CSIRO

www.csiro.au

Point of Truth Calibration: Putting science into scoring systems

Simon Barry and Xunguo Lin

Report Number: EP10378, 1 November 2010

Australian Centre of Excellence for Risk Analysis (ACERA)

Enquiries should be addressed to:

Dr Simon Barry
CSIRO Mathematical and Information Sciences
CSIRO, GPO Box 664, Canberra, ACT 2601, Australia
Telephone : +61 2 6216 7157
Fax : +61 2 6216 7111
Email : Simon.Barry@csiro.au

Distribution List

Client	(1)
Publications Officer	(0)
Stream Leader	(0)
Authors	(2)

Copyright and Disclaimer

Important Notice

Contents

1 Introduction 4
1.1 Point scoring systems 4
1.2 Related approaches to using expert judgements 5
1.3 Our approach and the organisation of the paper 6

2 Motivation 6

3 Methods 7

4 Examples 9
4.1 Weed eradication study 9
4.2 Designated ballast water exchange areas 13

5 Discussions and Conclusions 15

6 Acknowledgements 22

ABSTRACT

Scoring systems are commonly employed in risk analysis as a way of integrating information about a range of variables (called risk attributes) into a single risk score suitable for decision making. The construction of scoring systems is often criticised as being arbitrary as there is no transparent basis for the selection of weightings and integration methods. This paper describes an alternative approach which overcomes these difficulties. The method we detail is called Point of Truth Calibration. It is based on expert judgements on constructed risk scenarios rather than individual scores of risk attributes. We argue that it is an improved form of expert elicitation for many complex risk assessment problems. The proposed method will 'automatically' calibrate weightings (scores) of risk attributes from the overall risks given by experts using well developed statistical techniques. Several examples from biosecurity are presented to demonstrate the approach.

Keywords: Scoring system, risk assessment, risk calibration, expert elicitation.

1 Introduction

Decision making in biosecurity is a challenge and complex problems abound. In world trade it is common for a country to need to decide whether the importation of goods or organisms represents an acceptable risk. How do they determine the risk of a particular organism to an ecosystem or continent? Governments may need to decide particular management actions. For example, how do they determine the cost of eradicating a species that has invaded an ecosystem?

An additional challenge may be that directly relevant empirical data are not available. This is often the case in biosecurity assessments as they often deal with rare events in unique situations. For example, an application to import an organism into a country can be difficult to assess. As the organism has not been imported previously, direct data on its likelihood of establishment is not available. While data from analogous trade may be available, for example invasion success internationally, its interpretation typically requires significant assumptions.

In these cases it is usual to rely on expert judgements to underpin the assessment. While this can require considerable assumptions, it is often the case that a decision needs to be made, and interests to be traded off. In this case the choice of doing nothing is not viable and the expert represents the most trusted source of reliable information. There are of course a number of significant issues in using experts such as the expert's availability, overconfidence, and motivational bias (see Burgman (2005), Meyer and Booker (1990), Klayman et al. (1999) and Vose (2008)), which need to be considered carefully in all expert-based assessments.

1.1 Point scoring systems

While we may be satisfied that the use of expert opinions is appropriate in a particular case, the question of how we engage the experts remains. There is significant literature on a variety of issues in expert elicitation, for example, finding beliefs about particular parameters and related population values; see O'Hagan et al. (2006), Low Choy et al. (2009) and Kuhnert et al. (2010). We consider the complex task of constructing expert-based scoring systems to measure risk for particular scenarios. By scenarios we mean particular cases for which a decision needs to be made.

In an expert-based point scoring system, scores are normally allocated by experts for a number of risk attributes. A system of combining the scores is derived and applied to the individual scores to produce the overall risk score. The Australian weed risk assessment system (Pheloung et al., 1999), is a good example of a scoring system. It has been applied in Europe (Crosti et al. (2007)) and the U.S.A. (Gordon et al. (2008b)), and elsewhere (Gordon et al. (2008a)). Other examples of scoring systems used in biosecurity are Copp et al. (2005) and Copp et al. (2009). These scoring systems were generated by experts but were then available for use administratively by field staff.

Typically, the issue of weighting the individual scores to obtain an overall risk score is problematic. For example, a total score for the particular scenario could be obtained by summing over all individual scores, which is a type of averaging. Alternatively, the experts could construct a weighting system based on their expert knowledge. The difficulty with this approach is that there is no objective guarantee that the mapping from the attributes of the case under consideration to the final score is robust or defensible.

The advantage of point scoring systems is their use of explicit and consistent assessment of variables, which means that the process can be applied in a consistent manner. It may provide greater resolution in the ranking of risks than other methods such as rule sets or qualitative evaluations, thereby providing a means for allocating priorities that compete for scarce resources

(Burgman et al., 1999).

The main drawback of point scoring systems is that they are sensitive to the variables chosen and relative weighting of variables. As pointed out by Hubbard (2009), the scoring systems themselves add their own sources of error as a result of unintended consequences of their structure. The point scores and risk ranks are directly related to weights given to variables. In addition, the risk of a complex scenario may depend on a number of variables in quite complex ways. The expert's ability to decompose this complexity can be questioned.

1.2 Related approaches to using expert judgements

A concept called *bootstrapping* was first reported by Dawes (1971) and was extensively reviewed by Armstrong (2001). Because this name gained popularity in the statistics community for a different concept, Armstrong (2001) has renamed it *judgemental bootstrapping*. Judgemental bootstrapping derives a mathematical model using regression of the judgements from experts. The key idea is that a regression model, based on the attributes of a scenario and the expert's prediction for the outcome of a scenario, can perform better than the experts themselves in predicting new cases. For example, Simester and Brodie (1993) developed a judgemental bootstrapping model of criminal sentencing decisions from judges to predict sentence length and the outcome of sentencing appeals in New Zealand. They modelled sentencing length for sexual offences as a function of variables describing the character of the offender and the circumstances of the offence. The model outperformed both an equal weights approach and a naive mean projection in predicting sentence lengths.

To analyse an expert's prediction skills, Stewart (1990) combined a decomposition of the correlation coefficient and expert's bias originally developed by Murphy (1988) with the so called *lens* model introduced by Brunswik (1955). The lens model is closely linked to the assumption that the model between the forecast and the observed outcome is linear. The decomposition of correlation coefficient requires knowledge of the model attributes used by the experts. The lens model shows the relationships between the model attributes, the observed event and the expert's prediction.

Mazzuchi et al. (2008) reported the use of expert judgements in a risk assessment concerning aircraft wiring. Pairwise scenarios of different aircraft wiring environments were presented to a group of experts for their judgements on relative failure rates in the two environments. Statistical analyses of experts' responses were required to remove both random and inconsistent responses. Regression analysis was then applied to the failure pattern to relate it to environmental variables.

Research has been done to try to explore the best procedure for capturing the complexity of experts' rules, and to produce the most accurate predictions. Cook and Stewart (1975) compared seven different methods to obtain weights for attributes to predict a student's admission to graduate school. These included asking experts to divide 100 points among the attributes, rate attributes on a 100-point scale, make comparisons between a pair of attributes, etc. Cook and Stewart (1975) found that these methods produced similar results in terms of matching the experts' decisions. However, they all yielded more accurate predictions of actual judgements than an arbitrary policy of equal weights on all attributes. This conclusion was also confirmed by a partial replication study by Schmitt (1978).

