https://doi.org/10.1016/j.eswa.2012.07.026

A proposed validation framework for expert elicited Bayesian Networks

Jegar Pitchforth☆, Kerrie Mengersen

Queensland University of Technology

Abstract

The popularity of Bayesian Network modelling of complex domains using expert elicitation has raised questions of how one might validate such a model given that no objective dataset exists for the model. Past attempts at delineating a set of tests for establishing confidence in an entirely expert-elicited model have focused on single types of validity stemming from individual sources of uncertainty within the model. This paper seeks to extend the frameworks proposed by earlier researchers by drawing upon other disciplines where measuring latent variables is also an issue. We demonstrate that even in cases where no data exist at all there is a broad range of validity tests that can be used to establish confidence in the validity of a Bayesian Belief Network.

Keywords: expert, validation, bayesian network, sensitivity

☆*Phone:* +61 403 961 878

Email address: `jegar.pitchforth@qut.edu.au` ()

[1]Lvl 11, 126 Margaret St. , Brisbane 4000

1. Introduction

Bayesian Networks (BNs) are an increasingly popular tool for modelling complex systems, particularly in the absence of easily accessed data. A BN describes the joint probability distribution of a network of factors using a Directed Acyclic Graph (Pearl, 1988). Factors that influence the likelihood of the outcome node being in any given state are represented as nodes on the graph. If the state of one model factor influences the state of another a directional arc is drawn between the two nodes representing these factors in the model. The combination of the nodes and their relationships is the BN structure. Each node in the graph can adopt any one of a finite set of states. For example, a factor representing magnitude could be classified as 'high' or 'low'. While nodes do not strictly have to be discretised the practice is by far more commonly undertaken than not due to its computational convenience, and as such we do not discuss models that include non-discretised nodes in this paper. Finally, each node and relationship between nodes is quantified according to the likelihood of the node adopting a given state. In the case of input nodes these probabilities are seen as unconditional, whereas nodes internal to the model are dependent upon the states of the preceding nodes. The strength and direction of the relationship between model factors is defined in the conditional probability table associated with the child node.

BNs are often created through a process of expert elicitation, in which experts are asked to create a complex systems model by giving their opinions on the model structure, discretisation, and parameterisation. The validity of these models is generally tested through one of two procedures: by comparing the model predictions to data available for the subject matter, or by asking the experts who contributed to the model creation to comment on its accuracy. This paper argues that these tests are limited in their ability to accurately test the validity of BNs, and presents a framework for more thorough validity testing. The work presented here stems from questions raised during the creation of a BN from expert elicitation to model the inbound passenger processing time at Australian airports. The network was elicited in collaboration with managerial and operational experts from Australian Customs and Border Protection Service (ACBPS) for the purpose of gaining more informative reporting of key performance indicators. In particular, the modelling of critical infrastructure underlined the importance of establishing that both experts and modellers have confidence in the final model produced. The paper is structured as follows. First, the concept of validation as it applies to BNs is introduced in section 1.1. Second, the sources of confidence in BN validity are discussed, including network structure, discretisation, and parameterisation in section 1.2. Third, prior approaches to validating latent and expert elicited scales and models are introduced, drawing from psychometrics, system dynamics and other BN research in section sec:prevapproach. These principles are then applied to BNs with examples from the airport inbound passenger processing model in section 3.

1.1. Confidence in Bayesian Belief Network validity

Model validity is often conceptualised as a simple test of a model's fit with a set of data. However validity is a much broader construct: in essence, validity is the ability of a model to describe the system that it is intended to describe both in the output and in the mechanism by which that output is generated. In this paper we consider this broader definition of validity.

The need for an explicit set of validity tests for BNs over and above comparisons with data is clear. In current practice, where data are available on the phenomenon of interest, these data may be used to validate model predictions. Several tests of this nature exist, such as a variety of Normal Maximum Likelihood model selection criteria (Silander et al., 2009). However, a common reason for using BN models is a lack of available data. Examples of phenomena for which data are scarce include population characteristics in many developing countries (Shakoor et al., 1997), global epidemiological phenomena (Masoli et al., 2004), organised crime (Sobel and Osoba, 2009), conservation (Johnson, 2009) and biosecurity risk analysis (Barrett et al., 2010). In such cases, expert opinion can be elicited to create a Bayesian Belief Network (BBN). A common technique for validating BBNs based on expert opinion in the absence of data, is simply to ask the experts whether they agree with the model structure, discretisation, and parameterisation (see Korb and Nicholson (2010) for an excellent overview of BN applications and methods). This simple test is necessary, but not sufficient, to independently verify the validity of a complex model. Even where data are available, model fit is only a part of the model's overall validity. These considerations lead to this paper's proposition of a general validity framework for BNs.

