Diversity and Distributions, (Diversity Distrib.) (2008) **14**, 581–591

BIODIVERSITY REVIEW

Evaluating extreme risks in invasion ecology: learning from banking compliance

James Franklin[1], Scott A. Sisson[1], Mark A. Burgman[2] and Jennifer K. Martin[3]*

[1]*School of Mathematics and Statistics, University of New South Wales, New South Wales, Australia,* [2]*Australian Centre of Excellence for Risk Analysis, School of Botany, University of Melbourne, Parkville, Victoria 3010, Australia,* [3]*Department of Zoology, University of Melbourne, Parkville, Victoria 3010, Australia*

*Correspondence: Mark A. Burgman, Australian Centre of Excellence for Risk Analysis, School of Botany, University of Melbourne, Parkville, Victoria 3010, Australia.
E-mail: markab@unimelb.edu.au

ABSTRACT

Increasing international trade has exacerbated the risks of ecological damage due to invasive pests and diseases. For extreme events such as invasions of damaging exotic species or natural catastrophes, there are no or very few directly relevant data, so expert opinion must be relied on heavily. Expert opinion must be as fully informed and calibrated as possible – by available data, by other experts, and by the reasoned opinions of stakeholders. We survey a number of quantitative and non-quantitative methods that have shown promise for improving extreme risk analysis, particularly for assessing the risks of invasive pests and pathogens associated with international trade. We describe the legally inspired regulatory regime for banks, where these methods have been brought to bear on extreme 'operational risks'. We argue that an 'advocacy model' similar to that used in the Basel II compliance regime for bank operational risks and to a lesser extent in biosecurity import risk analyses is ideal for permitting the diversity of relevant evidence about invasive species to be presented and soundly evaluated. We recommend that the process be enhanced in ways that enable invasion ecology to make more explicit use of the methods found successful in banking.

Keywords

Biosecurity, extreme risks, extreme value theory, invasion, outliers, trade, uncertainty.

INTRODUCTION

A risk is called 'extreme' when it concerns an event that may happen very rarely or never. Typically, 'extreme' is used for events that are of high (negative) consequence as well as low probability, but in this review the emphasis is on events of very low probability rather than on consequences. Such events are at the edge of or outside the range of what has occurred, possibly far outside. Any data are unlikely to be reliably representative. The problem of evaluating extreme risks is therefore fundamentally different from the standard statistical approach of choosing a model to describe a quantitative problem, fitting the parameters of the model to the data available, then using the resulting tuned model for prediction.

Probabilities of extreme events must therefore be evaluated by combining disparate sources of evidence, none of which are reliable in isolation. Sources include whatever data there are, how far the event of interest is from the data, the opinion of experts (possibly in diverse disciplines), arguments from analogy (that is, from events whose similarity to the event in question is debatable), specialist scientific causal knowledge relevant to the case, and commonsense knowledge. There is no established methodology either for computing or eliciting the probabilities arising from these sources of knowledge, or for combining them once discovered. But the reasons for the difficulty of reaching a correct answer are the same as the reasons why it is important to succeed – because when data are scarce, neglecting any source of evidence or any method of interpretation may lead to the misevaluation of extreme risks and to substantial, avoidable environmental costs.

Recent developments in international trade agreements have accelerated the rates of entry, establishment, and spread of invasive species (Karatayev *et al.*, 2007; Meyerson & Mooney, 2007). Invasions by new species have low but non-negligible chances of causing catastrophic changes in ecosystems. Risk assessments are severely hampered by a lack of data on such things as entry pathways and rare, long-distance dispersal events, especially when such processes are critically important for prediction (Karatayev *et al.*, 2007; Mack *et al.*, 2007; Nehrbass *et al.*, 2007). Methods for risk assessment in such cases often rely on poorly grounded expert opinion confronting either a 'presumption of innocence' or a 'precautionary principle' (Simberloff, 2005; Suedel *et al.*, 2007).

At the same time the World Trade Organization, through the decisions of its Appellate Body, has enforced certain constraints on risk assessments of potential invasions through international trade. Its 1998 determination on the importation of salmon into

DOI: 10.1111/j.1472-4642.2008.00478.x
www.blackwellpublishing.com/ddi

Australia (WTO, 1998), in particular, ruled out risk assessments that relied on mere 'possibilities' of invasion, while allowing that the actual probabilities it required instead need not be numerical but could be based on substantial but qualitative evidence. In effect, it imposes on those arguing for a restriction on imports an onus to establish some substantial (but not necessarily numerical) probability of the establishment of a pest in the importing country.

We describe a case from biosecurity where the risk assessment was already conducted in a tribunal-like style – much more so than is normal in scientific practice. To find what further lessons can be learned from an approach based on semilegal methods, we survey the Basel II banking compliance regime as it applies to operational risk. We also survey a range of quantitative and semiquantitative data exploration and analysis methods that have proved particularly applicable to extreme risk analysis. Taken together they suggest an overall approach to extreme risk analysis that may have particular utility for evaluating the risks associated with trade-related invasive pests, pathogens and diseases. We conclude by making recommendations on how to improve current practice.

A BIOSECURITY CASE STUDY: DISEASE RISK FROM TRADE IN APPLES

In response to requests from New Zealand to permit the import of apples to Australia, Australian biosecurity agencies analysed risks that the apples would introduce disease (AQIS, 1998; Biosecurity Australia, 2006). The main stakeholders were strongly motivated by opposite concerns – the New Zealanders were concerned that the likelihood of disease had been overestimated and representatives of the Australian apple industry were concerned that it had been underestimated. Both sides presented detailed scientific analyses. A 1998 report (AQIS, 1998) recommended against import, while a 2006 report (Biosecurity Australia, 2006) recommended for it, but only after onerous inspection and disinfection measures.

The analyses looked at the possible chains of causes by which diseases might become established in Australia. (The analysis of causal chains is an important topic in biosecurity and ecological studies – e.g. Kilpatrick *et al.*, 2006 – but is problematic: a chain may include low-probability events and the chain may be repeated many times. Breaking the chain into many units for analysis of probabilities is desirable, but it is difficult to choose the correct unit, especially for the critical lowest probability, and to deal with correlations between errors at different points in the chain.) A particularly difficult point in the analysis, and the one most relevant to the study of extreme risks, came in trying to evaluate the probability of what was believed to be the most unlikely event in the most likely chain, the transfer of the pest fire blight from a discarded apple to an Australian plant. There were many imponderables in the scenario – including specifying the scenario with any exactitude, mechanisms of transfer, levels of infection of apples, the distribution of the possible host plants, and seasonal differences in the probability of transfer. Since the probability of transfer was believed to be of the order of one in a million, experimentation was not feasible – it would take several million experiments to achieve any moderately reliable estimate of the probability. The analyses therefore relied on an expert review of marginally relevant evidence (Roberts *et al.*, 1998; Biosecurity Australia, 2006).