There are other approaches which were based on so-called *predictive elicitation*; see Kadane et al. (1980), Kadane et al. (1996), Kadane and Wolfson (1998). During a predictive elicitation, questions about the expert's (probabilistic) view of the dependent variable were asked given various

values of the predictor variables. The predictive elicitation approach addresses the construction of priors for regression analysis. In this approach, experts will be asked to specify the range of each of the covariates in the model, then to elicit four hyper-parameters of the prior distribution, including the prior mean and the degrees-of-freedom, through a two-stage elicitation.

There have been recent developments in software for expert elicitation. For example, James et al. (2010) presented an expert elicitation tool for determining priors for use in Bayesian regression in ecology. It utilises an indirect elicitation approach to target expert knowledge and derives an expert-defined prior model including its hyper-parameters. This expert-defined prior, together with observed data, can then provide posterior estimates based on the Bayesian framework. This expert elicitation tool has been applied in a case study predicting the distribution of a species of Wallaby, a small marsupial, see Murray et al. (2009).

1.3 Our approach and the organisation of the paper

In this paper, we will describe a method called Point of Truth Calibration (PoTCal). It is based on expert judgements on constructed risk scenarios rather than individual scores of risk attributes. Our method will then empirically calibrate weightings (scores) of risk attributes from the overall risks given by experts using regression techniques. This methodology forms a bridge between best practice approaches to constructing scoring systems when objective data are available and the expert based approaches needed in the absence of data.

Our approach develops methodology in this area in several ways. First, it promotes the use of general regression approaches to empirically estimate the relationship between predictions and attributes. These can potentially remove biases. Second, it models variation in expert opinion so that inference over the population of experts can be considered. Third, it considers a general framework for these problems. Previous literature has either been one-off or examined simple cases. Fourth, it describes these techniques to a wider audience.

The examples presented in the paper will demonstrate the method and provide an opportunity for comparison to conventional expert-based scoring systems.

The paper is arranged as follows. In Section 2 we consider a motivating example. In Section 3 we define the PoTCal approach. In Section 4 we consider two examples from biosecurity and in Section 5 we discuss the links of the proposed method to existing techniques and explore the strengths and weaknesses of the approach.

2 Motivation

To motivate the problem consider a simple example. Assume decision makers are deciding whether particular plant species should be imported to a country and they are concerned about the possible environmental or economic impacts if these plants become weeds (see, for example, Pheloung et al. (1999)). Scientific advisors may believe that the attributes of the plant species such as its life history and growth rate are related to the probability that it becomes a weed. Thus there may be input data for the ith plant species as $A_i = [a_{i1}, a_{i2}, \ldots, a_{ik}]$ where a_{ij} is the jth attribute for the ith plant species and k represents the number of attributes.

Traditional approaches would consider using weights to determine a risk score. First, construct a design vector X_i from the A_i. This is done using the standard methods in linear modelling, converting categorical variables to the associated indicators. For example, if there were $A_i = [a_{i1}, a_{i2}]$ with $a_{\cdot 1}$ having three levels and $a_{\cdot 2}$ having 2 levels then it would have the design matrix $[I(a_{\cdot 1} = \text{level } 1), I(a_{\cdot 1} = \text{level } 2), I(a_{\cdot 1} = \text{level } 3), I(a_{\cdot 2} = \text{level } 1), I(a_{\cdot 2} = \text{level } 2)]$, where

$I()$ is the indicator function taking the value 1 when the condition in parentheses holds and zero otherwise.

The risk score is calculated as

$$S_i = X_i^T \beta \tag{2.1}$$

where β is a vector of parameters. The weights β are elicited from the experts. The experts can then construct cutoffs for the S_i to prompt particular management actions. For example, if $S_i < K$, where K is some constant, the risk may be considered acceptable. An issue with this approach is that the cutoffs may be considered arbitrary and not empirically justified.

A possible development of this system is to attempt to calibrate this cutoff to achieve optimal performance. Hughes and Madden (2003) considered this for a weed risk assessment system. They calibrated it as follows. Consider having a sample of plant species with associated attributes and using the scoring system to generate a set of risk scores. If we know whether each plant species i is considered a significant weed or not, we have a binary outcome y_i. Being significant in this context could mean that its impacts were severe enough to cause government action. Knowledge of y_i may be empirical but more importantly it may be expert based. To characterise the relationship between risk scores and the outcome y_i we can use logistic regression. This would have the linear predictor

$$\log(\frac{p_i}{1-p_i}) = \alpha + S_i\beta^* = \alpha + X_i^T\beta\beta^* \tag{2.2}$$

with p_i being the probability that the ith plant species is a weed and β^* is the slope of the logistic regression line relating the scores to the response. An important insight is that this calibration is a simple rescaling of the expert weights. The direction of the vector β does not change.

The approach advocated in this paper is as follows. We argue that it is better to calculate the regression

$$\log(\frac{p_i}{1-p_i}) = \alpha + X_i^T\beta' \tag{2.3}$$

based on the expert assessments y_i rather than letting the experts estimate the point scores for individual risk attributes. The important point to note is that we have used the experts to elicit the compound scenario y_i. This is compound in the sense that any number of the variables in A have impact on y_i. We then use regression to find the weights (i.e. scores) and explore the nature of the relationship between the assessments expressed by the experts and the attributes. We have made it into a calibration problem where the experts have provided the point of truth, which is the component that we accept as reflecting the reality we wish to model.

3 Methods

We assume that we have "scenarios" with associated attributes X. The ith scenario has attribute X_i. We wish to calculate a score Y_i for each scenario. This score is typically for use in decision making and could be binary, categorical, or continuous. We assume that there is a relationship between Y_i and X_i and we now consider the statistical characterisation of this relationship.

Assume that we have a large population of scenarios. In this case the conditional mean $E[Y|X]$ is well defined as well as the conditional variance $Var[Y|X]$. Now we can write $E[Y|X] = f(X, \alpha)$ where α is a set of parameters. The function $f()$ is potentially extremely complicated. It can have discontinuities, non linearities, complex interactions and other difficulties. Figure 3.1 shows some possible forms for this function in the simple case that there is a single

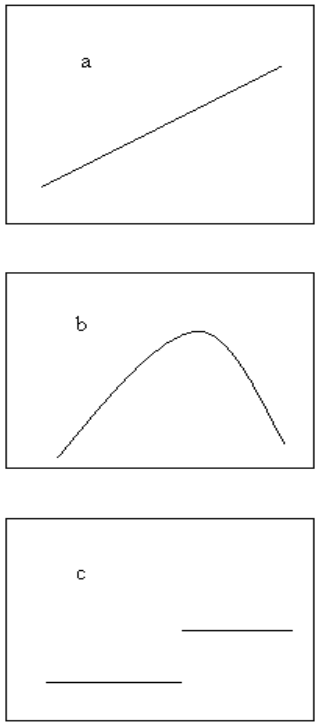

Figure 3.1: Examples of function $f()$: (a) linear; (b) non-linear; (c) discontinuous.

variable x. Its shape is a simple reflection of the complexities of the situation and cannot be simplified by further analysis.

