RIVM/MNP Guidance for Uncertainty Assessment and Communication:

Tool Catalogue for Uncertainty Assessment

J.P. van der Sluijs, P.H.M. Janssen, A.C. Petersen, P. Kloprogge, J.S. Risbey, W. Tuinstra, J.R. Ravetz

Report nr: NWS-E-2004-37
ISBN 90-393-3797-7

Copernicus Institute for Sustainable Development and Innovation
Utrecht University
Heidelberglaan 2
3584 CS Utrecht
The Netherlands

Netherlands Environmental Assessment Agency
National Institute for Public Health and the Environment (RIVM)
P.O. Box 1
3720 BA Bilthoven
The Netherlands

This report contains the Tool Catalogue of the RIVM/MNP Guidance for Uncertainty Assessment and Communication. The Guidance has been developed under the direction of Peter Janssen (RIVM/MNP) and Jeroen van der Sluijs (Utrecht University) as part of the strategic research project 'Uncertainty Analysis' (S/550002) at RIVM.

The *RIVM/MNP Guidance for Uncertainty Assessment and Communication* Series contains the following volumes:

1. Mini-Checklist & Quickscan Questionnaire, A. C. Petersen, P. H. M. Janssen, J. P. van der Sluijs et al., RIVM/MNP, 2003

2. Quickscan Hints & Actions List, P. H. M. Janssen, A. C. Petersen, J. P. van der Sluijs et al., RIVM/MNP, 2003

3. Detailed Guidance, J. P. van der Sluijs, J. S. Risbey et al., Utrecht University, 2003

4. Tool Catalogue for Uncertainty Assessment, J.P. van der Sluijs, P.H.M. Janssen, A.C. Petersen, P. Kloprogge, J.S. Risbey, W. Tuinstra, J.R. Ravetz, 2004.

All four volumes are available online and can be downloaded from www.nusap.net

Title: *RIVM/MNP Guidance for Uncertainty Assessment and Communication: Tool Catalogue for Uncertainty Assessment* (RIVM/MNP Guidance for Uncertainty Assessment and Communication Series, Volume 4)

Authors: J.P. van der Sluijs, P.H.M. Janssen, A.C. Petersen, P. Kloprogge, J.S. Risbey, W. Tuinstra, J.R. Ravetz

Report nr: NWS-E-2004-37
ISBN 90-393-3797-7

Copernicus Institute for Sustainable Development and Innovation
Department of Science Technology and Society
Utrecht University
Heidelberglaan 2
3584 CS Utrecht
The Netherlands

Netherlands Environmental Assessment Agency
National Institute for Public Health and the Environment (RIVM)
P.O. Box 1
3720 BA Bilthoven
The Netherlands

Table of contents

Introduction

This document provides an overview of a selection of tools for uncertainty assessment that can be applied to gain insights in nature and size of different sorts of uncertainties in environmental assessments that may occur at different locations.

The tools covered in this document are:

- Sensitivity Analysis (screening, local, global)
- Error propagation equations ("Tier 1")
- Monte Carlo Analysis ("Tier 2")
- Expert Elicitation
- NUSAP (Numeral Unit Spread Assessment Pedigree)
- Scenario Analysis
- PRIMA (Pluralistic fRamework of Integrated uncertainty Management and risk Analysis)
- Checklist for Model Quality Assistance
- Critical Review of Assumptions in models

This toolbox is under development and does not pretend to be exhaustive. The tools described may in literature and practice exist in many different flavours, not all of them being covered in this document. The selection is made in such a way that the set of tools covers all sorts and locations of uncertainties distinguished in the uncertainty typology presented in the guidance. Also it matches current practice and recent Research and Development and Demonstration activities within RIVM in the fields of uncertainty assessment and management.

To assist in selecting tools for conducting uncertainty assessment in a given case, table 1 presents the uncertainty typology used in the guidance and shows what tools can be used to address each of the sorts and locations of uncertainty distinguished.
Some of the tools are hard to map. For instance, the PRIMA approach is a meta-method for uncertainty assessment integrating many of the other tools depending on its particular implementation, and hence covering much more of the table than is suggested at first glance. We have listed the PRIMA in those boxes of the table where we consider it particularly strong. The same holds true for the NUSAP method, which generally includes some quantitative tool (sensitivity analysis or Monte Carlo analysis) in combination with systematic critical review and pedigree analysis.
Further it should be noted that the use of many of the tools is not limited to the boxes in which they are listed. For instance, sensitivity analysis could also be applied to assess sensitivity to different model structures and scenario analysis and sensitivity analysis (screening) may overlap. In a Monte Carlo assessment one could address model structure uncertainty by introducing a switch-parameter to switch between different model equations representing different conceivable model structures and sample for that switch-parameter from for instance a uniform distribution. So the table should not be interpreted too strict, it gives a rough overview of the basic scope of application of each tool.

The remainder of this document provides a tool-by-tool description. For each tool we give a brief description of what it does and how it works and we provide the following information:

- Brief description of the tool
- What are the goals and use of the tool? (Including some guidance on the application of the tools and hints on complementarity with other tools
- What sorts and locations of uncertainty does this tool address?
- What resources are required to use the tool?
- Strengths and limitations of each tool
- Typical pitfalls of each tool
- References to handbooks, user-guides, example case studies, web sites and in-house and external experts who have knowledge on and experience with each tool and who may be consulted by RIVM for further advice.

Table 1 Correspondence of the tools with the sorts and locations of uncertainty distinguished in the uncertainty typology from the hints and actions section of the quick scan. Entries printed in italics are not described in this toolbox because there are no standard methods to perform these tasks.

Type → / **Location ↓**			**Level of uncertainty** *(From determinism, through probability and possibility, to ignorance)*			**Nature of uncertainty**		**Qualification of knowledge base (backing)**	**Value-ladenness of choices**
			Statistical uncertainty (range+ probability)	**Scenario-uncertainty ('what-if' option)**	**Recognized Ignorance**	**Knowledge related uncertainty**	**Variability related uncertainty**		
Context	Ecological, technological, economic, social and political representation		SA *QA* EE	Sc *QA* *SI* EE	Sc MQC *QA* *SI* NUSAP/EP EE	NUSAP / EP MQC *QA* EE		NUSAP / EP MQC *QA* *PR* *EPR* EE	CRA, PRIMA Sc, *AA, SI,* EE *PR, EPR*
Data (in general sense)	Measurements+ Monitoring data; Survey data		SA, Tier 1 MCA EE	Sc EE	Sc *QA* NUSAP MQC *DV* *MV* EE	NUSAP MQC *DV* *QA* EE		NUSAP MQC *QA* *PR* *EPR* EE	CRA PRIMA Sc *PR* *EPR* *SI*
Model	Model Inputs	Measurements monitoring data; survey data							
	Model Structure	Parameters							
		Relations	SA, *MMS*, EE, MQC, *MC*	Sc, *MMS*	NUSAP, MQC, *MC, MV*	MQC, NUSAP, *QA*, EE		MQC, NUSAP, MC, *MV, PR, EPR,* EE	CRA, PRIMA, *MMS, PR, EPR, SI*
	Technical Model	Software& hardware-implement.	*QA* SA	*QA* SA	*QA* SA	*PR*		*PR*	SA *PR*
Expert Judgement		Narratives; storylines; advices	SA, *QA* EE	Sc, *QA, SI*, EE	Sc, MQC, *QA, SI,* NUSAP/EP, EE	NUSAP / EP MQC, *QA*, EE		NUSAP / EP, MQC, *QA, PR, EPR*, EE	CRA, PRIMA, Sc, *AA SI, PR, EPR*, EE
Outputs		(indicators; statements)	Sc, SA, Tier1, *MC*, EE	Sc, SA, EE	NUSAP, EE	NUSAP, MQC, *PR, EPR*, EE		NUSAP, MQC, QA, *PR, EPR*, EE	CRA, PRIMA, *PR, EPR*

Explanation of abbreviations in table 1:

AA	Actor Analysis
CRA	Critical Review of Assumptions
DV	Data Validation
EE	Expert Elicitation
EP	Extended Pedigree scheme
EPR	Extended Peer Review (review by stakeholders)
MC	Model Comparison
MCA	Tier 2 analysis / Monte Carlo Analysis
MMS	Multiple Model Simulation
MQC	Model Quality Checklist
MV	Model validation
NUSAP	NUSAP
PR	Peer Review
PRIMA	PRIMA
QA	Quality Assurance
SA	Sensitivity Analysis
Sc	Scenario Analysis
SI	Stakeholder Involvement
Tier 1	Tier 1 analysis (error propagation equation)

Sensitivity Analysis

Description

Sensitivity analysis (SA) is the study of how the variation in the output of a model (numerical or otherwise) can be apportioned, qualitatively or quantitatively, to different sources of variation, and of how the given model depends upon the information fed into it (Saltelli *et al.*, 2000).

Goals and use

The goal of sensitivity analysis is to understand the quantitative sources of uncertainty in model calculations and to identify those sources that contribute the largest amount of uncertainty in a given outcome of interest.
Three types of sensitivity analysis can be distinguished:

- *Screening*, which is basically a general investigation of the effects of variation in the inputs but not a quantitative method giving the exact percentage of the total amount of variation that each factor accounts for. The main purpose of screening methods is to identify in an efficient way a short list of the most important sensitive factors, so that in a follow-up uncertainty analysis the limited resources can be used in the most efficient way.
- *Local SA*, the effect of the variation in each input factor when the others are kept at some constant level. The result is typically a series of partial derivatives - or an approximation thereof-, one for each factor, that defines the rate of change of the output relative to the rate of change of the input.
- *Global SA*, the effects on the outcomes of interest of variation in the inputs, as all inputs are allowed to vary over their ranges. This can be extended to take into account the shape of their probability density functions. This usually requires some procedure for sampling the parameters, perhaps in a Monte Carlo form, and the result is more complex than for local SA. In their book, Saltelli *et al.* (2000) describe a range of different statistics describing how this type of information can be summarized. Global SA is a variance-analysis based method, using indices expressing the contribution of parameters to the variance in the output (e.g. standardized rank correlation coefficients and partial rank correlation coefficients) (cf. Saltelli *et al*. 2000).

There is one particular (global) screening method for sensitivity analysis that we consider state of the art and recommend for its computational efficiency: the Morris algorithm (Morris, 1991). The typical case to apply this tool is if there are many parameters and available resources do not allow to specify probability density functions for a full Monte Carlo analysis. The description of Morris given here is taken from Potting et al., (2001): "The Morris method for global sensitivity analysis is a so-called one step-at-a-time method, meaning that in each run only one input parameter is given a new value. It facilitates a global sensitivity analysis by making a number r of local changes at different points $x(1 \rightarrow r)$ of the possible range of input values. The method starts by sampling a set of start values within the defined ranges of possible values for all input variables and calculating the subsequent model outcome. The second step changes the values for one variable (all other inputs remaining at their start values) and calculates the resulting change in model outcome compared to the first run. Next, the values for another variable are changed (the previous variable is kept at its changed value and all other ones kept at their start values) and the resulting change in model outcome compared to the second run is calculated. This goes on until all input variables are changed. This procedure is repeated r times (where r is usually taken between 5 and 15), each time with a different set of start values, which leads to a number of $r*(k+1)$ runs, where k is the number of input variables. Such number is very efficient compared to more demanding methods for sensitivity analysis (Campolongo et al. 1999).
The Morris method thus results in a number of r changes in model outcome from r times changing the input value of a given variable. This information is expressed in so-called elementary effects. These elementary effects are approximations of the gradient $\delta y/\delta x$ of the model output y with respect to a specific value for input variable x. The resulting set of r

elementary effects is used to calculate the average elementary effect (to lose dependence of the specific point at which each measure was taken) and the standard deviation. The average elementary effect is indicated by μ, and the standard deviation by σ. The σ expresses whether the relation between input variable and model outcome has a linear ($\sigma = 0$) or a curvi-linear ($\sigma > 0$) character. (Campolongo et al. 1999) Curvi-linearity will be caused by curvi-linear (main) effects and interaction effects from the analysed input variable with other ones."

In summary, the Morris method applies a sophisticated algorithm for global SA where parameters are varied one step at a time in such a way that if sensitivity of one parameter is contingent on the values that other parameters may take, the Morris method is likely to capture such dependencies.

Sorts and locations of uncertainty addressed

Sensitivity Analysis typically addresses statistical uncertainty (inexactness) in inputs and parameters. It is however also possible to use this technique to analyse sensitivity to changes in model structure. It does not treat knowledge uncertainty separately from variability related uncertainty. It provides no insight in the quality of the knowledge base nor in issues of value loading.

Required resources

Skills:

- Basic computer skills
- Basic knowledge of statistical concepts

Computer requirements:
Software for sensitivity analysis will run on an average PC. The precise requirements depend on the model to which you apply the sensitivity analysis.

Strengths and limitations

Typical strengths of Sensitivity Analysis are:

- Gives insight in the potential influences of all sorts of changes in inputs
- Helps discriminating across parameters according to importance for the accuracy of the outcome
- Software for sensitivity analysis is freely available
  (e.g. SIMLAB: http://sensitivity-analysis.jrc.cec.eu.int/default2.asp?page=SIMLAB)
- Easy to use

Typical weaknesses of Sensitivity Analysis are:

- Has a tendency to yield an overload of information.
- Sensitivity analysis does not require one to assess how likely it is that specific values of the parameters will actually occur.
- Sensitivity testing does not encourage the analyst to consider dependencies between parameters and probabilities that certain values will occur together.
- (Morris:) interactions and non-linearity are hard to distinguish with the Morris method.

These weaknesses can be partly overcome by a skilled design of the SA experiments, taking into account dependencies and restrictions and by being creative in structuring, synthesizing and communicating the information captured in the large amount of numbers produced by the sensitivity analysis.

Guidance on application

- If a likely range is not known one can use for instance the point values plus or minus 50% or a factor 2 (half the point value to double the point value), depending on the nature of the variable, as a first go
- Make sure that the ranges do not include physically impossible values
- Explore possible dependencies

Pitfalls

Typical pitfalls of SA can be:

- Forgetting that SA takes the model structure and boundaries for granted
- Wasting time on finding out likely ranges for unimportant parameters
  This could be avoided by using a two-step approach applying a default range (e.g. a factor 2) on all parameters to find out which parameters appear to be sensitive at all. In such a case one should however also go over the list of variables identified as insensitive and include for the second step also those variables where one has doubts as to whether one is sure that the default range used in that calculation captures the full conceivable range for that parameter.
- Ignoring dependencies between parameters
- Exploring irrelevant or physically unrealistic parts of the parameter hyper space

References

Handbooks:
Andrea Saltelli, Karen Chan, Marian Scott*, Sensitivity Analysis* John Wiley & Sons publishers, Probability and Statistics series, 2000.

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, Marco Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models,* John Wiley & Sons publishers, 2004 (Where the Saltelli et al. 2000 book provides the theoretical basis, this book is a comprehensive practical compendium of recommended methods tailored to specified settings, built around a set of examples and the freely available SIMLAB software.).

Papers
Campolongo, F., S. Tarantola and A. Saltelli. Tackling quantitatively large dimensionality problems. Computer Physics Communication, Vol. 1999, Issue 117, pp75-85.

Janssen, P.H.M., P.S.C. Heuberger, & R.A. Sanders. 1994. UNCSAM: a tool for automating sensitivity and uncertainty analysis. Environmental Software 9:1-11.

Morris, M.D. Factorial sampling plans for preliminary computational experiments. Technometrics, Vol. 33 (1991), Issue 2.

RIVM example of application of Morris:
Jose Potting, Peter Heuberger, Arthur Beusen, Detlef van Vuuren and Bert de Vries, Sensitivity Analysis, chapter 5 in: Jeroen P. van der Sluijs, Jose Potting, James Risbey, Detlef van Vuuren, Bert de Vries, Arthur Beusen, Peter Heuberger, Serafin Corral Quintana, Silvio Funtowicz, Penny Kloprogge, David Nuijten, Arthur Petersen, Jerry Ravetz. 2001. Uncertainty assessment of the IMAGE/TIMER B1 CO2 emissions scenario, using the NUSAP method Dutch National Research Program on Climate Change, Report no: 410 200 104 (2001), 227 pp. (downloadable from http://www.nusap.net)

Websites
http://sensitivity-analysis.jrc.cec.eu.int/default.htm

Software
Available software for sensitivity analysis includes:

- SIMLAB: http://sensitivity-analysis.jrc.cec.eu.int/default2.asp?page=SIMLAB

- USATOOL: Currently under development at RIVM.
- See also the software tools listed under Monte Carlo Analysis.

Experts
RIVM: Peter Janssen (UNCSAM), Peter Heuberger (UNCSAM, Morris)
National: Ad Seebregts (ECN), Roger Cooke (TUD), Prof. Kleijnen (KUB), M. Jansen (Alterra, PRI)
International: Andrea Saltelli (JRC) Stefano Tarantola (JRC), John van Aardenne

Error propagation equations ("TIER 1")

Description

The Intergovernmental Panel on Climate Change (IPCC) has issued "Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories" (IPCC, 2000). In this report IPCC distinguishes two levels of comprehension for quantitative uncertainty assessment in emissions monitoring, which they named TIER 1 and TIER 2. TIER 1 uses the error propagation equation (Mandel 1984, Bevington and Robinson 1992) to estimate error propagation in calculations whereas TIER 2 consists of a full Monte Carlo analysis. Since this influential report, the method using the classic analytical equations for error propagation (well known by most students of the experimental sciences) has now become widely referred to as the 'TIER 1' approach.

Goals and use

The goal of the error propagation equations is to assess how the quantified uncertainties in model inputs propagate in model calculations to produce and uncertainty range in a given model outcome of interest. For the most common operations, the error propagation rules are summarized in box 1.

Box 1 Error propagation rules using standard deviation (σ)

Addition and Subtraction: z = x + y + .. or z = x - y - ..

$$\sigma_z = \sqrt{(\sigma_x{}^2) + (\sigma_y{}^2) + ..}$$

Multiplication by an exact number: z = c x

$$\sigma_z = c\sigma_x$$

Multiplication and Division: z = x y or z = x/y

$$\frac{\sigma_z}{z} = \sqrt{\left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 + ..}$$

Products of powers: $z=x^m y^n$

$$\frac{\sigma_z}{z} = \sqrt{\left(\frac{m\sigma_x}{x}\right)^2 + \left(\frac{n\sigma_y}{y}\right)^2}$$

For instance, in the case of emission monitoring where emissions are estimated by multiplying activity data by emission factors the error propagation equation can be written as:

$$\sigma_E^2 = \sigma_A^2 F^2 + \sigma_F^2 A^2$$

Where σ_E^2 is the emission variance, σ_A^2 is the variance of the activity data, σ_F^2 is the variance of the emission factor, A is the expected value of the activity data, and F is the expected value of the emission factor.

