18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Expert opinion in statistical models

Kuhnert, P.M. [1], **K.Hayes**[1], **T. G. Martin**[2] **and M. F. McBride**[3]

[1] *CSIRO Mathematical and Information Sciences, Brisbane and Hobart Laboratories*
[2] *CSIRO Sustainable Ecosystems, QBP, St Lucia*
[3] *Australian Centre of Excellence in Risk Analysis, University of Melbourne*
Email: Petra.Kuhnert@csiro.au

Abstract:

The incorporation of expert knowledge in ecological models is gaining prominence in ecology as it provides a quick and inexpensive alternative to experimental data and in situations where both data and expert opinion have been obtained, expert opinion can help to moderate the experimental data, offering insights into the problem that could have otherwise gone unnoticed.

The use of expert knowledge usually arises in one of two forms: (1) knowledge regarding the model structure for understanding the system process; and, (2) prior information concerning the parameters of the model which may arise from an elicitation process involving one or more experts. More effort is generally expended on the latter than the former, particularly for statistical models where the nature of any data that has been collected usually dictates the form of the model. In most cases, however, data collection designs are guided by a conceptual model, hence data-dictated statistical models are not necessarily immune from the effects of model structure error. Few studies attempt to address both structural and parametric issues in their models.

Expert opinion influences all stages of model conception, construction and parameterisation. Expert opinion is prone to a range of cognitive limitations but the effects of this are not always obvious. Careless use of experts during these modelling phases can result in inaccurate predictions stemming from a biased or inaccurate perception of what one or more experts believe to be the truth. Issues such as overconfidence, representativeness, translation and linguistic uncertainty can lead experts to provide false or misleading opinions. We argue that this usually occurs not because the expert intends to provide a false or misleading response but because insufficient care has been taken to avoid these issues during the elicitation process. Misunderstandings in the question being asked and the interpretation of what is fed back to the expert can result in an expert providing an inaccurate view of real world processes. Furthermore, in many situations it is not immediately clear whether an expert's interval around an estimate reflects their knowledge of the parameter's true variability or simply the expert's ignorance of its value.

Through three real examples, we investigate the effects of different elicitation procedures and highlight issues that lead to a biased response from the expert/s. These examples highlight the need to:

- conduct the elicitation process in a non-threatening manner so that experts feel comfortable when responding to questions;
- use multiple experts as opposed to a single expert (where possible) to avoid some potential biases such as overconfidence;
- include within the elicitation process a feedback, comparison and reflection stage that allows experts to discuss, consider and revise their initial opinions;
- include a calibration mechanism to ensure the translation of opinions was interpreted correctly in the context of the statistical model;
- where multiple experts are used, ensure pooling is conducted such that variability and ignorance are identified and separated; and
- design the elicitation process to ensure what is elicited can be incorporated into the structure of the model used to solve the problem and the impact of prior information in models is examined.

Keywords: *Bias; Elicitation; Priors; Bayesian modelling.*

1. INTRODUCTION

The incorporation of expert knowledge in ecological models is gaining prominence in ecology. Recent examples by Low-Choy et al. (2009), O'Leary et al. (2008), Griffiths et al. (2007), Kuhnert et al. (2005), Martin et al. (2005) and Kynn (2004) show how expert opinion is being incorporated into models to support natural resource management decisions. The reason for this rise in popularity is that expert opinion offers a relatively inexpensive, quick and efficient alternative to experimental data when constructing and parameterising a model. Furthermore, expert opinion can moderate the experimental data, and provide posterior estimates that could not have been observed using the data alone, particularly where data is limited.

Expert information enters model-based decision support systems in two ways: (i) in the development of the conceptual model of the problem in hand; and (ii) when the conceptual model is translated into a quantitative model, choosing the relationships between variables in the model and parameterising these relationships. These are two important steps that are often overlooked when using expert opinion in statistical models. Sometimes the structure of the model is considered without much thought to its parameterisation or how expert information can be appropriately translated into priors for that model. Alternatively, potential errors in the conceptual model are overlooked because the analyst focuses on how to parameterise the statistical model, ensuring it conforms well to (for example) the experimental data and statistical assumptions inherent in the model itself. Sometimes errors and inconsistencies in the expert's opinions and interpretations of the problem are not identified until the analyst attempts to translate them and incorporate them into the model. This usually occurs well after the elicitation process and may not therefore be readily corrected. Despite these potential problems few studies concentrate on addressing all of these issues in an elicitation and statistical modelling context.