In simple cases we will have a sample of the $[Y_i, X_i]$ pairs and the problem is a standard regression problem and a range of techniques are available to explore this relationship and form good predictors of Y_i. The case of interest here is when Y_i is not known. In this case the scoring approach considers a function $g(X_i, \varepsilon)$ where $g()$ and the parameters ε are defined by the experts. For example $g()$ can simply be the sum of the scores, and ε are the user defined weights.

Bias for scenario i in this measure is the discrepancy

$$B_i = f(X_i, \alpha) - g(X_i, \varepsilon). \tag{3.1}$$

We note that with the expert-based approach you cannot calculate $Var(Y_i|X_i)$. This provides an estimate of the variations in the scenario outcome given the attributes, and expresses uncertainty.

The alternative PoTCal approach is to get the experts to provide their views about the values Y_i. Denote these as $\hat{Y}_i$. We then use regression techniques to estimate the relationship $f(X, \alpha)$ by regressing the $\hat{Y}_i$ against the X_i. We can also estimate $Var[Y_i|X_i]$ if a well defined technique is used. In addition, by incorporating experts in the model as either fixed and/or random effects, the effect of the experts' can be estimated and incorporated into decision making. For example, we can make inference over the population of experts if we believe we have a representative sample of experts.

There are a number of issues that need to be considered in this approach. The first is what happens if there are no experts to estimate Y_i. If this is the case then the Y_i is unknown, and it is unlikely that a hypothesis could be formed about $g(X_i, \varepsilon)$. Thus if experts cannot estimate Y_i then a scoring system is not possible. The second issue is that the experts will not be able to

score Y_i exactly, but can estimate it by $\hat{Y}$ such that

$$\hat{Y}_i = Y_i + \phi_i \tag{3.2}$$

In this case the ϕ_i is the judgement error. Consider the simple linear case

$$Y_i = X_i\beta + \gamma_i. \tag{3.3}$$

In this case γ_i is the residual error, the component of Y_i that cannot be predicted by X_i. If we initially assume that the judgement error has mean 0 and variance σ^2_ϕ and the residual error also has mean 0 but variance σ^2_γ, then we have

$$\hat{Y}_i = X_i\beta + (\gamma_i + \phi_i) \tag{3.4}$$

where the combined error term $(\gamma_i + \phi_i)$ has mean 0 and variance $(\sigma^2_\phi + \sigma^2_\gamma)$ assuming these two errors are independent.

Thus the modelled error term based on the expert's input incorporates a component of uncertainty relating to their ability to predict the scenario. The error variance will bound the actual uncertainty, assuming unbiasedness. As the error becomes more complicated, different biases will result, and the magnitude and pattern of them will also depend on the true model $g(X_i, \alpha)$ and the regression technique used.

4 Examples

In the following section we consider two examples where the PoTCal approach has been applied to real world problems in biosecurity management. Each approach follows the general principles given in Section 3. The technical development of the methodology was lead by the first author of this report in both cases.

4.1 Weed eradication study

Weeds are one of the major natural resource management problems in Australia and are considered by farmers to be one of the highest priority land degradation issues. The cost of weeds to Australian agriculture has been estimated at $3.3 billion per year (Jones et al., 2002; Groves, 2002) compared to the $2.4 billion estimated for salinity, sodicity and soil acidity combined (CRCAWM, 2002). In Australia, 335 weeds are listed as noxious according to the National Weeds Strategy (www.weeds.org.au/noxious.htm).

Weed eradication is one of the strategic approaches to weed management. Weed eradication can be defined as '*the complete and permanent removal of all wild populations from a defined area by a time-limited campaign*'; see Bomford and O'Brien (1995). Weed eradication involves a finite investment compared to the indefinite commitment of resources to an ongoing containment strategy, or simply living with the costs of the weed.

Weed eradication may be both desirable and feasible on so called "sleeper weeds". A sleeper weed is a naturalised exotic plant species that is currently only present in a small area but that has the potential to spread widely and have a major negative impact on agriculture; see Cunningham et al. (2003). Groves et al. (2003) identified that 29 naturalised plant species have thought to have been eradicated from Australia and there were 156 cases where eradication has been unsuccessfully attempted at a State/Territory scale.

Given this failure rate, identifying and prioritising sleeper weeds to be eradicated is a complex but important task. Eradication should not be attempted if it is unlikely to succeed. Prioritisation involves weed short-listing, risk assessment, potential impact assessment and eradication feasibility assessment. In this example we consider the weed eradication feasibility assessment.

While eradication may be desirable for many weeds, it is not always feasible. The aim of the feasibility assessment is to assess the selected species for the feasibility of eradication (or the cost-benefit analysis) based on factors such as their current area of infestation and the relative merits of eradication and control. In this example, 17 species were selected for the eradication feasibility assessment, see Table 4.1.

Table 4.1: List of 17 potential agricultural sleeper weeds [source: Cunningham et al. (2003)].

Species	**Common name**	**State(s)**
Aeschynomene paniculata	pannicle jointvetch	Queensland
Brillantaisia lamium		Queensland
Froelichia floridana	snakecotton	Queensland
Gmelina elliptica	badhara bush	Queensland
Asystasia gangetica ssp. *micrantha*	Chinese violet	New South Wales
Eleocharis parodii		New South Wales
Baccharis pingraea	chilquilla	Victoria
Hieracium aurantiacum	orange hawkweed	Victoria, Tasmania
Hypericum tetrapterum	square-stalked St Johns wort	Victoria, Tasmania
Nassella charruana	Uruguay needle grass	Victoria
Oenanthe pimpinelloides	meadow parsley	South Australia
Onopordum tauricum	Taurian thistle	Victoria
Piptochaetium montevidense	Uruguayan ricegrass	Victoria
Rorippa sylvestris	yellow creeping cress	Tasmania, South Australia
Centaurea eriophora	mallee cockspur	South Australia
Crupina vulgaris	common crupina	South Australia
Cuscuta suaveolens	Chilean dodder	South Australia

The weed profiles for each of these 17 species were made up of 13 attributes which describe the current weed geographic distribution (four attributes), weed control effort, that is, access, detection and tolerance (four attributes) and weed persistence, that is, ability to recover and spread (five attributes). For example the weed geographic distribution are the categories "≤ 1 hectare", "1-10 hectares", "10-50 hectares", "50-100 hectares" and "100-3000 hectares".

Nine experts were asked to examine each of the 15 developed hypothetical weed profiles and gave assessments of eradication feasibility. For each scenario the expert was asked to assess the probability of success for each of a fixed set of expenditures ($25,000, $50,000, ...). This was done as the experts were uncomfortable giving an exact cost for a particular level of eradication probability.

The elicited probability of eradication data for each expert and scenario was analysed by fitting a logistic curve to it by regressing it on the cost data. This allowed estimation of the cost for a successful (95% probability) eradication for each scenario considered by the expert, providing a standardisation of the responses. We analysed the data as follows.

We modelled the probability of eradication, p, using the following relationship

$$\log(\frac{p}{1-p}) = \alpha + C\beta \tag{4.1}$$

where C is the cost and α and β are parameters to be estimated. We modelled the variation in p as

$$\text{var}(p) = p(1-p)D \tag{4.2}$$

where a scale factor D was estimated to account for over-dispersion.