1.2. Sources of confidence in Bayesian Network validity

In order to approach a validation framework for BNs, a short discussion of the background assumptions of this framework is required. First, we assume there exists a latent, unobservable 'true' model (or set of acceptable 'true' models) for the phenomenon of interest against which the expert elicited model can be compared. Second, for the purposes of the validity framework presented in this paper, we consider a BN model to consist of four elements: model structure (section 1.2.1), node discretisation(section 1.2.2), and discrete state parameterisation(section 1.2.3). Each of these elements has been raised as a source of uncertainty in BN modelling. We provide a discussion of each element and consider the importance of validity within each model element, and within the model as a whole. The model elements are summarised in figure 1.

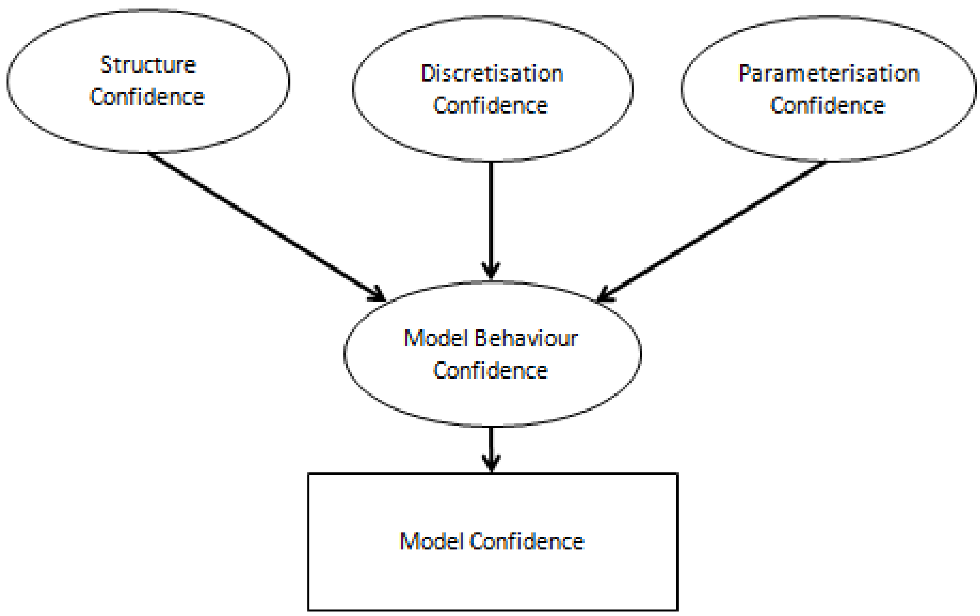

Figure 1: Sources of confidence in Bayesian Network validity

1.2.1. Structure

There are a number of questions when creating the structure of a BN. The first is the appropriate number of nodes to include which is a question of the modelling domain, level and scope. It is widely acknowledged that networks with a large number of nodes can easily become computationally intractable, as can networks with a large number of arcs between nodes (Koller and Pfeffer, 1997). The BN creator should ensure that the model is neither too simple nor too complex in its explanation of the system.

1.2.2. Discretisation

The discretisation process allows us to model systems probabilistically by taking continuous factors and assigning them intervals, ordinal states or categories, then modelling over the discrete domain. In more recent research, Uusitalo (2007) pointed out that such discretisation is a major disadvantage of BN modelling if it is necessary for the model, and Myllymaki et al. (2002) outlines how the process has the potential to destroy useful information. Given the information loss inherent in the discretisation process, ensuring that the states are a valid interpretation of the state space of the node is critical for a defensible network.

1.2.3. Parameterisation

Parameterisation refers to adding the values elicited from experts to the belief network (Woodberry et al., 2005). Much work has been conducted on controlling this stage of the process (Renooij, 2001), but little has been written about how to validate expert responses post-elicitation.

