Toward a Synthesis of Qualitative and Quantitative Uncertainty Assessment: Applications of the Numeral, Unit, Spread, Assessment, Pedigree (NUSAP) System

***Jeroen van der Sluijs*[a], *Penny Kloprogge*[a], *James Risbey*[b], *and Jerry Ravetz*[c]**

[a] Copernicus Institute for Sustainable Development and Innovation, Department of Science Technology and Society, Utrecht University, The Netherlands (j.p.vandersluijs@chem.uu.nl).

[b] School of Mathematical Sciences, Monash University, Clayton, Australia

[c] Research Method Consultancy (RMC), London

Abstract: A novel approach to uncertainty assessment, known as the NUSAP method (Numeral Unit Spread Assessment Pedigree) has been applied to assess qualitative and quantitative uncertainties in three case studies with increasing complexity: (1) the monitoring of VOC emissions from paint in the Netherlands, (2) the TIMER energy model, and (3) two environmental indicators from the Netherlands 5th Environmental Outlook. The VOC monitoring involves a simple calculation scheme with 14 parameters. The TIMER model is a complex non-linear dynamic system model, which consists of over 300 parameters. The indicators in the Environmental Outlook result from calculations with a whole chain of soft-linked model calculations, involving both simple and complex models. We show that the NUSAP method is applicable not only to simple but also to complex models in a meaningful way and that it is useful to assess not only parameter uncertainty but also (model) assumptions. The method provides a means to prioritize uncertainties and focus research efforts on the potentially most problematic parameters and assumptions, identifying at the same time specific weaknesses in the knowledge base. With NUSAP, nuances of meaning about quantities can be conveyed concisely and clearly, to a degree that is quite impossible with statistic methods only.

Keywords: uncertainty; pedigree; NUSAP; quality; environmental assessment; assumption ladenness

Introduction

In the field of environmental modeling and assessment, uncertainty studies have mainly involved quantitative uncertainty analysis of parameter uncertainty. These quantitative techniques provide only a partial insight into what is a very complex mass of uncertainties. In a number of projects, we have implemented and demonstrated a novel, more comprehensive approach to uncertainty assessment, known as the NUSAP method (acronym for Numeral Unit Spread Assessment Pedigree). This paper presents and discusses some of our experiences with the application of the NUSAP method, using three case studies with increasing complexity.

NUSAP and the Diagnostic Diagram

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. 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.

NUSAP acts as a heuristic for good practice in science for policy by promoting reflection on the various dimensions of uncertainty and making these explicit. It provides a diagnostic tool for assessing the robustness of a given knowledge base for policymaking and promotes criticism by clients and users of all sorts—expert and lay—and will thereby support extended peer review processes.

NUSAP yields insights on two independent properties related to uncertainty in numbers, namely spread and strength. Spread expresses inexactness, whereas strength expresses the quality of the underlying knowledge base, in view of its methodological and epistemological limitations. The two metrics can be combined in a diagnostic diagram mapping strength and sensitivity of model parameters. 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 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. Uncertainty then is a property of the modeled system and does not stem from imperfect knowledge on that system. Mapping model parameters in a diagnostic 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.

Case I: A Simple Model

Emissions of VOCs (Volatile Organic Compounds) from paint in the Netherlands are monitored in the framework of VOC emission reduction policies. The annual emission figure is calculated from a number of inputs: national sales statistics of paint for five different sectors, drafted by an umbrella organization of paint producers; paint import statistics from Statistics Netherlands (lump sum for all imported paint, not differentiated to different paint types); an assumption on the average VOC percentage in imported paint; an assumption on how imported paint is distributed over the five sectors; and expert guesses for paint-related thinner use during application of the paint.

We developed and used a NUSAP-based protocol for the assessment of uncertainty and strength in emission data (Risbey *et al.*, 2001), which builds *inter alia* on the Stanford Protocol (Spetzler and von Holstein, 1975) for expert elicitation of probability density functions to represent quantifiable uncertainty and extends it with a procedure to review and elicit parameter strength, using a pedigree matrix. The expert elicitation systematically makes explicit and utilizes unwritten insights in the heads of experts on the uncertainty in emission data, focusing on limitations, strengths, and weaknesses of the available knowledge base.