The conditions imposed for use of the error propagation equation are:
• The uncertainties are relatively small, the standard deviation divided by the mean value being less than 0.3;
• The uncertainties have Gaussian (normal) distributions;
• The uncertainties have no significant covariance.
Under these conditions, the uncertainty calculated for the emission rate is appropriate (IPCC, 2000). The method can be extended to allow non-Gaussian distributions and to allow for covariances (see e.g.: http://www.itl.nist.gov/div898/handbook/mpc/section5/mpc55.htm).

For a more comprehensive description of the TIER 1 approach we refer to annex 1 of the IPCC good practice guidelines (IPCC, 2001)

Sorts and locations of uncertainty addressed

TIER 1 addresses statistical uncertainty (inexactness) in inputs and parameters and estimates its propagation in simple calculations. It does not treat knowledge uncertainty separately from variability related uncertainty. It provides no insight in the quality of the knowledge base or in issues of value loading.

Required resources:

The error propagation equations require no specific hardware or software and can typically be applied on the back of the envelope or on an ordinary scientific calculator, or using a spreadsheet.
Most of the time will be consumed by quantifying the uncertainties in the parameters and inputs, which can be derived from statistics if available or otherwise can for instance be obtained by means of expert elicitation.

Strengths and limitations

Typical strengths are:
- Requires very little resources and skills (but the choice of the aggregation level for the analysis is an important issue that does require skills)
- Quick (but can be dirty)

Typical weaknesses are:
- Has a limited domain of applicability (e.g. near-linearity assumption)
- The basic error propagation equations cannot cope well with distributions with other shapes than normal (but the method can be extended to account for other distributions).
- Leads to a tendency to assume that all distributions are normal, even in cases where knowledge of the shape is absent and hence a uniform distribution would be reflecting better the state of knowledge.
- Can not easily be applied in complex calculations

Guidance on application

Do not use the error propagation equation if you do not have good reasons to assume that parameter uncertainty is distributed normally. Use Monte Carlo analysis instead.

For further guidance we refer to standard handbooks on statistics and measurement error analysis.

Pitfalls

Typical pitfalls in the use of the error propagation equation are:

- Forgetting that TIER 1 takes the model structure and boundaries for granted
- Bias towards assuming all parameter uncertainty to be distributed normally.
- Ignoring dependencies and covariance

References

Bevington, P. R. and D.K. Robinson, D. K. (1992) Data Reduction and Error Analysis for the Physical Sciences, WCB/McGraw-Hill Boston USA, p.328.

IPCC, Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories, IPCC, 2000.

Harry Ku (1966). Notes on the Use of Propagation of Error Formulas**,** *J Research of National Bureau of Standards-C. Engineering and Instrumentation*, Vol. 70C, No.4, pp. 263-273.

Mandel, J. (1984) The Statistical Analysis of Experimental Data, Dover Publications New York, USA, p.410.

Monte Carlo Analysis ("TIER 2")

Description

Monte Carlo Simulation is a statistical numerical technique for stochastic model-calculations and analysis of error propagation in (model) calculations.

Goals and use

The goal of Monte Carlo analysis is to trace out the structure of the distributions of model output that results from specified uncertainty distributions of model inputs and model parameters. In it's simplest form this distribution is mapped by calculating the deterministic results (realizations) for a large number of random draws from the individual distribution functions of input data and parameters of the model. To reduce the required number of model runs needed to get sufficient information about the distribution in the outcome (mainly to save computation time), advanced sampling methods have been designed such as Latin Hyper Cube sampling. The latter makes use of stratification in the sampling of individual parameters; like in random Monte Carlo sampling, pre-existing information about correlations between input variables can be incorporated. Monte Carlo analysis requires the analyst to specify probability distributions of all inputs and parameters, and the correlations between them. Both probability distributions and correlations are usually poorly known.

A number of software packages are available to do Monte Carlo analysis. Widely used are the commercial packages @Risk (http://www.palisade.com) and Crystal Ball (http://www.decisioneering.com/crystal_ball). Both are packages that are designed as fully integrated MS-Excel add-in programs with its own toolbar and menus. These packages can be used with minimal knowledge on the sampling and calculations techniques itself, which makes Monte Carlo Assessment easy (but tricky because it allows incompetent use). Another commercial package is Analytica (http://www.lumina.com), which is a quantitative modelling environment with built-in Monte Carlo algorithms.

If your model is not built in Excel you can use the SimLab package, which is freely available from the JRC (http://sensitivity-analysis.jrc.cec.eu.int/default2.asp?page=SIMLAB). SimLab can also be interfaced with Excel, but this requires some programming skills. For the UNIX and MS-Dos environments you can use the UNSCAM (Janssen *et al.*, 1994) software tool. RIVM is presently developing a new tool for Monte Carlo analysis, USATOOL, which will run under Windows.

Additionally most Monte Carlo analysis software offers the possibility to determine the relative contribution of uncertainty in each parameter to the uncertainty in a model output, e.g. by sensitivity charts, and can be used for a sophisticated analysis of trends in the presence of uncertainty.

Sorts and locations of uncertainty addressed

Monte Carlo analysis typically addresses statistical uncertainty (stochastic inexactness) in inputs and parameters. Although it is rarely used this way, it is possible to use Monte Carlo analysis also for assessing model structure uncertainty, by introducing one or more “switch parameter” to switch between different model structures with probabilities attached for each position of the switch.
Two-dimensional Monte Carlo Analysis allows for a separate treatment of knowledge related uncertainty and variability related uncertainty (see below under guidance for application). In this two-dimensional mode, Monte Carlo Analysis provides some insight in the quality of the knowledge base. It does not address issues of value loading.

Required resources

Computer

Monte Carlo software packages such as Crystal Ball and @Risk run on standard PCs with Pentium II, 200 MHz or faster, 32 MB RAM as officially recommended minimum configuration. On the basis of our experiences we recommend 500 Mhz processor or equivalent with 256 MB RAM as minimum configuration.

Training
Packages such as Crystal Ball are very easy to learn. If you are familiar with Excel it takes less than one hour to get proficient with Crystal Ball.

SimLab takes more time to get proficient with and requires more skills because one has to interface SimLab with one's own model. The forum on the SimLab website has however al lot of useful tips making the task easier. We recommend the book “Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models” by Saltelli et al. 2004. It has many practical examples of the use of SimLab.

Strengths and limitations

Typical strengths of Monte Carlo simulation

- Provides comprehensive insight in how specified uncertainty in inputs propagates through a model.
- Forces analysts to explicitly consider uncertainty and interdependencies among different inputs.
- Is capable to cope with any conceivable shape of PDF and can account for correlations.
- Can be used in 2-dimensional mode to separately assess variability and epistemological uncertainty.

Typical weaknesses of Monte Carlo simulation

- Monte Carlo assessment is limited to those uncertainties that can be quantified and expressed as probabilities.
- One may not have any reasonable basis on which to ascribe a parameterised probability distribution to parameters
- May take large run-time for computational intensive models. This can partly be remedied by using more efficient sampling techniques (e.g. Latin Hypercube Sampling).
- The interpretation of a probability distribution of the model output by decision makers is not always straightforward; there is no single rule arising out of such a distribution that can guide decision-makers concerning the acceptable balance between for instance expected return and the variance of that return.

Guidance on application

In their report "Guiding Principles for Monte Carlo Analysis" (EPA, 1997) the EPA presents 16 good practice guidelines for doing Monte Carlo assessment. These guidelines are (we have modified the phrasing slightly to keep terminology consistent within the guidance documents):

Selecting Input Data and Distributions for Use in Monte Carlo Analysis

1. Conduct preliminary sensitivity analyses or numerical experiments to identify model structures, model input assumptions and parameters that make important contributions to the assessment and its overall uncertainty.

2. Restrict the use of probabilistic assessment to significant parameters.

3. Use data to inform the choice of input distributions for model parameters.

- Is there any mechanistic basis for choosing a distributional family?
- Is the shape of the distribution likely to be dictated by physical or biological properties or other mechanisms?
- Is the variable discrete or continuous?
- What are the bounds of the variable?

- Is the distribution skewed or symmetric?
- If the distribution is thought to be skewed, in which direction?
- What other aspects of the shape of the distribution are known?

4. Proxy data can be used to develop distributions when they can be appropriately justified.

5. When obtaining empirical data to develop input distributions for model parameters, the basic tenets of environmental sampling should be followed. Further, particular attention should be given to the quality of information at the tails of the distributions.

6. Depending on the objectives of the assessment and the availability of empirical data to estimate PDFs, expert elicitation can be applied to draft probability density functions. When expert judgment is employed, the analyst should be very explicit about its use.

Evaluating variability and knowledge limitations

7. It is useful to distinguish between uncertainty stemming from intrinsic variability and heterogeneity of the parameters on the one hand and uncertainty stemming from knowledge limitations on the other hand. Try to separate them in the analysis where possible to provide greater accountability and transparency. The decision about how to track them separately can only be made on a case-by-case basis for each variable.

8. Two dimensional Monte Carlo techniques allow for the separate treatment of variability and epistemological uncertainty. There are methodological differences regarding how uncertainty stemming from variability and uncertainty stemming from knowledge limitations are addressed in a Monte Carlo analysis.

- Variability depends on the averaging time, averaging space, or other dimensions in which the data are aggregated.
- Standard data analysis tends to understate uncertainty from knowledge limitations by focusing solely on random error within a data set. Conversely, standard data analysis tends to overstate variability by implicitly including measurement errors.
- Various types of model errors can represent important sources of uncertainty. Alternative conceptual or mathematical models are a potentially important source of uncertainty. A major threat to the accuracy of a variability analysis is a lack of representativeness of the data.

9. Methods should investigate the numerical stability of the moments and the tails of the distributions.

- Data gathering efforts should be structured to provide adequate coverage at the tails of the input distributions.
- The assessment should include a narrative and qualitative discussion of the quality of information at the tails of the input distributions.

10. There are limits to the assessor's ability to account for and characterize all sources of uncertainty. The analyst should identify areas of uncertainty and include them in the analysis, either quantitatively or qualitatively.

Presenting the Results of a Monte Carlo Analysis

11. Provide a complete and thorough description of the model or calculation scheme and its equations, including a discussion of the limitations of the methods and the results.

12. Provide detailed information on the input distributions selected. This information should identify whether the input represents largely variability, largely uncertainty, or some combination of both. Further, information on goodness-of-fit statistics should be discussed.

A PDF plot is useful for displaying:

- The relative probability of values;
- The most likely values (e. g., modes);
- The shape of the distribution (e. g., skewness, kurtosis); and
- Small changes in probability density.

A CDF plot is good for displaying:
- Fractiles, including the median;
- Probability intervals, including confidence intervals;
- Stochastic dominance; and
- Mixed, continuous, and discrete distributions.

13. Provide detailed information and graphs for each output distribution.

14. Discuss the presence or absence of dependencies and correlations.

15. Calculate and present point estimates.

16. A progressive disclosure of information style in presentation, in which briefing materials are assembled at various levels of detail, may be helpful. Presentations should be tailored to address the questions and information needs of the audience.

- Avoid excessively complicated graphs. Keep graphs intended for a glance (e. g., overhead or slide presentations) relatively simple and uncluttered. Graphs intended for publication can include more complexity.
- Avoid perspective charts (3-dimensional bar and pie charts, ribbon charts), pseudo-perspective charts (2-dimensional bar or line charts).
- Color and shading can create visual biases and are very difficult to use effectively. Use color or shading only when necessary and then, only very carefully. Consult references on the use of color and shading in graphics.
- When possible in publications and reports, graphs should be accompanied by a table of the relevant data.
- If probability density or cumulative probability plots are presented, present both, with one above the other on the same page, with identical horizontal scales and with the location of the mean clearly indicated on both curves with a solid point.
- Do not depend on the audience to correctly interpret any visual display of data. Always provide a narrative in the report interpreting the important aspects of the graph.
- Descriptive statistics and box plots generally serve the less technically oriented audience well. Probability density and cumulative probability plots are generally more meaningful to risk assessors and uncertainty analysts.

For a full discussion of these 16 guidelines we refer to the EPA report (EPA, 1997).

The EPA report also gives some guidance on the issue of constructing adequate probability density functions using proxy data, fitting distributions, using default distributions and using subjective distributions. Important questions in this process are:

- Is there Prior Knowledge about Mechanisms?
- Are the proxy data of acceptable quality and representativeness to support reliable estimates?
- What uncertainties and biases are likely to be introduced by using proxy data?
- How are the biases likely to affect the analysis and can the biases be corrected?

In identifying plausible distributions to represent variability, the following characteristics of the variable should be taken into account:
- Nature of the variable (discrete or continuous)
- Physical or plausible range of the variable (e. g., takes on only positive values)
- Symmetry of the Distribution. (E.g. is the shape of the distribution likely to be dictated by physical/ biological properties such as logistic growth rates)

- Summary Statistics (Frequently, knowledge on ranges can be used to eliminate inappropriate distributions; If the coefficient of variation is near 1.0, then an exponential distribution might be appropriate etc.)

Pitfalls

Typical pitfalls of Monte Carlo Analysis are:

- Forgetting that Monte Carlo analysis takes the model structure and boundaries for granted
- Ignoring correlations
- Hyper precision: Often the PDFs on the inputs used have the status of educated guesses. The output produced by the software packages usually come out the computer with a high number of digits, which are certainly not significant. Also the shapes of the input distributions are usually not well known, therefore one should not attribute too much meaning to the precise shape of the distribution as it comes out of the calculation.
- Glossy reports: Present day software packages for Monte Carlo Analysis can be used easily without requiring prior knowledge of Monte Carlo analysis or prior theoretical knowledge of probability distributions theory. The somewhat glossy results produced by the computer look very professional even if the experiment was poorly designed. We therefore recommend not using these packages without understanding the basics of probability distributions theory, correlations and Monte Carlo analysis. The handbooks that go with the software provide good primers on these issues. We particularly recommend the Crystal Ball handbook in this respect.
- Note that several software packages for Monte Carlo Analysis (*inter alia* SimLab, and Crystal Ball) give false results if Windows is configured to use a comma as decimal separator rather than a dot.

References

Handbooks:

EPA, Risk Assessment Forum, Guiding Principles for Monte Carlo Analysis, EPA/630/R-97/001, 1997.

M.G. Morgan and M. Henrion, Uncertainty, A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis, Cambridge University Press, 1990.

Crystal Ball 2000, User Manual Decision Engineering Inc., Denver, 2000.

Palisade Corporation (2000): *Guide to Using @RISK - Risk Analysis and Simulation Add-in for Microsoft Excel*, Version 4, March 2000.

Andrea Saltelli, Karen Chan, Marian Scott*, Sensitivity Analysis* John Wiley & Sons publishers, Probability and Statistics series, 2000.

Andrea Saltelli, Stefano Tarantola, Francesca Campolongo, Marco Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models,* John Wiley & Sons publishers, 2004

Vose D. (2000): *Risk Analysis – A quantitative guide*, 2nd edition. John Wiley & Sons, Ltd. Chichester.

Papers and reports

Burmaster, D.E., and Anderson, P.D.: *Principles of Good Practice for the Use of Monte Carlo Techniques in Human Health and Ecological Risk Assessments*, Risk Analysis, Vol. 14, No. 4, 1994.

IPCC, Good Practice Guidance and Uncertainty Management in National Greenhouse Gas Inventories, IPCC, 2000.

P.H.M. Janssen, P.S.C. Heuberger, & R.A. Sanders. 1994. UNCSAM: a tool for automating sensitivity and uncertainty analysis. Environmental Software 9:1-11.

NUSAP

Description

NUSAP is a notational system proposed by Funtowicz and Ravetz (1990), which aims to provide an analysis and diagnosis of uncertainty in science for policy. It captures both quantitative and qualitative dimensions of uncertainty and enables one to display these in a standardized and self-explanatory way. It promotes criticism by clients and users of all sorts, expert and lay and will thereby support extended peer review processes.

Goals and use

The goal of NUSAP is to discipline and structure the critical appraisal of the knowledge base behind quantitative policy relevant scientific information. The basic idea is to qualify quantities using the five qualifiers of the NUSAP acronym: Numeral, Unit, Spread, Assessment, and Pedigree. By adding expert judgment of reliability (Assessment) and systematic multi-criteria evaluation of the production process of numbers (Pedigree), NUSAP has extended the statistical approach to uncertainty (inexactness) with the methodological (unreliability) and epistemological (ignorance) dimensions. By providing a separate qualification for each dimension of uncertainty, it enables flexibility in their expression. By means of NUSAP, nuances of meaning about quantities can be conveyed concisely and clearly, to a degree that is quite impossible with statistical methods only.

We will discuss the five qualifiers. The first is *Numeral*; this will usually be an ordinary number; but when appropriate it can be a more general quantity, such as the expression "a million" (which is not the same as the number lying between 999,999 and 1,000,001). Second comes *Unit*, which may be of the conventional sort, but which may also contain extra information, as the date at which the unit is evaluated (most commonly with money). The middle category is *Spread*, which generalizes from the "random error" of experiments or the "variance" of statistics. Although *Spread* is usually conveyed by a number (either ±, % or "factor of") it is not an ordinary quantity, for its own inexactness is not of the same sort as that of measurements. Methods to address *Spread* can be statistical data analysis, sensitivity analysis or Monte Carlo analysis possibly in combination with expert elicitation.

The remaining two qualifiers constitute the more qualitative side of the NUSAP expression. *Assessment* expresses qualitative judgments about the information. In the case of statistical tests, this might be the significance level; in the case of numerical estimates for policy purposes, it might be the qualifier "optimistic" or "pessimistic". In some experimental fields, information is given with two ± terms, of which the first is the spread, or random error, and the second is the "systematic error" which must estimated on the basis of the history of the measurement, and which corresponds to our Assessment. It might be thought that the "systematic error" must always be less than the "experimental error", or else the stated "error bar" would be meaningless or misleading. But the "systematic error" can be well estimated only in retrospect, and then it can give surprises.

