Empirical Evaluation of Decision Support Systems: Concepts and an Example for Trumpeter Swan Management

Richard S. Sojda

Empirical Evaluation of Decision Support Systems: Concepts and an Example for Trumpeter Swan Management

Richard S. Sojda

Northern Rocky Mountain Science Center, United States Department of the Interior - Geological Survey, 212 AJM Johnson Hall - Ecology Department, Montana State University, Bozeman, Montana 59717, USA (sojda@usgs.gov)

Abstract: Decision support systems are often not empirically evaluated, especially the underlying modelling components. This can be attributed to such systems necessarily being designed to handle complex and poorly structured problems and decision making. Nonetheless, evaluation is critical and should be focused on empirical testing whenever possible. Verification and validation, in combination, comprise such evaluation. Verification is ensuring that the system is internally complete, coherent, and logical from a modelling and programming perspective. Validation is examining whether the system is realistic and useful to the user or decision maker, and should answer the question: "Was the system successful at addressing its intended purpose?" A rich literature exists on verification and validation of expert systems and other artificial intelligence methods; however, no single evaluation methodology has emerged as preeminent. Under some conditions, modelling researchers can test performance against a preselected gold standard. Often in natural resource issues, such a standard does not exist. This is particularly true with near real-time decision support that is expected to predict and guide future scenarios while those scenarios are, in fact, unfolding. When validation of a complete system is impossible for such reasons, examining major components can be substituted, recognizing the potential pitfalls. I provide an example of evaluation of a decision support system for trumpeter swan (*Cygnus buccinator*) management that I developed using interacting intelligent agents, expert systems, and a queuing model. Predicted swan distributions over a 13 year period were tested against observed numbers. Finding such data sets is key to empirical evaluation.

Keywords: Decision support system; Verification; Validation; Empirical evaluation; Model; Trumpeter swan

1. INTRODUCTION

Decision support systems use a combination of models, analytical techniques, and information retrieval to help develop and evaluate appropriate alternatives [Adelman 1992; Sprague and Carlson 1982]. Because such systems handle complex and poorly structured problems, they are difficult to empirically evaluate. However, it is still easy to argue that evaluation of all decision support systems is important. For example, in the case of trumpeter swans, there are ecological and public policy reasons that increase the importance of ensuring that the right system has been built and been built correctly. In this paper, I focus on the modelling components of decision support systems and the integration of those components. Evaluation of the overall acceptance among natural resource managers of decision support systems, or other socioeconomic measures of their success and failure, are important but are not addressed.

2. DISCERNING DIFFERENCES BETWEEN VERIFICATION AND VALIDATION

Definitions of verification and validation in relation to computer software and modelling [Fishmann and Kiviat 1968; Mihram 1972; Adrion et al. 1982] have changed little over the years. These definitions are not absolute, but their use is becoming more definite over time. The following are from O'Keefe et al. [1987] and were adapted from Boehm [1981]: "Validation means building the right system. Verification means building the system right." These have been frequently referenced by others [e.g., D'Erchia 2001; Mosqueira-Rey and Moret-Bonillo 2000; Plant and Gamble 2003; Santos

2001]. A combined definition of verification and validation of software, provided by Wallace and Fujii [1989], was the analysis and testing "to determine that it performs its intended functions correctly, to ensure that it performs no unintended functions, and to measure its quality and reliability." Verification has been defined [Adrion et al. 1982] as "demonstration of consistency, completeness, and correctness of the software." The simplicity and completeness of Mihram's [1972] definition of validation in relation to simulation is attractive: "…the adequacy of the model as a mimic of the system which it is intended to represent." There is a plethora of discussions about the semantics of evaluating models, and Johnson [2001] provides a summary related to natural resource management.