Pedigree 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. The pedigree criteria used in this case are proxy, empirical basis, methodological rigor, and validation. 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. Table 1 presents the pedigree matrix we used in this case study.

Code	Proxy	Empirical	Method	*Validation*
4	Exact measure	Large sample direct measurements	Best available practice	Compared with independent measurements of same variable
3	Good fit or measure	Small sample direct measurements	Reliable method, commonly accepted	Compared with independent measurements of closely related variables
2	Well correlated	Modeled/ derived data	Acceptable method, limited consensus on reliability	Compared with measurements not independent
1	Weak correlation	Educated guesses/ rule-of-thumb estimate	Preliminary methods, unknown reliability	Weak/indirect validation
0	Not clearly related	Crude speculation	No discernible rigor	No validation

Table 1. Pedigree matrix for emission monitoring. Note that the columns are independent.

The expert elicitation interviews start with an introduction of the task of encoding uncertainty and a discussion of pitfalls and biases associated with expert elicitation (such as motivational bias overconfidence, representativeness, anchoring, bounded rationality, lamp-posting, and implicit assumptions).

	Proxy	*Empirical*	*Method*	*Validation*	*Strength**
NS-SHI	3	3.5	4	0	0.7
NS-B&S	3	3.5	4	0	0.7
NS-DIY	2.5	3.5	4	3	0.8
NS-CAR	3	3.5	4	3	0.8
NS-IND	3	3.5	4	0.5	0.7
Th%-SHI	2	1	2	0	0.3
Th%-B&S	2	1	2	0	0.3
Th%-DIY	1	1	2	0	0.25
Th%-CAR	2	1	2	0	0.3
Th%-IND	2	1	2	0	0.3
Imported paint	3	4	4	2	0.8
VOC % imp.	1	2	1.5	0	0.3

Table 2. Pedigree scores for input parameters.
*****The *Strength* column averages and normalizes the scores on a scale from 0 to 1.

Note: NS=National Sales, Th%=Thinner use during application of paint (SHI, B&S, DIY, CAR, and IND refer to each of the five sectors.)

Next, the expert is asked to indicate strengths and weaknesses in the knowledge base available for each parameter. This starts with an open discussion and then moves to the pedigree criteria that are discussed one by one for each parameter, ending with a score for each criterion (Table 2).

The protocol is designed to stimulate creative thinking on conceivable sources of error and bias. We identified 5 disputable basic assumptions in the monitoring calculation, and 15 sources of error and 4 conceivable sources of motivational bias in the data production.

In a next step in the interview, the expert is asked to quantify the uncertainty in each parameter as a PDF using a simplified version of the Stanford protocol (see *Risbey et al.*, 2001 for details). We used the PDFs elicited as input for a Monte Carlo analysis to assess propagation of parameter uncertainty and the relative contribution of uncertainty in each parameter to the overall uncertainty in VOC emission from paint. We found that a range of ±15% around the average for total 1998 VOC emission from paint (52 ktonne) captures 95% of the calculated distribution.

We further analyzed the uncertainty using a NUSAP diagnostic diagram (Fig. 1) to combine results from the sensitivity analysis (relative contribution to variance, Y-axis) and pedigree (strength, X-axis). Note that the strength axis is inverted, left-hand corresponds to a strong and right-hand to a weak knowledge base.