1.2.4. *Model Behaviour*

Finally, the behaviour of the model can be seen as the joint likelihood of the entire network as well as its sub-networks and relationships, hence confidence in model behaviour is founded upon the validity of the other three dimensions of the model. It is important to note that in the case of BNs, we are not only interested in whether the model can tell us what a system is doing under certain conditions, but also the factors and relationships that bring about this behaviour. This makes the problem of validating the model incredibly complex when attempted wholesale and justifies the need for partitioning the dimensions of uncertainty for BNs. As such it is recommended that the structure, discretisation and parameterisation are tested for validity before any model behaviour tests can be run.

2. Previous approaches to validity

2.1. *Psychometrics*

The discipline of psychometrics arose as a counterpart to the field of psychology, which at its foundation attempts to measure latent, unobserved, 'true' variables such as intelligence. Due to this rich tradition, the foundations of measurement validation in psychometry are particularly solid, and serve as a useful base to begin discussion of a similar framework for BNs. Psychometrics first identified four types of validity (Cronbach and Meehl, 1955); more recent research has reclassified and added dimensions of validity to establish a full validation framework (Trochim, 2001). Based on the framework depicted in figure 2, a psychometric test can pass all these tests of validity to varying degrees, providing a multidimensional measure of how well a particular test measures a latent variable. In psychometric testing there are seven commonly tested dimensions of validity: nomological validity, face validity, content validity, concurrent validity, predictive validity, convergent validity, and discriminant validity. In psychometrics, before any other tests

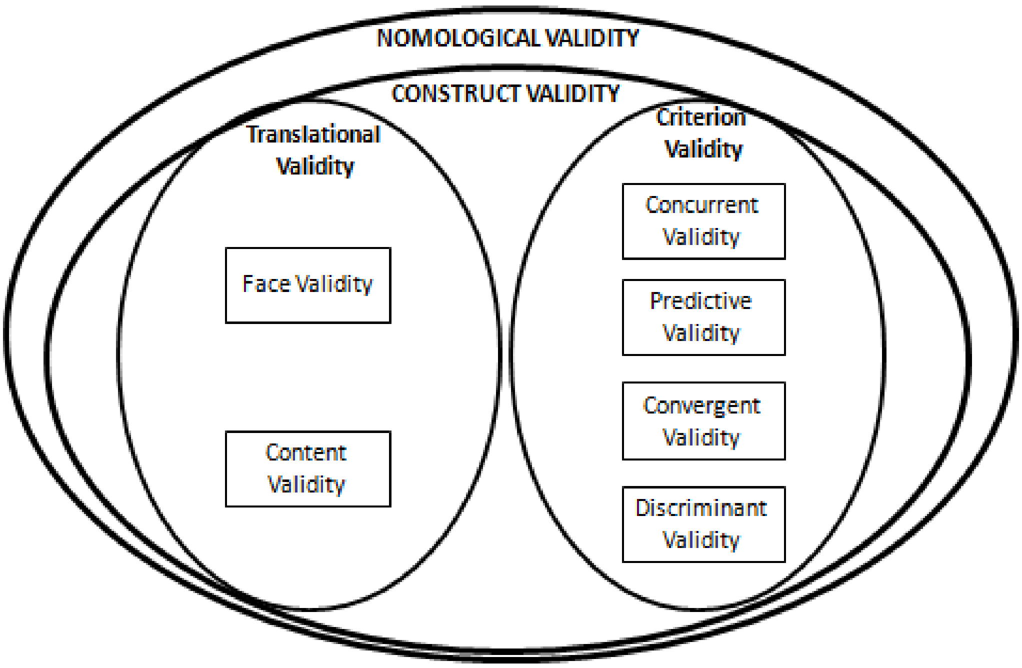

Figure 2: The psychometric validity testing framework adapted from Trochim (2001).

of validity can be undertaken, the nomological validity of the validity domain should be established. High nomological validity indicates that the measurement sits well within current academic thought on the subject. Face validity refers to the heuristic interpretation of a measure as a valid representation of the underlying psychometric construct. Content validity describes both the inclusion of all variables believed to be within a domain and the relevance of the factors included in the scale. Concurrent validity refers to the behaviour of a measurement scale; specifically, that the measure varies at the same point in time as another theoretically related measure taken on the same sample. Convergent validity refers to the criterion that scores on the measure to be validated (e.g. intelligence) should match scores on another, theoretically related measure (e.g. school grades) in the same sample. Finally, discriminant validity refers to the criterion that scores on the measure to be validated should be different from scores on tests that measure constructs that are theoretically unrelated. While this is a useful paradigm upon which to base our exploration, the differences between judging the validity of a complex model and the validity of a score of a single construct are significant enough to necessitate further exploration into other approaches.