My specifications for verification and validation in reference to decision support systems draw almost entirely from the above authors. Verification is ensuring that the system is internally complete, coherent, and logical from a modelling and programming perspective. Have the algorithm, knowledge, and other structures been correctly encoded? Validation is examining whether the system achieved the project's stated purpose related to helping the user(s) reach a decision(s). Validation of a particular model can also have the more limited meaning of whether the model is an adequate representation of the system it represents. This is sometimes described as black-box testing: do the inputs result in correct and useful outputs? Whether model or decision support system is being tested, I agree with Mihram [1972] that verification must occur before validation. This avoids the inadvertent situation where software provides expected outputs simply via calibration and correlation of input and outputs rather than via logical relationships. I use the term evaluation to encompass both verification and validation, but distinguish between them when used independently. I agree with Adelman [1992] that both should be part of the development process, and evaluators should specifically be part of the development team to foster iterative improvements. This is not to ignore the need for independent verification and validation of models and systems to ensure that the development team does not inadvertently err in their work.

3. POTENTIAL METHODS FOR EMPIRICAL EVALUATION

3.1 An Overview

Stuth and Smith [1993] followed the ideas of Eason [1988] and recommended iterative prototyping methods for decision support system development. Verification and validation are part of that iterative process. Verification should be performed prior to any delivery of a working system, even if a prototype. General validation might be done at this stage as well, with detailed efforts performed later. If one agrees that software development can be a living process, then verification and validation are part and parcel to that process and need to continue as system refinements and redeployments continue [Carter et al. 1992; Stuth and Smith 1993].

Sprague and Carlson [1982] recommend that organizations building their first decision support system recognize that it essentially is a research activity, and that evaluation should center on a general, "value analysis". Since then, it has become imperative that analytic and quantitative rigor be added beyond "soft testimonials" [Adelman 1991; Adelman 1992; Andriole 1989; Cohen and Howe 1989]. Sensitivity analysis can be a validation tool, especially for heuristic-based systems, and for systems where few or no test cases are available for comparison [Bahill 1991; O'Keefe et al. 1987]. Whenever validation is conducted, it is important to recognize to where, in space and time, inferences can be drawn from the validation data set. Another issue is the need to show not only how well a system performs, but also that it can avoid a catastrophic recommendation [Rushby 1988]. This is important in many natural resource venues because of the great concern for irretrievable and long term ecological changes.

It is my sense that validation is often the more neglected part of evaluation, so I will focus there. However, I do not wish to slight verification as it is critical to build decision support systems based on sound cause-effect relationships and not on poorly understood relationships between input and output.

3.2 Analogous Concepts From Artificial Intelligence

Successful implementation of decision support and expert systems hinges on incorporating three evaluation procedures [Adelman 1992]: (1) examining the logical consistency of system algorithms (verification), (2) empirically testing the predictive accuracy of the system (validation), and (3) documenting user satisfaction.

Verification and validation of knowledge-based and other decision support systems are known to be more problematic than in general modelling for many reasons [Gupta 1991]. A few difficulties in verifying multiagent systems [O'Leary 2001] are noteworthy, such as rule

conflict, circularity, non-used or unreachable antecedents, and agent isolation. Plus, not only is it important for a system to handle common cases, it ought to be able to deal with extreme events. This latter ability is one characteristic often only found with human experts. However, extreme events are not only common in, but often drive, ecological systems.

Wallace and Fujii [1989] provide a matrix of 41 techniques and tools that can be applied to 10 software verification and validation issues. Cohen and Howe [1989] take a slightly different approach specific to artificial intelligence methods, and they, too, discuss evaluation from the perspective of the software development life cycle. They emphasize empirical studies for such evaluation, whether focusing on verification or validation. For testing knowledge-based systems, Murrell and Plant [1997] provide a categorization of 145 automated techniques.

3.3 Alternative Validation Methods

3.3.1 Gold Standard

Under some conditions, modelling researchers can test performance against a preselected gold standard. Mosqueira-Rey and Moret-Bonillo [2000] describe this for intelligent systems as having test cases with known, prior outcomes. Virvou and Kabassi [2004] actually had such a set of cases based on expert opinion that they used for testing an intelligent graphical user interface. Often in natural resource issues, such a standard does not exist. This is particularly true with near real-time decision support that is expected to predict and guide future scenarios while those scenarios are, in fact, unfolding. Although this approach is theoretically desirable, I am not aware of an actual implementation in an environmental decision support system. This is not surprising in a domain where problems tend to be ill-defined and the associated knowledge uncertain.

