Lessons from modelling an airport terminal as a complex system:

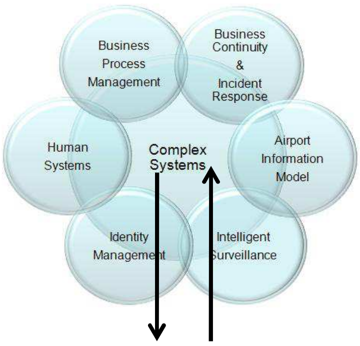
A proposed validation framework for expert elicited Bayesian Networks

Jegar Pitchforth
Kerrie Mengersen
Paul Wu
Sandra Johnson

Queensland University of Technology

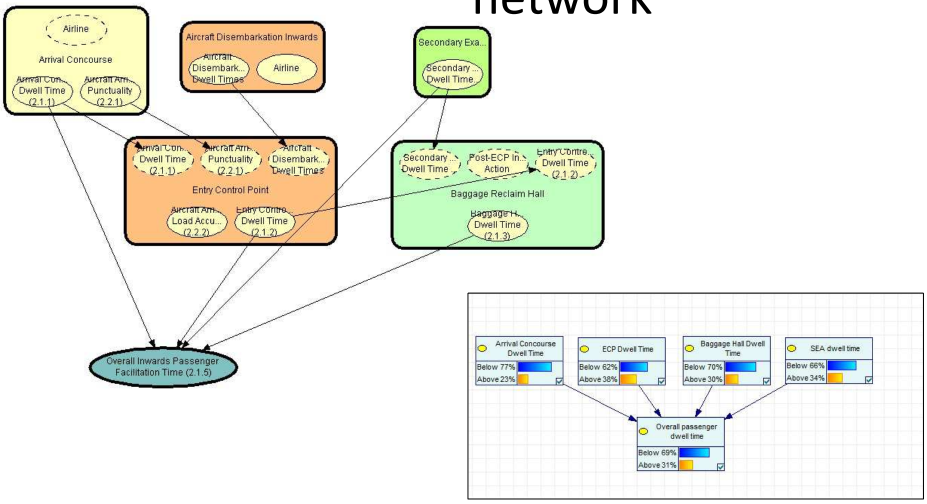
Airports of the Future

- 4 year ARC linkage project aiming to “improve airport effectiveness and cultivate flexibility for the sustained growth of airport operations.” (AOTF website)
- Includes 6 universities, 2 major airlines, 12 Australian airports, 5 government agencies and 5 service providers.
- This research is a part of the Complex Systems program to incorporate research from other groups.


Business Process Management
Business Continuity & Incident Response
Human Systems
Complex Systems
Airport Information Model
Identity Management
Intelligent Surveillance

Inbound passenger process

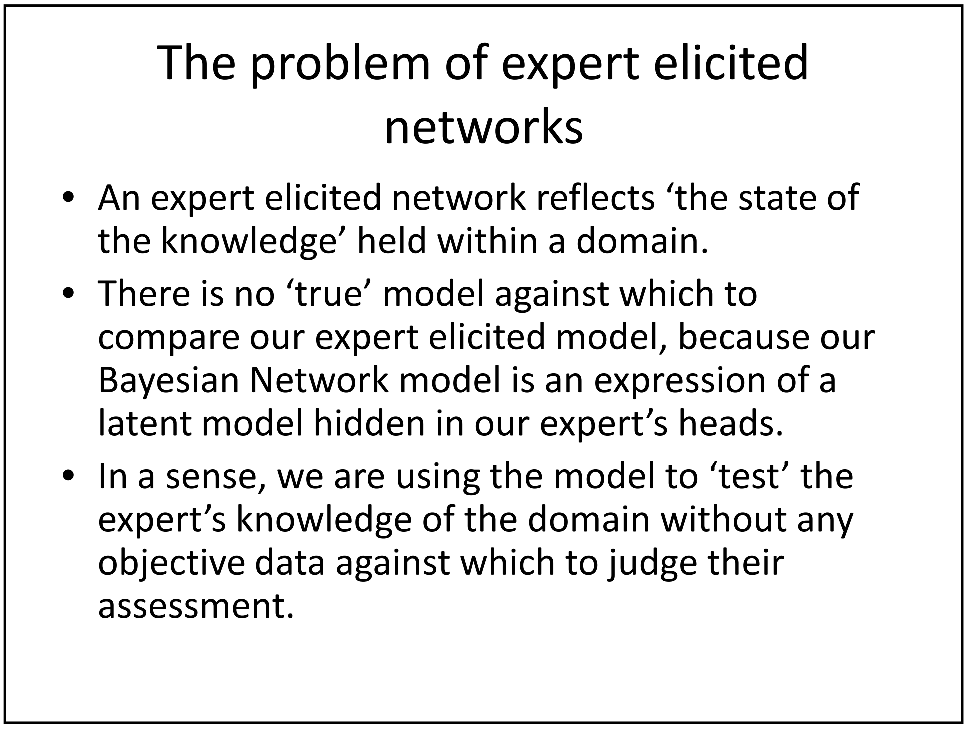
The inbound passenger facilitation network