The complexity of the analyses in this case study meant they had answers to many potential questions. That made them robust in the politically charged atmosphere of import controversies, which included grilling of the regulator's (the Australian Quarantine and Inspection Service) Executive Director by a Senate committee on the possible motives of New Zealand scientists (Senate Hansard, 1997), direct recommendations by the Australian Senate Committee that the regulator should conduct its risk assessment more quantitatively (Senate, 2005), and comment by the New Zealand Minister for Agriculture that 'the concept of honest science has no meaning [in Australia]' (Knight, 2005). In addition to bilateral debate, countries need to comply with the guidelines of the International Plant Protection Convention and the decisions of the WTO's Appellate Body, and scientists naturally desire to show to the international scientific community that their results are not swayed by political pressures. Such pressures are stressful, in much the same way as it is stressful to be cross-examined in court by experienced legal counsel. From the point of view of the 'advocacy' model that we outline below for evaluating extreme risks, however, that is not necessarily a bad thing. Pressures from different directions are integral to the process and (at least if the pressures are reasonably balanced) they can encourage care and transparency in the risk evaluation process.

BANK OPERATIONAL RISK IN THE BASEL II COMPLIANCE REGIME

Bank operational risk is a rapidly developing area in which massive resources have been committed to the study of, in part, the quantification of extreme risks. A powerful international body, the Committee on Banking Supervision of the Bank for International Settlements in Basel, enforces the Basel II standards (Bank for International Settlements, 2002, 2004; Marrison, 2002). A bank's business is to take in funds and lend most of them out for profit while reserving some against risks. Credit and market risk are rich in data and statistically tractable. Operational risk, on the other hand, is a grab-bag of unusual and extreme events ranging from massive internal fraud to tsunamis, typing errors in crucial places, incompetent CEOs, and major technological change (King, 2001; Bank for International Settlements, 2002; Rosen & Coreggia, 2004).

It is difficult for a bank to quantify its operational risks. The diversity of hazards and the lack of data are major challenges. Internal frauds, for example, are rarely reported publicly by individual banks unless they are catastrophic. Therefore, most banks have very little data on past events of the sort that may impact on them severely in the future. The paucity of data means that it is essential to combine what data there are with expert opinion (O'Hagan *et al.*, 2006).

Basel II permits larger banks to evaluate their risks using any models and statistical technology they wish, provided they disclose them to the (national) regulator (in the USA, The Federal

Reserve, in the UK, the Bank of England) and the regulator approves. That naturally allows free rein for statistical expertise, including the use of sophisticated methods relevant to extreme values, both on the side of banks and on the side of the regulator. It promises to improve risk evaluation greatly, by encouraging improvements in data collection and in risk analysis methods appropriate for the context.

There is a fundamental conflict between the perspective of the bank and that of the regulator. The bank wishes to minimize its calculated risks so as to be able to reserve less funds against them, allowing the bank to lend out and make profit on as much money as possible. The regulator, on the contrary, wishes to ensure that the bank fully states its risks and reserves against them, so that the bank and the whole banking system remain stable. In operational risk in particular, where unusual 'one-off' major events have occurred or may occur, there is potential for the results of argument about particular cases to make a large difference to the amount of funds that a bank is required to hold in reserve and thus make no profit from (Franklin, 2005). It is that conflict of perspectives and inherent disputability of individual data points that has led the banking industry to develop a package of mathematical and legally inspired methods, from which other areas such as biosecurity can learn.

Extreme risk analysis under Basel II is essentially inspired by the familiar 'adversary' model of reaching decisions in (Anglo-American) legal cases, but has adaptations to suit the more quantitative nature of the data and the more cooperative relation that exists between the regulator and regulatee than exists between opposing counsel in a court of law. We suggest the name 'advocacy model' for the result.

It is mandated that larger banks at least should quantitatively model the probability of losses of various sizes in each of 56 cells – eight 'loss types' (such as external fraud, damage to physical assets) in seven 'business lines' (such as retail banking, asset management). An individual bank may have no or very few data points (over say the last five years) in some cells but hundreds in others. It is also mandated that the loss models should take into account four types of evidence: internal data, relevant external data (that is, aggregated data on other banks, possibly in other countries), scenario analysis (that is, 'what-if' analyses conducted by teams of experts on situations of financial stress), and 'factors reflecting the business environment and internal control systems'. The models are expected to use state-of-the-art statistical methods such as Extreme Value Theory (EVT), with justifications of the distributional assumptions used. Correlations between the losses in different cells should also be modelled.

That provides a rigid and demanding framework for the format in which loss probabilities must be reported, but it is recognized that there are many points at which informed human judgement must come into play. They include borderline cases as to which losses should be classified into which cells (or divided among cells), the time to which a loss should be attributed, the likelihood of a previously experienced large loss recurring now that precautions against it have been taken, the relevance of external industry-wide data to the individual bank's case, and the judgements reached about the correlations between extreme losses in different cells (for example, estimating the impact of an IT 'meltdown' on the bank's various lines of business). The bank's internal modellers and the regulator both understand that the outcome of the process – the figure that the regulator requires the bank to hold in reserve – is very sensitive to both individual large-loss data points and to assumptions about distributions and scenarios. Thus, the quantitative models are regarded as an essential starting point but are also taken 'with a grain of salt'; they form the starting point for negotiations between modellers and regulators, often mediated by consultants. The consultants, specialists in operational risk from an independent firm, look at the modellers' attempts and advise on changes needed to meet the regulator's standards, while assuring the regulator that the modellers are reasonable in their assumptions and conclusions (or soon will be). Feedback proceeds up and down the line in a generally cooperative atmosphere.

The essential lesson that can be learned from the advocacy model as practised in bank operational risk assessment is that the normally cooperative but potentially adversarial relationship between quantitatively astute parties on either side encourages the utmost use of sophisticated quantitative methods like EVT to make the most of data, but at the same time permits honesty in allowing all parties to understand and admit exactly where expert judgement goes beyond the data.

OUTLIERS, CRITICISM OF INDIVIDUAL DATA POINTS, AND EXTREME VALUE THEORY

We first review the methods that have been found in bank operational risk and elsewhere to be particularly useful in making sense of any available extreme data points. One must identify the data points, then examine any relevant knowledge of them individually, and lastly apply the EVT formalism where appropriate.

First, one must distinguish extreme values from outliers. Extremes are the data points at the edge of the distribution and thus are the most important points for predicting what will occur near the edge and beyond the range of the data. Outliers, in contrast, are not part of the distribution at all. Typically an outlier is a mistake and it should be deleted from the data. But an outlier may also indicate contamination of the data, which could indicate an event deserving further investigation. For example, an outlier of a measure of an environmental pollutant may indicate an illegal discharge – the normal range of data comes from natural processes, but the illegal discharge is from a different cause or distribution.