3.3.2 Real-time and Historic Data Sets

In an ideal world, one could construct a decision support system and test its performance against actual scenarios as they unfold. This is not often possible because implementation of systems may need to be immediate. One alternative is to build the system using data, information, and knowledge from one set of situations and validate using an independent set, as done for crop yields [Priya and Shibasaki 2001], for a bass bioenergetics model [Rice and Cochran 1984], and for timber harvest [Wang and LeDoux 2003]. Prior versus post testing is another example of

this, and a decision support system for credit management was so validated by Kanungo et al. [2001]. When a data-driven model is a significant part of the decision support system, sometimes the data can be randomly separated into two parts, one for model development and one for validation. Pretzch et al. [2002] illustrate this using an extensive data set with a forest management simulator. Haberlandt et al. [2002] also took this approach for water quality assessments in river basins. A third option, when the decision support system is not data-based but rather knowledge-based, is to empirically evaluate predictions (outputs) from the system against a historic data set. This does assume that the logic underlying the system is constant over time. An example of this latter case is more fully developed in Section 4. (See tests 1 and 3A in Table1.)

3.3.3 Panel of Experts

It is sometimes possible to test performance against an independent panel of experts [O'Keefe et al. 1987]. This is a relatively common technique in the field of artificial intelligence and recent examples include multiagent web mining [Chau et al. 2003] and graphical user interface development [Virvou and Kabassi 2004]. Two concerns must be addressed, however. First, the panel of experts needed for such an evaluation must not be connected to system development. To do so would be so confounding that no reasonable experimental design would be feasible. Second, one of the basic tenets of using decision support systems for complex issues is that such questions can be beyond the capability of single persons to conceptualize and solve [Boland et al. 1992; Brehmer 1991].

3.3.4 Sensitivity Analysis

Sensitivity analysis is often more important in model validation than decision support system evaluation. This stems from the typical decision support system being highly complex, and it being difficult to isolate individual inputs, or small enough groups of inputs, to perform sensitivity analysis. Plus, some sort of gold standard or data set is still needed with which to work. (See test 7A in Table1.)

3.3.5 Component Testing

Sometimes it is not possible to validate a complete system, but one can test individual components. It is not uncommon, for example, to have multiple expert systems embedded in one

decision support system. When one validates each component separately, however, the interactions of the components and evolutionary behavior of the full system are not known. When testing of components is the only option, it is important to acknowledge this shortcoming. Often, when separate components of a system are validated, it can be argued that this is a form of system verification, as described by Rusu [2003]. (See test 6A in Table1.)

4. AN EXAMPLE: DECISION SUPPORT SYSTEM FOR TRUMPTER SWAN MANAGEMENT

4.1 Background

A multiagent system of interacting intelligent agents [Weiss 1999], expert systems, and a queuing model was developed to assist waterfowl managers simulate the effect of management actions on swan distributions [Sojda 2002]. This decision support system was evaluated at three levels: (1) verification of individual components, as well as the overall system, (2) soft validation of the expert systems, and (3) validation of the whole system.

It was decided not to evaluate the system against a team with expertise in flyway management of swans, primarily because it was not feasible to assemble such a panel that was independent of the people used in knowledge engineering. This was true for two related reasons. First, the total number of workers in the domain is small. Second, the cadre of such workers is closely interrelated institutionally and academically.

4.2 Verification of Components and Whole System

A key part of designing the individual expert systems was developing flowcharts of the ecological logic and using them to consult with experts for changes and refinement. Similarly, the "planeditor" facility in the multiagent software, DECAF, [Graham and Decker 2000; Graham 2001] allowed me to develop graphical representations of the logic underlying each agent

and consult with specialists in multiagent system design. When running the multiagent system, DECAF provided information about how each agent was functioning and about failed communications among agents. Utilities within the expert system development shell were used for verification of logical consistency of each expert system, including a static check for problems such as incomplete rules and trees. For example, an error would be detected if more than one rule tried to set a value for a single-valued variable, or if the consequent portion of a rule was inadvertently not provided. The utilities also dynamically checked the system with stochastic runs, and the final system was checked using 500,000 simulated runs with no problems detected.