Finally there is P for Pedigree, which conveys an evaluative account of the production process of information, and indicates different aspects of the underpinning of the numbers and scientific status of the knowledge used. Pedigree is expressed by means of a set of pedigree criteria to assess these different aspects. Assessment of pedigree involves qualitative expert judgment. To minimize arbitrariness and subjectivity in measuring strength, a pedigree matrix is used to code qualitative expert judgments for each criterion into a discrete numeral scale from 0 (weak) to 4 (strong) with linguistic descriptions (modes) of each level on the scale. Each special sort of information has its own aspects that are key to its pedigree, so different pedigree matrices using different pedigree criteria can be used to qualify different sorts of information. Table 1 gives an example of a pedigree matrix for emission monitoring data. An overview of pedigree matrices found in the literature is given in the pedigree matrices section of http://www.nusap.net. Risbey et al. (2001) document a method to draft pedigree scores by

means of expert elicitation. Examples of questionnaires used for eliciting pedigree scores can be found at http://www.nusap.net.

Table 1 Pedigree matrix for emission monitoring data (Risbey *et al.*, 2001; adapted from Ellis *et al.*, 2000a, 2000b).

Score	Proxy representation	Empirical basis	Methodological rigour	Validation
4	An exact measure of the desired quantity	Controlled experiments and large sample direct measurements	Best available practice in well established discipline	Compared with independent measurements of the same variable over long domain
3	Good fit or measure	Historical/field data uncontrolled experiments small sample direct measurements	Reliable method common within est. discipline Best available practice in immature discipline	Compared with independent measurements of closely related variable over shorter period
2	Well correlated but not measuring the same thing	Modelled/derived data Indirect measurements	Acceptable method but limited consensus on reliability	Measurements not independent proxy variable limited domain
1	Weak correlation but commonalities in measure	Educated guesses indirect approx. rule of thumb est.	Preliminary methods unknown reliability	Weak and very indirect validation
0	Not correlated and not clearly related	Crude speculation	No discernible rigour	No validation performed

We will briefly elaborate the four criteria in this example pedigree matrix.

Proxy representation
Sometimes it is not possible to measure directly the thing we are interested in or to represent it by a parameter, so some form of proxy measure is used. Proxy refers to how good or close a measure of the quantity that we measure or model is to the actual quantity we seek or represent. Think of first order approximations, over simplifications, idealizations, gaps in aggregation levels, differences in definitions, non-representativeness, and incompleteness issues.

Empirical basis
Empirical basis typically refers to the degree to which direct observations, measurements and statistics are used to estimate the parameter. Sometimes directly observed data are not available and the parameter or variable is estimated based on partial measurements or calculated from other quantities. Parameters or variables determined by such indirect methods have a weaker empirical basis and will generally score lower than those based on direct observations.

Methodological rigour
Some method will be used to collect, check, and revise the data used for making parameter or variable estimates. Methodological quality refers to the norms for methodological rigour in this process applied by peers in the relevant disciplines. Well-established and respected methods for measuring and processing the data would score high on this metric, while untested or unreliable methods would tend to score lower.

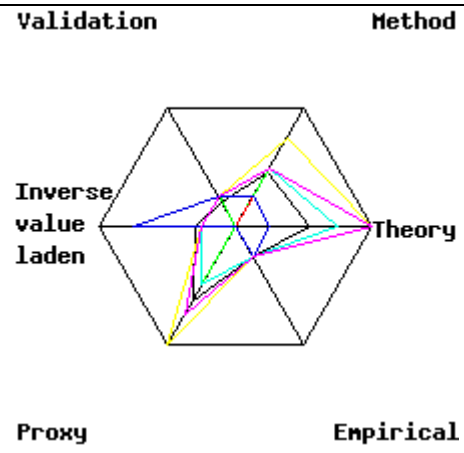
Validation
This metric refers to the degree to which one has been able to crosscheck the data and assumptions used to produce the numeral of the parameter against independent sources. In many cases, independent data for the same parameter over the same time period are not

available and other data sets must be used for validation. This may require a compromise in the length or overlap of the data sets, or may require use of a related, but different, proxy variable for indirect validation, or perhaps use of data that has been aggregated on different scales. The more indirect or incomplete the validation, the lower it will score on this metric.

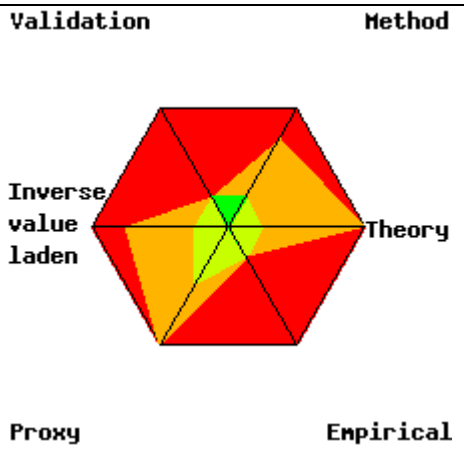
Visualizing pedigree scores

In general, pedigree scores will be established using expert judgements from more than one expert. Two ways of visualizing results of a pedigree analysis are discussed here: radar diagrams and kite diagrams. (Risbey et al, 2001; Van der Sluijs et al, 2001a). An example of both representations is given in figure 2.

Figure 2 Example of representations of same results by radar diagram and kite diagram (Van der Sluijs et al, 2001a)



Example of radar diagram representation of the pedigree scores for the gas depletion multiplier in the TIMER model as assessed by 6 experts.



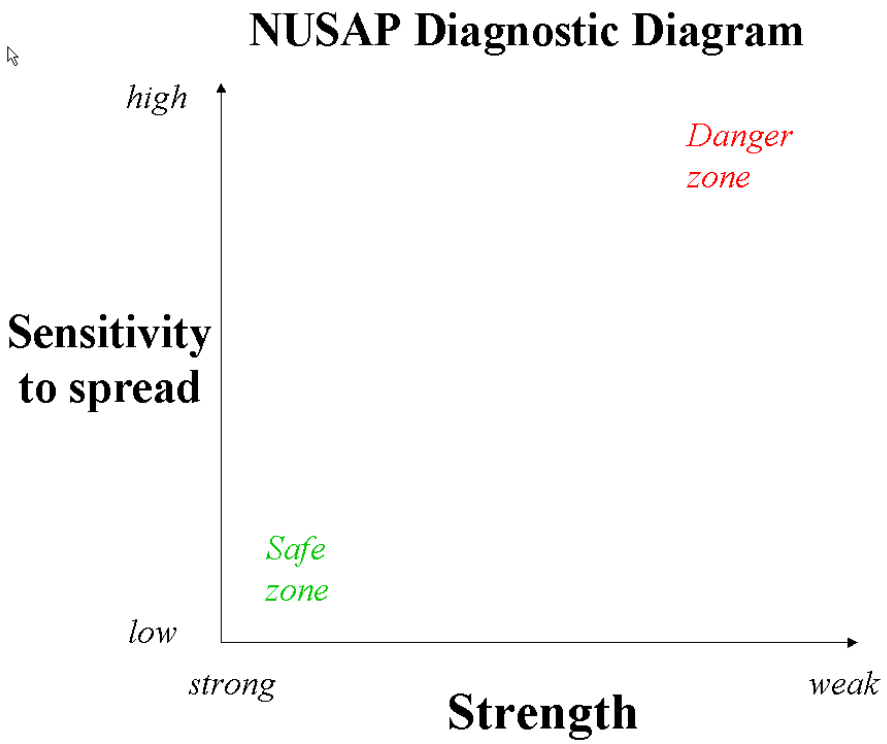
Same pedigree scores as kite diagram.

Both representations use polygons with one axis for each criterion, having 0 in the center of the polygon and 4 on each corner point of the polygon. In the radar diagrams a colored line connecting the scores represents the scoring of each expert, whereas a black line represents the average scores.
The kite diagrams follow a traffic light analogy. The minimum scores in each group for each pedigree criterion span the green kite; the maximum scores span the amber kite. The remaining area is red. The width of the amber band represents expert disagreement on the pedigree scores. In some cases the size of the green area was strongly influenced by a single deviating low score given by one of the experts. In those cases the light green kite shows what the green kite would look like if that outlier had been omitted. Note that the algorithm for calculating the light green kite is such that outliers are evaluated per pedigree criterion, so that outliers defining the light green area need not be from the same expert.
A web-tool to produce kite diagrams is available from http://www.nusap.net.

The kite diagrams can be interpreted as follows: the green colored area reflects the (apparent minimal consensus) strength of the underpinning of each parameter. The greener the diagram the stronger the underpinning is. The orange colored zone shows the range of expert disagreement on that underpinning. The remaining area is red. The more red you see the weaker the underpinning is (all according to the assessment by the group of experts represented in the diagram).
A kite diagram captures the information from all experts in the group without the need to average expert opinion. Averaging expert opinion is a controversial issue in elicitation methodologies. A second advantage is that it provides a fast and intuitive overview of parameter strength, preserving key aspects of the underlying information.

Propagation of pedigree in calculations
Ellis et al. (2000) have developed a pedigree calculator to assess propagation of pedigree in a calculation in order to establish pedigree scores for quantities calculated from other quantities. For more information we refer to http://www.esapubs.org/archive/appl/A010/006/default.htm

Diagnostic Diagram
The method chosen to address the spread qualifier (typically sensitivity analysis or Monte Carlo analysis) provides for each input quantity a quantitative metric for uncertainty contribution (or sensitivity), for instance the relative contribution to the variance in a given model output. The Pedigree scores can be aggregated (by dividing the sum of the scores of the pedigree criteria by sum of the maximum attainable scores) to produce a metric for parameter strength. These two independent metrics can be combined in a NUSAP Diagnostic Diagram. The Diagnostic Diagram is based on the notion that neither spread alone nor strength alone is a sufficient measure for quality. Robustness of model output to parameter strength could be good even if parameter strength is low, provided that the model outcome is not critically influenced by the spread in that parameter. In this situation our ignorance of the true value of the parameter has no immediate consequences because it has a negligible effect on calculated model outputs. Alternatively, model outputs can be robust against parameter spread even if its relative contribution to the total spread in model is high provided that parameter strength is also high. In the latter case, the uncertainty in the model outcome adequately reflects the inherent irreducible uncertainty in the system represented by the model. In other words, the uncertainty then is a property of the modelled system and does not stem from imperfect knowledge on that system. Mapping model parameters in the assessment diagram thus reveals the weakest critical links in the knowledge base of the model with respect to the model outcome assessed, and helps in the setting of priorities for model improvement.



Sorts and locations of uncertainty addressed

The different qualifiers in the NUSAP system address different sorts of uncertainty. The Spread qualifier addresses statistical uncertainty (inexactness) in quantities, typically in input data and parameters. The Assessment qualifier typically addresses unreliability. The Pedigree criterion further qualifies the knowledge base in such a way that it explores the border with ignorance by providing detailed insights in specific weaknesses in the knowledge base that underpins a given quantity.

Most of the pedigree assessments in the literature so far have addressed uncertainties located in "inputs" and "parameters", thereby focusing on the *internal strength* of the knowledge base. Recently Corral (2000) in his Ph.D. thesis extended the pedigree scheme to also address uncertainties located in the "socio political context", focusing on the *external strength* (its relations with the worlds outside the science) of the knowledge base. Criteria that Corral used to assess the pedigree of the processes of knowledge utilization and institutional context of the analysts were *inter alia*: Accessibility, Terminology, Completeness, Source of information, Verification, Colleague consensus, Extended peer acceptance, Legitimation, Experience, Flexibility. NUSAP can also be used to assess issues of value ladenness (see the entry "A method for critical review of assumptions in model-based assessments" in this tool catalogue).

Required resources

Resources required for assessing the Spread qualifier depend on the method chosen (some form of Sensitivity Analysis or Monte Carlo analysis usually in combination with expert elicitation will be needed).

For the assessment of Pedigree, many resources (pedigree matrices, pedigree calculator, kite diagram maker, elicitation protocol and questionnaires) are freely available from http://www.nusap.net. Basic interviewing skills and awareness of motivational bias that may occur in any expert elicitation are required. See also the section on expert elicitation in this toolbox.

If one uses an expert workshop, basic skills for facilitating structured group discussions are needed. In addition, skills are needed to arrive at a balanced composition of the workshop audience to minimize biases.

Time required per expert elicitation in a one to one interview depends on the number of parameters and the complexity of the case. It may typically vary between 1 and 5 hours. A substantial amount of time may be needed for a good preparation of the interviews.

Recommended length for a NUSAP expert elicitation workshop is between one and one and a half day.

Strengths and limitations

Typical strengths of NUSAP are:

- NUSAP identifies the different sorts of uncertainty in quantitative information and enables them to be displayed in a standardized and self-explanatory way. Providers and users of quantitative information then have a clear and transparent assessment of its uncertainties.
- NUSAP fosters an enhanced appreciation of the issue of quality in information. It thereby enables a more effective criticism of quantitative information by providers, clients, and also users of all sorts, expert and lay.
- NUSAP provides a useful means to focus research efforts on the potentially most problematic parameters by identifying those parameters, which are critical for the quality of the information.
- It is flexible in its use and can be used on different levels of comprehensiveness: from a 'back of the envelope' sketch based on self elicitation to a comprehensive and sophisticated procedure involving structured informed in-depth group discussions on a parameter by parameter format, covering each pedigree criterion combined with a full blown Monte Carlo assessment.
- The diagnostic diagram provides a convenient way in which to view each of the key parameters in terms of two crucial attributes. One is their relative contribution to the sensitivity of the output, and the other is their strength. When viewed in combination on the diagram, they provide indications of which parameters are the most critical for the quality of the result.

Typical weaknesses of NUSAP are:

- The method is relatively new, with a limited (but growing) number of practitioners. There is as yet no system of quality assurance in its applications, nor settled guidelines for good practice.
- The scoring of pedigree criteria is to a certain degree subjective. Subjectivity can partly be remedied by the design of unambiguous pedigree matrices and by involving multiple experts in the scoring. The choice of experts to do the scoring is also a potential source of bias.
- The method is applicable only to simple calculations with small numbers of parameters. But it may be questioned whether complicated calculations with many parameters are capable of effective uncertainty analysis by any means.

Guidance on application

- For guidance on the application of NUSAP we refer to http://www.nusap.net
- When eliciting pedigree scores, always ask for motivation for the score given and document the motivation along with the pedigree scores.
- Expert disagreement on pedigree scores for a parameter can be an indication of epistemic uncertainty about that parameter. Find out whether there are different paradigms or competing schools of thought on that parameter.

Pitfalls

Typical pitfalls of NUSAP are:

- Misinterpreting low pedigree scores as indicating low-quality science. In relation to whole disciplines, this amounts to 'physics-envy'. Quality in science depends not on removing uncertainty but on managing it.
- Misinterpreting pedigree scores as an evaluation of individual items of information, with low scores indicating bad research. The pedigree analysis is of the characteristic limits of knowledge of areas of inquiry. The quality of individual items of information depends crucially on the craftsmanship of the work, requiring a closer analysis, which the pedigree does not undertake.
- Motivational bias towards high pedigrees in (self) elicitation, especially in case of numbers where one or one's institute was involved in the knowledge production. This pitfall is avoided by the use of trained facilitators in an open process for the construction and assignment of pedigrees.
- Falsely thinking that pedigree and spread are correlated: *In principle these are independent dimensions*.

References

Handbooks:
Funtowicz, S.O. and Ravetz, J.R., 1990. *Uncertainty and Quality in Science for Policy*. Dordrecht: Kluwer.

Websites:
http://www.nusap.net
A website devoted to the further development and dissemination of the NUSAP method with direct access to tutorials, tools, papers and the like.

Papers with a focus on methodological aspects
R. Constanza, S.O. Funtowicz and J.R. Ravetz, Assessing and communicating data quality in policy-relevant research. *Environmental Management*, 16, 1992, pp. 121-131.

Corral Quintana, Serafin A. 2000. *Una Metodología integrada de exploración y compensión de los procesos de elaboración de políticas públicas*. Ph.D. thesis, University of La Laguna

J.S. Risbey, J.P. van der Sluijs and J. Ravetz, 2001. *Protocol for Assessment of Uncertainty and Strength of Emission Data*, Department of Science Technology and Society, Utrecht University, report nr. E-2001-10, 22 pp.

J.P. van der Sluijs, *Tuning NUSAP for its use in Integrated Model Quality Assurance the Case of Climate Change*, Report in commission of European Commission Directorate General CCR, Joint Research Centre, Institute for Systems, Informatics and Safety, Ispra, Italy (contract no. 13970 – 1998 – 05 F1EI ISP NL), Department of Science Technology and Society, Utrecht University, Utrecht, March 1999. 36 pp.

Case studies:

Erle C. Ellis, Rong Gang Li, Lin Zhang Yang and Xu Cheng. 2000a. Long-term Change in Village-Scale Ecosystems in China Using Landscape and Statistical Methods. *Ecological Applications* 10:1057-1073.

Erle C. Ellis, Rong Gang Li, Lin Zhang Yang and Xu Cheng. 2000b. Long-term Change in Village-Scale Ecosystems in China Using Landscape and Statistical Methods. Ecological Applications 10:1057-1073. Supplement 1: Data Quality Pedigree Calculator. *Ecological Archives* A010-006-S1. (http://www.esapubs.org/archive/appl/A010/006/default.htm)

R. van Gijlswijk, P. Coenen, T. Pulles and J.P. van der Sluijs, *Uncertainty assessment of NOx, SO2 and NH3 emissions in the Netherlands*, 2004. TNO and Copernicus Institute Research Report (available from www.nusap.net).

ORNL and RFF. 1994. *Estimating Fuel Cycle Externalities: Analytical Methods and Issues, Report 2*, prepared by Oak Ridge National Laboratory and Resources for the Future for the U.S. Department of Energy.

M. Hongisto, 1997. Assessment of External Costs of Power Production, A Commensurable Approach? paper presented at *Total Cost Assessement – Recent Developments and Industrial Applications*, Invitational Expert Seminar, Nauvo, Finland, June 15-17.