There are many methods for outlier detection in fault monitoring, intrusion detection, and the like (Barnett, 2004; Hand & Bolton, 2004; Hodge & Austin, 2004). All involve modelling the data in some way, finding a distribution that fits (most of) the data well, and which, if true, implies that the outlier is very unlikely to have occurred. This process emphasizes the problem of having to know the distribution of the data, especially in cases where the data, including the outlier(s), are the only source of knowledge of the distribution. There is no good solution to this fundamental problem. There has been progress on it in the field of fraud detection, where the aim is to identify by automatic methods 'unusual' or 'fringe' data points in large data sets that warrant further investigation. Pattern recognition (Bolton & Hand, 2002) has been applied to such problems as money laundering detection and intrusion in computer systems, but far less to such potential areas as environmental monitoring, epidemic alerting, or quarantine inspection (although see ANZECC/ARMCANZ, 2000).

Second, having made the most of automatic methods to delete outliers, one must consider one by one the most extreme remaining cases with a critical eye. A characteristic of extreme risk is the existence of individual data points whose relevance is itself a matter of dispute. Methods for extrapolating extreme values (see below) are very sensitive to the few most extreme values in the data, so great care needs to be taken in determining if those values are from the same distribution, that is, fully relevant to the prediction problem at hand. Purely quantitative methods are not suitable for examining data points where there is extra knowledge of the particular case. An 'advocacy' model should include drilling down to a critical consideration of individual dubious data points. For example, usually steps are taken to prevent disasters or near-disasters recurring, so the relevance of a past incident to present evaluations is unclear. For example, the National Australia Bank (NAB) lost about $A330 million in rogue trading in 2004. It is one of the few large (known) operational risk losses in recent times by an Australian bank. Its relevance to present operational risk evaluations, whether for NAB or other banks, is unclear, since NAB has had to submit detailed evidence of the steps it has taken to prevent a recurrence (Anon, 2005).

Another instance comes from the 1998 Import Risk Assessment for New Zealand apples (AQIS, 1998). The assessment dealt mainly with the risk of imported apples causing an outbreak of fire blight, a disease that is not present in Australia. In 1997, while the report was being prepared, fire blight was discovered on two shrubs in the Royal Botanic Gardens, Melbourne, and the assessment was suspended, pending study and eradication. It never became clear how the disease entered the Gardens or for how long it had been there. The outbreak did not spread. It was considered unlikely that the outbreak was caused by the import of commercial fruit. The IRA concluded that nothing of relevance to the IRA could be learned from the episode (AQIS, 1998; pp. 9, 23).

Debate about such individual cases is essential for the analysis of extreme risks because often, the cases are well studied, so much may be learned from them. Reasonable judgements can be made as to whether the same could happen again (or something different from a similar cause). It is unsatisfactory simply to delete the data point as no longer relevant, without careful consideration of context.

Third, having reached agreement on what the data set of extreme values is, one should apply EVT. EVT is the study of the extrapolation of the tails of distributions beyond the range of existing data (Embrechts *et al.*, 1997; Embrechts, 2000). The statistical modelling of extremes has its origins in the analysis of problems such as predicting 'once in a hundred years' floods from observations of the annual maximum river heights (e.g. Kotz & Nadarajah, 2000; Coles, 2001; Beirlant *et al.*, 2004; see Appendix for some technical details).

Extreme value theory is a very powerful, mathematically justified mechanism for making inference on extreme levels of a process. It is capable of evaluating the likelihood of future extreme values occurring beyond both the levels and the time-span of the observed data. It is also easily adaptable to the Bayesian framework, thus allowing incorporation of any quantitative prior knowledge of the situation (Coles & Pericchi, 2003; Sisson *et al.*, 2006).

However, it has some obvious limitations, which means that despite its potential for making the most of what data there are, it is inadequate as a stand-alone method of evaluating extreme risks for biosecurity. To have some belief in the outcome, one is required to have belief in the underlying assumptions: that there are sufficient data for the limiting asymptotic models to be valid; that dependence in the data has been adequately modelled; that for predictive purposes the future state of the model (e.g. incorporating estimated trends, or explicitly modelled system change-points) is known or estimated. And even if the assumptions are true, EVT is no more capable than other statistical methods of magically extracting reliable predictions from tiny data sets. In the contexts of extreme risks that preoccupy bank operational and biosecurity risk analyses, data of the kind necessary to reliably parameterize extreme value distributions are unlikely to be available. In these situations EVT acts especially to guard against dangerous illusions of false precision in extreme risk estimates and underestimates of tail probabilities. It can distinguish between reasonable and unreasonable orders of magnitude in estimates of extreme risks, but honesty requires admitting the imprecision in the answers.

Extreme value theory is not the only (relatively) new statistical method that has particular relevance to extremes. A number of other new statistical (or marginally statistical) methods have emerged in recent years, which *prima facie* have good possibilities for application to extreme risk analysis. Data mining has shown the possibilities of extracting value from large data sets and has proved its value to business in understanding customer behaviour; its applications to fraud detection are relevant to extreme risks because of their ability to distinguish the main body of data from extremes, especially in hard-to-visualize multivariate data. Many risks are spatially variable (for example the chance of transfer of fire blight from discarded apple cores to hosts is very dependent on the spatial distributions of cores, hosts, and vectors), and the general inadequacy of coverage of the space by data means there is (or should be) strong interaction between the methods of spatial statistics and extreme risk analysis.

ROBUSTNESS: IMPRECISE PROBABILITIES, SENSITIVITY ANALYSIS, DECISION THEORY

In traditional statistics, probabilities are based on large data sets or on physical considerations such as symmetry and are thus quite precise. In extreme risk evaluation, that is not the case because of the lack of data, so attempts to impose numerical precision on probabilities results in (potentially dangerous) distortion. Therefore, methods of handling and communicating imprecision in probabilities are of special significance in extreme risk analysis (though useful more widely). In dealing with extremes, risk analysis should take its cue from the law, which has resisted quantifying the criminal standard of 'proof beyond reasonable doubt': any attempt to lay it down as a precise number will not lead to improved consistency in decision-making, as it is impossible to determine a numerical probability that a defendant is guilty on the evidence (even if it is clear that the person is guilty 'quite certainly' or only 'on the balance of probabilities') (Franklin, 2006).

There is always some awkwardness in dealing explicitly with imprecision in probabilities. Probability is a measure of uncertainty, so dealing with 'uncertainty in uncertainty' or 'probabilities of probabilities' can easily seem overly elaborate and too confusing, especially if one attempts to represent the imprecision by some formal method. However, false precision can be unnatural and costly. The ubiquity of fuzzy language in discussing probability, embodied in terms such as 'extreme risk', 'quite likely', and 'a remote chance', is a sign that people are comfortable with imprecision and find it adequate in representing their ideas on probability (e.g. Olson & Budescu, 1998). Since people often use words in preference to numbers in discussing risks, it is fortunate that probabilistic words can be reasonably well calibrated across individuals (Franklin, 2001; pp. 324–5).

There are four ways of dealing with imprecision in probabilities. All have value and one (or more) can be chosen according to the pragmatic needs of the problem. In increasing order of sophistication they are:

- Keeping to fuzzy natural language and studying its grounding in numerical probabilities.
- Restricting numerical probabilities to one significant figure.
- Representing imprecise probabilities in some simple way such as by probability bounds or triangular distributions and using them to conduct sensitivity analyses.
- Using decision theory to study directly the robustness of decisions to imprecision in the probabilities.