4.3 Soft Validation of the Expert System Components

Demonstrations of each expert system were made to waterfowl managers, biologists, and researchers. This involved meetings and telephone consultations where individuals ran actual scenarios and provided comments. In addition, the expert systems were available in stand-alone fashion on the World Wide Web, both in prototype and final versions. Such validation targeted the underlying ontologies, knowledge, and problem solving logic, but was not empirical.

4.4 Validation Using An Historic Data Set

Based on queuing theory [Dshalalow 1995; Hillier and Lieberman 1995], the DSS begins by using an observed number of swans at each of 27 areas for the breeding season of one year, and then simulates the number at each of those areas for the four subsequent seasons, concluding with a simulated number for the breeding season of the subsequent year. The system simulates breeding swan numbers in one year increments. It was a comparison of the simulated number for the subsequent year versus the observed number for that same year that was the basis of my empirical

Test	MVPTMP *p*-value	Interpretation from rejecting the null hypothesis
1	.0001	output from base queuing model similar to observed numbers
3A	.0001	output using all expert systems (3) and activating all (7) refuge agents similar to observed numbers
6A	-	output using 3 expert systems identical to that with only the flyway expert system
7A	-	output using alternate breeding threshold of 0.4 identical to that using the standard, 0.6

Table 1. Interpretation of MVPTMP analyses from 4 of 34 experimental runs of the decision support system for trumpeter swan management. Null hypotheses were developed *a priori* [Sojda 2002]. No *p*-value is reported when output between the two groups was identical.

testing. An observed number of swans was available only for the breeding season, and not the other seasons, so analysis was limited to data for that season. Comparisons of simulated and observed data could be made for 13 years, 1988-2000. Observed numbers were those collected by the member agencies of the Pacific Flyway Council and informally reported by the United States Fish and Wildlife Service on an annual basis [e.g., Reed 2000].

4.5 Data Analysis

Although all 27 areas were always used in the queuing model, swans had never been observed in seven areas during the breeding season and those areas were excluded from statistical analysis. In all such cases, the system did not simulate swans in those areas. This ensured that the consistent simulation of no swans where none were expected did not artificially inflate the evaluated accuracy and precision of the system.

Thirty-four black-box experiments were conducted to empirically validate the decision support system's ability to predict swan distributions in the flyway [Sojda 2002]. The results from four of the experiments are provided in Table 1. Multivariate Matched-Pairs Permutation Test (MVPTMP) statistical procedures [Mielke and Berry 2001] were used for the analyses. The first of the pair is simulated data, the second is either observed data or simulated data from a run of the system with a different configuration. To test the base model (a queuing system), predicted numbers of swans for 20 areas were compared against observed numbers for a series of 13 years. In such analyses, a small *p*-value is evidence of similarity of distributions of swans over both space and time between the two groups of data forming a pair. Because of the multidimensional structure of such comparisons of spatial data over time, it was difficult to provide visualizations. Accompanying departures of the simulated from the observed

numbers of swans were simply graphed [Sojda 2002].

5. DISCUSSION AND CONCLUSIONS

Validation is the process of determining whether the stated purpose of the system was achieved. I conclude that multiagent systems were an effective way to model movement of waterfowl in a flyway. Because models are abstractions of reality, it is inherent that they will have shortcomings from not being able to accurately represent all knowledge, logical relationships, and probabilistic intricacies. Overall, the evidence was strong that the base model (in the decision support system for trumpeter swan management) mimicked the observed pattern of swan distributions over time, as does the system run with the default configuration. Almost all experimental runs of the decision support system showed the same pattern.

It seems irresponsible to deliver a decision support system that has not been adequately evaluated, including both verification and validation. Empirical evaluation in some form is critical, and can range from experiments run against a preselected gold standard to more simple testing of system components. It is imperative to understand, from an experimental and logical perspective, to what extent inferences can be made as a result of the validation. In the end, the question to answer is: Was the system successful at addressing its intended purpose? Often, searching for the right database for empirical evaluation can be as important as adequate decision support system development, itself.