It is a well documented, but still debated fact that experts are prone to a range of cognitive limitations (Kynn, 2008; O'Hagan et al. 2006) and therefore the introduction of expert opinion in statistical models needs to be carefully considered. Among the many types of limitations (Table 1) overconfidence, whereby the expert systematically overestimates the accuracy of his/her beliefs, or rather systematically underestimates the uncertainty in a process or its inherent variability, is arguably the most dangerous in decision support contexts (Kynn 2008). This can lead to a very informative but inaccurate prior as demonstrated by Kuhnert et al. (2009) and Griffiths et al. (2007). Depending on how expert information is captured and incorporated into a statistical model, "overconfidence" can have a substantial impact on estimates from the model irrespective of how much data is collected and recent research efforts have therefore focused on elicitation processes that are specifically designed to minimise this, and other cognitive limitations (O'Hagan et al. 2006, Burgman 2005). Overconfidence in model structure, however, often takes a back seat in this process.

In this paper we focus on the elicitation process used to elicit the expert information and how careless use of this information can lead to biases in the modelling and interpretation of the results. We stress that experts are likely to "get it wrong" if the information is not carefully elicited, interpreted and incorporated into the model. Except in very rare instances, this is not because experts deliberately seek to provide false, misleading or biased response to the question being asked, but rather because of their inherent cognitive limitations. We begin with a discussion of three real examples and show how subjective judgment can alter the interpretation of a model if not incorporated carefully.

Issues	Interpretation
Overconfidence/Conservatism	Overestimating the accuracy of his/her beliefs or alternatively underestimating the uncertainty in a process. Conservatism relates to the process of an expert understating their belief.
Representativeness	Providing opinions that are based on situations that are (wrongly or rightly) perceived to be similar.
Availability	Basing a response on most recent available information and not considering past events.
Anchoring and Adjustment	The tendency for groups to anchor around (any) initial estimates and adjust their final estimate from this value irrespective of the initial estimates accuracy
Misunderstanding of conditional probabilities	Confusion regarding the definition of conditional probability and failure to adhere to the axioms of conditional probability.
Translation	Confusion regarding the translation of a response to another scale
Affect	Expert's emotions entering into the judgment making.
Hindsight Bias	Expert places too much emphasis on past events and outcomes.
Law of Small Numbers	Expert bases their opinion on small pieces of information and assumes that this extrapolates to the population.
Linguistic Uncertainty	Misunderstanding the question and/or applying different interpretations to the same term.

Table 1. A summary of some of the key heuristics, judgments or mental operations that can result in bias when eliciting information from experts.

2. ELICITATION, ANALYSIS AND COGNITIVE BIAS

Extensive reviews of elicitation methods are described elsewhere (Kuhnert et al. 2009, Low-Choy et al. 2009, Kynn 2008, O'Hagan et al. 2006 & Garthwaite et al. 2005). Of primary importance in these reviews is the potential bias in an elicited response and how careless use of expert information can impact a model. There is a large body of literature on the cognitive limitations of experts mainly within the psychology literature. There have been relatively few reviews on this subject in the specific context of elicitation for statistical models (see however, Kynn 2008). Here we focus on illustrating the effect of some of the cognitive limitations presented in Table 1 using 3 case studies and provide advice on overcoming these issues.

2.1 Eliciting the probability of capture: overconfidence and hindsight bias

The Problem: Estimating the abundance of pelagic fish (e.g. tunas and mackerels) is a challenging task because they are fast swimming visual pursuit predators that feed at many different trophic levels. Standard methods such as trawling are not appropriate for surveying fish of this type. Methods which use a passive form of capture through gillnets for example are more appropriate. However gillnets are highly selective and depend on the mesh size for trapping fish by their gills in the net. The effective area fished by the net is also unknown and as a result, the catch cannot be expressed as a density.