Airline
Arrival Concourse
Arrival Con... Dwell Time (2.1.1)
Aircraft Arri... Punctuality (2.2.1)
Aircraft Disembarkation Inwards
Aircraft Disembark... Dwell Times
Airline
Secondary Exa...
Secondary ... Dwell Time...
Arrival Con... Dwell Time (2.1.1)
Aircraft Arri... Punctuality (2.2.1)
Aircraft Disembark... Dwell Times
Entry Control Point
Aircraft Arri... Load Accu... (2.2.2)
Entry Contro... Dwell Time (2.1.2)
Secondary ... Dwell Time ...
Post-ECP In... Action
Entry Contro... Dwell Time (2.1.2)
Baggage Reclaim Hall
Baggage H... Dwell Time (2.1.3)
Overall Inwards Passenger Facilitation Time (2.1.5)
Arrival Concourse Dwell Time
Below 77%
Above 23%
ECP Dwell Time
Below 62%
Above 38%
Baggage Hall Dwell Time
Below 70%
Above 30%
SEA dwell time
Below 66%
Above 34%
Overall passenger dwell time
Below 69%
Above 31%

The Question

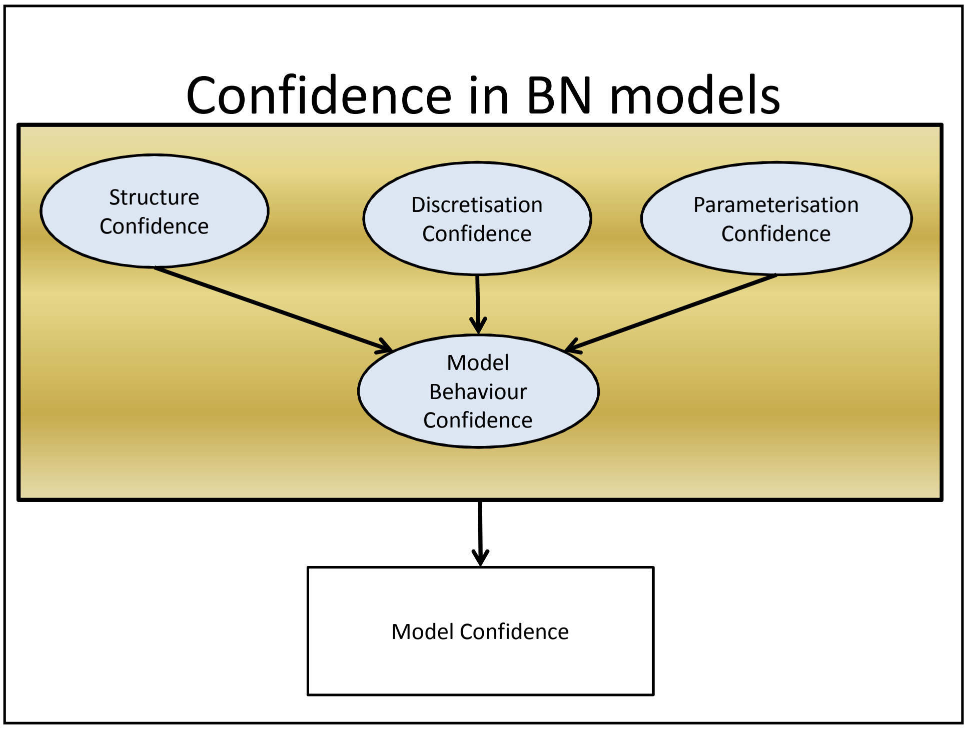
- The inbound passenger facilitation network model is expert elicited in part, and data driven in others.