The parameterisation process is the most similar to the psychometric discipline, as the parameters can be treated as scores denoting a given belief about the behaviour of that node. Using this approach, we can use the extensive literature on psychometrics and group behaviour to help validate the parameters we elicit from our experts.

2.2. System Dynamics

In his review of system dynamics validation tests Barlas (1996) describes a series of eight tests to validate system dynamics models; parameter confirmation, dimensional consistency, modified behaviour prediction, Turing tests, Qualitative Features analysis, extreme conditions testing, behaviour sensitivity tests and structure confirmation. Each of the tests can be classified in terms of the psychometric validity framework but can also be directly applied to specific sources of BN model uncertainty. For example, parameter confirmation can be seen as a special test of concurrent validity applied specifically to model parameterisation. The tests introduced in the Barlas (1996) paper are described in more depth in the following section with specific reference to BN modelling.

2.3. Machine Learning

It is worth mentioning the significant research that has been conducted in the field of machine learning, particularly regarding content validity of the network structure. Machine learning researchers often use BNs and Bayesian Belief Networks to discover true networks using full datasets (Heckerman et al. (1995) is a strong and widely cited example of this method). While this work is outside the scope of this paper, it is worth mentioning due to the minimalist approach used by machine learning researchers. In particular, the discipline is concerned with finding methods of excluding as many nodes and relationships from a BN as possible without losing explanatory power.

2.4. Bayesian Network specific tests

There are very few validity tests specific to BN modelling, but the few that are present are used commonly. Pollino et al. (2007) refers to the concepts of 'sensitivity to findings' and 'sensitivity to parameters' as methods of testing the predictive validity of expert-elicited networks. Other tests that have been introduced, such as d-separation analysis (Geiger et al., 1990) and causal independence-based tests (Cheng et al., 1997) are structural tests only, and are often used to establish internal consistency which is more elegantly defined as a reliability criterion.

2.5. Problem Statement

Unlike areas in which objective data are available, BNs built from expert elicitation cannot be validated using complete test datasets. As such, the concept of validity is not absolute but a question of additive strength. Often we cannot say whether a test has been conclusively passed or not, only take the weight of evidence over all the tests that have been applied. With this in mind we can begin to move toward a framework for validating all sources of uncertainty within the BN. While there are some tests introduced in previous research, these only test individual aspects of the network and can often only reflect the reliability rather than the validity of the model. For BN's based either entirely upon expert elicitation, or a combination of data and expert elicitation, to be judged as valid assessments of the knowledge around a domain, a more comprehensive and robust framework of validity measures needs to be established.

3. A validity testing framework for expert-elicited Bayesian Networks

The prior approaches to test and model validation are discussed and related to BNs in the following section, with examples from the airports inbound passenger processing network. When applying this validity testing framework to BNs, model structure, node discretisation, and overall model behaviour must be considered in addition to parameterisation. For this reason, in the following framework we consider the seven types of validity from psychometrics (including their special tests from system dynamics and BN modelling disciplines), and their application to the four sources of BN model uncertainty.

3.1. Nomological validity

In terms of an expert elicited BN, building nomological validity means establishing confidence that the model domain fits within a wider domain as established by the literature. For example, the passenger processing BN for ACBPS should sit within literature on airport terminals, way finding and security as well as other types of complex systems models and spatio-temporal model methods. If this test cannot be passed by the network, an argument must be made for why this model sits outside all current known research. This is very unusual, but may occur in fields such as advanced physics, where new information is shifting the entire paradigm of the discipline regularly. If this is the case, there may be an argument for a network having low nomological validity. Nomological validity is generally applied to the whole domain, but the nomological map serves as a reference for finding appropriate comparison models in later tests of specific sources of uncertainty. Given the power of nomological validity to place the research in a wider context, we begin the validation process with the questions:

- Can we establish that the BN model fits within an appropriate context in the literature?
- Which themes and ideas are nomologically adjacent to the BN model, and which are nomologically distant?

