

6. Validation Results

As noted above, many performance measures were analysed for the validation exercise. One representative measure is the number of releases annually at the rank of Lieutenant (Navy) (Lt(N)) from the Maritime Operations occupation. The rank of Lt(N) can be a career progression plateau for some

members, resulting in population sizes appropriate for the conduct of statistical tests. Figure 2 shows the number of Lt(N) releases observed across all replications for a single model run year. As expected, for any given run year there is variability in the results produced by the two model implementations.

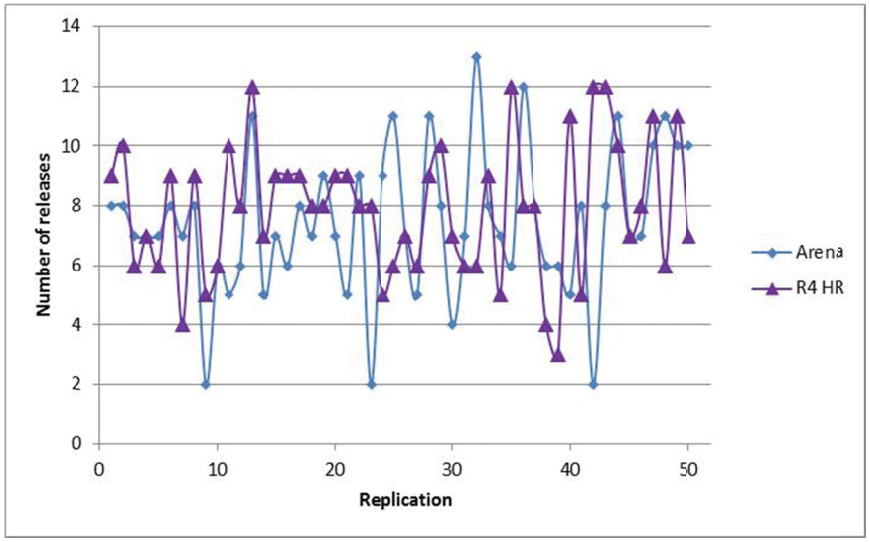

Figure 2: Number of Lt(N) releases observed in model run year 10

Applying the Kolmogorov-Smirnov two-sided test to the data samples of Lt(N) releases per run year generated by the two model implementations yields the results shown in Table 1. For model run year 10, the observed variability pictured above is not significant enough for the Kolmogorov-Smirnov test to reject the null hypothesis of a common population distribution for the two samples. The smallest p-value observed across the 20 multiple tests is 0.0171 in run year 15. As this value is greater than 0.05/20, the Holm-Bonferroni method does not reject any of the hypotheses. The same result was observed for all performance measures examined—in all cases the null hypothesis of a common population distribution for the Arena and R4 HR generated data samples could not be rejected. Failure to reject the hypotheses of common population distributions supports the conclusion of statistical equivalence of the two implementations of the military rank structure model.

Table 1: Kolmogorov-Smirnov two-sided test results for Lt(N) releases by model run year

Run year	Test statistic	p-value	Run year	Test statistic	p-value
1	0.08	0.9958	11	0.18	0.3584
2	0.0531	1	12	0.1	0.9541
3	0.08	0.9958	13	0.1	0.9541
4	0.12	0.8409	14	0.18	0.3584
5	0.1	0.9541	15	0.3	0.0171
6	0.08	0.9958	16	0.14	0.6779
7	0.14	0.6779	17	0.2	0.2408
8	0.08	0.9958	18	0.14	0.6779
9	0.16	0.5077	19	0.16	0.5077
10	0.16	0.5077	20	0.12	0.8409

The second component of the validation exercise took into account the time series nature of the data. Figure 3 shows the number of Lt(N) in the Maritime Operations occupation over time as generated by the two model implementations for a single replication. In this case the population is growing as the number

of individuals entering the rank annually exceeds the number leaving. The 5th, 50th and 95th percentiles of the observed autocorrelation coefficients for the two model implementations are shown in Figure 4. No visually noteworthy differences in the autocorrelation patterns between the two simulation output data sets are discernable. Similar results were observed for all performance measures analysed. These results provide further evidence to conclude that the R4 HR and Arena implementations of the military rank structure model are equivalent.

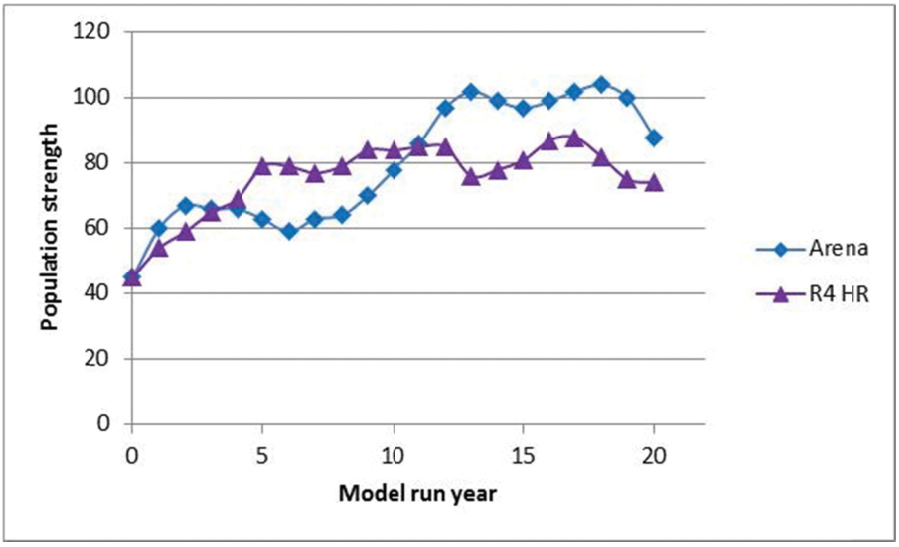

Figure 3: Number of Lt(N) in the Maritime Operations occupation over time for replication 5