J.P. van der Sluijs, J.S. Risbey and J. Ravetz, 2001, Uncertainty Assessment of VOC emissions from Paint in the Netherlands (available from www.nusap.net).

Jeroen P. van der Sluijs, Jose Potting, James Risbey, Detlef van Vuuren, Bert de Vries, Arthur Beusen, Peter Heuberger, Serafin Corral Quintana, Silvio Funtowicz, Penny Kloprogge, David Nuijten, Arthur Petersen, Jerry Ravetz. *Uncertainty assessment of the IMAGE/TIMER B1 CO2 emissions scenario, using the NUSAP method.* Dutch National Research Program on Climate Change, Bilthoven, 2002, 225 pp. (available from www.nusap.net)

Other references:

J. van der Sluijs, P. Kloprogge, J. Risbey, and J. Ravetz, Towards a Synthesis of Qualitative and Quantitative Uncertainty Assessment: Applications of the Numeral, Unit, Spread, Assessment, Pedigree (NUSAP) System. in: *International Workshop on Uncertainty, Sensitivity and Parameter Estimation for Multimedia Environmental Modeling*, (proceedings) Interagency Steering Committee on Multi Media Environmental Modeling, August 19-21 2003, Rockville MD, USA. p.81-86. (Available from www.nusap.net)

J.P. van der Sluijs, *Anchoring Amid Uncertainty, on the Management of Uncertainties in Risk Assessment of Anthropogenic Climate Change*, PhD thesis, Utrecht University, 1997.

NUSAP Experts:

RIVM: Mark van Oorschot, Arthur Petersen, Peter Janssen, Bert de Vries, Detlef van Vuuren.
National: Jeroen van der Sluijs, Penny Kloprogge, Jose Potting, Nelleke Honingh.
International: Silvio Funtowicz, Jerry Ravetz, Serafin Corral, James Risbey, Matthieu Craye, Heleen Groenenberg, Erle Ellis, Jean-Marc Douguet, Martin 'O Connor.

Expert Elicitation for Uncertainty Quantification

Description

Expert elicitation is a structured process to elicit subjective judgements from experts[1]. It is widely used in quantitative risk analysis to quantify uncertainties in cases where there is no or too few direct empirical data available to infer on uncertainty. Usually the subjective judgement is represented as a 'subjective' probability density function (PDF) reflecting the experts degree of belief.

Goals and Use

Expert elicitation in the context of uncertainty quantification aims at a credible and traceable account of specifying probabilistic information regarding uncertainty, in a structured and documented way. Typically it is applied in situations where there is scarce or insufficient empirical material for a direct quantification of uncertainty, and where it is relevant to obtain scrutable and defendable results (Hora, 1992).
Several elicitation protocols have been developed amongst which the much-used Stanford/SRI Protocol is the first (Spetzler and von Holstein, 1975; see also Merkhofer, 1987; Morgan and Henrion,1990; chapter 6 and 7). Related expert elicitation protocols have been employed by Sandia National Laboratories for uncertainty quantification in the nuclear energy risk assessment field (Hora and Iman, 1989; Keeney and von Winterfeldt, 1991; Ortiz et al. 1991; Hora, 1992; NCRP, 1996). As an outcome of a joint project of the European union and the US Nuclear Regulatory Commission, Cooke and Goossens (2000a,b) have developed a European guide for expert judgement on uncertainties of accident consequence models for nuclear power plants.

In the sequel we will discuss *two specific elicitation protocols*, briefly commenting on the steps involved.

[A] The **first** protocol is based for a large part on the Stanford/SRI protocol, but additionally provides an explicit assessment of the quality of the uncertainty information on basis of a pedigree analysis (see Risbey et al., 2001; van der Sluijs et al. 2002, and the NUSAP entry in this tool-catalogue). The following steps are involved:

Identifying and selecting experts
It is important to assemble an expert panel representing all points of view.

Motivating the Subject
Establish rapport with the expert. Explain to the expert the nature of the problem at hand and the analysis being conducted. Give the expert insight on how their judgements will be used. Discuss the methodology and explain the further structure of the elicitation procedure. Discuss the issue of motivational biases and try to let the expert make explicit any motivational bias that may distort his judgement.

Structuring
The objective is to arrive at a clear and unambiguous definition of the quantity to be assessed. Choose a unit and scale that is familiar to the expert in order to characterize the selected variable. Underlying conditions and assumptions that the expert is making should be clearly identified.

Elicit extremes
Let the expert state the extreme minimum and maximum conceivable values for the variable.

Extreme assessment
Ask the expert to try to envision ways or situations in which the extremes might be broader than he stated. Ask the expert to describe such a situation if he can think of one, and allow revision of the extreme values accordingly in that event.

[1] An expert is a person who has special skills or knowledge in a particular field. A judgement is the forming of an estimate or conclusion from information presented to or available to the expert.

Assessment of knowledge level and selection of distribution

Before letting the expert specify more detailed information about the distribution it is important that this is done in a way that is consistent with the level of knowledge about the variable. In particular, we seek to avoid specifying more about the distribution shape than is actually known.

Risbey et al. (2001) have proposed a heuristic for this, making use of aggregated normalized pedigree scores (see NUSAP entry in this tool-catalogue) to guide selection of distribution shape: If the aggregated normalized pedigree grade for the parameter is less than 0.3, use a uniform distribution. If it is between 0.3 and 0.7, use a triangular distribution. If it is greater than 0.7, use a normal distribution or other distributions as appropriate.

Specification of distribution

If the expert selected a uniform distribution you do not need to elicit any further values. If the expert selected a triangular distribution, let him estimate the mode. If he chooses another shape for the distribution (e.g. normal), you have to elicit either parameters (e.g. mean and standard deviation for normal distribution) or values of - for instance - the 5th, 50th, and 95th percentiles. Let the expert briefly justify his choice of distribution.

Check

Verify the probability distribution constructed against the expert's beliefs, to make sure that the distribution correctly represents those beliefs.

Aggregating expert distributions

In case that multiple experts have assessed PDFs, there is no single best way to combine their findings. It is recommended to run the Monte Carlo simulations of the model under study separately for each expert's uncertainty specification, and to compare their differences. If differences between experts are large, one should analyse where and why this happens. A conservative choice could be to select the broadest PDF from among the different experts, and use that, unless there are good reasons not to do so. In communicating the results one should explicitly address that there is expert disagreement, and mention that the choice of distribution is somehow indicative of the upper range of the spread amongst the disparate experts.

[B] The **second** protocol that we present is the one by Cooke and Goossens (2000a,b) which was further adapted by Van der Fels-Klerx et al (2002) for use in heterogeneous expert panels on broad or multidisciplinary issues. Major ground-rule in Cooke and Goossens set-up is that the experts should in principle only be questioned about *(potentially) observable* quantities within the area of their expertise[1]. Moreover the protocol aims to explicitly assess the expert's performance by letting the expert elicit so called 'performance' or 'seed' variables, the values of which are unknown to the expert, but known to the analyst. Furthermore performance based weights can be determined to aggregate the assessed PDFs of the individual experts into a combined assessment, which is supposed to reflect a kind of rational consensus on the PDF at hand. The various steps of Cooke and Goossens (2000a,b), protocol are as follows:

Preparation for elicitation:

(1) Definition of the 'case structure' document which clearly describes the field of interest for which expert judgements will be required; the document moreover discusses the aim

[1]Therefore they e.g. prefer to elicit/question on concentrations rather than on transfer function coefficients in compartmental models; the uncertainty information can then be translated back into uncertainty information on the coefficients (e.g. by probabilistic inversion cf. Cooke and Kraan, 2000; Bedford and Cooke, 2001).
Reason for this rather 'empirical stance' concerning questioning on (potentially) observable quantities, is that Cooke and Goossens view uncertainty – from a scientific and engineering viewpoint - essentially as 'that which is removed by observation'. Moreover they put forward that not all experts may subscribe to the particular model choices that have been made and that parameters may not necessarily correspond to the measurement material with which they are familiar. A further argument for their stance is to be found in the fact that direct specification/elicitation of correlations between variables in abstract spaces can be rather problematic and arbitrary.

of the elicitation, and provides background information on applied assumptions and on which issues are taken into account in the uncertainty assessment and which issues are excluded.

(2) Identification of the variables of interest or *'target'* variables for the uncertainty elicitation. Typically a certain pre-selection has to take place, to focus on the most important ones for expert elicitation, since the number of questions to be asked by the experts is limited.

(3) Identification of the '*query'* or *'elicitation'* variables: Target variables, which can in principle be measured by a procedure with which experts are familiar, can directly serve as query variables in an elicitation session. However, target variables for which no such measurement procedures exist cannot be quantified by direct elicitation, and for these variables other derived elicitation query variables (e.g. proxy's) must be found which – ideally - should be (potentially) observable quantities. Information on the uncertainty in these derived elicitation variables must then be translated back into the target variables (see step (14)).

(4) Identification of *performance variables* (or seed variables): These variables serve as a means to assess the expert's performance in rendering uncertainty information on the target variables. There must be experimental evidence on the seed variables, which is unknown to the experts, but known to the analyst, against which the expert's performance can be gauged somehow. Preferably the seed parameters are so-called 'domain variables', referring directly to the target variables. When this is not feasible, 'adjacent variables' may be used.

(5) Identification of experts: In this step an (as large as possible) list of names of 'domain' experts is collected

(6) Selection of experts: In general, the largest possible number of experts should be used, but at least four. Selection of experts may take place on basis of selection criteria (e.g. reputation in field of interest, experimental experience; diversity in background, balance of views etc.)

(7) Definition of an elicitation format document, which should contain clear questions, explanations, and remarks on what is to be included or excluded in the uncertainty assessments, as well as the specific format in which the assessments should be provided by the experts. The elicitation principally focuses on variables, which are (at least in theoretical sense) measurable. The other target variables parameters are deduced by probabilistic inverse modelling principle (see point (3) and (14).

(8) Dry run exercise: Performing a try out of the elicitation serves to find out where ambiguities and flaws need to be repaired, and whether all relevant information and questions are provided.

(9) Training experts for the quantitative assessment task: Typically experts are asked to provide their subjective PDFs in terms of quantiles of the cumulative distribution, for instance, 5%, 50% and 95% percentiles. They need to be trained in providing subjective assessments in probabilistic terms, and in understanding subjective probability related issues.

Elicitation:

(10) Expert elicitation session, where the experts are questioned individually by an analyst[1] to assess the PDF of the query variables (including the seed variables), referring to his field of expertise. As an aid in this process Van der Fels-Klerx et al. (2002) recommend the use of the interactive software package ELI[2] (van Lenthe, 1993), which makes the process of eliciting continuous PDFs easier and less prone to errors and biases.
In addition to the individual expert interviews, there will in some cases also be joint expert meetings, e.g. to discuss general starting points, or in an intermediate stage as a qualitative wrap-up reviewing of rationales behind the assessments, which can then be used as a shared information base for the next iteration in the individual expert elicitation.

[1] In complex situations two analysts will be recommended, a normative analyst (experienced in probability issues) and a substantive analyst (experienced in the expert's field of interest)

[2] Other examples of elicitation software are PROBES and HYPO, described in Lau and Leong (1999) and Li et al. (2002). Apparently these packages focus on the elicitation process for Bayesian networks.

Post-elicitation:

(11) Analysis of expert data, e.g. aggregating the individual experts assessment in one combined probability density function for each of the query variables, e.g. by weighing the experts according to their expertise as measured e.g. by the performance on the seed variables. Software for performing this task is Excalibur (http://ssor.twi.tudelft.nl/~risk/software/excalibur.html).

(12) Robustness and discrepancy analysis, e.g. by removing experts or seed variables from the data set one at a time, and recalculating the combined PDF, comparing it with the original one, which uses all information. Discrepancy analysis identifies items on which the uncertainty assessments of the experts differ most. These items should be reviewed to ascertain any avoidable causes of discrepancy.

(13) Feed back communication with the experts: In general results are treated anonymously, and each expert should have access to his/her assessment and performance weights and scores.

(14) Post-processing analyses (e.g. inverse probability mapping) using the methods for processing uncertainties of the combined expert assessments (see step (11)) of query variables (defined in step 3) into uncertainties on the target variables from step 2. See e.g. Cooke and Kraan (2000).

(15) Documentation of the results: All relevant information and data, including the rationales of the elicitations should be documented in formal reports, to be presented to the experts and to the decision makers.

The above-presented methods differ in a number of respects:

(i) In method [A] the *qualification of the elicited uncertainty information* has an explicit place and is done on basis of a pedigree analysis, which invites the expert to explicitly indicate the quality and underpinnings of his uncertainty statements. In method [B] this qualification is done on a more empirical basis, by measuring the performance scoring of the expert on basis of seed variables. If seed variables are not available, then in fact no explicit or systematic qualification of uncertainty information is undertaken. The best one can hope for is that the expert's elicitation rationale offers suitable information on the underpinnings of the uncertainty specifications, but this is not explicitly commanded in the protocol.
Finally, we must realize that in both cases, [A] as well as [B], one is confronted with the problem how 'valid' the established qualifications/scoring are. In [A], since the pedigree scoring is partly done on basis of subjective judgement, and in [B] since one can rightfully ask to what extent the performance scorings on the seed variables are representative for the measuring the performance on all the other target variables. Moreover the quality of the empirical information on the seed variables - which is ideally only known to the analysts - can also be a problematic factor in this context.

(ii) The second major difference is that method [B] is stricter on the choice of elicitation variables: only those variables are explicitly elicited for which there is (in principle) empirical evidence with which the expert is acquainted (query variables). Information on other target variables is deduced indirectly from the elicited information on the query variables by applying formal mathematical techniques as e.g. probabilistic inversion. Method [A] is less strict: an expert can e.g. be elicited directly on variables, which have no or very low empirical support (i.e. having a low score on the empirical or validation pedigree). Needless to say this can make the elicited PDF rather arbitrary or badly testable, unless there is a good proxy, which can serve as a suitable benchmark. It is therefore important to ask the expert explicitly to indicate how he makes his inference on the PDF; the reasoning involved will typically be more heuristic and less traceable than in the use of probabilistic inversion.

(iii) Thirdly, there is an apparent difference in the specification of the PDFs in both methods. [A] typically uses PDFs of specific and familiar form, while [B] primarily does not require an explicit distribution shape. It focuses instead on specifying a number of quantiles, e.g. the 5, 50 and 95 quantiles (see for instance van Oorschot et al. 2003 where additionally the 25 and 75 quantiles are elicited; note that van der Fels-Klerx, 2002 propose to use ELI in the elicitation process which applies Beta-distributions) and then

uses information theoretic arguments to further process this information. As such [B] only seems to use limited distributional information, and further supplies it by using information theoretic principles/assumptions. Potentially available information on the specific form of a distribution is not taken fully into account.

(iv) Finally, the treatment of multiple experts in method [A] is more heuristic and less formalized than under [B], where an explicit weighted averaging is applied on basis of the seed-variable.

It is difficult to say beforehand which method, [A] or [B], would be preferable, since both have their strengths and weaknesses. In practice we would recommend a judicious mix of both methods, depending on the availability and quality of data and information, and comparing the pros and cons mentioned in the foregoing.

Moreover in setting up a specific elicitation protocol for a particular case there will be additional points of attention to be dealt with. See the 'guidance on application'-entry listed below for a more comprehensive overview.

Sorts and locations of uncertainty addressed

The Stanford Protocol and its variants typically address inexactness in inputs and parameters, and are focused mainly on statistical uncertainty. The Risbey et al. (2001) protocol combines an elicitation on PDFs with an elicitation of the parameter pedigree (see also van der Sluijs et al. 2001, and the NUSAP entry in this tool catalogue), and therefore addresses also the unreliability of inputs, parameters and instruments (or model structure).

In principle expert elicitation techniques can be tailored and used to elicit and encode subjective expert judgements on any sort of uncertainty at any location distinguished in the uncertainty typology.

Required resources

- Typically performing a formal expert elicitation is a time and resource intensive activity. The whole process of setting up a study, selecting experts, preparing elicitation questions, performing expert training, expert meetings, interviews, analyses, writing rationales, documentation etc. can easily stretch over months or years, and involve. See Hora and Iman, 1989, Ortiz et al. (1991). The choice whether to perform a formal or a more informal elicitation (NCRP, 1996) depends on the price one is willing to pay for more scrutable and defensible results, and will be influenced by the relevance and controversies regarding the problem area.
- One needs to have good interviewing skills and needs to have a reasonable understanding of the field under consideration.
- Skill is needed to draft a good questionnaire or template for the elicitation.
- Training in elicitation techniques may be needed.

Strengths and limitations

Strengths

- It has the potential to make use of all available knowledge including knowledge that cannot be easily formalized otherwise.
- It can easily include views of sceptics and reveals the level of expert disagreement on certain estimates.

Weaknesses

- The fraction of experts holding a given view is not proportional to the probability of that view being correct.
- One may safely average estimates of model parameters, but if the expert's models were incommensurate, one may not average models (Keith, 1996).

- If differences in expert opinion are irresolvable, weighing and combining the individual estimates of distributions is only valid if weighted with competence of the experts regarding making the estimate. There is no good way to measure competence. In practice, the opinions are often weighted equally, although sometimes self-rating is used to obtain a weight-factor for the experts competence
- The results are sensitive to the selection of the experts whose estimates are gathered.

Although subjective probability is an imperfect substitute for established knowledge and despite the problems of aggregation of expert judgement, if nothing better is available it is better to use subjective probability distributions than deterministic point-values so that one has at least a first approximation of the uncertainty.

Guidance on application

A good primer on expert elicitation is Frey, 1998, which is available from: http://courses.ncsu.edu/classes/ce456001/www/Background1.html. See also Baecher, 2002 available from http:/www.glue.umd.edu/~gbaecher/papers.d/judge_prob.d/judge_prob.html. For books dedicated to expert elicitation see e.g. Meyer and Booker, 1991 and Ayyub, 2001. Cooke, 1991 addresses wider methodological and theoretical issues concerning expert judgement in uncertainty.
Below we discuss a number of extra points, next to the ones mentioned under the heading 'Goals and Use' which deserve special attention when setting up and performing an elicitation:

1. Preliminary screening
The amount of work can be reduced by performing some preliminary screening to select those variables whose uncertainty will affect the outcomes of interest most. Expert elicitation can then focus first on these variables, whilst the other variables are assessed e.g. in a less thorough way.