3.2. Face validity

Face validity is one of the most commonly used tests for expert-elicited BNs. For example, we can look at our passenger processing BN and check that baggage delivery time is part of the model and that it is related to the time spent picking up baggage to approximately the right level. However, despite the ease of establishing face validity it is considered the weakest form of validity within the psychometric framework. One of the primary dangers in establishing face validity is criterion contamination an issue that arises when the test dataset is the same as the validation set (Darkes et al., 1998). In our case, we might ask our set of experts whether they think the network looks the same as expected. Unsurprisingly, there are very few cases where the experts disagree with their own judgment. A more robust way of establishing face validity would be to split the population of experts into test and validation groups, and ask the validation group only about the face validity of the network (Johnson et al., 2010). In cases where few experts are available, we can undertake a number of other strategies normally used for elicitation, such as using different experts for different parts of the BN, asking experts to assess their answers from a rival's perspective, asking experts whether the model is applicable outside their domain and many others(Low Choy et al., 2009; James et al., 2010). In addition, often the entire model is tested at once (Korb and Nicholson, 2010). In order to learn as much as possible about the model through the validation process it is worthwhile to assess the face validity of the structure (including sub-networks), discretisation and parameterisation independently. We therefore suggest the second set of questions in this validation stage:

- Does the model structure (the number of nodes, node labels and arcs between them) look the same as the experts and/or literature predict?
- Is each node of the network discretised into sets that reflect expert knowledge?
- Are the parameters of each node similar to what the experts would expect?

3.3. Content Validity

To test for content validity of the structure we can check that all noted factors and relationships from the literature are included in the model, and discover which relationships are novel to the BN model. For example, in the passenger processing BN we could ensure that all the factors considered to important by the regulating bodies are included. To check the content validity of the discretisation of nodes within the model, we can ensure that all intervals implicated in the literature are included in the network. For example, if we were to discover that a node is generally classified at three levels in the literature, then a node with binary states would have low content validity. From a systems dynamics perspective, Barlas (1996) describes a dimensional consistency test which when applied to a BN paradigm could be defined as ensuring that all possible states of the node are included in the discrete states. For example, if a node were to include binary states of above twelve people and below twelve people, then the node would lack dimensional consistency as the possibility of there being exactly twelve people has been excluded. Finally, the content validity of the parameterisation can be checked through comparing expert elicited probabilities and relationships to analogous relationships in the literature. If parameters in the expert elicited model are significantly different, an argument should be made for the difference. To assess the content validity of a BN model, the following questions are suggested:

- Does the model structure contain all and only the factors and relationships relevant to the model output?
- Does each node of the network contain all and only the relevant states the node can possibly adopt?
- Are the discrete states of the nodes dimensionally consistent?
- Do the parameters of the input nodes and CPT reflect all the known possibilities from expert knowledge and domain literature?

3.4. Concurrent Validity

In the context of BNs, concurrent validity can refer to the possibility that a network or section of a network behaves identically to a section of another network, preferably driven by data. While this seems improbable, the nature of BN modelling seems to lend well to concurrent validity. For example, the passenger processing BN shares some sub networks and nodes with the customer satisfaction model for the same airport. In her introduction to Object Oriented Bayesian Networking, Koller and Pfeffer (1997) describes the technique as a way of capitalising on this high concurrent validity by building networks from instances, or nodes representing sub-networks that can be easily transposed to other networks. This method allows large and highly complex BNs to be built without the researcher repeating modelling work performed by other researchers in the same domain. To test the concurrent validity of the structure of a BN, we can check other networks in related domains for sub-networks that are similar to sub-networks in the network. A model with high concurrent validity would have sub-networks in common with networks that are theoretically related, with the same number of nodes and relationships, with the relationships in the same direction. Similarly, when similar sub-networks from theoretically related networks are identified, we can judge the validity of the discretisation of nodes and their parameterisation against the intervals of nodes and probabilities supplied in the comparison network. In the Barlas (1996) review of system dynamics tests, the application of concurrent validity criteria specifically to the parameters of the model factors is known as 'parameter confirmation'. Given these approaches, the following questions are suggested as tests of a BN's concurrent validity:

- Does the model structure or sub-networks act identically to a network or sub network modelling a theoretically related construct?
- In identical sub networks, are the included factors discretised in the same way as the comparison model?
- Do the parameters of the input nodes and CPTs in networks of interest match the parameters of the sub network in the comparison model?