We survey briefly what is achievable by each method. People operate naturally with language-based descriptions of probabilities such as 'very likely' and may prefer to use them for reporting so as to avoid false precision (or to maintain deniability). They are prevalent in scorecards and informal risk discussions (e.g. Table 1).

Such translation tables encounter the problem that natural language is in general highly context-sensitive (a small elephant is bigger than a big mosquito because being a small elephant is being small *for an elephant* – with reference, that is, to the mean in the appropriate, context-dependent, reference class). There is some consistency in how subjects translate verbal to numerical probabilities, but some individual variability (Wallsten *et al.*, 1993) and sensitivity to context (Fox & Irwin, 1998; Burgman, 2005; p. 77). Verbal probabilities and translation tables are usable in the elicitation and communication of risk judgements, but only with great care to ensure that experts and non-experts mean the same by such expressions as 'extreme risk' and that they mean the same in one risk setting as in another.

A particular contextual matter, often commented on in bank operational risk, is the need for clarity in the time period to which the risk refers: a loss that has a one-in-a-thousand chance of happening in a day is quite likely to happen in a year. It is much easier to clarify such matters with numbers than with words.

Reporting numerical probabilities to only one significant figure (for example, 0.4 or 2×10^{-6} but not 0.41 or 2.4×10^{-6}) is a common practice but one usually done unreflectively (Phillips & LaPole, 2003). It relies on the fact that it is quite rare for decisions to be sensitive to differences in probability of less than one significant figure: an annual chance of three-in-a-million may warrant higher precautions than a chance of one-in-a-million, but it is hardly likely that one will take much notice of the difference between one-in-a-million and 1.3-in-a-million, even if one is convinced that the difference in the chances is real and not just measurement error. However, the cognitive basis for reporting can have a significant effect on risk perception. Relative risks (e.g. the risk has increased by 30%) communicate a different message from absolute frequencies (e.g. the annual risk increased by 0.3 in a million) or natural frequencies (e.g. out of 10 million people, we expect an additional three people to die this year; Gigerenzer, 2002).

To report a probability to one significant figure is to make implicit use of an interval-valued probability, since by 'probability 0.4' one means 'probability in the range 0.35–0.45'. Explicit use of bounded probabilities (Walley, 1991; Ferson *et al.*, 2004) may involve intervals with the implicit assumption of a uniform distribution between the bounds, or triangular distributions with a 'midpoint' that is the best estimate of the probability and a (not necessarily symmetric) range of uncertainty on either side (cf. Burgman, 2005; pp. 78–9).

The use of interval probabilities encourages sensitivity analyses, since it is easy to calculate what would happen if the ends of the ranges were used. For example, Biosecurity Australia's

Table 1 Nomenclature for qualitative likelihoods, corresponding semiquantitative probability intervals (after Biosecurity Australia, 2006; Table 12).

Likelihood	Qualitative descriptors	Probability interval
High	The event would be very likely to occur	0.7 → 1
Moderate	The event would occur with an even probability	0.3 → 0.7
Low	The event would be unlikely to occur	$5 \times 10^{-2} \rightarrow 0.3$
Very low	The event would be very unlikely to occur	$10^{-3} \rightarrow 5 \times 10^{-2}$
Extremely low	The event would be extremely unlikely to occur	$10^{-6} \rightarrow 10^{-3}$
Negligible	The event would almost certainly not occur	$0 \rightarrow 10^{-6}$

apple risk analysis (Biosecurity Australia, 2006; pp. 114–5) concludes that 'a maximum value three times larger than the value agreed by the IRA team for every exposure value results in an overall risk ... that just exceeds Australia's appropriate level of protection'. There is, however, a conceptual difficulty with the idea of interval-valued, bounded or triangular probabilities – the ends themselves appear to be precise, whereas they are not (since if the probability itself is not known precisely, it is hardly likely that bounds on it will be).

Decision theory can provide a framework for more complete sensitivity analyses. For instance, info-gap decision theory (Regan *et al.*, 2005; Ben-Haim, 2006) deals with the sensitivity of decisions to uncertainty in the inputs by calculating outcomes for a continuous range of deviations from estimates of parameters and models. A decision maker can impose a range of acceptable possibilities for the output and map the range of inputs that lead to acceptable outputs.

All these techniques are needed in the toolkit of any risk analyst, but especially of the analyst of extreme risk who cannot fall back on traditional data-rich statistical methods.

STRENGTHS AND WEAKNESSES OF INTUITIVE REASONING UNDER UNCERTAINTY

Given that some reliance on expert judgement is inevitable in the analysis of extreme risks, it is essential to understand where expert judgement can be taken to be reasonably reliable and where it cannot. The story is a complex one.

The systematic errors of experts are well documented – their overconfidence, inability to know where their expertise ends, sensitivity to framing effects, confusion over base rates and conditional probabilities, sensitivity to the order of presentation of evidence and so on (see Gigerenzer, 2002; Burgman, 2005; Tetlock, 2005). Certainly, one cannot accept human judgements of probabilities uncritically. However, subjective assessments of risk expressed in words are reasonably accurate in many circumstances. In forecasting such quantities as stock prices, human 'judgemental forecasting' is still generally comparable to the best statistical methods (e.g. Lawrence & O'Connor, 1992). Risks expressed in words are reasonably accurate in circumstances such as stock prices where the system is stable and well known, data are plentiful, feedback is immediate, and there are strong incentives to improve performance.

Extreme risks do not satisfy those conditions, and people's reasoning and fuzzy representations begin to break down as events become rarer and their consequences become more extreme and visible. For example, people tend to over-weight small risks based on written information, but under-weight those based on (lack of) experience (Hertwig *et al.*, 2004). Workable probabilistic methods may exist for low-data, poor-feedback, extreme risk situations if suitable precautions are taken. It remains an important area for research.

Human judgement is superior to formal methods of inference in combining evidence from different sources – partly because of the inability of formal methods to offer any useful guidance at all. The simplest model of this problem is a form of the 'reference class problem', which asks how to combine statistical evidence from the different classes to which an individual belongs. The most basic evidence for probabilities is observation of a relative frequency. For example, the probability that Tex is rich, given that Tex is a Texan and 90% of Texans are rich, is 0.9. But typically, a case is a member of many classes, in which relative frequencies vary. There is no formal way to combine the probabilities arising from different 'reference' classes. For example, if the evidence is that Tex is a Texan philosopher, that 90% of Texans are rich and 10% of philosophers are rich, then standard statistical methods provide no guidance on how to combine these two numbers to achieve a numerical probability that Tex is rich, on the given evidence (Hájek, 2007).