2. Importance of providing rationale on choices, and assessment of quality (checks and balances)
The uncertainty assessments, as well as the underlying motivations, assumptions and information (measurements, models, reasoning, literature references etc.) that have been used to provide them should be well documented. Some discussion on the backing and quality of this material is also important, as well as an indication of which uncertainty aspects have been left aside. Applying a systematic analysis like e.g. pedigree analysis can be helpful for this purpose.

3. Variability vs. lack-of-knowledge related uncertainty
Uncertainty can partly be seen as an intrinsic property of the system (variability and diversity; but also sampling error), and partly as property of the analyst and knowledge base (e.g. lack of good-quality data, lack of expert knowledge; lack of consensus; controversy[1]).
Though there is often a certain degree of arbitrariness[2] in distinguishing between this variability-induced uncertainty ('aleatory') and lack-of-knowledge induced uncertainty ('epistemic')[3], depending on e.g. modelling purpose, choice of analysis and aggregation level, available information, tradition, (see e.g. Baecher and Christian, 2001), we think it is important to treat this distinction[4] explicitly and with care when eliciting on uncertainty. Hora,

[1] Though these latter aspects (lack of consensus, controversy) can also be partly seen as system properties, reflecting variability and diversity rather than lack-of-knowledge.
[2] Some researchers even argue that at a basic level it is questionable to make this distinction (see Winkler, 1996; Bedford and Cooke, 2001), but that from a practical viewpoint it can certainly be helpful to distinguish between aleatory and epistemic, in order to decompose and analyse the underlying issues leading to uncertainty in an adequate way.
[3] This distinction has been described under various names, e.g. stochastic, type A, irreducible, variability for aleatory, and subjective, type B, reducible, and state of knowledge for epistemic.
[4] See e.g. Hofer, 1996, for a clear discussion on the need to use this distinction. See also Hoffman and Hammonds, 1994.

1996, illustrates that a careful decomposition and treatment of aleatory and epistemic uncertainty, centered on the notion of conditional probabilities, is essential for the expert elicitation process and can highly influence the further processing and assessment of uncertainties. See also the recent study of van Oorschot et al., 2003, which underlines these findings.

4. Heuristics and biases

In providing information by expert elicitation one has to cope with judgemental 'heuristics' (mental rules of thumb) and the 'biases', which they produce in the expert judgement (see Kahneman et al. 1982, Griffin et al. 2002). This relates to biases due to cognitive processes as well as to motivational, social and cultural biases (see the subsequent entry on pitfalls). One can try to diminish these biases by training the expert in elicitation and its pitfalls and by setting up the elicitation process adequately (using individual and/or controlled group settings, applying e.g. Delphi technique; nominal group (Benarie, 1988, Cooke, 1991, Ayyub, 2001)) and formulating the elicitation questions in a judicious and unambiguous way (e.g. by applying counterfactual reasoning; asking first to specify the extremes). However, debiasing expert judgements before and during elicitation will stay a difficult goal. Ex post calibration can have some value, but it requires a reasonable amount of data, and moreover the quality and representativeness/covering should be adequate (e.g. Clemen and Lichtendahl (2002) present a Bayesian based calibration method to debias expert overconfidence, enabling also interrelationships between the inference of experts).

5. Selection of distributional form

When data is scarce, a test on goodness of fit will usually give no distinctive outcomes concerning the distribution types (see Cullen and Frey,1999, Morgan and Henrion 1990, Seiler and Alvarez 1996).

For situations with few data a more robust approach is to select distributions in such way that the uncertainty associated with the available information is maximized, i.e. not imposing extra assumptions that are not warranted by data (minimal information). As a first approximation the following distributions are suggested:

Available values	**Shape of distribution to use**
{min,max}	Uniform
{mean, standard deviation}	Normal
{min,max,mode}	Triangular
{min=0, mean}	Exponential
(min,max,mean,sd}	Beta
{min>0,quantile}	Gamma
{min,max,mean}	Beta
{min=0, quantile}	Exponential

The following rules of thumb can offer some guidance in this setting (an important issue which is not explicitly addressed, but should nevertheless be taken into account, is the quality (e.g.) accuracy of the available information on min, max etc.):

- If little or no relevant data exist, and information on min, max, or most probable value is not available, then it is recommended to carry out calculations with different PDFs in the parameter, to reflect whatever feasible information is available. The uncertainty-range in the corresponding outcomes gives a rough indication of the lack-of-knowledge in the parameter.
- If Min,max is given try uniform distributions; in case of a large range, try loguniform distribution; if additionally a mode is given, try triangular, or likewise log-triangular in case of a large range.
- If some relevant data exist, but cannot be represented by standard statistical distribution, then use piecewise uniform (empirical) distribution.
- If substantial amount of data exist, and can be reasonably well represented by standard distribution, use estimation to find the characteristic distributional parameters (e.g. maximum likelihood, method of moments; Bayesian etc.)

- In case the parameter can be expressed as quotient/product of other parameters, it is often feasible to approximate the PDF by a lognormal distribution (see also (Vose 2000; Cullen and Frey 1999; Morgan and Henrion 1900).

In the Cooke and Goossens protocol, the experts typically have to assess the 5, 50 and 95 quantiles (but other or more pinpoints can be chosen; cf. van Oorschot et al., 2003), as a basis for specifying the distribution. This information is further processed using minimal information theoretic arguments.
Van der Fels-Klerx et al. 2002 recommend the use of the graphical software tool ELI (van Lenthe, 1993) to elicit PDFs for continuous variables. It employs the general beta-distribution as a template, and makes elicitation easier, and less prone to errors and biases.
Needless to say, specific applications can require special forms of distributions (e.g. when modelling extreme events) and the above-given recommendations are therefore not necessarily the optimal ones.

6. Correlation or dependence specifications
The way in which dependence or correlation is taken into account and elicited can highly influence the outcomes of the uncertainty assessment. Often little is known on specific dependencies, and dependence elicitation is not an easy job. In the Cooke and Goossens methodology this elicitation is done using the notion of conditional probability, querying an expert to *'specify what the probability is that the value of Z will lay above its median, in case that Y was observed to lie above its median, in an experiment which involves both Z and Y'.* This probability-information can be easily translated into a specific rank-correlation between Z and Y (cf. e.g. Kraan, 2002; section 2.1.2). In order to prevent that an expert has to specify too many dependencies - which moreover can easily lead to incompatible correlation matrices - two parsimonious query procedures have been proposed by Cooke and his co-workers, use copulas[1] as a basis for the dependence structure. The first one employs a *tree* (i.e. an acyclic undirected graph) in which the rank correlations are specified for a (limited) selection of all possible correlations. Using minimal information theoretical arguments and bivariate copulas (Meeuwissen and Cooke 1994), a sample is constructed with the requested marginal distributions having a compatible correlation structure with the specified dependencies. The second approach is a generalization of the correlation tree method, and uses a *vine* as the basic structure for specifying desired rank correlations. A vine is a nested set of trees build on top of each other where the edges of tree j are the nodes of tree j+1 (see e.g. Bedford and Cooke, section 17.2.5). By using partial correlation specification associated to the vine edges, and using e.g. elliptical copula, a sample can be constructed which exhibits desired marginals and a specific rank correlation matrix. The advantage of the partial correlation based specification is that no conditions like positive definiteness need to be satisfied for the specification, and that the associated sampling 'works on the fly': i.e. one sample vector at a time is drawn, without a need to store large numbers of samples in memory. See e.g. Kurowicka and Cooke (2001). Part of these procedures have been implemented in UNICORN (Cooke, 1995).

Apart from the way in which correlation and dependence is expressed mathematically, also the structuring (decomposition, recomposing and aggregating) of the problem will to a large extend determine in which way dependence will be encountered. It makes for instance quite a difference whether the basic level of analysis and elicitation is an individual, a subpopulation or a complete population. Moreover, (unexpected) dependencies and correlations can be introduced when both aleatory and epistemic uncertainties are present (Hora, 1996): e.g. when the parameters which describe the variability are not completely known[2] this epistemic uncertainty in fact pervades all elements of the population in a similar way, rendering a certain dependence between the sampled individuals. The associated uncertainty which is introduced in this manner in fact reflects a systematic error, cf. also Ferson and 1996.
Part of these issues are illustrated by the recent study of van Oorschot et al. 2003 where an extensive uncertainty assessment of an emission inventory is reported. This assessment

[1] A copula is a joint distribution on the unit square having uniform marginals. It provides a suitable technique for modelling dependence, going beyond the pitfalls of correlation. (see e.g. Embregts, McNeil and Straumann, 1999; Clemen and Reilly, 1997)
[2] I.e. there is epistemic uncertainty concerning the precise form of the variability.

consisted of upgrading an expert elicitation on the individual level towards uncertainty statements on the total population level, taking due account of the presence of aleatory and epistemic uncertainty. Experiences with this study made clear that elicitation and analysis of aleatory and epistemic uncertainty and associated dependencies, remains a challenging field of research.

7. Aggregating expert judgements

In practice it can occur that expert opinions differ considerably on specific issues (cf. Morgan and Keith, 1995; Morgan et al. 2001, Morgan, 2003). There is no univocal strategy on how to handle these situations:

- One option is trying to combine the individual estimates into a kind of *group-PDF* (see e.g. Clemen and Winkler, 1999), which is supposed to 'summarize' the group opinion. However one has to be careful with drawing such a conclusion: a 'summary' in the form of one group PDF does not necessarily express a consensus, and moreover the summary may obscure or suppress differences among experts and thus over present the precision in the judgements (Hora, 1992). It has to be stressed that it always will be important - notwithstanding the focus on coming up with a group- PDF - to analyse and discuss the diversity in individual PDFs in some detail: e.g. where does it occur, what are its main reasons and what are its major effects on the final results of interest (compare e.g. the robustness and discrepancy analysis of Cooke and Goossens 2000a,b). Such kind of analysis will render relevant information for the decision maker on how to interpret and use the results and where to focus for potential improvements in uncertainty information.
- Another option is not to combine the individual PDFs in case of considerable diversity, but to present and discuss the diversity in the individual PDFs *separately* in its full scope, indicating its potential consequences for policy analysis. See e.g. Keith, 1996, who advocates that diversity can serve as a warning flag to seek for meaningful alternative modes of policy analysis, which may be highly relevant for the debate concerning the problem. He warns against adhering to one 'pseudo'-truth, on basis of an 'aggregated PDF', and thereby masking diversity which can be due to e.g. disparate values and interests. Especially in cases where there exist scientific controversies, we recommend to avoid combining expert judgements because it goes at the expense of transparency of the analysis and looses some insightful information on the level and nature of scientific controversy and expert disagreement. In case of policy problems, which require a post-normal science approach to risk assessment, such information is crucial and should not be obscured by aggregation.

Practical considerations will however often force one to work with *one PDF* for each source of uncertainty (e.g. it is often practically infeasible to work through and present all the multi-expert combinations on all sources of uncertainty etc.). Given the above caveats it is important to clearly indicate the assumptions and limitations in doing so, to prevent that the results will be wrongly interpreted.

Finally we will discuss two main ways to aggregate expert opinions (see Clemen and Winkler, 1999):

- Using *behavioural* approaches which try to establish a consensus-PDF by carefully-structured joint group meetings (using e.g. 'Delphi method', 'nominal group technique', 'decision conferencing', to exchange, discuss and process information and views etc.). Trying to avoid social and cognitive trappings in group discussion is an important issue, but no overall best method seems to exist. Often consensus cannot be reached, despite of repeated group-meetings. Mathematical aggregation is finally used as a rather arbitrary and artificial way of providing one PDF.
- Using *mathematical* aggregation approaches which can range from simply averaging the individual information on probabilities to a more sophisticated analysis of the information aggregation process, accounting for information on the performance-quality of the experts (Goossens, Cooke and Kraan, 1998) and the dependence among the experts' probabilities (Reichert and Keith, 2003). Clemen and Winkler, 1999, state that the more complex combination rules sometimes outperform the simple rules (e.g. simple averaging), but that they can be more sensitive, leading to poor performance in some instances (i.e. robustness is low).

The empirical evidence in Clemen and Winkler 1999 does not suggest a preference for one of these approaches, but suggests that a wise use of both methods will often be the best approach in practice. More future research is apparently needed.

Pitfalls

- Unfamiliarity of the expert with the wording and statistical terminology in elicitation questions (e.g. sometimes people are inclined to specify the mean when they are asked for the median, which is apparently incorrect for asymmetric distributions). Clear and unambiguous questions are of importance, especially concerning issues as variability- and 'lack-of-knowledge'-induced uncertainty and correlations and dependencies. Using frequency formats when talking on probabilities connects better to the way people reason and experience (Hoffrage et al. 2000), and is therefore suggested as the preferred way of communicating on chances, even in a 'uncertainty as degree-of-belief'-setting (Anderson, 1998a, 1998b). A brief training in background and use of statistical terminology for elicitation is recommended.
- The occurrence of groupthink or social bias in group settings during an elicitation process. It is recommended to use procedures or techniques to diminish this influence.
- The 'validity' of the obtained scores in assessing the quality of the provided uncertainty information can be low e.g. due to lack of representativity, scope and accuracy of the seed variables in a performance assessment or due to inherent subjectivity and limitations of personal scope in the self-rating and pedigree analysis process. Therefore explicit attention should be spend on these aspects, and potential weak spots should be mentioned.
- The outcomes of the probabilistic inversion are dependent on the model structure that is used. Ideally some level of 'model-structure validation' will be required to improve the confidence in the obtained results.
- In combining expert-opinions one runs the risk of masking expert disagreement and throwing away important information concerning the problem, especially if the major differences between the expert opinions are not explicitly discussed and explained. Moreover one should be cautious in interpreting a combined PDF: it by no means needs to represent a consensus view on uncertainty.
- Bias: The major pitfall in expert elicitation is expert bias. Experts and lay people alike are subject to a variety of potential mental errors or shortcomings caused by man's simplified and partly subconscious information processing strategies. It is important to distinguish these so-called cognitive biases from other sources of bias, such as cultural bias, organizational bias, or bias resulting from one's own self-interest (from Psychology of Intelligence Analysis, R.J. Heuer, 1999; http://www.cia.gov/csi/books/19104/index.html). Some of the sources of cognitive bias are as follows: overconfidence, anchoring, availability, representativeness, satisficing, unstated assumptions, coherence, and experts should be informed on the existence of these biases during the expert elicitation process. Below a brief explanation is given of these sources; see for more details e.g. Dawes (1988).

  Anchoring Assessments are often unduly weighted toward the conventional value, or first value given, or to the findings of previous assessments in making an assessment. Thus, they are said to be `anchored' to this value.

  Availability This bias refers to the tendency to give too much weight to readily available data or recent experience (which may not be representative of the required data) in making assessments.

  Coherence Events are considered more likely when many scenarios can be created that lead to the event, or if some scenarios are particularly coherent. Conversely, events are considered unlikely when scenarios cannot be imagined. Thus, probabilities tend to be assigned more on the basis of one's ability to tell coherent stories than on the basis of intrinsic probability of occurrence.

  Overconfidence Experts tend to over-estimate their ability to make quantitative judgements. This is often manifest with an estimate of a quantity and its uncertainty range that does not even encompass the true value of the quantity. This is difficult for an individual to guard against; but a general awareness of the tendency can be important.

Representativeness This is the tendency to place more confidence in a single piece of information that is considered representative of a process than in a larger body of more generalized information.
Satisficing This refers to the tendency to search through a limited number of solution options and to pick from among them. Comprehensiveness is sacrificed for expediency in this case.
Motivational People may have incentives to reach a certain conclusion or see things a certain way. Reasons for occurrence of motivational bias include: a) a person may want to influence a decision to go a certain way; b) the person may perceive that he will be evaluated based on the outcome and might tend to be conservative in his estimates; c) the person may want to suppress uncertainty that he actually believes is present in order to appear knowledgeable or authoritative; and d) the expert has taken a strong stand in the past and does not want to appear to contradict himself by producing a distribution that lends credence to alternative views.
Unstated assumptions A subject's responses are typically conditional on various unstated assumptions. The effect of these assumptions is often to constrain the degree of uncertainty reflected in the resulting estimate of a quantity. Stating assumptions explicitly can help reflect more of a subject's total uncertainty.

Gigerenzer (1991,1994) and Cosmides and Tooby (1996) argue that part of these biases are not so much caused by the limited cognitive abilities of the human mind, but more by the way in which information is presented or elicited. A thoughtful wording of questions can be helpful to avoid part of these biases. Performing dry run exercises (try-outs) can render important feedback on the suitability of the posed questions.

References

Handbooks:

B.M. Ayyub, (2001) *Elicitation of Expert Opinions for Uncertainty and Risks*, CRC Press, Florida.

T. Bedford, R. Cooke (2001). *Probabilistic Risk Analysis: Foundations and Methods.* Cambridge University Press.

R.M. Cooke. *Experts in uncertainty: Opinion and Subjective Probability in Science*. New York, Oxford University Press, 1991.

Cullen, A.C. and H.C. Frey, H.C. (1999). *Probabilistic Techniques in Exposure Assessment*, Plenum Publishing Corp., New York, USA.

R. Dawes, (1990) *Rational Choice in an Uncertain World.*

D. Griffin, T. Gilovich, D. Kahneman (eds.) (2002) *Heuristics and Biases: Psychology of Intuitive Judgment*.

D. Kahneman, A. Tversky, P. Slovic (eds.) [1982]. *Judgment under Uncertainty: Heuristics and Biases.* Cambridge University Press.

M.A. Meyer, J.M. Booker (1991). *Eliciting and Analyzing Expert Judgement: A practical Guide*, Academic Press, London

Papers and Reports:

J.L. Anderson (1998) Embracing Uncertainty: The Interface of Bayesian Statistics and Cognitive Psychology. Conservation Ecology Vol. 2 (cf. http://www.consecol.org/Journal/vol2/iss1/art2/)

J.L. Anderson [1998] Enhancing Communication About Uncertainty. Extension Note, British Columbia; July 1998;

G.B. Baecher, (2002). Expert elicitation in geotechnical risk assessments, (http:/www.glue.umd.edu/~gbaecher/papers.d/judge_prob.d/judge_prob.html)

G.B. Baecher, and J.T. Christian (2000). "Natural variation, limited knowledge, and the nature of uncertainty in risk analysis". Presented at '*Risk-based Decisionmaking in Water Resources IX*', Oct. 15-20, 2000, Santa Barbara. http://www.glue.umd.edu/~gbaecher/papers.d/Baecher_&_Christian Sta Barbara 2000.pdf

M. Benarie (1988) Delphi- and delphilike approaches with special regard to environmental standard setting. Technological Forecasting and Social Change, Vol. 33, pp. 149-158.