How can we tell if the interesting behaviours our model depicts are valid representations of the system, or simply mistakes in model design and elicitation?



The problem of expert elicited networks

- An expert elicited network reflects ‘the state of the knowledge’ held within a domain.
- There is no ‘true’ model against which to compare our expert elicited model, because our Bayesian Network model is an expression of a latent model hidden in our expert’s heads.
- In a sense, we are using the model to ‘test’ the expert’s knowledge of the domain without any objective data against which to judge their assessment.



Confidence in BN models

Structure Confidence

Discretisation Confidence

Parameterisation Confidence

Model Behaviour Confidence

Model Confidence

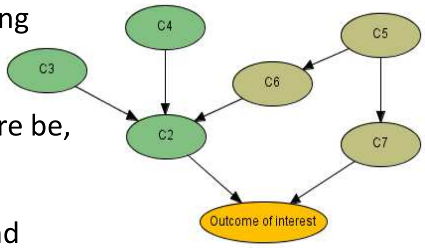
Structure

– How many nodes should there be?

– Have we included everything that is relevant?

– How many arcs should there be, and in which direction?

– How are nodes defined, and does this remain logically consistent throughout the network?



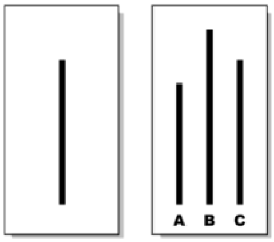
Discretisation

– How many states should we assign to each node?
  - Trade off between utility, representativeness, data and computational ability.
  - Even large datasets often cannot support a high number of intervals per variable - Uusitalo (2007)
  - *“No suitable automatic discretisation methods have been found” – Myllymaki (2002)*

– What are the most appropriate labels for node states?

Parameterisation

- How should we interpret the answers we've been given?
  - There are many known issues with human capabilities of judging probabilities (Renooij, 2001).
- How do we handle outlier inputs?
- How can we establish that the various expert opinions have been combined in a valid way?
  - Experiments almost 60 years ago show that individual opinions can be skewed by the group opinion in interesting but often predictable ways. (Crano, 2000).