The problem has caused a great deal of trouble in, for example, the law of evidence, where often there is evidence of different classes but it is of dubious legal relevance (Colyvan *et al.*, 2001; Tillers, 2005), and in attempts to construct medical diagnosis expert systems, where combining evidence from different symptoms is essential but how to do it is theoretically poorly understood. Nevertheless, humans successfully and intuitively combine evidence from different classes, weighting them in some way that is based on the comparative experience that has gone into each class. For example, they can learn enough about being Texan, being a philosopher, and being rich to have some sense of whether being Texan or being a philosopher is more likely to be relevant to being rich. This remarkable ability is very relevant to extreme risks, where disparate pieces of evidence need to be combined. There is no choice but to rely on human intuition to accomplish the task.

ADVERSARY AND ADVOCACY MODELS FOR EXTREME RISK ANALYSIS

Given that human probabilistic judgement has to be relied on to a considerable extent in extreme risk analysis and that it has potential, it needs to be asked how it can be appropriately constrained to take advantage of its strengths but avoid its weaknesses. We suggest that an 'advocacy model' is ideal. It replaces the ideal but unavailable feedback of real experience with the 'virtual' feedback provided by the scrutiny of experts' assessments, a neutral panel of 'judges', informed by the scenarios and reasoning put forward by potentially hostile stakeholders (Hagafors & Brehmer, 1983; Lee *et al.*, 1999; Fig. 1). It gives human intuition the last word in combining the evidence to reach a final conclusion, while allowing maximum space for the use of technical methods to support it. Figure 1 shows governance structures that would support an 'advocacy' model for the evaluation of extreme risks.

The advocacy model works because of the psychological pressure it applies. True accountability requires that the people to be held accountable fear their judges. They must be motivated by anxiety as to what the judges' views might be (Tetlock, 1983; Simonson & Staw, 1992; Siegel-Jacobs & Yates, 1996). The authority to which justification is submitted must be perceived as legitimate and itself having the expertise to evaluate the justification (Lerner & Tetlock, 1999). Many, but not all, cognitive

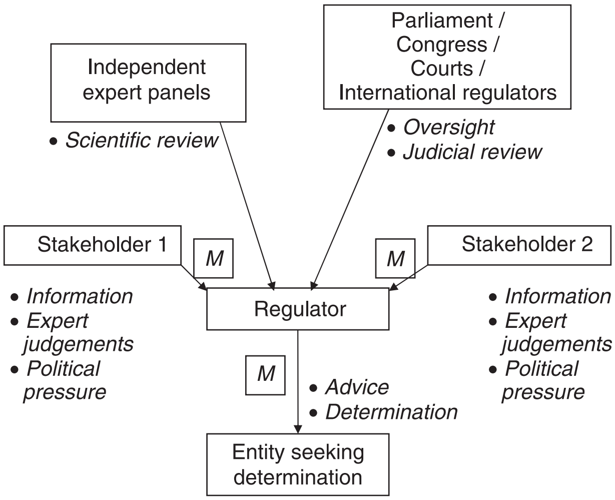

Figure 1 Representation of governance structures that would support an 'advocacy' model for the evaluation of extreme risks. The '*M*' stands for a potential role for independent mediators/facilitators to manage interactions between stakeholders and regulators. For Basel II, the entity seeking determination is a bank seeking approval from a national banking regulator that their risk analysis is adequate.

biases are attenuated by accountability, including hastiness in judgement, lack of awareness of one's own judgement processes, overconfidence, oversensitivity to the order in which information appears, pursuing sunk costs, and groupthink (accountability is not helpful with insensitivity to base rates and insensitivity to sample size).

The essential idea in the advocacy model for extreme risk analysis is to have an *adversary* look for the weaknesses in an assessment. Such methods have proved useful in, for example, software testing (Myers, 2004; p. 15) and computer security (Klevinsky *et al*., 2002). The most developed and best-known use is the system of legal trials in Anglo-American law. The two sides are represented by counsel who have wide discretion to put their cases as they think fit, though the judge moderates the process to some degree. The final decision is made either by a jury which acts as a 'black-box' fact evaluator, or by a judge or panel of judges who deliver reasons for their judgement. The model encourages effort to present a rational case that will be as convincing as possible, while leaving the final decision to disinterested parties.

That model can be problematic when there is a need to evaluate technical complexity. For example, in medical negligence or complicated financial cases, the evidence may be beyond the understanding of juries or legally trained professionals. It also tends to be impervious to some systematic errors, for example, psychological evidence on the low reliability of eyewitness identification evidence (Wells & Olson, 2003).

Compliance regimes that regulate industries such as banking have adopted models that have some of the qualities and advantages of a trial but are fundamentally different. Typically, a regulatory body oversees the compliance with published standards by the participants in the regulated industry. A party seeking a determination from the authority (for example, permission to import a commodity or to continue to lend money) submits extensive documentation, typically about risk measurement and mitigation. The documentation may be prepared by specialists, sometimes outside consultants who work with insiders. The documentation is examined by experts from the regulator, who can and typically do demand further documentation on matters they consider possibly suspicious or poorly described. In some cases, draft determinations may be published and comment from stakeholders may be invited. After some rounds of queries, investigations, data gathering, and inspections, a decision is reached. A generally cooperative attitude is maintained between the regulator and body regulated, except in extreme cases.

The degree of confidentiality of the process varies; in cases of accreditation confidentiality is normal during the process to encourage honesty in sharing of data, but a public report is issued at the end of the process. The regulator is responsible to some outside body such as Parliament, and is subject to embarrassment if a risk it has overlooked results in preventable deaths, irreversible environmental harm, or substantial financial or social losses.

The case studies described above in which an advocacy model was used in one form or another (bank operational risk and Biosecurity Australia's apple risk analysis) show how the model has acted to force the parties involved to work hard to identify and quantify all the risks and to lay them out for inspection. The success of these cases indicates why further study and implementation of such models is desirable.

In planning the implementation of an advocacy model, a number of administrative issues arise such as the exact locus of final judgement, incentives to improve technical analyses, compensation, security of tenure, financial arrangements for tribunals, stakeholders and consultants, and the like. These are important issues in ensuring the independence and credibility of the decisions reached by the process – indeed, there are a few cases of spectacular failures of semijudicial regulatory tribunals from problems in these areas (Franklin, 2007). Research on these questions needs to draw on expertise in public administration and corporate governance. The multidisciplinary nature of extreme risk analysis requires expertise in the social sciences as much as in mathematical statistics and psychology.

CONCLUSION AND RECOMMENDATIONS

In data-poor but decision-critical situations such as extreme risk evaluation, it is necessary to give human intuition the last word in risk assessment, while at the same time using formal quantitative and qualitative methods as a kind of prosthesis to supplement and control its known weaknesses. An 'advocacy' model, as implemented in bank operational risk and to some degree in biosecurity, provides a strong framework for allowing the interplay of formal methods and human intuition.

We recommend two ways in which the so far successful use of 'advocacy' models can be consolidated and extended.