After the model parameters α and β had been estimated (denoted by $\hat{\alpha}$ and $\hat{\beta}$), the eradication cost for a given probability p can also be estimated using the following formula which was derived by rearranging Equation (4.1):

$$\hat{C}(p) = \frac{\log(\frac{p}{1-p}) - \hat{\alpha}}{\hat{\beta}} \tag{4.3}$$

By choosing a fixed p we standardised the results across the scenarios. In this example, $p = 0.95$ was chosen, which represents a high likelihood that eradication would be successful.

The calibration model that was fitted is as follows

$$\log(\hat{C}(.95)_{ij}) = \alpha + X_i^T\beta + b_j + \varepsilon_{ij} \tag{4.4}$$

where $C(.95)_{ij}$ is the estimated eradication cost (in thousands of dollars) given by Equation (4.3) for the ith weed scored by the jth respondent; X_i is the column vector of attributes for the ith weed; β is the regression coefficients for the weed attributes; α is the intercept; b_j is the effect of the jth respondent which is assumed to be a random effect (i.e., drawn from the population of potential respondents) with mean 0 and variance σ^2; and finally ε_{ij} is the error term assumed independent Normal with mean 0 and variance σ^2. The model allows for a systematic (across scenarios) additive effect attributable to experts. The attributes where ordered categories each on an approximate exponential scale so the logarithm of the costs were reasonably assumed to be linearly related to category level (e.g. 1,2,3 for a three level category). The alternative approach of fitting unordered categories could not be attempted due to the available sample size. Standard regression diagnostics were used to ensure reasonable fit.

Maximum likelihood was used to fit Equation (4.4) and the effect of design variables was calculated using likelihood ratio statistics. Significant design variables were selected for the final model with the assistance of their p-values from t tests for the regression coefficients. The final model below gives the eradication cost with respect to the risk attributes and the estimated weights:

$$\hat{C} = \exp[9.43 + (-0.5X_1) + (-0.63X_2) + (-0.36X_3) + (-0.42X_4)] \times \text{AUD\$1000} \tag{4.5}$$

where $\hat{C}$ is the modelled eradication cost; X_1 is the area of the infestation; X_2 is the number of infestations; X_3 is the ease of access; and X_4 is the seedbank longevity.

Both X_1 and X_2 are from the original four attributes defining the weed geographic distribution. X_3 is one of the four attributes relating to the weed control effort. X_4 is one of the original five attributes describing the the weed persistence.

The PoTCal method took into account the variability between experts and estimated the individual expert effects, that is, the mean difference between the nine experts averaged over all 15 weeds are shown in Figure 4.1. The details of this study can be found in Cunningham et al. (2003).

If a traditional point scoring method was used for this example, instead of the proposed PoTCal method, then the experts would have to provide point scores for each of the 13 risk attributes.

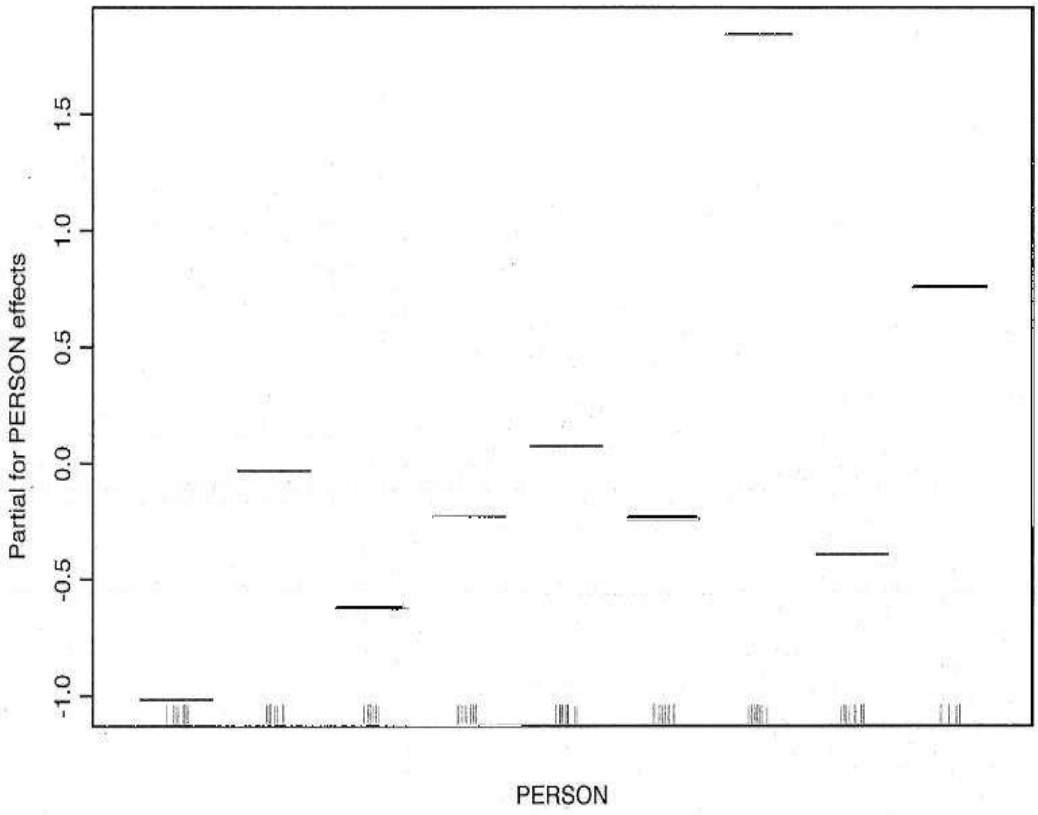

Figure 4.1: Plot of estimated person effects, i.e., mean difference between the nine experts averaged over all weeds, [source: Cunningham et al. (2003)].

After that, a method to combine these 13 scores including the relative weighting of each attribute would have to be determined. Finally, the score would have to be calibrated to the decision process.

4.2 Designated ballast water exchange areas

Invasive marine pests pose a threat to Australian marine environments and to industries dependent on these environments; see Bax et al. (2003). Exchanging ballast water is one mechanism for the introduction and translocation of marine pests. The Australian government has signed and ratified the International Convention for the Control and Management of Ships Ballast Water and Sediments (the Convention), and is working with its States and Northern Territory towards a single consistent national ballast water management system; see Knight et al. (2007).

The Convention provides for the management of ballast water through two main mechanisms - exchange and treatment. The designation of ballast water exchange areas is an interim solution for approximately 10 years. This required investigation of the possible locations for ballast water exchange areas to reduce the risk of translocating harmful aquatic organisms around Australia's marine environment until onboard treatments become available.

In this example, we estimate the biological risks posed by exchange of ballast water around the Australian coast. The Australian coastline and territorial waters are vast (over eight million square kilometres) and it is not feasible to carry out individual assessments at all locations. Instead, Knight et al. (2007) applied the PoTCal approach to assess the biological risks at 12 chosen locations around the Australian coast (see Figure 4.2). We note that at each location there are four scenarios (i.e., discharging points) so that there are four elicitation observations per location for each expert.