3.5. Convergent Validity

Convergent and discriminant validity are usually considered together, as they both reflect the relationship the BN has with other models. Convergent validity in BNs refers to how similar the model structure, discretisation, and parameterisation are to other models that are intended to describe a similar system. For example, we would expect our passenger processing BN to look similar to a network describing the processing of cargo at a seaport. The selection of comparison models is dependent upon the literature and knowledge of the domain at hand, but the original nomological map created in the first step of validation can be used as a reference for which sources may be of use. In particular, the comparison model for establishing convergent validity should be taken from an area as nomologically proximal as possible. In practise this could mean using a comparison model drawn from another complex systems discipline applied to the same domain, or alternatively using a BN drawn from a theoretically similar domain. As with the other types of validity, we can test the expert elicited BN regarding the convergent and discriminant validity of the structure, discretisation and parameterisation in isolation using the following questions:

- How similar is the model structure to other models that are nomologically proximal?
- How similar is the discretisation of each node to the discretisation of nodes that are nomologically proximal independent of their network domain.
- Are the parameters of nodes that have analogues in comparison models assigned similar conditional probabilities?

3.6. Discriminant Validity

The counterpart to convergent validity is discriminant validity, defined in this framework as the degree to which a model is different to models that should be describing a different system. For example, we would expect our passenger processing BN to look different to a model describing students' progression through school. As in the case of convergent validity, the comparison model can be chosen using the nomological map as a reference guide for useful sources. The ideal method for establishing good discriminant validity would be to select models from nomologically distal disciplines and work toward the construct of interest. Given that convergent validity has already been established, the ideal model would be one that is similar in most respects to the convergent comparison model, but dissimilar in all respects to the discriminant comparison model, which would be drawn from an area of research very close to the convergent validity comparison model.

A system dynamics test of experts' judgement of the discriminant validity of any source of uncertainty in a BN model is known as a Simulation Turing test (Schruben, 1980). The test requires many versions of the model to be shown to the researcher, only one of which is the expert-elicited model in every respect. Experts can be asked to choose the correct structure, discretisation or parameterisation from either a set of models of through binary choice experiments in which every model is compared to every other model. As in the case of face validity, the Turing test is ideally carried out on a separate set of experts to the set that originally created the model to avoid criterion contamination. The fewer differences in the final model chosen to the expert-elicited network, the higher the discriminant validity of that source of uncertainty. For this framework, the following questions are suggested as tests of the discriminant validity of the BN model:

- How different is the model structure to other models that are nomologically distal?
- How different is the discretisation of each node to the discretisation of nodes that are nomologically distal independent of their network domain?
- Are the parameters of nodes in the comparison models that have oppositional definitions to the node in question parameterised differently?
- When presented with a range of plausible models, can experts choose the 'correct' model or set of models?

3.7. Predictive Validity

In BNs, predictive validity can be considered to encompass both the model behaviour and the model output. This is the type of validity covered by traditional model and data fitting techniques.

When applying predictive validity tests within a complex systems and specifically a BN paradigm, the comparison model can be an alternative hypothesised model rather than a data-driven model. Such hypothesised models could be elicited using a number of techniques, such as case studies or formal walkthroughs (Barlas, 1996; Pollino et al., 2007). Luu et al. (2009) used case studies to formulate alternative hypothetical networks against which to compare the predictive validity of their BN model. While they did not specifically apply the tests presented in this paper, their work represents one of few papers to attempt to establish confidence in the predictive validity of an expert-elicited BN. Half of the special tests of system dynamics model validity presented by Barlas (1996) refer to the predictive validity of the model in that they test the model behaviour specifically. Of particular relevance to establishing confidence in the predictive validity of BN are behaviour sensitivity tests, Qualitative Features Analysis and the extreme conditions tests. When applied within a BN paradigm, the behaviour sensitivity test can be applied to the model structure and parameters by determining to which factors and relationships the model is sensitive, and comparing this to hypothetical models or alternative empirical models. The terms 'sensitivity to parameters' and 'sensitivity to findings' are used by Pollino et al. (2007) to describe the application of behaviour sensitivity tests to the parameters and model behaviour specifically, however it should be noted that this test can be just as easily applied to the structure and discretisation of nodes in the model as well. These tests are commonly used, and various versions of them can be executed using the GeNiE 2.0 (DSL, 2007), Hugin Expert (Andersen et al., 1989) or Netica (Norsys, 2007) software packages among others.