6. ACKNOWLEDGEMENTS

7. REFERENCES

Adelman, L., Experiments, quasi-experiments, and case studies: a review of empirical methods for evaluating decision support systems, *IEEE Transactions on Systems, Man, and Cybernetics*, 21(2), 293-301, 1991.

Adelman, L., Evaluating decision support and expert systems, John Wiley and Sons, New York, New York, 1992.

Andriole, S. J., Handbook of decision support systems, TAB Professional and Reference Books, Blue Ridge Summit, Pennsylvania, 248 pages, 1989.

Adrion, W. R., M. A. Branstad, and J. C. Cherniavsky, Validation, verification, and testing of computer software, *ACM Computing Surveys*, 14(2), 159-192, 1982.

Bahill, T. A., Verifying and validating personal computer-based expert systems, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 205 pages, 1991.

Boehm, B. W., Software engineering economics, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 767 pages, 1981.

Boland, R. J., A. K. Mahewshwari, D. Te'eni, D. G. Schwartz, and R. V. Tenkasi, Sharing perspectives in distributed decision making, pages 306-313, *in*: *Proceedings of the Conference on Computer-Supported Cooperative Work*, Association for Computing Machinery, New York, New York, 1992.

Brehmer, B. Distributed decision making: some notes on the literature, pages 3-14 *in:* Rasmussen, J., B. Brehmer, and J. Leplat, eds., Distributed decision making: cognitive models for cooperative work, John Wiley and Sons, Chichester, England, 1991.

Carter, G. M., M. P. Murray, R. G. Walker, and W. E. Walker, Building organizational decision support systems, Economic Press Inc., San Diego, California, 358 pages, 1992.

Chau, M., D. Zeng, H. Chen, M. Huang, and D. Hendriawan, Dwsign and evaluation of a multiagent collaborative Web mining system, *Decision Support Systems*, 35, 167-183, 2003.

Cohen, P. R., and A. E. Howe, Toward AI research methodology: three case studies in evaluation, *IEEE Transactions on Systems, Man, and Cybernetics*, 19(3), 634-646, 1989.

Dshalalow, J. H., An anthology of classical queuing models, pages 1-42 *in*: Dshalalow, J. H., ed., Advances in queuing: theory, methods, and open problems, CRC Press. Boca Raton, Florida, 1995.

D'Erchia, F., C. Korschgen, M. Nyquist, R. Root, R. Sojda, P. Stine, A framework for ecological decision support systems: building the right systems and building the systems right., U.S. Geological Survey, Biological Resources Division, Information and Technology Report USGS/BRD/ITR-2001-0002, Washington, DC, 50 pages, 2001.

Eason, K., Information technology and organizational change, Taylor and Francis Publishing, London, United Kingdom, 247 pages, 1988.

Fishmann, G. S., and P. J. Kiviat, The statistics of discrete-event simulation, *Simulation*, 10, 185-195, 1968.

Graham, J. R., Real-time scheduling in distributed multi agent systems, PhD Dissertation, University of Delaware. Newark, Delaware, 166 pages, 2001.

Graham, J. R. and Decker, K. S., Towards a distributed, environment-centered agent framework, pages 290-304 *in*: Jennings, Nicholas R. and Y. Lesperance, eds., *Proceedings of the Sixth International Workshop on Agent, Theories, Architectures, and Languages (ATAL-99)*, Springer-Verlag, Berlin, Germany, 2000.

Gupta, U., Validating and verifying knowledge-based systems, IEEE Computer Society Press, Washington, DC, 423 pages, 1991.

Haberlandt, U., V. Krysanova, and A. Bardossy, Assessment of nitrogen leaching from arable land in large river basins - Part II: regionalisation using fuzzy rule based modelling, *Ecological Modelling*, 150, 277-294, 2002.

Hillier, F. S. and G. J. Lieberman, Introduction to operations research, McGraw-Hill, Inc., New York, New York, 998 pages, 1995.