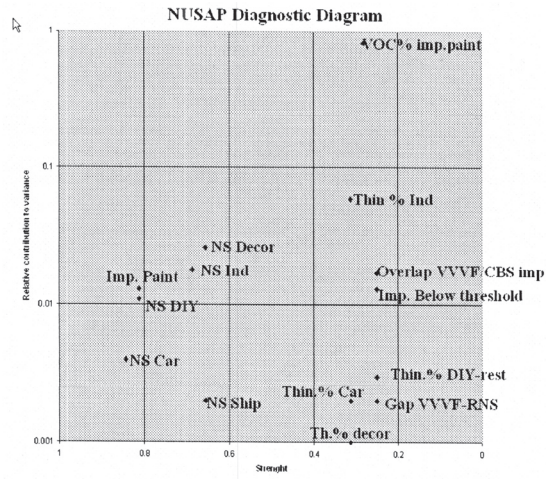

Figure 1 ***Diagnostic diagram for VOC from paint***

The diagnostic diagram identified uncertainty regarding the assumed VOC percentage of imported paint as the most problematic. Other input quantities in the VOC monitoring calculations whose uncertainty was diagnosed to be "important" are assumed percentage of additional thinner use for paint applied in industry, the overlap between the paint import statistics and the national paint sales statistics, and import in volumes below the import statistics reporting threshold. The case is documented in detail in Van der Sluijs *et al*. (2002a).

Case II: A Complex Model

The TIMER (Targets IMage Energy Regional model) model is part of RIVM's Integrated Model to Assess the Global Environment (IMAGE). TIMER is an energy model that, amongst others, was used in the development of the 2001 greenhouse gas emission scenarios from the Inter-Governmental Panel on Climate Change (IPCC). We used the so-called B1 scenario produced with IMAGE/TIMER for the IPCC Special Report on Emissions Scenarios as case study.

Using the Morris (1991) method for global sensitivity analysis we explored quantitative uncertainty in parameters in terms of their relative importance in influencing model results. TIMER is a non-linear model containing a large number of input variables. The Morris method is a sophisticated algorithm 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. TIMER contains 300 variables. Parameters were varied over a range from 0.5 to 1.5 times the default values. The method and full results are documented in Van der Sluijs *et al.* (2002b).

The analysis clearly differentiated between sensitive and less sensitive model components. Also, sensitivity to uncertainty in a large number of parameters turned out to be contingent on the particular combinations of samplings for other parameters, reflecting the non-linear nature of several parts of the TIMER model. The following input variables and model components were identified as most sensitive with regard to model output (projected CO_2 emissions):

- Population levels and economic activity;
- Variables related to the formulation of intra-sectoral structural change of an economy;
- Progress ratios to simulate technological improvements, used throughout the model;
- Variables related to resources of fossil fuels (size and cost supply curves);
- Variables related to autonomous and price-induced energy efficiency improvement;
- Variables related to initial costs and depletion of renewables;

We assessed parameter pedigree by means of a NUSAP expert elicitation workshop. 19 experts on the fields of energy economy and energy systems analysis and uncertainty assessment attended the workshop. We limited the elicitation to those parameters identified either as sensitive by the Morris analysis or as a "key uncertain parameter" in an interview with one of the modelers. Our selection of variables to address in the NUSAP workshop counted 39 parameters. To further simplify the task of reviewing parameter pedigree, we grouped together similar parameters for which pedigree scores might be to some extent similar. This resulted in 18 clusters of parameters. For each cluster a pedigree-scoring card was made, providing definitions and elaborations on the parameters and associated concepts, and a scoring part to fill out the pedigree scores for each parameter. We used the same criteria and pedigree matrix as in the VOC case (table 1), but added a fifth criterion: *theoretical understanding*. This is because the theoretical understanding of the dynamics of the energy system is in its early stage of development. The modes for this pedigree criterion are: Well-established theory (4); accepted theory partial in nature (3); partial theory limited consensus on reliability (2); preliminary theory (1); and crude speculation (0).

For the expert elicitation session, we divided the participants into three parallel groups. Each participant received a set with all 18 cards. Assessment of parameter strength was done by discussing each of the parameters (one card at a time) in a moderated group discussion addressing strengths and weaknesses in the underpinning of each parameter, focusing on, but not restricted to, the five pedigree criteria. Further, we asked participants to provide a characterization of value-ladenness. A parameter is said to be value-laden when its estimate is influenced by ones preferences, perspectives, optimism, or pessimism or co-determined by political or strategic considerations. Participants were asked to draft their pedigree assessment as an *individual* expert judgment, informed by the group discussion.

We used radar diagrams, and kite diagrams (Risbey *et al.*, 2001) to graphically represent results (Fig. 2). 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 line connecting the scores represents the scoring of each expert. 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. A kite diagram captures the information from all experts in the group without the need to average expert opinion.