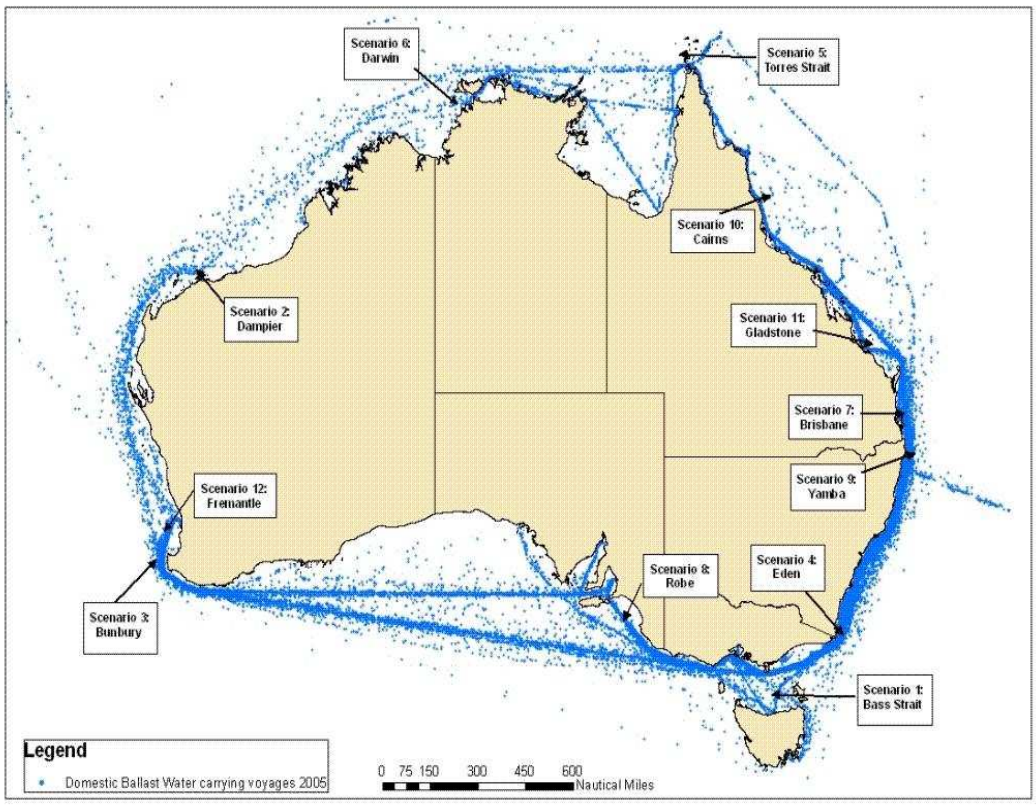

Figure 4.2: 12 locations and domestic shipping activities in 2005, [source: Knight et al. (2007)].

Associated with the scenarios per location was a range of data thought to be related to the risk of discharge of ballast water to the environment. These variables included the assumed exchange sites (i.e., discharging points; see for example Figure 4.3) with their coordinates, water depth, distance offshore and a range of contours describing the time it took discharged particles to reach shallow water.

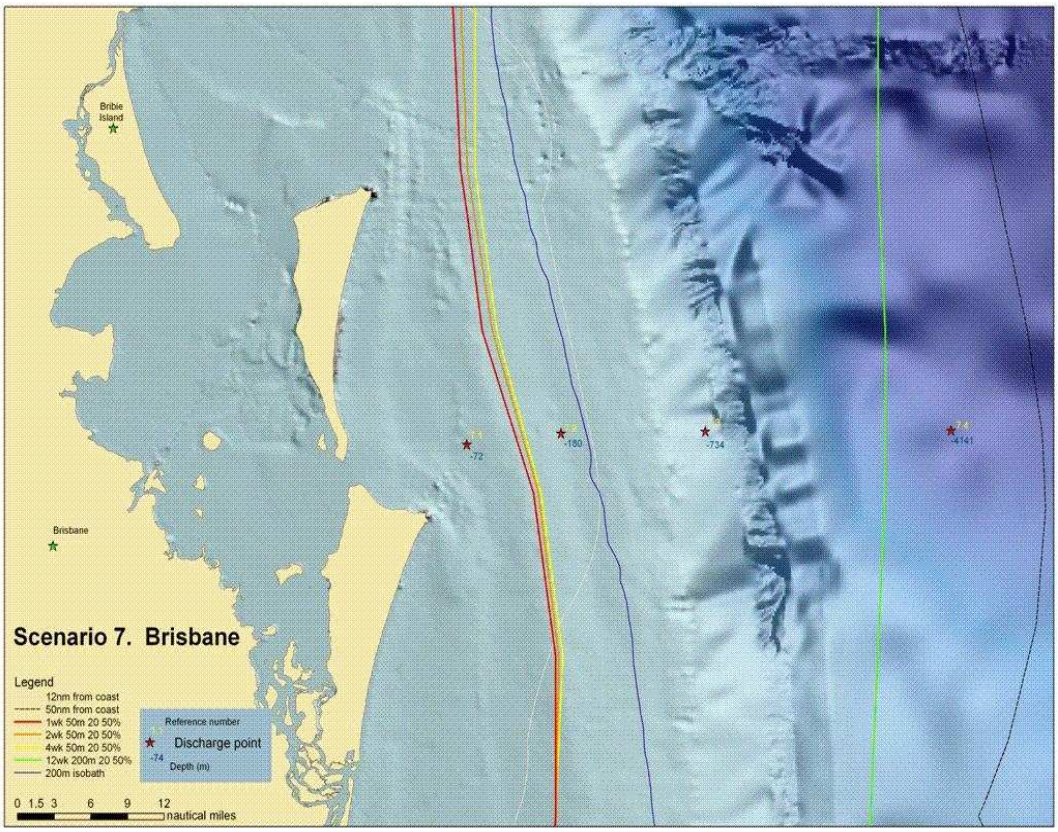

Figure 4.3: Location 7 - Brisbane coast where stars represent discharging points (i.e., scenarios) [source: Knight et al. (2007)].

For every scenario at each location, the experts were asked to consider the relative risk between the discharge at that point and discharge in a nearby port environment. They were asked to estimate the percentage of species that would establish if the ballast water was discharged at each site. For example, if the expert thought 10 species could establish in the port, while only five would establish based on discharge at the scenario point due to the different environmental conditions, they would rate it 0.5.

There were five risk attributes potentially related to the biological risk at each scenario. They are latitude, longitude, distance from coast, water depth and larval survivability. A logistic model was set up in the form of

$$\begin{aligned} \log(\frac{p}{1-p}) &= \alpha + b + f_1(\text{depth}) + f_2(\text{distance}) + f_3(\text{larval.baseline}) \\ &\quad + f_4(\text{latitude}) + f_5(\text{longitude}) \end{aligned} \tag{4.6}$$

where p is the risk estimated by the experts, b is a random effect associated with each expert and f_i represents a smooth function such as a spline. This was fitted to the elicited data therefore calibrating the risk attributes with respect to the relative risks estimated by the experts at each

discharging point in each location. The conditional variation in p was modelled with a scale factor to accommodate estimating dispersion as in the weed example.

The derived biological risk model was then extrapolated from those 12 locations to all Australian coastal waters. The final product of this study was an estimated spatial risk map of ballast water exchange around the Australian coastline, shown in Figure 4.4. The risk here was defined as the relative risk of discharging ballast water at the location compared with the nearest port. For more details of this study; see Knight et al. (2007).