The first concerns education. The strongly quantitative style of education in statistics, valuable as it is, can lead to a neglect of the more qualitative, logical, legal and causal perspectives needed to understand data intelligently. That is especially so in extreme risk analysis, where there is a lack of large data sets to ground solidly quantitative conclusions, and correspondingly a need to supplement the data with outside information and with argument on individual data points. So we suggest better education of statisticians in non-numerical methods including legal-style advocacy. At the same time, risk evaluators in general need better education in certain statistical methods (extreme value theory, Bayesian methods of combining expert opinion with data, and robust decision methods); and all parties need to understand the psychological findings on expert judgement.

The second recommendation is for the use of independent facilitators to mediate between the regulator/evaluator and the client/stakeholder. Business compliance protocols such as bank risk analyses make extensive use of independent mediators between the final risk evaluator (the regulator) and the client whose risk analysis has to pass inspection. They have the potential to become an important part of the advocacy model for extreme risk analysis, when applied to the management of the risks of invasive species associated with increasing international trade.

ACKNOWLEDGEMENTS

This work was supported by the Australian Centre of Excellence for Risk Analysis. We are grateful to Keith Hayes, Colin Thompson, and two anonymous reviewers for comments that improved earlier drafts.

REFERENCES

Australia and New Zealand Environment and Conservation Council & Agriculture and Resource Management Council of Australia and New Zealand (ANZECC/ARMCANZ). (2000) *Australian and New Zealand water quality guidelines*. ANZECC/ARMCANZ, Canberra, ACT.

Australian Quarantine Inspection Service (AQIS). (1998) *Final import risk analysis of the New Zealand request for the access of apples into Australia*. Commonwealth of Australia, Canberra, ACT. Available at http://www.affa.gov.au/corporate_docs/publications/pdf/market_access/biosecurity/plant/ACF133.pdf.

Bank for International Settlements, Basel Committee on Banking Supervision (2002) Sound practices for the management and supervision of operational risk, December 2001, revised July 2002.

Bank for International Settlements, Basel Committee on Banking Supervision (2004) Basel II: international convergence of capital measurement and capital standards: a revised framework, June 2004, Available at http://www.bis.org/publ/bcbs107.htm.

Barnett, V. (2004) *Environmental statistics: methods and applications*. Wiley, Chichester, UK.

Beirlant, J., Goegebeur, Y. & Teugels, J. (2004) *Statistics of extremes: theory and applications*. Wiley, Hoboken NJ.

Ben-Haim, Y. (2006) *Info-gap decision theory decisions under severe uncertainty*, 2nd edn. Oxford.

Biosecurity Australia (2006) *Final import risk analysis report for apples from New Zealand*, Part B. Department of Agriculture, Fisheries and Forestry, Government of Australia, Canberra, ACT.

Bolton, R.J. & Hand, D.J. (2002) Statistical fraud detection: a review. *Statistical Science*, **17**, 235–255.

Bottolo, L., Consonni, G., Dellaportas, P. & Lijoi, A. (2003) Bayesian analysis of extreme values by mixture modelling. *Extremes*, **6**, 25–47.

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

Coles, S.G. (2001) *An introduction to statistical modelling of extreme values*. Springer, London.

Coles, S.G. & Pericchi, L.R. (2003) Anticipating catastrophes through extreme value modelling. *Applied Statistics*, **52**, 405–416.

Coles, S.G. & Powell, E.A. (1996) Bayesian methods in extreme value modelling: a review and new developments. *International Statistical Review*, **64**, 119–136.

Colyvan, M., Regan, H.M. & Ferson, S. (2001) Is it a crime to belong to a reference class? *Journal of Political Philosophy*, **9**, 168–181.

Embrechts, P., ed. (2000) *Extremes and integrated risk management*. Risk Waters Group, London.

Embrechts, P., Klüppelberg, C. & Mikosch, T. (1997) *Modelling extremal events for insurance and finance*. Springer, New York.

Ferson, S., Nelsen, R.B., Hajagos, J., Berleant, D.J., Zhang, J., Tucker, W.W., Ginzburg, L.R. & Oberkampf, W.L. (2004) *Dependence in probabilistic modeling, Dempster–Shafer theory, and probability bounds analysis*. SANDIA report 2004-3072. Sandia National Laboratories, New Mexico.

Fox, C.R. & Irwin, J.R. (1998) The role of context in the communication of uncertain beliefs. *Basic and Applied Social Psychology*, **20**, 57–70.

Franklin, J. (2001) *The science of conjecture: evidence and probability before Pascal*. Johns Hopkins University Press, Baltimore, MD.

Franklin, J. (2005) Risk-driven global compliance regimes in banking and accounting: the new Law Merchant. *Law, Probability and Risk*, **4**, 237–250.

Franklin, J. (2006) Case comment: quantification of the 'proof beyond reasonable doubt' standard. *Law, Probability and Risk*, **5**, 159–165.

Franklin, J. (2007) International compliance regimes: a public sector without restraints. *Australian Journal of Professional and Applied Ethics*, **9**, 86–95.

Gigerenzer, G. (2002) *Reckoning with risk: learning to live with uncertainty*. Penguin, London.

Hagafors, R. & Brehmer, B. (1983) Does having to justify one's judgments change the nature of the judgment process? *Organizational Behavior and Human Performance*, **31**, 223–232.

Hájek, A. (2007) The reference class problem is your problem too. *Synthese*, **156**, 563–585.

Hand, D.J. & Bolton, R.J. (2004) Pattern discovery and detection: a unified statistical methodology. *Journal of Applied Statistics*, **31**, 885–924.

Hertwig, R., Barron, G., Weber, E.U. & Erev, I. (2004) Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, **15**, 534–539.

Hodge, V.J. & Austin, J. (2004) A survey of outlier detection methodologies. *Artificial Intelligence Review*, **22**, 85–126.

Karatayev, A.Y., Padilla, D.K., Minchin, D., Boltovskoy, D. & Burlakova, L.E. (2007) Changes in global economies and trade: the potential spread of exotic freshwater bivalves. *Biological Invasions*, **9**, 161–180.

Kilpatrick, A.M., Daszak, P., Goodman, S.J., Rogg, H., Kramer, L.D., Cedeño, V. & Cunningham, A.A. (2006) Predicting pathogen introduction: West Nile Virus spread to Galápagos. *Conservation Biology*, **20**, 1224–1231.

King, J.L. (2001) *Operational risk: measurement and modelling*. Wiley, New York.

Klevinsky, T.J., Laliberte, S. & Gupta, A. (2002) *Hack I.T.: security through penetration testing*. Addison-Wesley, Boston, MA.

Knight, J. (2005) Underarm bowling and Australia-New Zealand trade. *Australian Review of Public Affairs* 18 July 2005. Available at http://www.australianreview.net/digest/2005/07/knight.html.

Kotz, S. & Nadarajah, S. (2000) *Extreme value distributions: theory and applications*. Imperial College Press, London.

Lawrence, M. & O'Connor, M. (1992) Exploring judgemental forecasting. *International Journal of Forecasting*, **8**, 15–26.