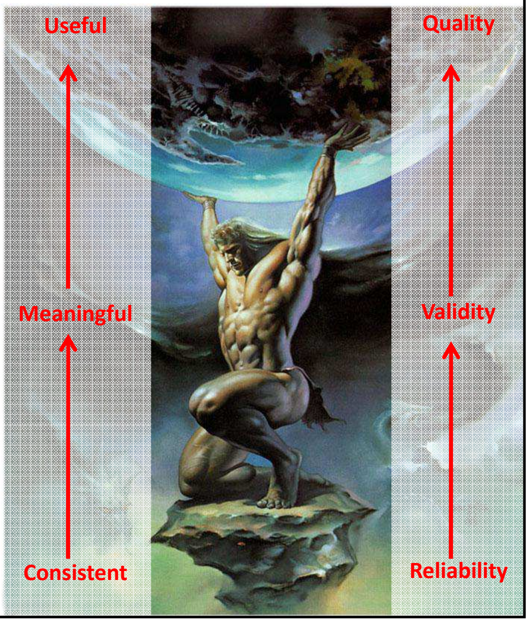


Model Behaviour

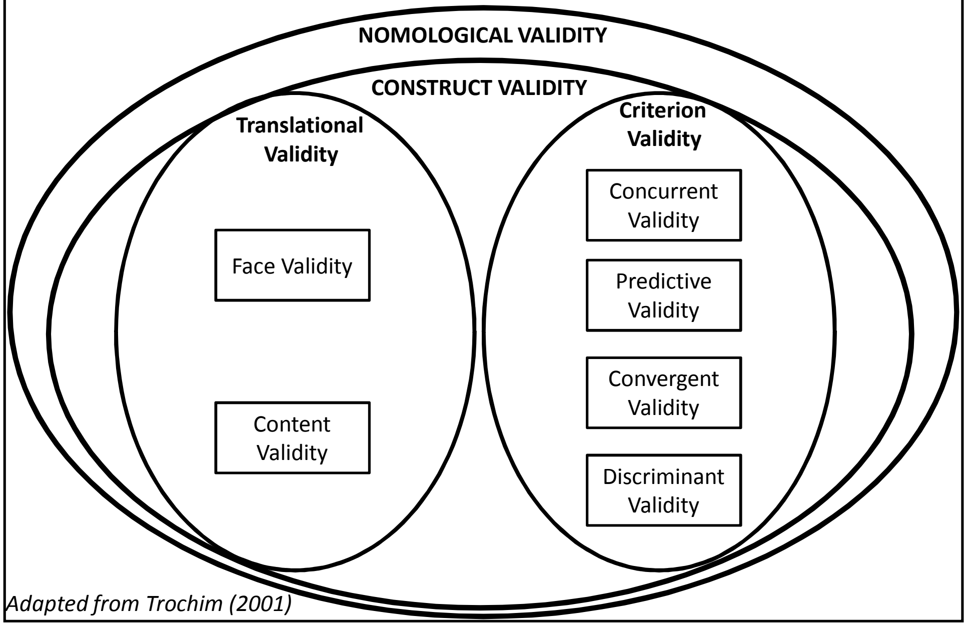
- How do we judge the quality of the model output?
- Where the first three areas are concerned with *how* a model works, this area is regarding *what* the model produces.
  - This is the fun bit, and as such we usually pay the most attention to it!
- Rather than being concerned with individual relationships, when validating BN's we are primarily interested in patterns of model behaviour.
  - Patterns of behaviour could be sharp rises and falls, undulating or cyclical patterns, stepwise rise and fall and more.

Reliability, Validity and Quality

- Reliability and validity are widely discussed, but in Bayesian Networking we are also interested in the quality of the model to the user.
- Before we can establish quality, we must establish confidence in the validity of the model.
- While the approach we take acknowledges that no ‘true’ model may exist, this doesn’t excuse us from establishing confidence in the network!
- In the context of a BBN, confidence in model quality is based on the weight of theoretical and logical evidence.



The psychometric validity framework



NOMOLOGICAL VALIDITY

CONSTRUCT VALIDITY

Translational Validity

Face Validity

Content Validity

Criterion Validity

Concurrent Validity

Predictive Validity

Convergent Validity

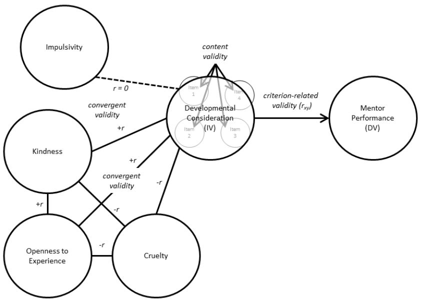
Discriminant Validity

Adapted from Trochim (2001)

Nomological validity (Messick 1980)

"Does the model fit logically within a wider domain?"

- A nomological network defines related themes and constructs from within the literature with or without causal or directional inference.
- If a model is nomologically valid, we can place the factors and relationships referenced within a wider network of literature.
- This not only helps us to determine whether the literature broadly supports our BN, but also helps outline criteria for other validity measures.



Face validity

"Does the model look like we expect it to look?"

-Structure: At a first glance, are all the nodes we consider important included, are all the relevant relationships represented by arcs?

-Discretisation: Do all the states of the variables look how they are expected to look? Do the number and names of the states look right?

-Parameterisation: Do all the input values look about right? Do they say what the expert meant to say?

Face validity is a weak indicator of model validity, but is commonly used exclusively, or as the primary method.

Content Validity

"Does the model represent the complete domain acknowledged in literature and theory?"

-Structure: Are all the factors suggested by the literature included in the model? Have we included all the relationships that are suggested by the literature or theory?

-Discretisation: Within each node, are the intervals included that are suggested by the theory?