R.T. Clemen, K.C. Lichtendahl, jr. 2002 Debiasing Expert Confidence: A Bayesian Calibration Model, presented at PSAM6-conference

R.T. Clemen, T. Reilly [1999] Correlations and copulas for decision and risk analysis. Management Science. 45: pp. 208-224.

R.T. Clemen, R.L. Winkler, 1999 Combining probability distributions from experts in risk analysis. Risk Analysis, Vol. 19, pp. 187-203

R.M. Cooke (1995). UNICORN; Methods and Code for Uncertainty Analysis. Published by the Atomic Energy Association, Delft University of Technology.

R.M. Cooke, L.J.H. Goosens, (2000a). Procedures Guide for Structured Expert Judgment. Technical report EUR 18820 EN, European Commission, Directorate-General for Research, Brussels, Belgium, 2000

R.M. Cooke, L.H.J. Goossens (2000b) Procedures Guide for Structure Expert Judgement in Accident Consequence Modelling. Radiation Protection and Dosimetry vol 90, no. 3 2000 pp 303-311

R. Cooke, B. Kraan (2000) Processing expert judgements in accident consequence modelling. Radiation Protection Dosimetry. Vol. 90, No. 3, pp. 311-315.

R. Cooke, A. Meeuwissen, 1994 Tree dependent random variables. Technical Report 94-28. Dept. of Mathematics, Delft University of Technology.

Cosmides L. en J. Tooby [1996] Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. Cognition, Vol. 58, pp. 1-73.

P. Embrechts, A. McNeil, D. Straumann (1999). Correlation and Dependence in Risk Management: Properties and pitfalls, To appear in Risk Management: Value at Risk and Beyond, ed. By M. Dempster, H.K. Moffatt. Cambrdige University Press, 2001

S. Ferson, L.R. Ginzburg, 1996. Different methods are needed to propagate ignorance and variability. *Reliability Engineering and System Safety* Vol. 54, pp. 133 – 144.

H.C. Frey, 1998, BRIEFING PAPER PART 1: Introduction to Uncertainty Analysis. (http://courses.ncsu.edu/classes/ce456001/www/Background1.html)

Gigerenzer, 1991 How to make cognitive illusions disappear: Beyond 'heuristics and biases'. European Review of Social Psychology, Vol. 2, 83-115.

Gigerenzer 1994, Why the distinction between single event probabilities and frequencies is relevant for psychology (and vice versa).

L. Goossens, R. Cooke, B. Kraan (1998) Evaluation of weighting schemes for expert judgment studies. PSAM 4 Proceedings, (Moselh and Bari eds) Springer 1998,1937-1942.

A O'Hagan (1998) Eliciting expert beliefs in substantial practical applications. The Statistician, Vol. 47, pp. 21-35

E. Hofer (1996) When to Separate Uncertainties and When Not to Separate. *Reliability Engineering and System Safety* Vol. 54, pp. 113 – 118

F.O. Hoffman, J.S. Hammonds (1994) Propagation of uncertainty in risk assessment: the need to distinguish between uncertainty due to lack of knowledge and uncertainty due to variability. *Risk Analysis*, Vol. 14, pp. 707-712.

Hoffrage U., Lindsey, S., Hertwig R., Gigerenzer G. [2000] Communicating Statistical Information. Science Vol. 290, pp. 2261 e.v.

S.C. Hora (1992) Acquisition of Expert Judgment: Examples from Risk Assessment. Journal of Energy Engineering, Vol. 118, pp. 136-148.

S.C. Hora (1996) Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management. Reliability Engineering and System Safety, Vol. 54, pp. 217-223

S.C. Hora and R.L. Iman (1989) Expert opinion in risk analysis: The NUREG-1150 Methodology. *Nuclear Science and Engineering*, Vol. 102, pp. 323-331.

J.B. Kadane, L.J.Wolfson, P.S. Craig, M. Goldstein, A.H. Seheult, J.A. Smith, A. O'Hagan (1998) Papers on 'Elicitation'. *The Statistician*. Vol. 47, Part 1, pp. 3-68

R.L. Keeney, D. van Winterfeldt (1991) Eliciting probabilities from experts in complex technical problems. *IEEE Trans. On Engineering Management*. Vol. 38, pp. 191-201.

D.W. Keith (1996). When is it appropriate to combine expert judgements? *Climatic Change*, Vol. 33, pp. 139-143.

B. Kraan, (2002). Probabilistic Inversion in Uncertainty Analysis. PH-D. Technical University Delft.

D. Kurowicka, R. Cooke, 2001 Conditional, partial and rank correlation for elliptical copula. In Dependence Modeling in Uncertainty Analysis, Proc. Of ESREL 2001, pp. 259-276.

Lau, A. H. and Leong, T. Y. (1999) PROBES: A Framework for probabilities elicitation from experts. In *Proceedings of the 1999 AMIA Annual Fall Symposium*, pages 301-305, AMIA. See http://www.amia.org/pubs/symposia/D005714.PDF

J. Li, A. Dekhtyar, J. Goldsmith, 2002: Efficiently eliciting many probabilities online. See http://www.cs.uky.edu/~dekhtyar/dblab/hypo.ldg.pdf

M.W. Merkhofer (1987) Quantifying judgmental uncertainty: methodology, experiences and insights. IEEE Trans. On Systems, Man and Cybernetics. Vol. SMC-17, pp.741-752.

M.G. Morgan M. Henrion (1990) *Uncertainty, A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press.

M.G. Morgan and D. Keith (1995), Subjective Judgments by Climate experts. *Environmental Science & Technology*, Vol. 29, pp. 468-476.

M.G. Morgan, L.F. Pitelka, E. Shevliakova (2001) Eastimates of Climate Change impacts on forest ecosystems. *Climatic Change*, Vol. 49, pp. 279-307.

M.G. Morgan (2003), Characterizing and Dealing with Uncertainty: Insights from the Integrated Assessment of Climate Change. *Integrated Assessment*. Vol. 4, pp. 46-55.

Nauta, M, (2001) Risk assessment of Shiga-toxin producing Escherichia coli O157 in steak tartare in the Netherlands. RIVM Report 257851003

NCRP, 1996 A Guide for Uncertainty Analysis in Dose and Risk Assessments related to Environmental Contamination NCRP Commentary no. 14.

W.D. Nordhaus (1994) Expert opinion on climate change. *American Scientist*, Vol. 82, pp. 45-51.

N.R. Ortiz, T.A. Wheeler, R.J. Breeding, S. Hora, M.A. Meyer, R.L. Keeney (1991) Use of expert judgment in NUREG-1150. *Nuclear Engineering and Design*, Vol. 126, pp. 313-331.

H. Otway, D. van Winterfeldt (1992) Expert judgment in risk analysis and management: process, context and pitfalls. *Risk Analysis*, Vol. 12, pp. 83-93.

P. Reichert, D.W. Keith [2003] Bayesian Combination of Expert Judgements: Reinterpretation of the Meaning and dependence of elicited distributions. Submitted for publication.

J.S. Risbey, J.P. van der Sluijs, and J. Ravetz, 2001, A Protocol for Assessment of Uncertainty and Strength of Emissions Data. Copernicus Institute, University Utrecht; NW&S E-2001-10

F.A. Seiler, J.L. Alvarez, 1996. On the selection of distributions for stochastic variables. Risk Analysis, Vol. 16. pp. 5-18.

J.P. van der Sluijs, J.S. Risbey and J. Ravetz, 2003, Uncertainty Assessment of VOC emissions from Paint in the Netherlands. Copernicus Institute, University Utrecht; NW&S E-2002-13

C.S. Spetzler, and S. von Holstein (1975). "Probability Encoding in Decision Analysis." Management Science, 22(3).

A. Tversky, D. Kaheneman (1974) Judgment under uncertainty: heuristics and biases. Science, Vol. 185, pp. 1124-1131.

I.H.J. Van der Fels-Klerx, Cooke, R.M., Nauta, M.J., Goossens, L.H.J., Havelaar, A.H., 2004. A structured expert judgment study for a model of campylobacter contamination during broiler chicken processing. Submitted to Risk Analysis

I.H.J. Van Der Fels-Klerx, L.H.J. Goossens, H.W. Saatkamp, S.H.S. Horst (2002) Elicitation of quantitative data from a heterogeneous expert panel: formal process and application in animal health. Risk Analysis. Vol. 22, pp. 67-81.

J. van Lenthe (1993) A blueprint of ELI: a new method for eliciting subjective probability distributions. *Behavior Research Methods, Instrum. & Computers*. Vol. 25, pp. 425-433.

M.M.D. van Oorschot, B.C.P. Kraan, R.M.M. van den Brink, P.H.M. Janssen, R.M. Cooke, 2003. Uncertainty Analysis for NO_x emissions from Dutch passenger cars in 1998. Applying structured expert elicitation and distinguishing different types of uncertainty. RIVM report 550002004.

Example case studies:

Van der Sluijs, et al. 2003: Uncertainty assessment of VOC-emissions from paint in the Netherlands.

Van Oorschot, et al. 2003: Uncertainty analysis for NO_x emissions from Dutch passenger cars in 1998.

Nauta (2001): Risk assessment for foodproduction and foodprocessing, case of steak tartare.

Van der Fels (2004): Risk assessment for campyylobacter contamination

Software:
Concerning software which can be an aid in probability elicitation:

- PROBES: A Framework for Probability Elicitation from Experts; see Lau and Leong (1999).
- HYPO: Software for elicitation of many probabilities in Bayesian network context; see Li, Dekhtyar, Goldsmith (2002)
- ELI: Software foreliciting subjective probability distributions; see van Lenthe, 1993.
- EXCALIBUR: Software for expert calibration.
- UNICORN: software for uncertainty analysis with correlations

For the latter two packages, see the website: http://ssor.twi.tudelft.nl/~risk/br/software.html

Websites:
http://ssor.twi.tudelft.nl/~risk/br/software.html: Software for expert calibration (EXCALIBUR)

and for uncertainty analysis with correlations (UNICORN)

http://courses.ncsu.edu/classes/ce456001/www/Background1.html: primer on expert elicitation

Experts at RIVM:
Mark van Oorschot, Peter Janssen, Maarten Nauta

Scenario analysis

Description

Scenario analysis is a method that tries to describe logical and internally consistent sequences of events to explore how the future might, could or should evolve from the past and present. The future is inherently uncertain. Through scenario analysis different alternative futures can be explored and thus uncertainties addressed. As such, scenario analysis is also a tool to deal explicitly with different assumptions about the future. Several definitions of scenarios can be found in the literature. In the definition of UNEP (2002), the uncertainty aspect is explicitly incorporated. "Scenarios are descriptions of journeys to possible futures. They reflect different assumptions about how current trends will unfold, how critical uncertainties will play out and what new factors will come into play". Another definition is the following: A scenario is a description of the present state of a social and or natural system (or a part of it), of possible and desirable future states of that system along with sequences of events that could lead from the present state to these future states (e.g. Jansen Schoonhoven and Roschar, 1992). Other definitions also include the purposes of the use of scenarios. Van Notten (2002) defines scenarios as "descriptions of possible futures that reflect perspectives on past, present, and upcoming developments in order to anticipate the future".
Different types of scenarios exist. Alcamo (2001) discerns baseline vs. policy scenarios, exploratory vs. anticipatory scenarios and qualitative vs. quantitative scenarios.

- *Baseline scenarios* (or reference-, benchmark- or non-intervention scenarios) present the future state of society and environment in which no (additional) environmental policies do exist or have a discernable influence on society or the environment. *Policy scenarios* (or pollution control-, mitigation- or intervention scenarios) depict the future effects of environmental protection policies.
- *Exploratory scenarios* (or descriptive scenarios) start in the present and explore possible trends into the future. *Anticipatory scenarios* (or prescriptive or normative scenarios) start with a prescribed vision of the future and then work backwards in time to visualise how this future could emerge.
- *Qualitative scenarios* describe possible futures in the form of narrative texts or so-called "story-lines". *Quantitative scenarios* provide tables and figures incorporating numerical data often generated by sophisticated models.

Finally scenarios can be surprise-free or trend scenarios, which extend foreseen developments, on the one hand or including surprises and exploring the extremes (e.g. best case / worst case) on the other hand.

Goals and use

Typical objectives of scenario analysis in environmental assessment are (Alcamo, 2001):

- Providing a picture of future alternative states of the environment in the absence of additional environmental policies (baseline scenarios).
- Illustrating how alternative policy pathways can achieve an environmental target
- Identifying the robustness of environmental policies under different future conditions
- Raise awareness about different (future) environmental problems and the connection between them
- Help stakeholders, policymakers and experts to take into account the large time and space scales of a problem
- Combine qualitative and quantitative information about the future evaluation of an environmental problem

Alternative *baseline scenarios* can be used to evaluate the consequences of current policies taking into account uncertainties in driving forces, such as economic and socio-cultural developments. Also alternative baseline scenarios can be used to take into account uncertainties about environmental processes occurring in nature and about impacts of

environmental conditions on human health. In the same way *policy scenarios* can be used to evaluate environmental and economic impacts of environmental policies or other policies taking into account uncertainties in e.g. societal driving forces and environmental processes.

Exploratory scenarios can be used when the objective is to explore the consequences of a specified future trend in driving forces, or the consequences of implementing a policy. *Anticipatory scenarios* can be used when the objective is to investigate the steps leading to a specified end state, such as an environmental target.

Qualitative scenarios analysis can be used when the objective is to stimulate brainstorming about an issue, when many views about the future have to be included or when an idea has to be formed about for example general social and cultural trends. *Quantitative scenario analysis* can be used for assessments that require data and numbers, for example on the magnitude of air pollutant emissions.
Combinations are also possible, e.g. the "Story-and-Simulation" (SAS) approach, which combines the development of qualitative "storylines" by a group of stakeholders and experts and the use of models to quantify the storylines (Alcamo, 2001).

The principal elements of typical scenarios used in environmental studies are (adapted from Jansen Schoonhoven and Roschar, 1992 and Alcamo 2001):

- Description of the present situation
- Several alternative views on future developments, e.g. by means of story lines
- Description of step-wise changes in the future state of society and the environment i.e. trajectories consisting of logical sequences of events that correspond with and are consistent with each view on future developments.
- Driving forces influencing the step-wise changes
- Base year
- Time horizon and time steps

The main methods for in developing scenarios are:

- Scenario writing (qualitative scenarios): policy exercises
- Modeling analysis (quantitative scenarios)

The University of Kassel developed a method to combine both qualitative and quantitative approaches: the SAS is approach as mentioned above. (Alcamo, 2001)

The SAS approach includes the following steps:

1. The scenario team and scenario panel are established
2. The scenario team proposes goals and outline of scenarios
3. The scenario panel revises goals and outline of the scenarios, and constructs zero order draft of storylines
4. Based on the draft story line the scenario team quantifies the driving forces of the scenarios
5. Based on the assigned driving forces , the modelling teams quantify the indicators of the scenarios
6. At the next meeting of the scenario panel , the modelling team report5s on the quantification of the scenarios and the panel revises the storylines
7. Steps 4,5 and 6 are repeated until an acceptable draft of storylines and quantification is achieved.
8. The scenario team and panel revise the scenarios based on results of the general review
9. The final scenarios are published ad distributed.

Sorts and locations of uncertainty addressed

Scenario Analysis typically addresses ignorance, value-ladenness of choices (assumptions) and "what-if" questions (scenario uncertainty) with regard to both the context of the (environmental) system considered in the assessment and assumptions about the

environmental processes involved. Furthermore Scenario Analysis addresses ignorance, value-ladenness of choices and scenario uncertainty associated with input data and driving forces used in models.

Required resources

Scenario analysis requires creativity and ability to think outside the scope of the familiar and the present. Further it requires insight in dynamics, relationships, en synergies of systems and their environment and thus it requires a broad knowledge of the field involved. Therefor scenarios analysis should take place in an interdisciplinary team.

In the case of a quantitative approach, computer models or spreadsheets or other software are needed to run/visualise scenarios. Access to relevant data is important in order to be able to construct the scenarios.

In the case of a qualitative approach, input has to be collected from experts, stakeholders or users in workshops with stakeholders to be able to develop storylines. Basic skills for facilitating groups.

Both approaches are time and resource consuming.

Strengths and weaknesses

Strengths of scenario analysis:

- Scenarios are often the only way to deal with the unknown future;
- Assumptions about future developments are made transparent and documented
- Gives insight in key factors that determine future developments;
- Creates awareness on alternative development paths, risks, and opportunities and possibilities for policies or decision-making.

Weaknesses of scenario analysis are:

- The analysis is limited to those aspects of reality that can be quantified (quantitative scenarios)
- Difficult to test underlying assumptions (qualitative scenarios)
- Frequently scenarios do not go beyond trend extrapolation (quantitative scenarios);
- Frequently scenarios are surprise-free;
- Frequently models used contain only one view, which will make the outcomes narrow in scope, thus not doing justice to the wish to explore fundamentally different futures

Guidance on application

- Define very well what the objectives of the scenario analysis are: adjust the scenario development and analysis according to the objectives.
- Make sure the scenarios are transparent and well documented
- Scenario should not be implausible (they should be recognizable and internally consistent)
- Be aware and explicit about the limited scope of a certain model used
- Present as many scenario as possible and as few as necessary: it is important to represent many views and possibilities, however with too many scenarios it will be difficult to communicate results (analyse many, report few)

Pitfalls

Typical pitfalls of scenario analysis are:

- Undue suggestion of objectivity and completeness: a presentation of e.g. 4 different does not mean that there are only 4 possible ways for the future to develop

- Quantitative scenarios might suggest numerical exactness and suggest more certainty about the future than we have ;
- Analysts and users often forget that outcomes are often based on models which themselves contain assumptions on the future already
- Scenarios often reflect more our present expectations and beliefs then future developments, and therefore have a tendency to be rather conservative.
- Scenarios do not forecast what will happen in the future, rather they indicate what might happen under certain well-specified conditions (what-if). Typically scenarios are used in situations where there is lack of information on underlying mechanisms and developments, and therefore it is usually impossible to adhere probabilities to scenarios. Statements on the likelihood of scenarios therefore should be considered with due care.
- Presenting an uneven number of scenarios may lead users to assume that the middle scenario is the most probable scenario

References

Alcamo J. (2001) Scenarios as tools for international environmental assessments. Environmental issues report. Experts corner report. Prospects and Scenarios No.5, European Environment Agency, Copenhagen

P. Jansen Schoonhoven and F.M. Roschar, Werken met scenario's; ook kwalitatieve informatie is te verwerken. *Beleidsanalyse* 1, p. 146-153, 1992.