Lee, H., Herr, P.M., Kardes, F.R. & Kim, C. (1999) Effects of choice accountability, issue involvement, and prior knowledge on information acquisition and use. *Journal of Business Research*, **45**, 75–88.

Lerner, J.S. & Tetlock, P.E. (1999) Accounting for the effects of accountability. *Psychological Bulletin*, **125**, 255–275.

Mack, R.N., Von Holle, B. & Meyerson, L.A. (2007) Assessing invasive alien species across multiple spatial scales: working globally and locally. *Frontiers in Ecology and the Environment*, **5**, 217–224.

Marrison, C. (2002) *Fundamentals of risk management*. McGraw-Hill, Boston, MA.

Martin, J.K. (2006) Den-use and home range characteristics of bobucks (*Trichosurus cunninghami*) resident in a forest patch. *Australian Journal of Zoology*, **54**, 225–234.

Martin, J.K. & Handasyde, K.A. (2007) Comparison of bobuck (*Trichosurus cunninghami*) demography in two habitat types in the Strathbogie Ranges, Australia. *Journal of Zoology*, **271**, 375–385.

Meyerson, L.A. & Mooney, H.A. (2007) Invasive alien species in an era of globalization. *Frontiers in Ecology and the Environment*, **5**, 199–208.

Myers, G.J. (2004) *The art of software testing*, 2nd edn. Wiley, Hoboken, NJ.

Nehrbass, N., Winkler, E., Mullerova, J., Pergl, J., Pysek, P. & Perglova, I. (2007) A simulation model of plant invasion: long-distance dispersal determines the pattern of spread. *Biological Invasions*, **9**, 383–395.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.E., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E. & Rakow, T. (2006) *Uncertain judgements – eliciting experts' probabilities*. Wiley, London.

Olson, M.J. & Budescu, D.V. (1998) Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making*, **10**, 117–131.

Phillips, C.V. & LaPole, L.M. (2003) Quantifying errors without random sampling. *BMC Medical Research Methodology*, **3**, 9.

Regan, H.M., Ben-Haim, Y., Langford, B., Wilson, W.G., Lundberg, P., Andelman, S.J. & Burgman, M.A. (2005) Robust decision-making under severe uncertainty for conservation management. *Ecological Applications*, **15**, 1471–1477.

Risk Management Magazine (2005) Inside NAB's mightmare. 21 September 2005. Available at http://www.riskmanagementmagazine.com.au/articles/25/0c036625.asp.

Roberts, R.G., Hale, C.N., van der Zwet, T., Miller, C.E. & Redlin, S.C. (1998) The potential for spread of *Erwinia amylovora* and fire blight via commercial apple fruit; a critical review and risk assessment. *Crop Protection*, **17**, 19–28.

Rosen, R.A. & Coreggia, A. (2004) The New Basel Capital Accord: Part I: Environmental risks for banks. *Environmental Claims Journal*, **16**, 93–101.

Senate (2005) *Australian Parliament, Senate Rural and Regional Affairs and Transport Legislation Committee*. Administration of Biosecurity Australia: Revised draft import risk analysis for apples from New Zealand, March 2005. Available at http://www.aph.gov.au/senate/committee/rrat_ctte/apples04/report/report.pdf.

Senate Hansard (1997) *Australian Parliament, Senate Rural and Regional Affairs and Transport Legislation Committee*. 11 June 1997. Available at http://www.aph.gov.au/hansard/senate/commttee/s1423799.pdf.

Siegel-Jacobs, K. & Yates, J.F. (1996) Effects of procedural and outcome accountability on judgment quality. *Organizational Behavior and Human Decision Processes*, **65**, 1–17.

Simberloff, D. (2005) The politics of assessing risk for biological invasions: the USA as a case study. *Trends in Ecology and Evolution*, **20**, 216–222.

Simonson, I. & Staw, B.M. (1992) De-escalation strategies: a comparison of techniques for reducing commitment to losing courses of action. *Journal of Applied Psychology*, **77**, 419–426.

Sisson, S.A., Pericchi, L.R. & Coles, S. (2006) A case for a reassessment of the risks of extreme hydrological hazards in the Caribbean. *Stochastic Environmental Research and Risk Assessment*, **20**, 296–306.

Smith, R.L. (1989) Extreme value analysis of environmental time series: an example based on ozone data (with discussion). *Statistical Science*, **4**, 367–393.

Suedel, B.C., Bridges, T.S., Kim, J., Payne, B.S. & Miller, A.C. (2007) Application of risk assessment and decision analysis to aquatic nuisance species. *Integrated Environmental Assessment and Management*, **3**, 79–89.

Tetlock, P.E. (1983) Accountability and complexity of thought. *Journal of Personality and Social Psychology*, **45**, 74–83.

Tetlock, P.E. (2005) *Expert political judgment: How good is it? How can we know?* Princeton University Press, Princeton, NJ.

Tillers, P. (2005) If wishes were horses: discursive comments on attempts to prevent individuals being unfairly burdened by their reference classes. *Law, Probability and Risk*, **5**, 33–49.

Walley, P. (1991) *Statistical reasoning with imprecise probabilities*. Chapman & Hall, London.

Wallsten, T.S., Budescu, D.V. & Zwick, R. (1993) Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, **39**, 176–190.

Wells, G.L. & Olson, G.A. (2003) Eyewitness testimony. *Annual Review of Psychology*, **54**, 277–295.

World Trade Organization (WTO) (1998) Measures affecting importation of salmon. Report of the Appellate Body AB-1998-5, Australia.

Editor: Prof David Richardson

Appendix Extreme value theory

Suppose we have a sequence of independent random variables $X_1, X_2, \ldots, X_n$ drawn from a common distribution function F. Classical extreme value theory models focus on the statistical behaviour of the 'block maxima' over blocks of n observations:

$M_n = \max\{X_1, X_2, \ldots X_n\}$.

The X_i usually represent (continuous) values of a process observed on a regular timescale, such as daily rainfall amounts or log daily returns of some stock. Biosecurity applications have not been much studied but could include daily levels of a contaminant or the distances travelled by dispersers.

It can be shown that as n grows large, the probability distribution of M_n (after some minor rescaling) approaches the generalized extreme value (GEV) distribution, with distribution function

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]_+^{-1/\xi}\right\} \qquad (1)$$

where $[a]_+ = \max(0, a)$. The parameters μ, σ and ξ correspond to location, scale, and tail shape parameters, respectively. The GEV distribution incorporates into a single formulation three families of extreme value distributions: the Weibull, Gumbel, and Fréchet distributions.