To solve this problem we needed to incorporate net length, soak time (length of time the net was set in the water), fish swim speed and net selectivity to estimate abundance. Using 208 gillnet sets and making some geometric assumptions about the potential domain of interaction (an area that fish can swim in to have any chance of reaching the net) we were able to construct a statistical model to estimate the number of fish per unit area. See Griffiths et al. (2007) for specific details relating to the model. In constructing this model, two priors needed to be incorporated. The first was a prior for the swimming speed of fish which was based on previously published studies. The second and most controversial was a prior representing net selectivity which needed to be elicited.

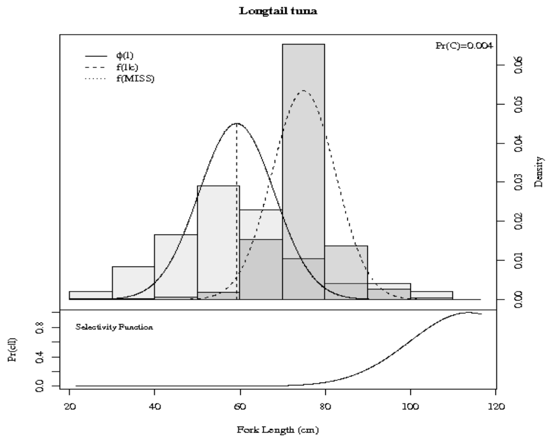

Figure 1. Prior information elicited from a fish biologist showing the observed fork length distribution (---) of longtail tuna, the population of fork lengths (—) and the density of missed fish (…). [Source: Griffiths et al. (2007)]

The Elicitation Process: An expert fish biologist was engaged to provide information about the population abundance density, $\phi(l)$ for different pelagic fish in an attempt to estimate the probability of capture, p_c. As outlined in Griffiths et al. (2007), the probability of capture given the fork length[1], can be expressed as $\Pr(c \mid l) = p_C \times f(l \mid c)/\phi(l)$ where $f(l \mid c)$ represents the density of observed fork lengths from captured fish. If we make an assumption that $\Pr(c \mid l) = 1$ when the ratio of the density of observed fish size and that in the population is at a maximum, we can rearrange to form an expression for the probability of capture: $p_C = \phi(\tilde{l}) / f(\tilde{l} \mid c)$. The expert fish biologist was asked to provide an estimate for the mean and standard deviation for $f(\tilde{l} \mid c)$ (dashed line) and $\phi(\tilde{l})$ (solid line) for longtail tuna, which are both assumed to be normally distributed (Figure 1). With these elicited moments, we were able to construct a selectivity function and provide an estimate for the probability of capture, p_c.

Cognitive Limitations and Impacts: The information provided by the expert (Figure 1) highlighted some classic cognitive limitations. This is indicated by the dotted line in the figure that is superimposed over the solid line. This density represents the fish in the population that were missed by the expert's specification of $\phi(l)$ and is represented by $f(MISS)$. The selectivity function shown beneath these density plots confirms this inconsistency and shows that only larger fish are captured by gillnets, with smaller fish avoiding the net. Overall the probability of capture was estimated as 0.004. Using this result in further calculations would lead to an estimate of abundance of 137.9 fish per square kilometer, a very unrealistic and biased estimate. After some discussion and feedback of results, it became clear that the expert's initial estimates were based on a

[1] Fork length represents the length from the fork in a fish's tail to its mouth and is considered a standard way of measuring length.

large number of Taiwanese catch records that the expert was most familiar with. In this example, the expert was clearly driven by past events and outcomes (*hindsight bias*) and exhibited *overconfidence* due to the fact that there was a large volume of Taiwanese data that he based his opinion on. Correcting for this bias resulted in a second elicitation, where the feedback resulting from Figure 1 was used to arrive at a revised estimate of abundance: 1.81 fish per square kilometer (σ_μ = 0.499) and p_c = 0.271.

Overcoming Bias and Lessons Learnt: Although it was quite clear that the estimate of abundance for longtail tuna was incorrect, having a graphical aid to feed back to the expert in this elicitation exercise played an important role in obtaining a more accurate and unbiased estimate for this problem. This example illustrates how easy it is for an expert to become confused with the statistical terminology used to elicit the priors and highlights the need for detailed explanation of what is to be elicited, the incorporation of graphical aids and feedback to circumvent hindsight bias, overconfidence and linguistic uncertainty. Having access to a number of experts may have avoided some of these issues and this is always recommended, where possible.