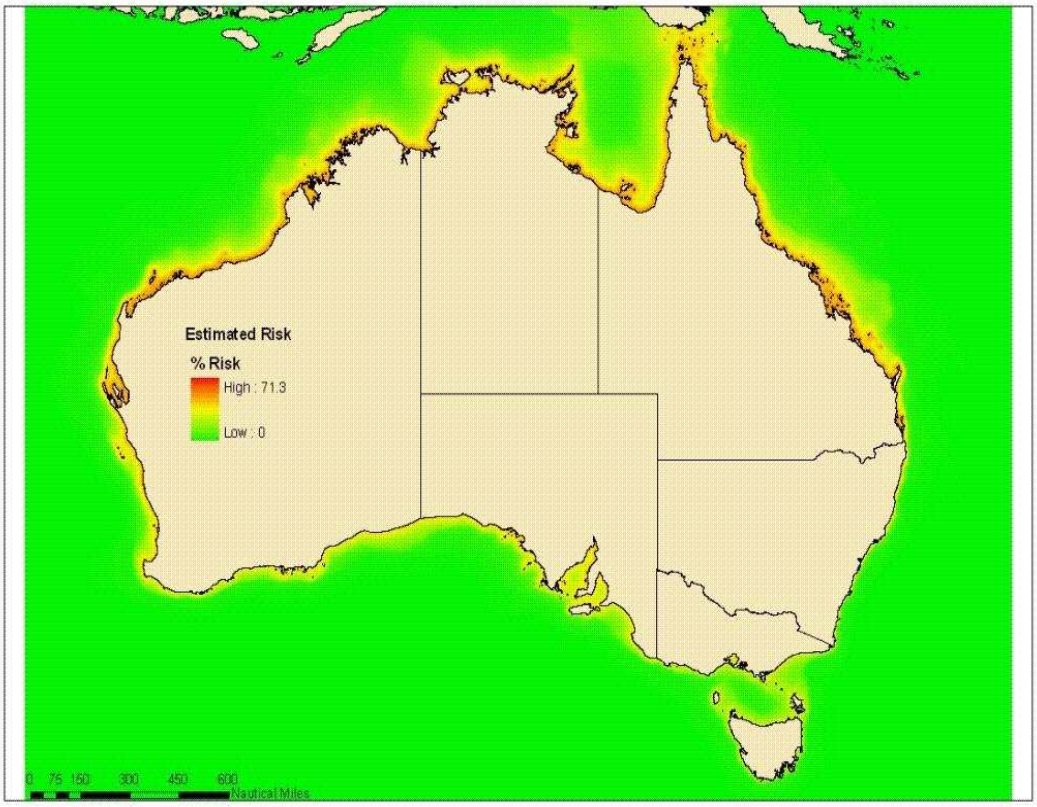

Figure 4.4: Map of the estimated risk from ballast water exchange, [source: Knight et al. (2007)].

5 Discussions and Conclusions

There are strong linkages between the PoTCal approach and a number of other techniques already in the literature. Judgemental bootstrapping was originally proposed by Dawes (1971) and was recently reviewed by Armstrong (2001). Similarly to PoTCal, judgemental bootstrapping involves experts rating scenarios and typically using multiple linear regression to define scoring rules. The key claim of practitioners of judgemental bootstrapping is that the regression models produced outperform the experts in many cases, being able to smooth out idiosyncrasies in the experts performance. There is a significant literature that demonstrates this (see Armstrong (2001)). Judgement bootstrapping also recommends fitting separate models to different individuals or groups (Armstrong, 2001). The PoTCal approach explicitly considers how to model variation in opinion over scenario attributes. It explicitly considers how to incorporate, estimate and report variation between experts without expending large numbers of degrees of freedom on individual models. It also explicitly considers modelling the non-linearities in the relationship

between scenario outcomes and attributes.

A number of studies have considered the relative accuracy of expert defined weighting schemes and models derived by regression techniques (Hamm, 1991; Cook and Stewart, 1975). Hamm (1991) studied traffic engineers' ability to predict the aesthetics, capacity and safety of roads. He found that models fitted to the engineer's judgements performed better when the variables available for modelling were fixed by the researchers before the start of the analysis. When the variables were defined by the experts as part of the analysis, the results were mixed. Cook and Stewart (1975) compared seven methods for obtaining expert judgements concerning graduate student admission to the University of Colorado. While not statistically significant in all cases, the use of optimal linear weights outperformed the subjective approach in the sample in both cases considered. We note that in these studies simple linear models were used with different models fitted to each expert, different to the approach in this paper.

In the meteorological literature, Stewart (1990) has considered the application of the lens model in forecasting. The lens model considers decomposing forecast skill into components relating to the forecast and the actual outcome. As described by Stewart (1990), the lens model of Brunswik describes the relationship between cues, forecast and observation. This is considering similar issues to PoTCal but it is not developed as a predictive analysis but rather as a technique to decompose correlations related to forecasting skill.

A slightly different approach is probabilistic inversion (Du et al., 2006). Probabilistic inversion is used to parameterise stochastic models of processes. There are often occasions when the parameters of a model may not be available because the process might be new or there is no opportunity to perform measurements (Cooke et al., 2006). In these situations it might still be possible to obtain information about other variables predicted by the model. Probabilistic inversion considers how to infer the stochastic nature of the unknown variables from the structure of the model and the observed information. This approach requires strong assumptions about the model formulation to perform its inference. PoTCal is targeted at less structured problems. It relies only on defining the empirical relationship between the cues and the expert's opinions.

James et al. (2010) have developed techniques for eliciting priors for Bayesian regression models. There is related work discussed in Kadane and Wolfson (1998). While these techniques elicit compound judgements of scenarios from experts, their focus on the derivation of priors and the use of Bayesian techniques is different from PoTCal.

The last example to consider is the use of pairwise comparisons. Cooke (1991) reviewed a number of these approaches. Pairwise comparisons do not in themselves address the calibration problem but they can be effective techniques for elicitation. A relevant example of its application is in Mazzuchi et al. (2008) who considered estimating the safety of electrical wiring. They used a Bradley-Terry model to estimate relative failure rates and applied a multiple linear regression based approach to generalise this over environmental variables. They then used empirical data on a subset of cases to convert the relative failure rates to absolute failure rates. The PoTCal approach solves the empirical calibration problem. If there is additional empirical data available, a number of analyses could also be considered.

In summary, the proposed PoTCal method has a number of potential advantages over the conventional expert-based scoring systems, particularly:

1. In some cases it is more straightforward for the experts to consider scenarios rather than risk components which are dependent on implicit modelling. It only needs a single elicitation of the risk of the scenarios under assessment. Therefore, the elicitation is direct and transparent, while the conventional point scoring methods performed on the attributes of

risk are indirect and less transparent.

2. Assessment of scenarios allows the expert to apply all of their knowledge about the scenario, and uncertainties in the ability to assess risk which can be quantified in the analysis.
3. Variations between experts can be quantified and incorporated into decision making.
4. It avoids experts having to consider formulation of complex interactions between individual risk components and/or attributes.
5. Using modern statistical or machine learning methods, such as regression or classification trees, means that the weights are automatically calibrated to real world scenarios.
6. It forms a logical bridge to traditional regression based approaches to scoring when data is available.
7. The method provides consistent results derived from expert judgements, in the sense that they can be reproduced administratively .