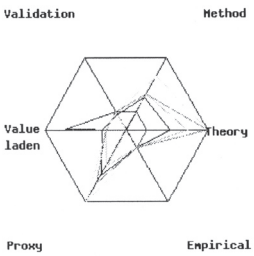

Figure 2a. Example of radar diagram of the gas depletion multiplier assessed by six experts.

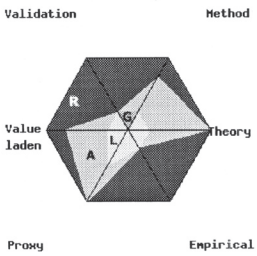

Figure 2b. same, but represented as kite diagram. G=green, L=light green, A=amber, R=red

Results from the sensitivity analysis and strength assessments were combined in Figure 3 to produce a diagnostic diagram.

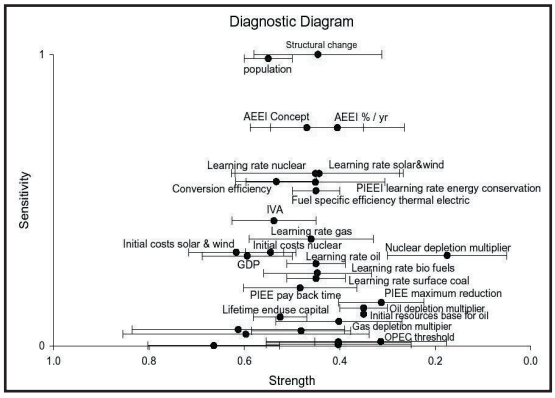

Figure 3. Diagnostic diagram for key uncertainties in TIMER model parameters.

The diagram shows each of the reviewed parameters plotted. The sensitivity axis measures (normalized) importance of quantitative parameter uncertainty. The strength axis displays the normalized average pedigree scores. Error bars indicate one standard deviation about the average expert value, to reflect expert disagreement on pedigree scores. The strength axis has 1 at the origin and zero on the right. In this way, the more "dangerous" variables are in the top right quadrant of the plot (high sensitivity, low strength).

We identified three parameters as being close to the danger zone: Structural change, B1 population scenario, and Autonomous Energy Efficiency Improvement (AEEI). These variables have a large bearing on the CO_2 emission result, but have only weak to moderate strength as judged from the pedigree exercise.

When variables are particularly low in strength, the theory, data, and method underlying their representation may be weak and we can then expect that they are less perfectly represented in the model. With such high uncertainty on their representation, it cannot be excluded that a better representation would give rise to a higher sensitivity. An example of such a variable could be the nuclear depletion multiplier, which has a strength from almost none to weak and a moderate sensitivity contribution.

Case III: Chains of Models

As input for the Netherlands Environmental Policy Plan, the Netherlands Environmental Assessment Agency (EAA/RIVM) prepares every 4 years an assessment of key environmental indicators outlining different future scenarios for a time period of 30 years: the National Environmental Outlook (EO). It presents hundreds of indicators reflecting the pressure on or state of the Dutch, European, or global environment. Model calculations play a key role in the assessments. In a "model chain" of soft-linked computer models—varying in complexity—effects regarding climate, nature, and biodiversity, health and safety, and the living environment are calculated for different scenarios. The total of model and other calculations and operations can be seen as a "calculation chain." Often, these chains behind indicators involve many analysts from several departments within the RIVM. 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).

We developed a NUSAP-based method to systematically identify, prioritize and analyze importance and strength of assumptions in these model chains including potential value-ladenness. We demonstrated and tested the method on two EO5 indicators: "change in length of the growth season" and "deaths and emergency hospital admittances due to tropospheric ozone."

We identified implicit and explicit assumptions in the calculation chain by systematic mapping and deconstruction of the calculation chain, based on document analysis, interviews and critical review. The resulting list of key assumptions was reviewed and completed in a workshop. Ideally, importance of assumptions should be assessed based on a sensitivity analysis. However, a full sensitivity analysis was not attainable because varying assumptions is much more complicated than, for instance, changing a parameter value over a range; it often requires construction of a new model. Instead, we used the expert elicitation workshop not only to review pedigree of assumptions but also to estimate their quantitative importance.