2.2 Assessment of military working dogs: affect, translation and overconfidence

The Problem: The Military Working Dog (MWD) program of the Royal Australian Air Force (RAAF) breeds 80 German shepherd and Belgian Malinois puppies each year with the aim of developing each pup as a MWD or guard dog for Defence Force bases around Australia. At present, the success rate of dogs in this program is between 30-40% (Julie Herbert, RAAF, *pers. comm.*). The program consists of 5 distinct phases, one of which is *foster care*, where the pups are first exposed to a range of different environments before commencing their formal military training. It is the overall aim of this project to determine factors across the 5 phases of development that might increase a dog's chance of successfully graduating from the program. We investigate one of these phases to examine potential bias that may result from how assessors grade each pup.

The Elicitation Process: Foster care assessments were conducted as part of the program between 30/05/06 through to the 1/10/08 resulting in multiple assessments performed on 109 dogs. The majority of dogs (79%) had either two or three assessments performed during foster care while the remaining 21% of dogs had only one assessment recorded. Assessments were conducted using one or two assessors out of a pool of six, each having varying levels of experience.

The assessment itself was based on four separate tasks: (1) *recall* - ability to gain the dog's attention and have it run to the assessor; (2) *retrieve* - ability of the dog to retrieve an object e.g. ball, when thrown; (3) *boldness* - behaviour of the dog when exposed to different environments; and (4) *bite response* - ability of the dog to pull on a rag or toy and/or its response when provoked. Each assessment was conducted on a five point scale with 5 representing the worst and 1 representing the best outcome. "Word pictures" were used by the assessors to aid in the assessment of each dog and these represented a mechanism for keeping responses comparable, although no formal training or calibration is routinely performed in this type of assessment.

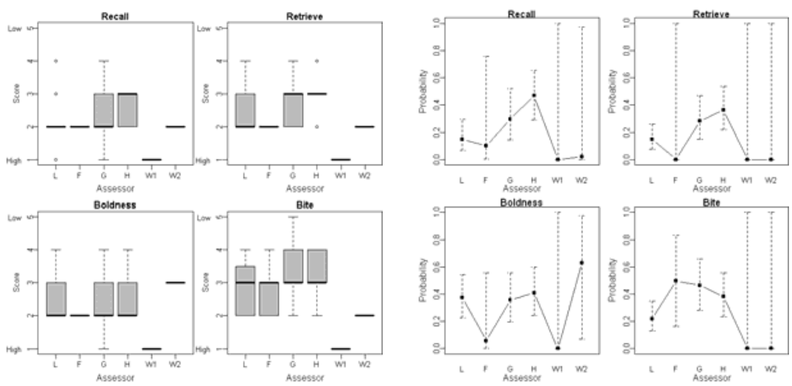

Figure 2. Military working dog characteristics as captured by *recall*, *retrieve*, *boldness* and *bite* assessments summarized by exploratory boxplots of the raw scores (left plot) and the predicted probability of assigning a high score (or worst assessment) (right plot). 95% confidence intervals are indicated by dotted lines.

Cognitive Limitations and Impacts: To investigate any potential bias that could result from this form of subjective assessment, we conducted a simple analysis of the data and compared the ability and consistency of each assessor with respect to a dog's ability to perform each of the 5 tasks described above. The analysis consisted of fitting a proportional-odds logistic regression (McCullagh 1980) to each of the response variables (recall, retrieve, boldness and bite) with the assessor representing the explanatory variable, *x* in the model to investigate differences in rating abilities and potential biases. The model can be written as $P_j = \sum_{s=1}^{j} p_s,\ j = 1,\ldots,r-1.,P_r = 1$ where $\mathrm{logit}(P_j) = \theta_j + \beta' x$ and P_j represents the cumulative probability of up to the *j*-th response for the response of interest, θ_j represents the cut points (on the logistic scale) and β represents a vector of coefficients for the explanatory vector, *x*, which describe how the logit of the cumulative probability relates to *x*.