Advantage 1 is key. Instead of a circular process of iterating through variables, weighting and functional form, the PoTCal approach only requires a single elicitation about a number of real world scenarios. This frees the expert from being a subject matter specialist, statistician and decision theorist at the same time. This ensures that the expert is used for their expertise, that is, knowledge of the system of interest. When seen as a regression problem, the design of scoring systems requires understanding of conditional and marginal relationships in high dimensional space. It is a technical area even for appropriately trained specialists. The subject matter expertise is still invaluable in selecting the scenarios and analysing the results, for example, considering the plausible relationships.

Previous experiments that have only compared linear models with expert derived models (e.g. Cook and Stewart (1975)) have not fully explored this issue. The simple linear model will typically be biased because of non-linearities. Therefore demonstrating that an expert defined scoring system outperforms a linear model has not necessarily demonstrated that it will outperform all models.

We do not currently possess the empirical data to demonstrate when one approach is better than the other. The application of either technique to credible examples will require significant resources. At this stage we are comfortable that logically the approach is an improvement in many cases. This is consistent with the findings of Hamm (1991) and Cook and Stewart (1975). We do not argue that it will be better in all cases but are confident that it will improve performance in many situations. For example, Cook and Stewart (1975) found that the regression based estimate was a greater improvement as the number of variables increased.

Advantage 2 is slightly subtle but an important consideration. When an expert is considering a scenario, they can factor in whatever knowledge they may possess. This promotes realism in the analysis, and the additional variation from the experts can be factored into the uncertainty about our ability to replicate the expert's views from the scenario attributes. This is also a significant difficulty. If the scenario does not contain variables the expert considers are significant, then guidance needs to be given about how to accommodate this. In our work we have asked the experts to mentally average over these variables (thus forming an expectation) but it is a significant complication. We note that the usual expert-based scoring systems are also impacted by these issues. We acknowledge that the examples presented here have not explored the issue of uncertainty. In the applications considered, the decision makers were satisfied in using the expected risk in decision making. In other analyses more detailed assessment may be warranted.

Advantage 3 is a significant improvement. The existence of variation between experts is well

recognised. While this causes difficulties for decision makers it is more transparent to quantify it and find a principled way of incorporating it into decision making. For example the decision maker could use the average response or the worst case or the best case as defined by the experts. Examples of modelling this variation is demonstrated in the two examples presented in this paper. The use of statistical models provides a rigorous framework for incorporating these variations. It allows inference about the population of experts. Note that this is different from the approach used in judgemental bootstrapping (Armstrong, 2001) were different models are fitted to different experts.

Advantage 4 is clear. When posed as a regression problem the data can be used to explore the nature of relationships between the scenario attributes and the expert's views. Nonlinearities, interactions and non significant variables can be identified and quantified. Trying to do this in an *ad hoc* way is difficult and it is hard to see how it could be done effectively otherwise. The expert views are still vital in considering the possible forms of the relationship, and the analysis still relies on the empirical relationships rather than any unverifiable assumptions. We note that, in practice, with limited experts and scenarios, the data may be too sparse to fully explore this issue. We also note that in almost all cases we have seen in the literature such as (Armstrong, 2001), Hamm (1991) and Cook and Stewart (1975) simple multiple linear regression was the approach used. If this model cannot fit the underlying data significant biases could result.

Advantage 5 arises because of the empirical nature of the weightings. Provided a set of scenarios is generated which spans the attributes of scenarios for which the model will be used, the model fitting will typically ensure that predictions are sensible. Thus there is less possibility of unrealistic scores. Note that it is important not to oversell this. Either approach will be susceptible to error if extrapolation occurs outside the domain considered by the experts. Extrapolation for regression models can lead to poor predictions so care needs to be taken.

Advantage 6, though theoretical, is still an important contribution to the development of these methods. Regression based approaches to developing scoring systems are well known in medicine. In these cases they typically have large data bases of cases (scenarios) with records of disease or condition development. They can then use regression methods based on the related attributes to construct scores. The PoTCal approach is identical, except we use the experts to assess the scenarios. This logical linkage to the empirical approach is reassuring.

The final advantage 7 is important because experts are inconsistent in making their judgements, which has been pointed out by many researchers, see e.g., Dawes (1971), Armstrong (2001) and (Hubbard, 2009, pg. 111). These researchers have proved, through their individual studies, that the mathematical models derived from expert judgements will often work better than experts themselves in terms of consistency, unbiasedness, reliability and repeatability. Obviously if there are large variations in experts' views, caution should used in using an averaged response.

We have applied the technique in several potentially contentious areas. We have been encouraged by its ease of implementation. While there is still significant work to be performed to ensure the experts are well directed and that the results are appropriate for decision making, we have found that the approach can be implemented to produce credible results. As previously mentioned, experts will often want information about complex scenarios that is not contained in the available attributes. In these cases a strategy, such as asking them to average over missing variables, is needed.

We have also been surprised about the acceptance of the results of a PoTCal analysis in contentious issues. Choosing contingency deballasting zones in Australia had the potential to impact a number of jurisdictions and industries. The acceptance of results occurs because of the

transparency of the approach and the limitation of subjective assessment to those components where there was no other way to approach the problem. By choosing the weightings objectively, an area of significant contention is removed. Provided the decision makers have confidence in the experts used, the extension to confidence in the results is not too difficult.

The approach is not a replacement for sound analysis. If the experts are misguided or wrong, then the PoTCal approach cannot redeem this. We note that expert-based scoring systems will suffer from the same shortcomings. PoTCal should not be oversold to imply it can correct these effects. It also relies on having enough data points to estimate parameters adequately.

The PoTCal approach can be applied to a wide number of areas beyond the biosecurity examples here. For example, in environmental monitoring the concept of ecosystem health is often of interest and ways of measuring this compound phenomenon with simple environmental measurement are required. The PoTCal approach would use experts to assess a sample of locations for ecosystem health and then apply regression to derive appropriate weightings to map the simple indicators to the complex phenomena. Any situation where experts are used in developing scoring systems is a candidate for the PoTCal approach.

References

Armstrong, J. S. (2001). *Principles of forecasting: a handbook for researchers and practitioners*, chapter Judgemental bootstrapping: inferring experts rules for forecasting, pages 171–192. Kluwer Academic Publishers, Boston.

Bax, N., Williamson, A., Aguero, M., Gonzalez, E., and Geeves, W. (2003). Marine invative alien species: a threat to globe biodiversity. *Marine Policy*, 27:313–323.

Bomford, M. and O'Brien, P. (1995). Eradication or control for vertebrate pests? *Wildlife Society Bulletin*, 23:249–255.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62:193–217.

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

Burgman, M. A., Keith, D. A., and Walshe, T. V. (1999). Uncertainty in comparative risk analysis for threatened Australian plant species. *Risk Analysis*, 19:585–598.

Cook, R. L. and Stewart, T. R. (1975). A comparison of seven methods for obtaining subjective descriptions of judgmental policy. *Organizational Behavior and Human Performance*, 13:31–45.