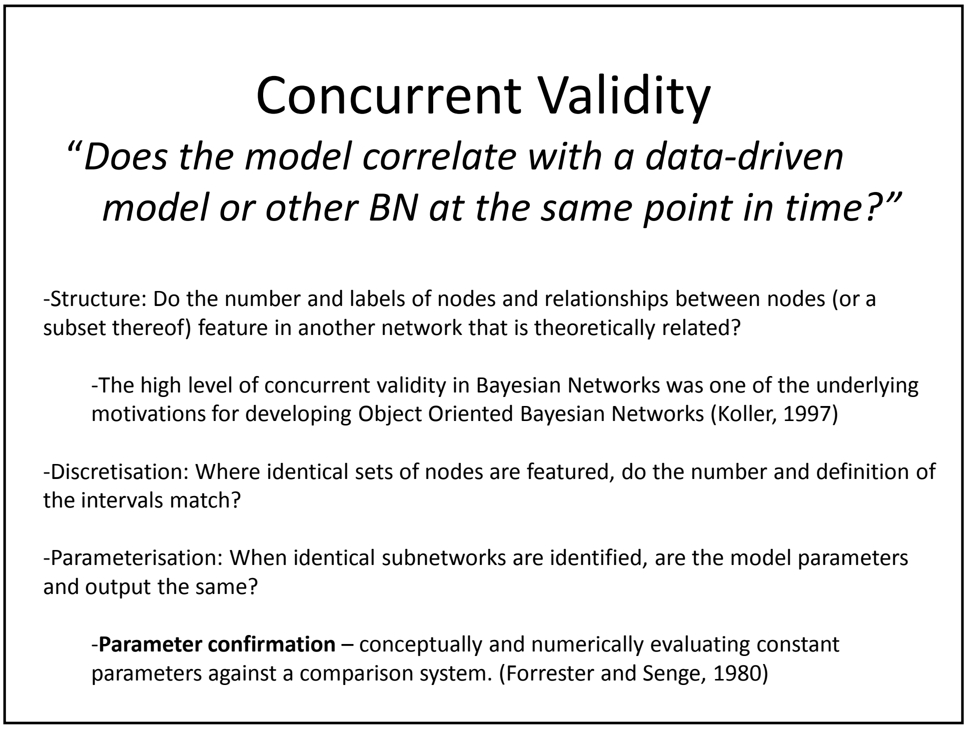
- **Dimensional consistency** – testing that all intervals in the state space are described within the model node.

-Parameterisation: How close are the input parameters of the nodes to those suggested by the literature?

-Refers to both the relevance and the coverage of the factors in the model.

The lessons from machine learning

- Machine learning approaches are advanced in the area of *learning Bayesian Networks* (Korb & Nicholson, 2010).
  - When there is a true dataset (either observed or simulated), expert elicited models can be treated as prior belief models to generate a posterior probability that the resulting network B_i is a representation of the true network.
  - Using this process as an inspiration, we can work backwards and use a version of sensitivity testing to identify factors that contribute the least to the model.



Concurrent Validity

"Does the model correlate with a data-driven model or other BN at the same point in time?"

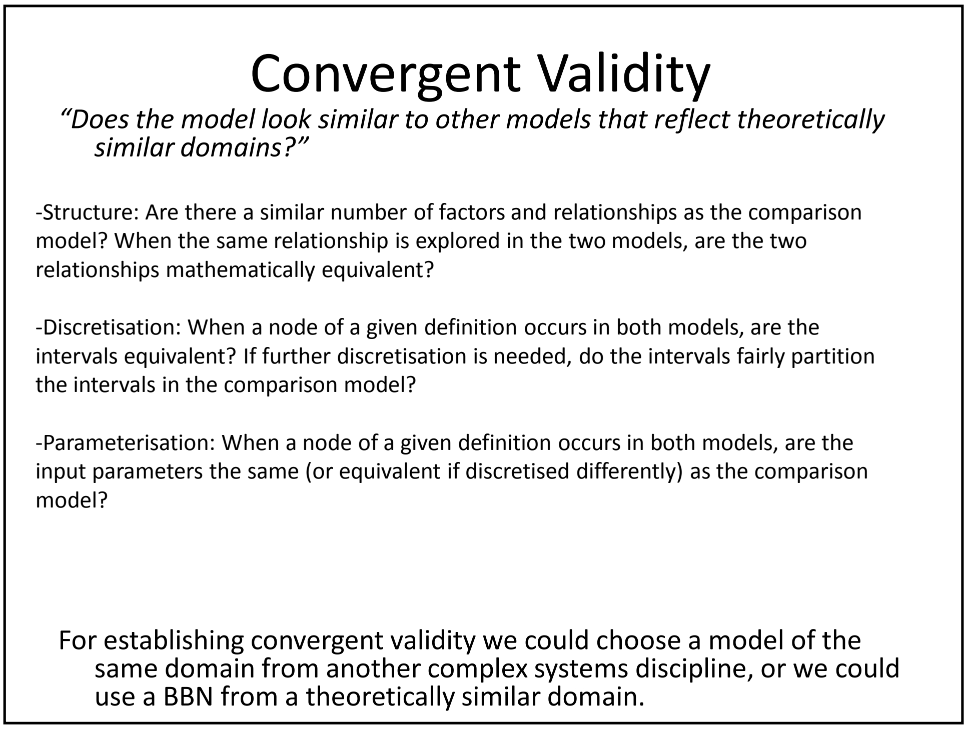
-Structure: Do the number and labels of nodes and relationships between nodes (or a subset thereof) feature in another network that is theoretically related?