The results from this analysis are shown in Figure 2 and indicate some variability in assessments. Assessor W1 in particular shows a marked difference in assessment across all 4 behaviours tested (Figure 2-left). Where the other assessors responded with scores between 2 and 4, this particular assessor consistently scored a 1, corresponding to a high assessment for the dogs evaluated. Figure 2 (right) also illustrates this point but indicates high uncertainty in the predicted probability of assigning a high score, resulting from having only assessed one dog in the program. This pattern was also observed for assessor W2. After some discussion with military staff, it was revealed that assessor W1 had been asked to rate the dog he/she fostered, thereby giving a very biased and *overconfident* view of the pup's performance. This assessor was not involved in regular assessments of MWDs during foster care and therefore this assessor's rating may have also suffered from an *affect* bias due to the close interaction he/she had with the pup as well as *translation* problems, due to not having had any training or calibration in the assessment process. The variability amongst the other assessors also hints to *translation* problems with this part of the program. In addition to highlighting inconsistencies amongst the raters, Figure 2 (right plot) also highlights some consistencies which may also form a type of bias. Assessors G and H tend to rate dogs similarly. The estimated 95% confidence intervals for both assessors also indicate that the variability around their assessments were somewhat consistent. It was revealed that both assessors work closely in the foster care program and have been assessing dogs for a lengthy period of time. The closeness in their assessments could be due to being well calibrated and trained across time but could also be due to other factors and therefore requires further investigation.

Overcoming Bias and Lessons Learnt: Many of the inconsistencies noted above are not new in the context of modelling data, and could be handled using multilevel or random effects models. The observed variability, however, clearly impacts the assessment of the pups during foster care. Furthermore, if this information were to be incorporated in a global assessment of a dogs ability across the 5 phases of the program it could either provide an inaccurate assessment if the inconsistencies are ignored or alternatively, if they are properly accommodated, it could lead to an uninformative and useless assessment. This example illustrates the potential pitfalls that could be experienced with multiple raters that use a multi-point scale for assessment and how easy it is for assessors to interpret and apply terminology and rating scales in an inconsistent fashion. It therefore highlights the need for improvements in the rating system used and the level of training supplied to ensure consistencies with expert responses.

2.3 Import risk assessments: linguistic uncertainty, variability and ignorance

The Problem: Import risk assessments are mandated under several international legal instruments. Their aim is to assess the incidence of pests and diseases in the exporting nation, the extent to which these will survive during importation, arrive unnoticed in the importing nation, and consequently establish and spread. Further details can be found in the published examples of quantitative (Yamamura et al. 2001, Venette and Gould 2006) and qualitative (USDA, 1997; Kahn et al. 1999) assessments. As with many other natural resource problems, there is often very little observational data for many of the steps modelled in the risk assessment, and for this reason quantitative assessments are often eschewed in favour of qualitative assessments (Hayes 2003).

The Elicitation Process: Figure 3 shows an excerpt from an elicitation exercise designed to improve the transparency of a risk assessment for mango seed weevils in mangoes imported from India to Western Australia. The elicitation followed a "normative group elicitation" method (O'Hagan 2006) where seven experts where asked *inter alia* to quantify the proportion of orchards in India infested with seed weevils. The experts were asked to provide a best guess, an upper bound and a lower bound, together with an estimate of how confident they were that the true value lies between their bounds. They were allowed to compare and discuss their initial answers (grey lines, left panel, Figure 3) – to minimise the effect of *linguistic uncertainty* - before providing their final response (not discussed), which were translated to 80% confidence intervals (black lines, panel a, Figure 3).

In this example we explored three ways to pool expert's elicited intervals. The first approach uses linear pooling (O'Hagan et al. 2006 and references therein), assuming that the 7 expert intervals are normally distributed on the logit scale (blue lines, Fig. 3). The second approach fits a beta distribution to each of the intervals prior to linear pooling (green lines, Fig. 3) while the third approach again assumes the intervals are normally distributed on the logit scale and uses a Bayesian hierarchical approach to pool the responses similar to a meta-analysis (Hedges and Olkin 1985) (red lines, Fig. 3).

Cognitive Limitations and Impacts: This example illustrates an important difficulty that can occur when experts are uncertain. The group normative theory helps eliminate *linguistic uncertainty* (e.g. misunderstanding the question) so that the elicited intervals can reasonably be assumed to represent knowledge uncertainty (ignorance) or variability. The difficulty in this context is that the elicitation and subsequent linear pooling does not allow the analyst to separate ignorance from variability. Experts 5 and 7 may know that the proportion of infested orchards in India is highly variable, or they may simply be expressing ignorance on this particular question. Linear pooling methods can include expertise weights to discount the beliefs of less knowledgeable experts, but is otherwise is unable to distinguish these two types of uncertainty.