Johnson, D. H., Validating and evaluating models, pages 105-119 *in*: Shenk, T. M., and A. B. Franklin, eds., Modelling in Natural Resource Management: Development, Interpretation, and Application, Island Press. Washington, DC, 2001.

Kanungo, S., S. Sharma, and P. K. Jain, Evaluation of a decision support system for credit management decisions, *Decision Support Systems*, 30, 419-436, 2001.

Mielke, P. W. and K. J. Berry, Permutation methods: a distance function approach.

Springer-Verlag, New York, New York, 352 pages, 2001.

Mihram, G. A., Some practical aspects of the verification and validation of simulation models, *Operational Research Quarterly*, 23(1), 17-29, 1972.

Mosqueira-Rey, E, and V. Moret-Bonillo, Validation of intelligent systems: a critical study and a tool, *Expert Systems with Applications*, 18, 1-16, 2000.

Murrell, S. and R. T. Plant, A survey of tools for the validation and verification of knowledge-based systems: 1985-1995, Decision Support Systems, 21, 307-323, 1997.

O'Keefe, R. M., O. Balci, and E. P. Smith, Validating expert system performance, *IEEE Expert*, 2(4), 81-90, 1987.

O'Leary, D. O. Verification of multiple agent knowledge-based systems, *International Journal of Intelligent Systems*, 16, 361-376, 2001.

Plant, R., and R. Gamble, Methodologies for the development of knowledge-based systems, 1982-2002, *The Knowledge Engineering Review*, 81(1), 47-81, 2003.

Pretzsch, H., P. Biber, and J. Dursky, The single tree-based stand simulator SILVA: construction, application and evaluation, *Forest Ecology and Management*, 162,3-21, 2002.

Priya, S., and R. Shibasaki, National spatial crop yield simulation using GIS-based crop production model, *Ecological Modelling*, 135, 113-129, 2001.

Reed, T., 2000 Fall trumpeter swan survey. Unpublished report, U.S. Fish and Wildlife Service, Lakeview, Montana, 28 pages, 2000.

Rice, J. A., and P. A. Cochran, Independent evaluation of a bioenergetics model for largemouth bass, *Ecology*, 65(3), 732-739, 1984.

Rushby, J., Validation and testing of knowledge-based systems: how bad can it get?, pages 77-83 *in*: Gupta, U., ed., Validating and Verifying Knowledge-Based Systems, IEEE Computer Society Press, Los Alamitos, California, 1988.

Rusu, V., Combining formal verification and conformance testing for validating reactive systems, *Software Testing, Verification and Reliability*, 13, 157-180, 2003.

Santos, E., Jr., Verification and validation of Bayesian knowledge-bases, Data and Knowledge Engineering, 37, 307-329, 2001.

Sojda, R. S., Artificial intelligence based decision support for trumpeter swan management, PhD Dissertation, Colorado State University, Fort Collins, Colorado, 193 pages, 2002.

Sprague, R. H. Jr., and E. D. Carlson, Building effective decision support systems, Prentice-Hall, Englewood Cliffs, New Jersey, 329 pages, 1982.

Stuth, J. W. and M. S. Smith, Decision support for grazing lands: an overview, pages 1-35 in J.W. Stuth and B.G. Lyons, eds., Decision support systems for the management of grazing lands, Man and the Biosphere Series Volume 11: Papers from the International Conference on Decision Support Systems for Resource Management, The Parthenon Publishing Group, Pearl River, New York, 1993.

Virvou, M., and K. Kabassi, Evaluating an intelligent graphical user interface by comparison with human experts, Knowledge-Based Systems, 17, 31-37, 2004.

Wallace, D. R. and R. U. Fujii, Software verification and validation: an overview, *IEEE Software*, 6(3), 10-17, 1989.

Wang, J., and C. B. LeDoux, Estimating and validating ground-based timber harvesting production through computer simulation, *Forest Science*, 49(1), 64-76, 2003.

Weiss, G., Prologue: multiagent systems and distributed artificial intelligence, pages 1-23 in: Weiss, G., ed., Multiagent systems: a modern approach to distributed artificial intelligence, MIT Press, Cambridge, Massachusetts, 1999.