-The high level of concurrent validity in Bayesian Networks was one of the underlying motivations for developing Object Oriented Bayesian Networks (Koller, 1997)

-Discretisation: Where identical sets of nodes are featured, do the number and definition of the intervals match?

-Parameterisation: When identical subnetworks are identified, are the model parameters and output the same?

-**Parameter confirmation** – conceptually and numerically evaluating constant parameters against a comparison system. (Forrester and Senge, 1980)



Convergent Validity

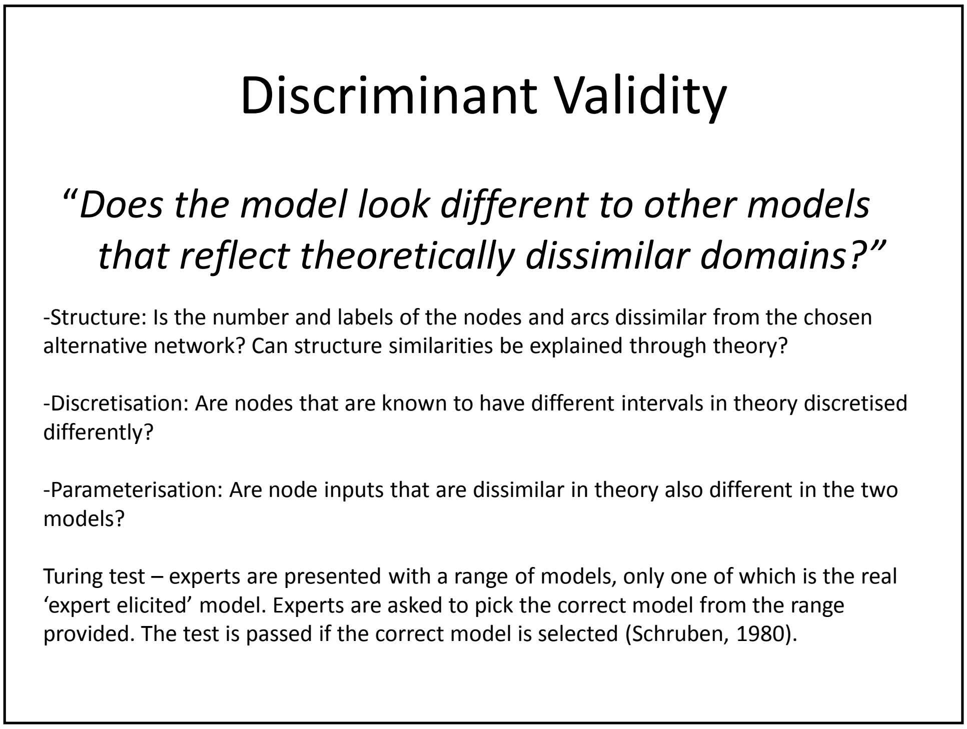
"Does the model look similar to other models that reflect theoretically similar domains?"

-Structure: Are there a similar number of factors and relationships as the comparison model? When the same relationship is explored in the two models, are the two relationships mathematically equivalent?

-Discretisation: When a node of a given definition occurs in both models, are the intervals equivalent? If further discretisation is needed, do the intervals fairly partition the intervals in the comparison model?

-Parameterisation: When a node of a given definition occurs in both models, are the input parameters the same (or equivalent if discretised differently) as the comparison model?

For establishing convergent validity we could choose a model of the same domain from another complex systems discipline, or we could use a BBN from a theoretically similar domain.