Qualitative features analysis (Carson and Flood, 1990) is a case of predictive validity testing where behaviour in a hypothetical model is compared to the behaviour of individual pairs of nodes, sub-networks and the entire model. As in the case of predictive validity, the hypothetical models can be achieved through a number of formal strategies; however in this case, we are interested in the comparison of simulation output rather than comparison of model features directly. It is for this reason that model behaviour is outlined as the fourth source of model uncertainty. While this area is the product of the uncertainty of its component features, predictive validity requires that model behaviour be simulated from the model for tests to occur. For this reason, predictive validity should be the final type of validity to be tested. Finally, the extreme conditions test can be seen as a special case of qualitative features analysis, as it sets the hypothetical model to extreme conditions where the behaviour of the model is more predictable (Forrester and Senge, 1980). For example, if the number of passengers is set to 0 then the model should reflect that there is a probability of 1 that 0 passengers are processed within the time range of interest. The direct extreme conditions test examines the behaviour of individual pairs of nodes and sub-networks under such extreme conditions, while the indirect extreme conditions test examines the behaviour of the entire network against such hypotheses. The range of tests to establish confidence in the predictive validity of a model is notable considering the issue at hand that true objective data on the model are not available, and suggests that the lack of data available does not preclude predictive validity testing, as hypothesis-driven models can be used in place of data-driven models. From examination of the various techniques associated with assessing predictive validity, we arrive at the following set of questions:

- Is the model behaviour predictive of the behaviour of the system being modelled?

- Once simulations have been run, are the output states of individual nodes predictive of aspects in the comparison models?

- Is the model sensitive to any particular findings or parameters to which the system would also be sensitive?

- Are there qualitative features of the model behaviour that can be observed in the system being modelled?
- Does the model including its component relationships predict extreme model behaviour under extreme conditions?

4. Conclusions and Recommendations

In this paper we have outlined a broad range of conceptual tests that can be applied to validate BNs. These validity tests incorporate standard model-data fit comparisons, but expand the construct of validity to the broader definition of whether or not a model describes the system it is intended to describe, and produces output it is intended to produce. Many of these validity tests can be used where no objective data exist.

By combining existing research from BN validation with validation tests from psychometrics as well alternative complex systems disciplines, this paper introduces a starting point for discussing a framework for building confidence in the validity of BNs. The presented framework is not intended to be comprehensive; instead, the aim is to establish that the validity of a BN can be tested, and should be tested, independent of the model fit to available data or expert confirmation. Disciplines such as psychometrics, with a history of measuring latent constructs, can provide a useful perspective on the problem.

The framework presents a sequence of steps that can be followed to establish confidence in model validity, beginning with creating a nomological map of the literature surrounding the domain, then gradually building confidence in six types of model validity, using both general and specific tests.

The application of this framework to the BN developed in conjunction with

ACBPS will to our knowledge be a novel practical demonstration of such an approach to BN validation. The framework presented in this paper is intended to be domain-general, and there would be great value in establishing the versatility of the tests by applying them to complex models in other domains. Future work will extend to formalising and quantifying many of the tests in the context of BN modelling, and obtaining perspectives on model validity from other disciplines that deal with unobserved variables and complex systems.

5. References

S.K. Andersen, K.G. Olesen, F.V. Jensen, and F. Jensen. Hugin - a shell for building bayesian belief universes for expert systems. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pages 1080–1085, United States of America, 1989. MIT press.

Y. Barlas. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, 12(3):183–210, 1996.

S. Barrett, P. Whittle, K. Mengersen, and R. Stoklosa. Biosecurity threats: the design of surveillance systems, based on power and risk. *Environmental and ecological statistics*, 17:503–519, 2010.

E.R. Carson and R.L. Flood. Model validation: philosophy, methodology and examples. *Transactions of the Institute of Measurement and Control*, 12:178–185, 1990.