Cooke, R. M. (1991). *Experts in uncertainty: opinion and subjective probability in science*. New York, NY: Oxford University Press.

Cooke, R. M., Nauta, M., Havelaar, A. H., and van der Fels, I. (2006). Probabilistic inversion for chicken processing lines. *Reliability Engineering and System Safety*, 91(10-11):1364–1372.

Copp, G. H., Garthwaite, R., and Gozlan, R. E. (2005). Risk identification and assessment of non-native freshwater fishes: Concepts and perspectives on protocols for the UK. Technical Report no. 129, Cefas Lowestoft, UK.

Copp, G. H., Vilizzi, L., Mumford, J., Fenwick, G. V., Godard, M. J., and Gozlan, R. E. (2009). Calibration of FISK, an invasiveness screening tool for nonnative freshwater fishes. *Risk Analysis*, 29(3):457–467.

CRCAWM (2002). Cooperate Research Centre for Australian Weed Management: Annual Report 2001-2002. CRC for Australian Weed Management (CRCAWM), South Australia.

Crosti, R., Cascone, C., and Testa, W. (2007). Towards a weed risk assessment for the italian peninsula: preliminary validation of a scheme for the central mediterranean region in italy. In Rokich, D., Wardell-Johnson, G., Yates, C., Stevens, J., Dixon, K., McLellan, R., and Moss, G., editors, *Proceedings of the International Mediterranean Ecosystems (MEDECOS XI) Conference Perth, Western Australia, September 2-5*, pages 53–54.

Cunningham, D. C., Woldendorp, G., Burgess, M. B., and Barry, S. C. (2003). Proioritising sleeper weeds for eradication: Selection of species based on potential impacts on agriculture and feasibility of eradication. Australian Bureau of Rural Sciences, Canberra.

Dawes, R. M. (1971). A case study of graduate admission: Application of 3 principles of human decision making. *American Psychologist*, 26:180–188.

Du, C., Kurowicka, D., and Cooke, R. (2006). Techniques for generic probabilistic inversion. *Computational Statistics & Data Analysis*, 50:1164–1187.

Gordon, D. R., Onderdonk, D. A., Fox, A. M., , and Stocker, R. K. (2008a). Consistent accuracy of the Australian weed risk assessment system across varied geographies. *Diversity and Distributions*, 14:234242.

Gordon, D. R., Onderdonk, D. A., Fox, A. M., Stocker, R. K., and Gantz, C. (2008b). Predicting invasive plants in Florida using the Australian Weed Risk Assessment. *Invasive Plant Science and Management*, 1:178–195.

Groves, R. H. (2002). *Biological Invasions: Economic and Environmental Costs of Alien Plant, Animal, and Microbe Species*, chapter The impacts of alien plants in Australia. CRC Press, New York.

Groves, R. H., Hosking, J. R., Batianoff, G. N., Cooke, D. A., Cowie, I. D., Johnson, R. W., Keighery, G. J., Lepschi, B. J., Mitchell, A. A., Moerkerk, M., Randall, R. P., Rozefelds, A. C., Walsh, N. G., and Waterhouse, B. M. (2003). Weed categories for natural and agricultural ecosystem management. Australian Bureau of Rural Sciences, Canberra.

Hamm, R. M. (1991). Accuracy of alternative methods for describing experts' knowledge of multiple influence domains. *Bulletin of the Psychonomic Society*, 29(6):553–556.

Hubbard, D. W. (2009). *The Failure of Risk Management: Why it's broken and how to fix it*. John Wiley, New Jersey.

Hughes, G. and Madden, L. V. (2003). Evaluating predictive models with application in regulatory policy for invasive weeds. *Agricultural Systems*, 76(2):755–774.

James, A., Low Choy, S., and Mengersen, K. (2010). Elicitator: An expert elicitation tool for regression in ecology. *Environmental Modelling & Software*, 25(1):129–145.

Jones, R. N., Hennessy, K. J., Kenny, G. J., Suppiah, R., Walsh, K. J. E., Wet, N. D., and Whetton, P. H. (2002). Scenarios and projected ranges of change for mean climate and climate variability for the South Pacific. *Asia Pacific Journal on Environment and Development*, 9(1 & 2):1–42.

Kadane, J. B., Chan, N. H., and Wolfson, L. J. (1996). Priors for unit root models. *Journal of Econometrics*, 75:99–111.

Kadane, J. B., Dickey, J. M., Winkler, R. L., Smith, W. S., and Peters, S. C. (1980). Interactive elicitation of opinion for a Normal linear model. *Journal of the American Statistical Association*, 75(372):845–854.

Kadane, J. B. and Wolfson, L. J. (1998). Experiences in elicitation. *The Statistician*, 47(3):3–19.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., and Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79(3):216–247.

Knight, E., Barry, S., Summerson, R., Cameron, S., and Darbyshire, R. (2007). *Designated Exchange Areas Project - Providing informed decisions on the discharge of Ballast Water in Australia (Phase 2)*. Australian Bureau of Rural Sciences.

Kuhnert, P. M., Martin, T. G., and Griffiths, S. P. (2010). A guide to eliciting and using expert knowledge in Bayesian ecological models. *Ecology Letters*, In Press.

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(1):265–277.

Mazzuchi, T. A., Linzey, W. G., and Bruning, A. (2008). A paired comparison experiment for gathering expert judgment for an aircraft wiring risk assessment. *Reliability Engineering and System Safety*, 93:722–731.

Meyer, M. A. and Booker, J. M. (1990). *Eliciting and analyzing expert judgement: A practical guide*. Washington, DC: U.S. Nuclear Regulatory Commission.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116:2417–2424.

Murray, J. V., Goldizen, A. W., OLeary, R. A., McAlpine, C. A., Possingham, H. P., and Low Choy, S. (2009). How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? a case study using brush-tailed rock-wallabies petrogale penicillata. *Journal of Applied Ecology*, 46:842–851.

O'Hagan, A., Buck, C., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Expert's Probabilities*. Wiley, UK.

Pheloung, P. C., Williams, P. A., and Halloy, S. R. (1999). A weed risk assessment model for use as a biosecurity tool evaluating plant introductions. *Journal of Environmental Managament*, 57:239–251.

Schmitt, N. (1978). Comparison of subjective and objective weighting strategies in changing task situations. *Organizational Behavior and Human Performance*, 21(2):171–188.

Simester, D. I. and Brodie, R. J. (1993). Forecasting criminal sentencing decisions. *International Journal of Forecasting*, 9(1):49–60.

Stewart, T. R. (1990). A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *Weather and Forecasting*, 5:661–666.

Vose, D. (2008). *Risk analysis: a quantitative guide*. John Wiley & Sons, 3rd edition.

6 Acknowledgements

This report is the product of the Australian Centre of Excellence for Risk Analysis (ACERA). In preparing this report, the Authors acknowledge the financial and other support provided by the Department Of Agriculture, Fisheries and Forestry(DAFF) and the University of Melbourne.

This report has been improved by discussions with Keith Hayes, Greg Hood and Mark Burgman and the comments on an earlier draft from Mark Burgman and Mick McCarthy. Any remaining errors are our own.

Contact Us

Phone: 1300 363 400
+61 3 9545 2176

Email: enquiries@csiro.au

Web: www.csiro.au

Your CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.