The most important parameter is the tail shape ξ, which determines the general behaviour of the probability for the extreme values, whose prediction is the focus of interest. For $\xi < 0$, the GEV yields Weibull tails with a finite upper endpoint of $\mu - \sigma/\xi$. Realizing a Weibull distribution is often of practical importance, as there is a clear maximum bound for the process under study. For $\xi > 0$, the GEV gives Fréchet tails with no upper endpoint. This is a polynomial decay, where the larger the value of ξ, the heavier the tail. Fréchet tails are common in environmental applications, such as precipitation (Smith, 1989; Sisson *et al.*, 2006), and in financial applications, such as insurance (Smith, 1989) (where values of ξ between 2 and 4 are typical). Polynomial tails decay much more slowly than exponential tails such as those in the normal (Gaussian) distribution – that explains why the naïve procedure (still sometimes followed) of fitting normal distributions to data will give dangerously low estimates of the probability of extreme events. In the limit as $\xi \rightarrow 0$, the GEV reduces to the Gumbel distribution, with exponential decay tails.

If the common distribution function F is known, the distribution of M_n may be calculated explicitly. But in many applications, the distribution F is unknown. Usually, one takes advantage of the single GEV formulation of the Weibull, Gumbel, and Fréchet distributions by statistically fitting this distribution to the observed sample maxima. The estimated value of the tail shape parameter will determine which family the sample maximum belongs to. Various methods of fitting are available. In applying this theory, the practitioner must make a conscious decision when partitioning observed data into blocks of size n, amounting to the trade-off between bias and variance: blocks that are too small mean that approximation by the limit model is likely to be poor, leading to bias in estimation; large blocks generate few block maxima leading to large estimation variance. In practice, pragmatic considerations often lead to the adoption of blocks of length of 1 year. For example, daily temperatures are likely to vary according to season, violating the assumption that the X_i have a common distribution F *(*daily temperatures are also not independent; see Coles, 2001; Bottolo *et al.*, 2003), but taking blocks of a year means that it is plausible that the maxima of the blocks should have the same distribution.

Once estimates of model parameters are obtained, the GEV distribution function (1) may be inverted to give the 'return level' associated with the return period *1/p*, the level that is expected to be exceeded on average once every *1/p* years. It provides an easily communicated measure of how unlikely a given possible extreme event is.

Block maxima analyses are arguably wasteful of data. Only the largest value in each block is used to fit the model. An alternative formulation is based on threshold exceedance models (Smith, 1989). They regard as extreme events those of the X_i that exceed some high threshold u. It can be shown that the distribution of events exceeding a high threshold is approximately distributed as a generalized Pareto distribution (Coles, 2001).

EVT also has the advantage of being readily adaptable to the Bayesian framework, allowing any expert prior knowledge to be incorporated as a prior distribution over the parameters, which is then updated in the light of the data (by Bayes' formula). Bayesian inference is particularly suited to extreme value theory (Coles & Powell, 1996). The requirement of prior specification means that the natural scarcity of extreme data may be supplemented through an informative prior formulation from a subject matter expert.

As a simple illustration of how Extreme Value Theory can cast light on invasion biology, we consider the prediction of dispersal events larger than those observed in a data set.

The data are daily movements of bobucks. Bobucks, or mountain brushtail possums, *Trichosurus cunninghami*, are medium-sized (2.6–4.2 kg), nocturnal, semi-arboreal marsupials that depend on tree hollows for diurnal shelter. We trapped and individually marked every bobuck resident in a forest patch at Boho South (36°48′S, 145°45′E) in the Strathbogie Ranges, north-eastern Victoria. Furthermore, we fitted all adults and sub-adults ($n = 37$) in the population with radio-transmitters (Martin, 2006). We conducted radio-tracking regularly on foot between June 1999 and November 2003 and located possums during daylight while they were in their den-trees (mean 309 locations per adult). It was always possible to determine the exact tree in which an individual was located. For each individual, we also had a record of sex and tooth-wear class (a relative estimate of age; see Martin & Handasyde, 2007). We then calculated the distance (m) between the den-tree(s) used by each individual on consecutive days of radio-tracking (ignoring the distance moved by that individual during the nightly foraging period in between).

There are a total of 4996 observations on 37 individuals. There are many zero values (where the possum returns to the same den-tree as was used the previous day after nightly foraging), and the mean run (including the zeros) is about 56 m. But there are a small number of much larger values. The largest 10 observations (rounded to the nearest metre) are: 1256, 1488, 1723, 2011, 2523, 2587, 5492, 5525, 7024, and 7152.

(The closeness of the two largest values is coincidental, as they come from different individuals on different dates.)

For application to problems such as predicting the spread of a disease carried by possums, it would be desirable to be able to estimate approximately the probability of much larger values than those seen in the data, for example the probability of a run greater than say 10,000 m. Although the data do not appear to come from any model of the types used in deriving EVT, fitting an EVT model just to the larger values can give a means of allowing the data in the tail to give information as to what lies beyond the range of the data.

A threshold exceedance EVT model was fitted to the extreme data. A run was considered as 'extreme' if it exceeded 400 m. This judgement was based on diagnostics that suggested that a higher threshold would give unstable predictions of model parameters and of probabilities (e.g. Coles, 2001). This yielded 122 extreme data observations.

The model was fitted using both classical (maximum likelihood) and Bayesian frameworks. For the Bayesian framework a prior proportional to 1 was assumed for all unknown parameters, indicating maximal prior ignorance as to the distribution of the parameters. The probabilities predicted for a run over 10,000 m are very similar in the classical and Bayesian analyses. They are: Probability of observing a single observation over 10,000 m:

Classical (maximum likelihood)	0.00028
Bayesian	0.00024

(The probabilities refer to the chance that a single observation from the data set gives a reading of greater than 10,000 m – that is, probabilities are per possum-day.)

These probabilities are larger than 1/4996 = 0.00020, the reciprocal of the number of data points. That means that it would be expected that in a data set of that size, there should be on average a value greater than 10,000, so that the occurrence of no values beyond 7152 in the actual data set is strongly coincidental. This is not inconsistent with the observed data as quantile–quantile plots indicate an adequate model/data fit, with the exception of the largest observation, which is simply observed to be smaller than that predicted by the model (that is, it does not follow the rate of tail decay exhibited by the other extreme observations). This effect is picked up in the Bayesian analysis, which by its nature permits improved flexibility in its predictions. Here the slightly smaller probability of exceeding 10,000 m reflects that it is attempting to give higher probability to the most extreme events. (In general, an alternative explanation for a lack of observed very high values could be data censorship, but that is not the case here as radio tracking gives accurate and complete data.)

The tail shape parameter ξ is about 1 or a little more, indicating Fréchet (polynomial) decay, that is, a heavy tail. The analysis thus warns that fitting standard models such as a normal or exponential decay to the data would seriously underestimate the probability of extreme events.

The applicability of the analysis for prediction is subject to various matters of human judgement concerning causal knowledge of possums and characteristics of the particular data set. For example, it is clear from the data that all of the more extreme values were from a small number of dispersing sub-adult male possums (whereas females are all philopatric), so one may want to separate out this subpopulation for further analysis, and inquire as to whether other populations to which prediction is applied have similar mixes of age and sex. Any information about possible changes in the environment and the possibility of more informative priors could also be considered, subject to expert judgement. However, the simple analysis of the whole data that we have described is sufficient to illustrate the utility of Extreme Value Theory in making predictions of extremes considerably beyond what is visible in the data.