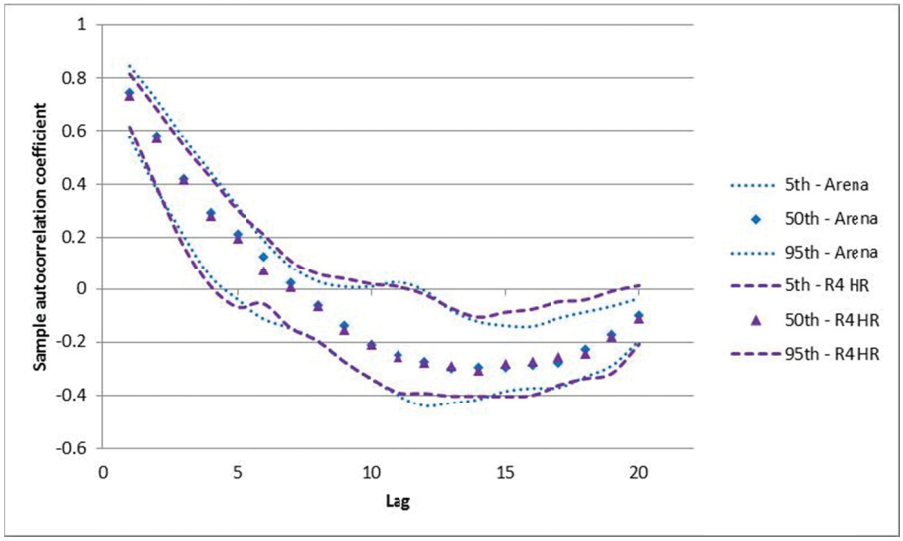

Figure 4: Select percentiles of the sample autocorrelation coefficients at lag *k* for Lt(N) population strength

7. Conclusion

The development of the R4 HR simulation environment is enabling DRDC to update its workforce modelling and simulation technology. This paper provided brief overviews of the R4 HR application and a well-established and trusted military rank structure model that has been rebuilt in the new environment. In

order to establish confidence in the new model implementation, a validation process was undertaken. As described in this paper, the old (Arena) and new (R4 HR) model implementations were used to analyse the same fictional military HR scenario. Non-parametric statistical tests and time series analysis techniques were used to test statistical equivalence of the simulation outputs, and by association, the logical equivalence of the two model implementations. The results of Kolmogorov-Smirnov tests and examinations of autocorrelation patterns for each of several measures of system performance did not detect any statistically significant differences in the simulation outputs. We thus conclude that the R4 HR model implementation is indeed logically equivalent to that of the Arena model, which will assist with building user and client confidence in the new R4 HR simulation environment. The approaches presented here can be deployed readily by analytics practitioners facing the challenge of statistically validating the equivalence of independently-developed simulation models of the same system.

References

Corbett, N. (2013) *Modelling the Production and Absorption of Pilots: The Development of the Production, Absorption and Retention Simulation (PARSim)*. DRDC Technical Report 2013-023, Centre for Operational Research and Analysis, Ottawa, Canada.

Hollander, M. and Wolfe, D. (1973) *Nonparametric Statistical Methods*. John Wiley & Sons, New York.

Holm, S. (1979) "A simple sequentially rejective multiple test procedure*", Scandinavian Journal of Statistics,* Vol 6, No. 2, pp. 65 – 70.

Latchman, S. and Hunter, C. (2002) *Preliminary Results from the Pilot Production/Absorption/Retention Simulation (PARSim) Model*. 1 CAD/CANR Headquarters Research Note 0202, Centre for Operational Research and Analysis, Ottawa, Canada.

Moorhead, P., Wind, A. and Halbrohr, M. (2008) "A Discrete Event Simulation Model for Examining Future Sustainability of Canadian Forces Operations", *Proceedings of the 2008 Winter Simulation Conference*, S.J. Mason, R.R. Hill, L. Mönch, O.Rose, T. Jefferson, and J.W. Fowler (ed.), pp. 1164 – 1172. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Okazawa, S. (2013) "A Discrete Event Simulation Environment Tailored to the Needs of Military Human Resources Management", *Proceedings of the 2013 Winter Simulation Conference*, R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M.E. Kuhl (ed.), pp. 2784 – 2795. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Scales, C., Okazawa, S. and Ormrod, M. (2011) "The Managed Readiness Simulator: A Force Readiness Model", *Proceedings of the 2011 Winter Simulation Conference*, S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu (ed.), pp. 2519 – 2529. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Seguin, R. (2011) "1 Canadian Forces Flying Training School (1 CFFTS) Resource Allocation Simulation Tool", *Proceedings of the 2011 Winter Simulation Conference*, S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu (ed.), pp. 2495 – 2506. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Straver, M., Okazawa, S. and Wind, A. (2009) *Training Pipeline Modelling Using the Production Management Tool*. DRDC Technical Memorandum 2009-019, Director General Military Personnel Research and Analysis, Ottawa, Canada.

Zegers, A. and Isbrandt, S. (2010) "The Arena Career Modelling Environment – A New Workforce Modelling Tool for the Canadian Forces", *Proceedings of the 2010 Summer Computer Simulation Conference*, G. Wainer (ed.), pp. 94 – 101. Curran Associates, Inc.