Discriminant Validity

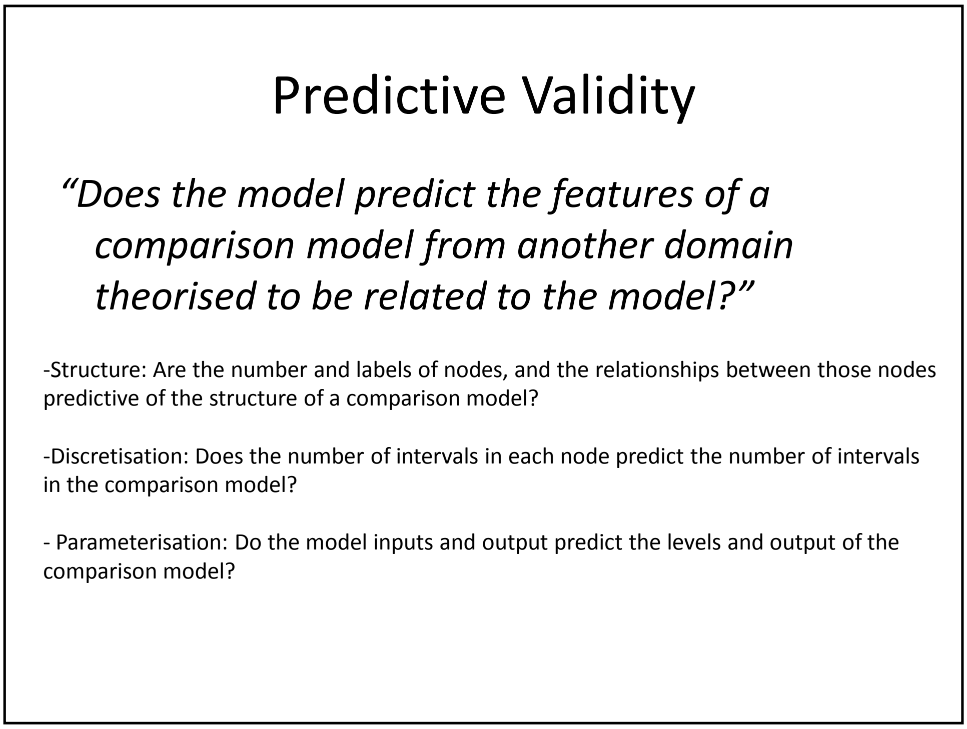
"Does the model look different to other models that reflect theoretically dissimilar domains?"

-Structure: Is the number and labels of the nodes and arcs dissimilar from the chosen alternative network? Can structure similarities be explained through theory?

-Discretisation: Are nodes that are known to have different intervals in theory discretised differently?

-Parameterisation: Are node inputs that are dissimilar in theory also different in the two models?

Turing test – experts are presented with a range of models, only one of which is the real 'expert elicited' model. Experts are asked to pick the correct model from the range provided. The test is passed if the correct model is selected (Schruben, 1980).



Predictive Validity

"Does the model predict the features of a comparison model from another domain theorised to be related to the model?"

-Structure: Are the number and labels of nodes, and the relationships between those nodes predictive of the structure of a comparison model?

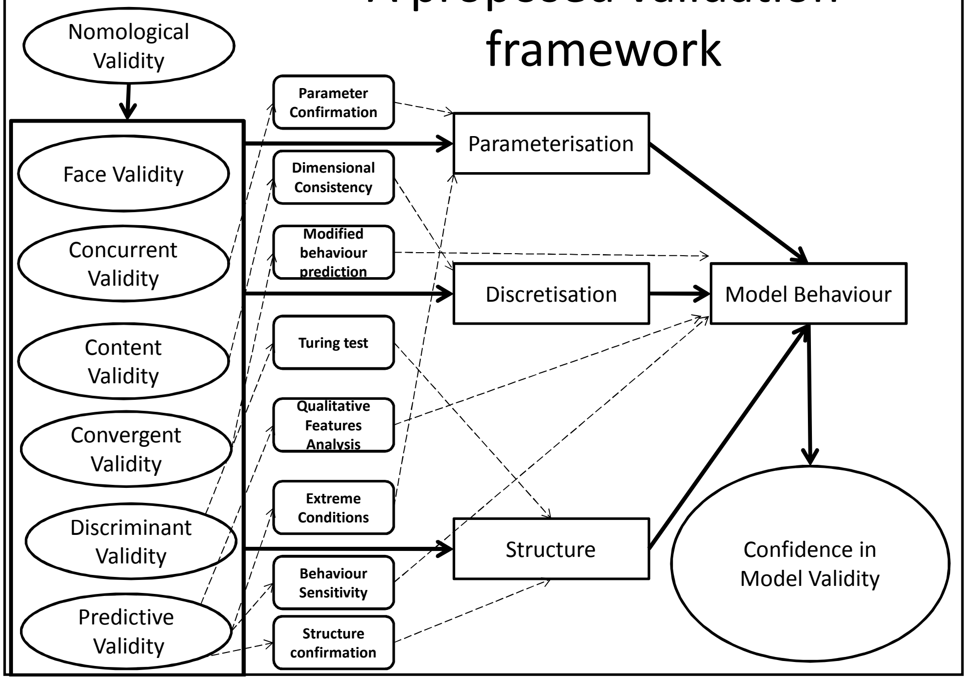
-Discretisation: Does the number of intervals in each node predict the number of intervals in the comparison model?