Van Notten, P. (2002) Foresight in the face of scenario diversity. Paper presented at the international conference Probing the Future: developing organizational foresight in the knowledge economy, 11-13 July 2002 Glasgow.

UNEP (2002) Global Environmental Outlook 3: Past, present and future perspectives, Earthscan.

Handbooks:
Kees Van Der Heijden, *Scenarios: The Art of Strategic Conversation*, John Wiley & Sons; 1996 (ISBN: 0471966398),

Example case studies:
IPCC SRES

Websites:

Experts at RIVM:
Rob Swart, Bert de Vries, Rik Leemans, Jos Olivier, Petra van Egmond, Jan Bakkes, Leon Jansen

PRIMA: A framework for perspective-based uncertainty management

by Marjolein B.A. van Asselt

Description

PRIMA is an acronym for Pluralistic fRamework of Integrated uncertainty Management and risk Analysis. PRIMA is not a tool in the classic sense, but a meta approach (organising framework) to structure the process of uncertainty management (van Asselt, 2000). The guiding principle is that uncertainty legitimates different perspectives on policy issues and that, as a consequence, uncertainty management should explicitly take these different perspectives into account. In doing so, different legitimate interpretations of uncertain values and causal relationships are explored in a systematic manner, which enables to tell a story behind the various outcomes/outlooks.

Goals and use

Central to the PRIMA approach is the determination of the most policy-relevant uncertainties that play a role in controversies surrounding complex issues. Subsequently, in the process of assessing the policy problem these uncertainties are explicitly 'coloured' according to various perspectives. Starting from these perspective-based interpretations, various legitimate and consistent narratives are developed to serve as a basis for perspective-bases model routes (example TARGETS, Rotmans and de Vries) or as a basis for selected model experiments (example with water management, van Asselt, Middelkoop et al. 2001). These model routes are then used for systematic experiments to assess robust strategies and potential risks. It is at least an alternative for ad hoc based scenario experiments with models.

Sorts and locations of uncertainty addressed

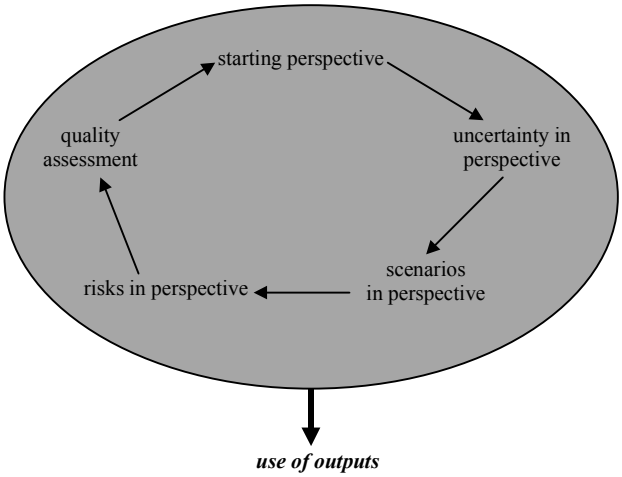
PRIMA is especially suited for uncertainties, which can be interpreted differently from normative standpoints. In practice this usually means that PRIMA is useful for uncertainties of the sort of *scenario uncertainty* and *recognized ignorance*. The method of perspective-based multiple model routes, which is the main PRIMA technique by definition involves both *model, input* and *parameter* uncertainties and to a lesser extent involves the *context*.

Required resources

- The most important requirement is real *interdisciplinary teamwork*; especially learning each other's language and taking disciplinary boundaries takes time, patience and social skills.
- The major investment in PRIMA is spending time on interpreting uncertainties from a particular perspective, first in a qualitative narrative and then in the translation in model terms (quantities, parameter values and mathematical equations). This concerns not just time which must be invested in the actual interpretation exercises (hours–weeks), but also time which is needed to enable the interpretations to stabilize and mature (months at least). Searching the literature for values and equations that can be used to quantify the perspectives in a scientifically sound manner takes time (dependent on the actual expertise and experience of the experts involved). The actual full throughput time for building perspective-based model routes in the TARGETS model (i.e. five submodules with completely different topics) was about one year. The analysis of the perspective-based experiments for the model as a whole took another half year. In the case of the application of PRIMA to water management, in which participatory perspective workshops were held (which is an additional investment), it took one year to arrive at stable qualitative interpretations and another year for quantification and model experiments.
- The idea of perspective-based multiple model routes can be applied in any software environment in which it is possible to program switches. In the TARGETS model, technically speaking a switch parameter "PERSPECTIVE" was used, which had three 'values': hierarchist (1), egalitarian (2) and individualist (3). Dependent on the value of this switch parameter particular parameter value sets or model routes are activated.

Guidance on application

PRIMA is an iterative and cyclical approach that has been set up in a modular manner:



Key characteristics of module:

- **starting perspective**: to what type of uncertainties PRIMA will be applied and which controversy or dilemma being assessed. ***tools***: discussion (also with clients/stakeholders), reflection
- **uncertainties in perspective**: scan and selection of key uncertainties and perspective-based interpretation of these uncertainties. ***tools***: iterative process which can includes brainstorming exercises, ranking techniques, interviews with experts, participatory workshops with role-playing, desk/literature study,:
- **scenarios in perspective**: scanning the future from a wide variety of perspectives; this can be done both in a qualitative and quantitative manner. ***tools:*** participatory techniques for scenario analysis; perspective-based model routes; Monte-Carlo can be used as complementary tool to assess statistical uncertainties in values used for each perspective
- **risk-in-perspective**: assess risks and developing robust strategies, taking into account the variety of perspective-based assessments gathered. ***tools:*** sensible risk comparisons, critical reflection and synthesis; *Note*: most difficult (!) and immature module, for which not many concrete hints can be given.
- **quality assessment**: test the quality of the associated robust insights by reflecting on the previous steps that means evaluating whether the uncertainties have been considered in an adequate manner. ***tools***: quality checklists, pedigree matrices, uncertainty review

Hints: It is not necessary to carry out the full cycle, modules can to a certain extent be used in a stand-alone version.

Strengths and limitations

Typical strengths of PRIMA are:

- The major innovative feature is to see pluralism not as part of the problem, but as part of the solution. From that perspective, PRIMA is the only approach so far that advances and provides structure to the systematic use of multiple values, paradigms, perceptions, judgements, etc in assessment processes.
- The major advantage of PRIMA is that more than one perspective made explicit is, in view of the inherent bias of interpretations of uncertainty better than an at first sight objective model with a hidden perspective.
- The PRIMA approach is by definition a group process approach. An experienced spin-off is the interdisciplinary team-building potential, because the process (facilitated by an experienced facilitator) offers a safe environment to share what is usually regarded as ‘non-knowledge’.

Typical weaknesses, pitfalls and limitations associated with PRIMA are:

- PRIMA is in the first place a framework that aims to structure the process of perspective-based uncertainty management. Although it provides a logic for using existing tools and techniques in a complementary way, and while it proposes a novel approach (i.e. perspective-based model routes), PRIMA is not (and was never meant as) a rigid recipe.
- Any interpretation of the (ideal types of) perspectives is to a certain degree subjective. The only advantage is that the use of stereotypes helps to make interpretations and structures of argumentation at least more explicit.
- In the prototypical versions (TARGETS, water management) the perspectives from Cultural Theory have been used. This theory has received (and still receives) strong criticism. Cultural theory has been criticised for[1] (see, for example, Trisoglio, 1995):
  - providing an oversimplification of reality;
  - being too static in time and place (in reality one can be a hierarchist at work, a fatalist in leisure time, and an egalitarian at home);
  - undue universal claims (whereas, in reality, one can for instance act as a hierarchist when confronted with one problem and as an egalitarian when confronted with another problem) ;
  - not taking account of complex systems of myths of nature.

  However, alternative social scientific theories/perspective frameworks that are both sound and have the same strengths are not available.
- Methods for articulating perspectives of real actors or from behavioural patterns are underresearched and underdeveloped.
- Because of the previous two bullets the key weakness of PRIMA is the perspective framework used.
- As with all scenario exercises, another limitation is that almost no explicit heuristics/approaches/methods have been developed for how to arrive insights from a set of fundamentally different outcomes / futures. That is a general weakness of Integrated Assessment.

Pitfalls

See weaknesses under previous heading.

References

van Asselt, M. B. A., and Rotmans, J. (1996). "Uncertainty in Perspective." *Global Environmental Change*, 6(2), 121-157.

van Asselt, M. B. A., Beusen, A. H. W., and Hilderink, H. B. M. (1996). "Uncertainty in Integrated Assessment: A Social Scientific Approach." *Environmental Modelling and Assessment*, 1(1/2), 71-90.

Rotmans, J., and Vries, B. d. (1997). *Perspectives on global change: the TARGETS approach*, Cambridge University Press, Cambridge.

Rotmans, J., and van Asselt, M. B. A. (1999). "Perspectives on a sustainable future." *International journal for sustainable development*, 2(2), 201-230.

van Asselt, M. B. A. (2000). *Perspectives on Uncertainty and Risk: The PRIMA approach to decision support*, Kluwer Academic Publishers, Dordrecht, The Netherlands.

van Asselt, M. B. A., Langendonck, R., van Asten, F., van der Giessen, A., Janssen, P., Heuberger, P., and Geuskens, I. (2001). "Uncertainty & RIVM's Environmental Outlooks. Documenting a learning process.", ICIS/RIVM, Maastricht/Bilthoven, The Netherlands.

[1] It is important to realize that especially in the United States Cultural Theory is also criticized because of the controversial political role and standpoints of one of the founding fathers of Cultural Theory, i.e. Wildavsky.

van Asselt, M. B. A., Middelkoop, H., van 't Klooster, S. A., van Deursen, W. P. A., Haasnoot, M., Kwadijk, J. C. J., Buiteveld, H., Können, G. P., Rotmans, J., van Gemert, N., and Valkering, P. (2001). "Integrated water management strategies for the Rhine and Meuse basins in a changing environment : Final report of the NRP project 0/958273/01.", ICIS, Maastricht/Utrecht.

van Asselt, M. B. A., and Rotmans, J. (2002). "Uncertainty in Integrated Assessment modelling: From positivism to pluralism." Climatic Change (54), 75-105.

van Asselt, M. B. A., and Rotmans, J. (forthcoming). "Uncertainty management in regional integrated assessment." Chapter for START-CIRA-IHDP book on Regional Integrated Assessment, J. Jäger and G. Knight, eds.

Key references on Cultural Theory:

Thompson, M., Ellis, R., and Wildavsky, A. (1990). *Cultural Theory*, Westview Press, Boulder, USA.

Thompson, M., Grendstad, G., and Selle, P. (1999). *Cultural Theory as Political Science*, Routledge, London.

Rayner, S. (1984). "Disagreeing about Risk: The Institutional Cultures of Risk Management and Planning for Future Generations." Risk Analysis, Institution and Public Policy, S. G. Hadden, ed., Associated Faculty Press, Port Washington, USA.

Rayner, S. (1991). "A Cultural Perspective on the Structure and Implementation of Global Environmental Agreements." *Evaluation Review*, 15(1), 75-102.

Rayner, S. (1992). "Cultural Theory and Risk Analysis." Social Theory of Risk, G. D. Preagor, ed., Westport, USA.

Rayner, S., and Cantor, R. (1987). "How fair is safe enough? The cultural approach to societal technology choice." *Risk Analysis*, 7(1), 3-9.

Rayner, S., and Malone, E. (1996). "Human Choice and Climate Change: An International Social Science Assessment." Battle Press, Colunbus, Ohio, 429.

Douglas, M., and Wildavsky, A. (1982). *Risk and Culture: Essays on the Selection of Technical and Environmental Dangers*, University of California Press, Berkley, USA.

Trisoglio, A., *Plurality and Political Culture, A Response to Thompson and Drake*, paper Presented at workshop "Risk, Policy and Complexity", IIASA, Laxenburg, 7-9 August 1995, 18 pp.

***experts within RIVM**:*
Bert de Vries, Henk Hilderink, Arthur Beusen

National experts:
m.vanasselt@tss.unimaas.nl; s.vantklooster@tss.unimaas.nl; j.rotmans@icis.unimaas.nl; H.Middelkoop@geog.uu.nl; wvandeursen@carthago.nl; Jaap.Kwadijk@wldelft.nl; Marjolein.haasnoot@wldelft.nl

Checklist for model quality assistance

Description

The Checklist for Model Quality Assistance is an instrument for internal use to assist modellers and users of models in the process of quality control.

Goals and use

The goal of the Checklist for Model Quality Assistance is to assist in the quality control process for environmental modelling. The point of the checklist is not that a model can be classified as 'good' or 'bad', but that there are 'better' and 'worse' forms of modelling practice. The idea behind the checklist is that one should guard against poor practice because it is much more likely to produce poor or inappropriate model results. Further, model results are not 'good' or 'bad' in general (it is impossible to 'validate' a model in practice), but are 'more' or 'less' useful when applied to a particular problem. The checklist is thus intended to help guard against poor practice and to focus modelling on the utility of results for a particular problem. That is, it should provide insurance against pitfalls in process and irrelevance in application.

Large, complex environmental models present considerable challenges to develop and test. To cope with this, there has been a lot of effort to characterize the uncertainties associated with the models and their projections. However, uncertainty estimates alone are necessarily incomplete on models of such complexity and provide only partial guidance on the quality of the results. The conventional method to ensure quality in modelling domains is via model validation against observed outcomes. Unfortunately, the data are simply not available to carry out rigorous evaluations of many models (Risbey et al., 1996).

Lack of validation data is critical in the case of complex models spanning human and natural systems because they require: socio-economic data which has frequently not been collected; data related to value dimensions of problems that is hard to define and quantify; data on projections of technical change which must often be guessed at; data on aggregate parameters like energy efficiency which is difficult to measure and collect for all the relevant economies; geophysical data on fine spatial and temporal scales worldwide that is not generally available; data pertinent to non-marginal changes in socio-economic systems which is difficult to collect; and experience and data pertaining to system changes of the kind simulated in the models for which we have no precedent or access.

Without the ability to validate the models directly, other forms of quality assessment must be utilized. Unfortunately, there are few ready-made solutions for this purpose. For complex coupled models there are many pitfalls in the modelling process and some form of rigour is all that remains to yield quality. Thus, a modeller has to be a good craftsperson (Ravetz, 1971; 1999). Discipline is maintained by controlling the introduction of assumptions into the model and maintaining good `practice'. What is needed in this case is a form of heuristic that encourages self-evaluative systematisation and reflexivity on pitfalls. The method of systematisation should not only provide some guide to how the modellers are doing; it should also provide some diagnostic help as to where problems may occur and why.
Risbey *et al.,* (2001) have developed a model quality assistance checklist for this purpose to be used in the project).

The philosophy underlying the checklist is that there is no single metric for assessing model performance and that, for most intents and purposes, there is no such thing as a `correct' model or at least no way to determine whether it is correct. Rather, models need to be assessed in relation to particular functions. Further, that assessment is ultimately about quality -- where quality relates a process/product (in this case a model) to a given function. The point is not that a model can be classified as `good' or `bad', but that there are `better' and `worse' forms of modelling practice, and that models are `more' or `less' useful when applied to a particular problem. The checklist is thus intended to help guard against poor practice and to focus modelling on the utility of results for a particular problem. That is, it should provide some

insurance against pitfalls in process and irrelevance in application. The questions in the checklist are designed to uncover at least some of the more common pitfalls in modelling practice and application of model results in policy contexts. The output from the checklist is both indirect, via reflections from the modeller's self-assessment, and direct in the form of a set of potential pitfalls triggered on the basis of the modeller's responses.

The checklist is structured as follows. First there is a set of questions to probe whether quality assistance is likely to be relevant to the intended application. If quality is not at stake, a checklist such as this one serves little purpose. The next section of the checklist aims to set the context for use of the checklist by describing the model, the problem that it is addressing, and some of the issues at stake in the broader policy setting for this problem. The checklist then addresses `internal' quality issues, which refers to the processes for developing, testing, and running the model practiced within the modelling group. A section on `users' addresses the interface between the modelling group and outside users of the model. This section examines issues such as the match between the production of information from the model and the requirements of the users for that information. A section on `use in policy' addresses issues that arise in translating model results to the broader policy domain, including the incorporation of different stakeholder groups into the discussion of these results. The final section provides an overall assessment of quality issues from use of the checklist.
The automated version of the checklist also contains an algorithm to produce a list of pitfalls based on the answers given.

Sorts and locations of uncertainty addressed

The checklist for model quality assistance addresses all sorts of uncertainties at all locations distinguished in the uncertainty typology. The focus is mainly on unreliability and ignorance and the different sections of the checklist address the different locations where uncertainty may be manifest. The sections on internal strength address inputs and model structure, the sections on external strength address system boundary and socio-political context.

Required resources

It takes between two and four hours to complete the checklist for a given model, depending on the nature and the complexity of the model.
The checklist is freely available from http://www.nusap.net, both as a interactive web tool and as a downloadable pdf file.

Strengths and weaknesses

Typical strengths of the checklist for model quality assistance are:

- It provides diagnostic help as to where problems with regard to quality and uncertainty may occur and why
- It raises awareness of pitfalls in the modelling process;

Typical limitations of the checklist for model quality assistance are:

- It is not a panacea for the problem that models of complex systems cannot be validated.

Guidance on application

- Make sure that the one who completes the checklist has sufficient knowledge on the model and its use.
- It is recommendable that several members of a modelling team complete the checklist independently of one another and then afterwards discuss backgrounds of eventual differences in results.