J. Cheng, D.A. Bell, and W. Liu. An algorithm for bayesian belief network construction from data. In *Proceedings of Conference on Artificial Intelligence and Statistics*, pages 83–90, United States, 1997.

L.J. Cronbach and P.E. Meehl. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302, 1955.

J. Darkes, P.E. Greenbaum, and M.S. Goldman. Sensation seekingdisinhibition and alcohol use: Exploring issues of criterion contamination. *Psychological Assessment*, 10:71–76, 1998.

DSL. Genie and smile, 2007. Bayesian Network Modelling software package and decision platform.

J.W. Forrester and P.M. Senge. Tests for building confidence in system dynamics models. *TIMS studies in the management sciences*, 14:209–228, 1980.

D. Geiger, T. Verma, and J. Pearl. d-separation: From theorems to algorithms. In *Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence*, UAI '89, pages 139–148, The Netherlands, 1990. North-Holland Publishing Co.

D. Heckerman, D. Geiger, and D.M. Chickering. Learning bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20: 197–243, 1995.

A. James, S. Low Choy, and K. Mengersen. Elicitator: An expert elicitation tool for regression in ecology. *Environmental modelling and Software*, 25: 129–145, 2010.

S. Johnson. *Integrated Bayesian network frameworks for modelling complex ecological issuesy*. PhD thesis, Queensland University of Technology, Australia, 2009.

S. Johnson, F. Harding, G. Hamilton, and K. Mengersen. An integrated bayesian network approach to lyngbya majuscula bloom initiation. *Marine Environmental Research*, 69:27–37, 2010.

D. Koller and A. Pfeffer. Object-oriented bayesian networks. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence*, JMLR workshop and conference proceedings, pages 302–313, United States, 1997.

K.B. Korb and A.E. Nicholson. *Bayesian Artificial Intelligence*, page 462 pp. CRC Press, United Kingdom, 2010.

S. Low Choy, R. O'Leary, and K. Mengersen. Elicitation by design in ecology: using expert opinion to inform priors for bayesian statistical models. *Ecology*, 90:265–277, 2009.

V. Luu, S. Kim, N. Tuan, and S. Ogunlana. Quantifying schedule risk in construction projects using bayesian belief networks. *International Journal of Project Management*, 27:39–50, 2009.

M. Masoli, D. Fabian, S. Holt, and R. Beasley. The global burden of asthma: executive summary of the gina dissemination committee report. *Allergy*, 59(5):469–478, 2004.

P. Myllymaki, T. Silander, H. Tirri, and P. Uronen. B-course: A web-based tool for bayesian and causal data analysis. *International Journal on Artificial Intelligence*, 11(3):369–387, 2002.

Norsys. Netica, 2007. Proprietary Bayesian Network Modelling software package.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

C.A. Pollino, O. Woodberry, A. Nicholson, K. Korb, and B.T. Hart. Parameterisation and evaluation of a bayesian network for use in an ecological risk assessment. *Environmental Modelling and Software*, 22:1140–1152, 2007.

S. Renooij. Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review*, 16(3):255–269, 2001.

L.W. Schruben. Establishing the credibility of simulations. *Simulation*, 34:101–105, 1980.

O. Shakoor, R.B. Taylor, and R.H. Behrens. Assessment of the incidence of substandard drugs in developing countries. *Tropical Medicine and International Health*, 2(9):839–845, 1997.

T. Silander, T. Ross, and P. Myllymaki. Locally minimax optimal predictive modeling with bayesian networks. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS) 2009*, JMLR workshop and conference proceedings, pages 504–511, United States, 2009.

R.S. Sobel and B.J. Osoba. Youth gangs as pseudo-governments implications for violent crime. *Southern Economic Journal*, 75(4):996–1018, 2009.

W.M. Trochim. Research methods knowledge base, 2001. URL `http://www.socialresearchmethods.net/kb/index.htm`.

L. Uusitalo. Advantages and challenges of bayesian networks in environmental modelling. *Ecological Modelling*, 203(3), 2007.

O. Woodberry, A. Nicholson, K. Korb, and C Pollino. Parameterising bayesian networks. In Geoffrey Webb and Xinghuo Yu, editors, *AI 2004: Advances in Artificial Intelligence*, volume 3339 of *Lecture Notes in Computer Science*, pages 711–745. Springer Berlin / Heidelberg, 2005.