- Parameterisation: Do the model inputs and output predict the levels and output of the comparison model?

Complex Systems Tests of Predictive Validity

- When discussing expert elicited complex systems models generally, the comparison model can be a hypothesised model derived from another set of experts from the same domain or derived from a logical foundation.
- Hypothesised models could be derived from a number of techniques, such as case studies and formal walkthroughs (Barlas, 1990).
  - Luu et. al. (2009) used case studies as their method of validating a construction industry BBN.
- **Qualitative Features Analysis** from Systems Dynamics (Carson and Flood, 1990) identifies behaviour from the hypothesised model and compares it to behaviour simulated by the elicited model.
- The **direct and indirect extreme conditions tests** (Forrester and Senge, 1980) are powerful tests where we can hypothesise a model perfectly under extreme conditions.
- **Behaviour sensitivity tests** determine to which factors the model is highly sensitive, and compare this to a set of hypothesised models.
- **Structure Confirmation** compares the form of the equations in the model with the hypothesised set of models (Forrester and Senge, 1980).
- **Modified-behaviour prediction**: if data or better knowledge exists on a modified version of the system, then the BBN behaviour can be compared to the modified behaviour (Barlas 1989).

A proposed validation framework



Nomological Validity

Face Validity

Concurrent Validity

Content Validity

Convergent Validity

Discriminant Validity

Predictive Validity

Parameter Confirmation

Dimensional Consistency

Modified behaviour prediction

Turing test

Qualitative Features Analysis

Extreme Conditions

Behaviour Sensitivity

Structure confirmation

Parameterisation

Discretisation

Structure

Model Behaviour

Confidence in Model Validity

Conclusions

- Despite the proposed framework being less than comprehensive, the broad range of theoretical and mathematical tests to establish confidence in the validity of a BBN model has been demonstrated.
- Rather than waving off validation of BBNs as 'too difficult', we can look to disciplines with similar problems and approaches for ideas.
- Each of the presented types of validity has demonstrable applications to all four points of uncertainty in the model, as well as containing special tests for specific points.
- In future we could look at further strategies for improving model confidence using these tests as criterion.
- The lack of research in this area suggests that there are further validation tests that can be designed for use in any of the four areas for measurement of uncertainty in BN modelling.

References

- S. Messick. 1980. Test Validity and the Ethics of Assessment, *American Psychologist*, 35(11), 1012-1027.
- J.W. Forrester and P.M. Senge. 1980. Tests for building confidence in system dynamics models. *TIMS studies in the management sciences, 14,* 209--228.
- S. Renooij. 2001. Probability elicitation for belief networks: issues to consider. *The Knowledge Engineering Review,* 16(3), 255 – 269.
- L. Uusitalo. 2007. Advantages and challenges of bayesian networks in environmental modelling. *Ecological Modelling, 203(3).*
- Myllymaki, P., Silander, T., Tirri, H. & Uronen, P. 2002. B-Course: A web-based tool for Bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools.* 11(3), 369 – 387.
- W. Crano & R. Prislin. 2006. Attitudes and Persuasion.*Annual Review of Psychology*, 57, 345—374.
- W.M. Trochim. 2001. Research methods knowledge base. http://www.socialresearchmethods.net/kb/index.htm.
- D. Koller and A. Pfeffer. 1997.Object-oriented bayesian networks. In *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence, JMLR workshop and conference proceedings,* 302 -- 313, United States.
- Y. Barlas. 1990. Formal aspects of model validity and validation in system dynamics. *System Dynamics Review, 12(3),*183--210.
- E.R. Carson and R.L. Flood. 1990. Model validation: philosophy, methodology and examples. *Transactions of the Institute of Measurement and Control, 12,* 178--185.
- L.W. Schruben. 1980. Establishing the credibility of simulations. *Simulation,* 34, 101-- 105.
- V. Luu, S. Kim, N. Tuan, and S. Ogunlana. 2009. Quantifying schedule risk in construction projects using bayesian belief networks. *International Journal of Project Management, 27,* 39--50.