Pitfalls

Typical pitfalls of the checklist for model quality assistance are:

- The checklist is a tool for self-elicitation. Consequently the pitfalls of expert bias apply. For a full description we refer to the pitfalls listed under the entry "Expert Elicitation" in this toolbox document.
- Note that quality assistance is not the same as quality control: Running the checklist on your model does not warrant quality.
- Be aware that running the checklist does not make model evaluation or validation efforts superfluous.

References

The checklist:
Risbey, J., van der Sluijs, J., Ravetz, J., and P. Janssen 2001: *A checklist for quality assistance in environmental modelling*. Research Report E-2001-11 (ISBN 90-73958-66-0), Department of Science Technology and Society, Utrecht University (http://www.nusap.net/download.php?op=viewdownload&cid=1)

An interactive fully automated version of the checklist is available in the interactive tools section of the nusap.net: http://www.nusap.net/sections.php?op=viewarticle&artid=15

Underlying theory:
Ravetz, J. 1971: *Scientific knowledge and its social problems*. Oxford. Clarendon Press.

Risbey, J., M. Kandlikar, and A. Patwardhan 1996: Assessing Integrated Assessments. *Climatic Change* 34, 369-395.

Application:
James Risbey, Jeroen van der Sluijs, Penny Kloprogge, Jerry Ravetz, Silvio Funtowicz, and Serafin Corral Quintana, Application of a Checklist for Quality Assistance in Environmental Modelling to an Energy Model, *Environmental Modeling and Assessment,(forthcoming, accepted for publication)*

Experts:
RIVM: Peter Janssen, Arthur Petersen, Detlef van Vuuren.
National: Jeroen van der Sluijs, Penny Kloprogge
International: James Risbey, Jerry Ravetz, Silvio Funtowicz, Bruna De Marchi

A method for critical review of assumptions in model-based assessments

Description

This method enables to systematically identify and prioritize critical assumptions in (chains of linked) models and provides a framework for the critical appraisal of model assumptions.

Goals and use

This method aims to systematically identify, prioritise and analyse importance and strength of assumptions in models, such as those used for the quantification of Environmental Indicators under various scenarios (such as in the Netherlands Environmental Outlook). These indicators are typically based on chains of soft-linked computer model calculations that start with scenarios for population and economic growth. The models in the chain vary in complexity. Often, these calculation chains behind indicators involve many analysts from several disciplines. Many assumptions have to be made in combining research results in these calculation chains, especially since the output of one computer model often does not fit the requirements of input for the next model (scales, aggregation levels). Assumptions are also frequently applied to simplify parts of the calculations. Assumptions can be made explicitly or implicitly.
Assumptions can to some degree be value laden. This method distinguishes 4 types of value-ladenness of assumptions: value-ladenness in a socio-political sense (e.g., assumptions may be coloured by political preferences of the analyst), in a disciplinary sense (e.g., assumptions are coloured by the discipline in which the analyst was educated), in an epistemic sense (e.g., assumptions are coloured by the approach that the analyst prefers) and in a practical sense (e.g., the analyst is forced to make simplifying assumptions due to time constraints).
The method can be applied by the analysts carrying out the environmental assessment. However, each analyst has limited knowledge and perspectives with regard to the assessment topic, and in consequence will have some 'blind spots'. Therefore preferably other analysts (peers) are involved in the method as well. Stakeholders, with their specific views and knowledge, can be involved as well. This can, for instance, be organised in the form of a workshop. The group of persons involved in the assumption analysis will be referred to as 'the participants'

The method involves 7 steps:
ANALYSIS
1. Identify explicit and implicit assumptions in the calculation chain
2. Identify and prioritise key-assumptions in the chain
3. Assess the potential value-ladenness of the key-assumptions
4. Identify 'weak' links in the calculation chain
5. Further analyse potential value-ladenness of the key-assumptions

REVISION
6. Revise/extend assessment
   - sensitivity analysis key assumptions
   - diversification of assumptions
   - different choices in chain

COMMUNICATION
7. Communication
   - key-assumptions
   - alternatives and underpinning of choices regarding assumptions made
   - influence of key-assumptions on results
   - implications in terms of robustness of results

All steps will be elaborated on below.

Step 1 - Identify explicit and implicit assumptions in the calculation chain
In the first step implicit and explicit assumptions in the calculation chain are identified by the analyst by systematic mapping and deconstruction of the calculation chain, based on document

analysis, interviews and critical review. The resulting list of assumptions is then reviewed and completed in a workshop.
The aggregation level of the assumptions on the assumption list may vary. An assumption can refer to a specific detail in the chain ("The assumption that factor x remains constant'), as well as refer to a cluster of assumptions on a part of the chain ("Assumptions regarding sub-model x").

Step 2 - Identify and prioritise key-assumptions in the chain
In step 2 the participants identify the key-assumptions in the chain. The assumptions identified in step 1 are prioritised by taking into account the influence of the assumptions on the end results of the assessment. Ideally, this selection is based on a quantitative sensitivity analysis. Since such an analysis will often not be attainable, the participants are asked to estimate the influence of the assumptions on outcomes of interest of the assessment. An expert elicitation technique can be used in which the experts bring forward their opinions and argumentation on whether an assumption is of high or low influence on the outcome. Based on the discussion the participants then can indicate their personal estimate regarding the magnitude of the influence, informed by the group discussion. A group ranking is established by aggregating the individual scores.

Step 3 - Assess the potential value-ladenness of assumptions
To assess potential value ladenness of assumptions, a 'pedigree matrix' is used that contains criteria by which the potential value-ladenness of assumptions can be reviewed. The pedigree matrix is presented in Table 1 and will be discussed in detail later on.
For each key-assumption all pedigree criteria are scored by the participants. Here, again a group discussion takes place first, in order for the participants to remedy each other's blind spots and exchange arguments.

The order in which the key-assumptions are discussed in the workshop is determined by the group ranking established in step 2 of the method, starting with the assumption with the highest rank.

Step 4 - Identify 'weak' links in the calculation chain
The pedigree matrix is designed such that assumptions that score low on the pedigree criteria have a high potential for value-ladenness. Assumptions that, besides a low score on the criteria, also have a high estimated influence on the results of the assessment can be viewed as problematic weak links in the calculation chain.

Step 5 - Further analyse potential value-ladenness key-assumptions
In step 5, the nature of the potential value-ladenness of the individual key-assumptions is explored. Based on inspection of the diagrams visualizing the pedigree scores (or the table of pedigree scores), it can be analysed:
- what types of value-ladenness possibly play a role and to what extent
- to what extent there is disagreement on the pedigree scores among the participants
- whether changing assumptions is feasible and desirable

Step 6 - Revise/extend assessment
Based on the analysis in step 5, it can be decided to change or broaden the assessment. As a minimum option, the assessment can be extended with a sensitivity analysis, which gives more information on the influence of weak links in the assessment.
Besides a sensitivity analysis, specific assumptions can be revised or diversified. In the case of revising an assumption, the assumption is replaced by a different assumption. In some cases however, it will be difficult or undesirable to choose between alternative assumptions, since there might be differing views on the issue. If these assumptions have a high influence on the assessment as a whole, it can be decided to diversify the assumptions: the calculation chain is 'calculated' using several alternative assumptions in addition to the existing ones. In this way several assessments are formed, with differing outcomes, depending on what assumptions are chosen.

Step 7 - Communication

It is important to be explicit about potential value-ladenness in the chain and the effects of potentially value-laden assumptions on the outcomes of the assessment. Analogous to a patient information leaflet, the presentation of the assessment results should be accompanied by information on:

- what are the key-assumptions in the calculation chain
- what are the weak links in the chain
- what were the alternatives and what is the underpinning of the choices that were made regarding assumptions
- what is the robustness of the outcomes of interest in view of the key assumptions

The pedigree matrix for assessing the potential value-ladenness of assumptions is presented in Table 1. For a general introduction to the concept of pedigree matrix, we refer to the description of the NUSAP system in this tool catalogue. The criteria are discussed below.

Type of value-ladenness	**Score→ Criteria↓**	4	3	2	1	0
Practical	**Influence situational limitations**	choice assumption hardly influenced		Choice assumption moderately influenced		totally different assumption had there not been limitations
Epistemic	**Plausibility**	the assumption is plausible		the assumption is acceptable		the assumptions is fictive or speculative
Epistemic	**Choice space**	hardly any alternative assumptions available		limited choice from alternative assumptions		ample choice from alternative assumptions
Disciplinary, epistemic	**Agreement among peers**	many would have made the same assumption		Several would have made the same assumption		few would have made the same assumption
Socio-political	**Agreement among stakeholders**	many would have made the same assumption		Several would have made the same assumption		few would have made the same assumption
Socio-political	**Sensitivity to view and interests of the analyst**	choice assumption hardly sensitive		Choice assumption moderately sensitive		Choice assumption sensitive
	Influence on outcomes of interest	the assumption has little influence on the outcome of interest		the assumption has a substantial influence on an intermediate variable but/or has moderate influence on the outcome of interest		the assumption has a large influence on the outcome of interest

Table 1: Pedigree matrix for the assessment of the potential value-ladenness of assumptions

Influence of situational limitations

The choice for an assumption can be influenced by situational limitations, such as limited availability of data, money, time, software, tools, hardware, and human resources. In absence of these restrictions, the analyst would have made a different assumption.
Although indirectly these limitations might be of a socio-political nature (e.g., the institute the analyst works for has other priorities and has a limited budget for the analyst's work), from the analyst's point of view these limitations are given. It can therefore be seen as primarily producing value-ladenness in a practical sense.

Plausibility
Although it is often not possible to assess whether the approximation created by the assumption is in accordance with reality, mostly an (intuitive) assessment can be made of the plausibility of the assumption.
If an analyst has to revert to fictive or speculative assumptions, the room for epistemic value-ladenness will often be larger. To some extent a fictive or speculative assumption also leaves room for potential disciplinary and socio-political value-ladenness. This is, however, dealt with primarily in the criteria 'agreement among peers' and 'agreement among stakeholders' respectively.

Choice space
The choice space indicates to which degree alternatives were available to choose from when making the assumption. In general, it can be said that a large choice space leaves more room for the epistemic preferences of the analyst. Often, the potential for value-ladenness in an epistemic sense is larger in case of a larger choice space. A large choice space will to some extent also leave more room for disciplinary and socio-political value-ladenness. These are however primarily dealt with in the criteria 'agreement among peers' and 'agreement among stakeholders' respectively.

Agreement among peers
An analyst makes the choice for a certain assumption based on his or her knowledge and perspectives regarding the issue. Other analysts might have made different assumptions. The degree to which the choice of peers is likely to coincide with the analyst's choice is expressed in the criterion 'agreement among peers'. These choices may be partly determined by the disciplinary training of the peers, and by their epistemic preferences. This criterion can thus be seen connected to value-ladenness in a disciplinary sense and in a epistemic sense.[1]

Agreement among stakeholders
Stakeholders, though mostly not actively involved in carrying out assessments, might also choose a different assumption in case they were asked to give their view. The degree to which it is likely that stakeholders agree with the analyst's choice is expressed in the criterion 'intersubjectivity among stakeholders'. This will often have to do with the socio-political perspective of the stakeholders on the issue at hand and this criterion can therefore be seen as referring to value-ladenness in a socio-political sense.

Sensitivity to view and interests of the analyst
Some assumptions may be influenced, consciously or unconsciously, by the view and interests of the analyst making the assumption. The analyst's epistemic preferences, and his cultural, disciplinary and personal background may influence the assumption that is eventually chosen. The influence of the analyst's disciplinary background on the choices regarding an assumption and the influence of his epistemic preferences are taken into account in the criteria 'agreement

[1] There is a link to controversy, as not all peers would agree to the same assumption if there was controversy regarding the issue of the assumption. However, if the majority of peers would choose the same assumption, still the score would be 2 ('many peers would have made the same assumption'). The occurrence of controversies in the scientific field thus is not always visible in the score. Reasoned the other way around, a score of 0 ('few peers would have made the same assumption') does not imply that there are controversies surrounding the assumption: it is possible that all peers agree on the issue, yet that the analyst for some reason has chosen a different assumption. The same applies to the criterion 'agreement among stakeholders'.

among peers', 'plausibility' and 'choice space'. In this criterion the focus is on the room for value-ladenness in a socio-political sense.

Influence on results
In order to be able to pinpoint important value-laden assumptions in the calculation chain it is not only important to analyse the potential value-ladenness of the assumptions, but also to assess the influence on the end result of the assessment. Ideally, a sensitivity analysis is carried out to assess the influence of each of the assumptions on the results. In most cases, however, this will not be attainable because it requires the building of new models. This is why the pedigree matrix includes a column 'influence on results'.

The modes for each criterion are arranged in such a way that the lower the score, the more value-laden the assumption potentially is.

When all participants have scored the assumptions on the criteria, the scores can be presented in a table. In order to facilitate a quick overview of the results, diagrams can be used that aggregate the scores of the individual experts without averaging them, and in such a way that expert disagreement on the scores is visualised.

One diagram is made for each assumption. The diagram is divided into 6 triangular segments, each segment representing one criterion (fig 1). The scale in each segment is such that zero is in the center of the diagram and two on the border. For each criterion, the area of the corresponding segment from the center of the diagram up to the minimum score given in the group is colored green. If there is no consensus on the score for a given criterion, the area in each segment spanned up between the minimum and the maximum score in the group for that criterion is colored amber. The remaining area (from the maximum score to the outside border of the diagram) -if any- is colored red.

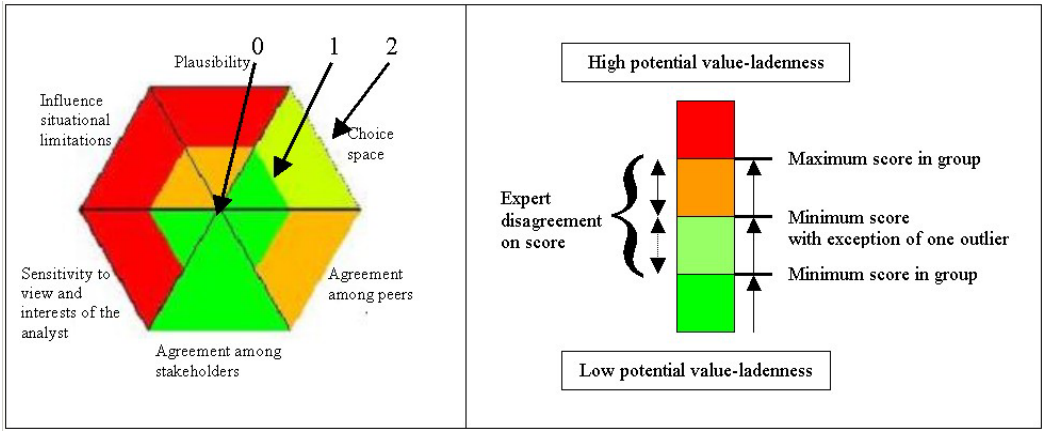


Figure 1 Left: an example diagram. Right: explanation of the colours. Note that this example used a 3 point scale. Based on our experiences we recommend to use a 5 point scale as in table 1 (Kloprogge et al. 2004).

The convention follows a traffic-light analogy and is such that would an assumption on all criteria score 0 unanimously, the entire diagram will be red. If scores are better, more and more green comes into the diagram, whereas expert disagreement on scores is reflected in amber. On the other extreme, if an assumption scores 2 unanimously for all criteria, the entire diagram will be green. The scores for each criterion are such that in all cases more green in the diagram corresponds to lower potential value-ladenness and more red to higher potential value-ladenness.
A further nuance has been made to account for outliers: in some cases a single outlier score in the group distorts the green area in the diagram. In these cases, a light-green area indicates what the green area would look like if that outlier were omitted.

By looking at the red areas, the extent to which the different types of value-ladenness may have played a role in the production process of the assumption can be assessed. Green areas indicate that the participants think value-ladenness with regard to the criteria at hand played a small role in the production process, red areas that they think value-ladenness played a large role. In case of orange areas it can be concluded that there is disagreement among the participants on these matters.

Sorts and locations of uncertainty addressed

The method presented here for critical review of assumptions in assessments typically addresses value-ladenness of choices. The locations that are addressed by the method basically include all locations that contain implicit or explicit assumptions.

Required resources

- The time required for this method is variable. Firstly, it depends on the number of calculation chains in the assessment that are analysed and on the complexity of the models in these chains. Secondly, the method can be applied by the analysts carrying out the assessment alone or can be applied by the analysts, peers and stakeholders.
- For the workshop, basic facilitator skills are needed.

Strengths and weaknesses

Typical strengths of the method are:

- The method enables a well-structured discussion on potentially value-laden assumptions among scientists and stakeholders. In this discussion not only the politically controversial assumptions are addressed (as is often the case when assessment results are discussed in public), but also other assumptions that turn out to be important for assessment results.
- The method acknowledges that also pragmatic factors may play a role in the colouring of assumptions

Typical weaknesses of the method are:

- The results may be sensitive to the composition of the group of participants (both the number of persons and the persons' backgrounds).
- The results may be sensitive to procedure details as determined by the group facilitator.
- The method does not offer a clear answer to how to deal with extensive disagreement on the pedigree scores of assumptions.

Guidance on application

The method can both be applied while the environmental assessment is being carried out, and *ex post*. Application during the assessment is preferable, since an iterative treatment of assumptions can improve the environmental assessment.

Pitfalls

- Potential value-ladenness should not be confused with actual value-ladenness. Assessing the actual value-ladenness of an assumption is impossible, since it would require exact and detailed knowledge on what factors contributed to what extent to the analyst's choices.
- The scores 0 and 2 in the pedigree matrix should not be seen as extremes or ideal types. If the 0 and 2 score are interpreted as two extremes, the tendency will be to only use the '1' score. This is an issue that the facilitator should address.
- It is the facilitator's job to make sure that the discussions among the participants do not slide off to a quick group consensus, but that there is an open discussion promoting critical review.

References

The method:

Kloprogge, P., J.P. van der Sluijs, A. Petersen, 2004, *A method for critical review of potentially value-laden assumptions in environmental assessments*. Utrecht University, Department of Science, Technology and Society.

Experts:
RIVM: Arthur Petersen, Peter Janssen
National: Penny Kloprogge, Jeroen van der Sluijs, Nelleke Honingh
International: Matthieu Craye