The credible intervals of the posterior pooled mean allows the analyst to construct a probability box (Ferson et al. 2004) (dotted red lines, right panel, Fig. 3). Theoretically this enables the analyst to separate variability and ignorance. Here variability is measured via the standard deviation of the posterior pooled mean, and ignorance via the credible interval. In this example, however, the experts were not given the opportunity to comment on this separation, either during the elicitation or subsequent analysis, and this technique may not accurately reflect their beliefs.

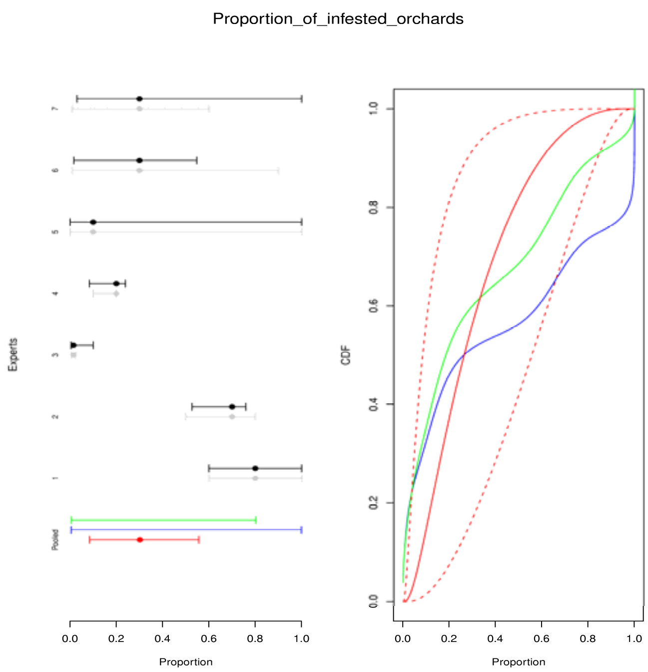

Figure 3. Synthesis of experts' belief about the proportion of mango orchards in India that are infested with mango seed weevil. The plot in the left panel shows the 7 expert intervals and pooled responses based on 3 different pooling methods: linear (blue), beta (green), Bayesian (red). The second panel shows the cumulative density function of these methods in addition to a probability box.

Overcoming Bias and Lessons Learnt: Group elicitation must be performed carefully. Clearly translation and feedback is essential to minimise linguistic uncertainty within the expert group, but it is also important that during the feedback phase experts are not overly influenced by the initial estimates of others in the room (*Anchoring*). Questions and analysis methods that can separate variability from ignorance should also be investigated and ideally explicitly incorporated into both the elicitation process and the pooling methodology.

3. DISCUSSION AND CONCLUSIONS

We have shown examples where experts' prior beliefs and their inherent cognitive limitations, have dramatic effects on parameter estimates. We argue that instances where the expert "gets it wrong", are more likely to occur because of the limitation of the elicitation process, and the subsequent pooling of data where there are multiple experts, rather than the expert/s intending to provide a false or misleading response. This highlights the importance of carefully designing the elicitation process to minimise the effect of linguistic uncertainty and thinking carefully about how this information can be incorporated into a statistical model and interpreted such that the effect of the expert knowledge (encoded in a prior) on the posterior, for a particular likelihood structure is realised.

In practice, to avoid these issues we recommend where possible: (1) the use of multiple experts in a normative setting to avoid overconfidence which is sometimes experienced with one expert (case study 1); (2) pooling of expert beliefs with a mechanism for separating variability from ignorance (case studies 2 and 3); (3) calibration to ensure experts report what they actually mean (case study 3); (4) a feedback and comparison process that allows experts to discuss and revise their opinions, and compare the assumptions of the analysis method (e.g. pooling) with their beliefs (case studies 1 and 3); (5) a methodology that allows the expert to respond in a natural and non-threatening manner about the question being asked (case studies 1-3); and, (6) designing the elicitation process around the statistical methods that are subsequently used to analyse and pool the information that is elicited (case study 2 and 3) and investigating the impact of this information in the model. This latter point is extremely important because no matter how well you have attempted to eliminate bias from the elicitation approach, if the statistical model has difficulty incorporating the information or approximations need to be devised then the effort that has gone into the elicitation can be in vain. If expert opinion is elicited with this care and incorporated in a transparent manner it can become a powerful source of information in models that can effectively assist in the decision making process.

ACKNOWLEDGMENTS

We thank Mark Burgman for earlier discussions regarding this topic and for facilitating the mango weevil elicitation and Bill Venables and two anonymous reviewers for kindly reviewing this manuscript. We also acknowledge the RAAF and in particular Julie Herbert and Alan Grossman for supplying the MWD data.

REFERENCES

Burgman, M.A. (2005) Risks and decisions for conservation and environmental management, Cambridge University Press, Cambridge, UK.

Ferson, S., Nelsen, R.B., Hajagos, J., Berleant, D.J., Zhang, J., Tucker, W.T., Ginzberg, L. and Oberkampf, W.L. (2004), Dependence in probabilistic modeling, Dempster-Shafer theory and probability bounds analysis. Sandian National Laboratories, Alburquerque, USA, 151 pp.

Garthwaite P.H., Kadane J.B. and O'Hagan A. (2005) Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association*, 100, 680-700.

Griffiths S.P., Kuhnert P.M., Venables W.N. and Blaber S.J.M. (2007) Estimating abundance of pelagiv fishes using gillnet catch data in data-limited fisheries: a Bayesian approach, *Canadian Journal of Fisheries and Aquatic Sciences*, 64, 1019-1033.

Hayes KR (2003), Biosecurity and the role of risk-assessment, pp. 382-414 in Ruiz GM and Carlton JT, (Eds), Bioinvasions: Pathways, Vectors, and Management Strategies. Island Press, Washington, D.C., USA.

Hedges, L.V. and Olkin, I. (1985) Statistical methods for meta-analysis, New York: Academic Press.

Kahn, S. A., Beers, P. T., Findlay, V. L., Peebles, I. R., Durham, P. J., Wilson, D. W. and Gerrity, S. E. (1999), Import Risk Analysis on Non-viable Salmonids and Non-salmonid Marine Finfish. Australian Quarantine and Inspection Service, Canberra, Australia, 409 pp.

Kuhnert P.M., Martin T.G. and Griffiths, S.P. (2009) Expert elicitation and incorporation in Bayesian Ecological Models, CMIS Technical Report, Cleveland Australia.

Kuhnert P.M., Martin T.G., Mengersen K. and Possingham H.P. (2005) Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion, *Environmetrics*, 16, 717-747.

Kynn M. (2008) The 'heuristics and biases' bias in expert elicitation, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 239-264.

Kynn M. (2004) Eliciting expert knowledge for Bayesian logistic regression in species habitat modelling, PhD, Department of Statistics, Queensland University of Technology, Brisbane.

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

Martin T.G., Kuhnert P.M., Mengersen K. and Possingham H.P. (2005) The power of expert opinion in ecological models using Bayesian methods: impact of grazing on birds, *Ecological Applications*, 15, 266-280.

McCullagh P. (1980) Regression Models for Ordinal Data, *Journal of the Royal Statistical Society, Series B*, 42, 109-142.

O'Hagan A., Buck C.E., Daneshkhah A., Eiser J.R., Garthwaite P.H., Jenkinson D.J., Oakley J.E. and Rakow T. (2006) Uncertain Judgements: Eliciting Expert Probabilities, Wiley, United Kingdom.

O'Leary R., Murray J., Low-Choy S. and Mengersen K. (2008) Expert elicitation for Bayesian classification trees, *Journal of Applied Probability and Statistics*, 3, 95-106.

United States Department of Agriculture (USDA) (1997), Importation of fresh citrus fruit (Sweet Orange, Citrus sinensis, Lemon, C. limon, and Grapefruit, C. paradisi) from Argentina into the continental United States: Supplemental plant pest risk assessment. United States Department of Agriculture, Riverdale, MD, USA, 107 pp.

Venette RC and Gould JR (2006), A pest risk assessment for Copitarsia spp., Insects associated with importation of commodities into the United States. *Euphytica*, 8: 165–183.

Yamamura, K., Katsumata, H. and Watanabe, T (2001), Estimating invasion probabilities; a case study of fire blight disease and the importation of apple fruits. *Biological Invasions*, 3: 373-378.