Score	2	1	0
Plausibility	plausible	acceptable	fictive or speculative
Inter-subjectivity peers	many would make same assumption	several would make same assumption	few would make same assumption
Inter-subjectivity stakeholders	many would make same assumption	several would make same assumption	few would make same assumption
Choice space	hardly any alternative assumptions available	limited choice from alternative assumptions	ample choice from alternative assumptions
Influence situational limitations (time, money, etc.)	choice assumption hardly influenced	choice assumption moderately influenced	totally different assumption when no limitations
Sensitivity to view and interests of the analyst	choice assumption hardly sensitive	choice assumption moderately sensitive	choice assumption sensitive
Influence on results	only local influence	greatly determines the results of link in chain	greatly determines the results of the indicator

Table 3. Pedigree matrix for reviewing the knowledge base of assumptions

Table 3 presents the pedigree matrix used in this study. In the workshop, the experts indicated on scoring cards (one card for each assumption) how they judge the assumption on the pedigree criteria and how much influence they think the assumption has on results. An essential part of our method

is that a moderated group-discussion takes place in which arguments for high or low scores per criterion are exchanged and discussed. In this way experts in the group remedy each other's blind spots, which enriches the quality of the individual expert judgments. We deliberately did not ask a consensus judgment of the group, because we consider expert disagreement a relevant dimension of uncertainty.

Assumptions that have a low score on both influence on the results and on the pedigree criteria can be qualified as "weak links" in the chain of which the user of the assessment results needs to be particularly aware.

Analysis of the calculation chain of the indicator "change in length of the growth season" yielded a list of 23 assumptions. The workshop participants selected seven assumptions as being the most important ones. These were reviewed using the pedigree matrix and prioritized according to estimated influence. Combining the results, the weakest links (high influence, low strength) in the calculation chain turned out to be the choice for a GCM (General Circulation Model, projecting time series of geographic patterns of temperature change as a function of greenhouse forcing) and the assumption that the scenarios used for economic development were suitable for the EO5 analyses for the Netherlands and that the choice for the range in global greenhouse gas emission scenarios used was suitable for the global analysis.

Analysis of the calculation chain of the indicator 'deaths and hospital admittances due to exposure to ozone' yielded a list of 24 assumptions. 14 key-assumptions were selected by the workshop participants as the most important ones, and prioritized. Combining the results of pedigree analysis and estimated influence, the following assumptions showed up as the weakest links of the calculation chain: Assumption that uncertainty in the indicator is only determined by the uncertainty in the Relative Risk (RR is the probability of developing a disease in an exposed group relative to those of a non-exposed group as a function of ozone exposure) and the assumption that the global background concentration of ozone is constant over the 30 year time horizon. The full EO5 case and method for the review of assumptions is documented in Kloprogge *et al.* (2003).

Conclusion

We have implemented and demonstrated the NUSAP method to assess qualitative and quantitative uncertainties in three case studies with increasing complexity: a simple model, a complex model, and environmental indicators stemming from calculations with a chain of models.

The cases have shown that the NUSAP method is applicable not only to simple but also to complex models in a meaningful way and that it is useful to assess not only parameter uncertainty but also (model) assumptions. A diagnostic diagram synthesizes results of quantitative analysis of parameter sensitivity and qualitative review (pedigree analysis) of parameter strength. It provides a useful means to prioritize uncertainties according to quantitative and qualitative insights.

The task of quality control in complex models is a complicated one and the NUSAP method disciplines and supports this process by facilitating and structuring a creative process and in depth review of qualitative and quantitative dimensions of uncertainty. It helps to focus research efforts on the potentially most problematic parameters and assumptions, identifying at the same time specific weaknesses in the knowledge base.

Similar to a patient information leaflet alerting the patient to risks and unsuitable uses of a medicine, NUSAP enables the delivery of policy-relevant quantitative information together with the essential warnings on its limitations and pitfalls. It thereby promotes the responsible and effective use of the information in policy processes. With NUSAP, nuances of meaning about quantities can be conveyed concisely and clearly, to a degree that is quite impossible with statistic methods only.

References

Funtowicz, S.O., and J.R. Ravetz, *Uncertainty and Quality in Science for Policy*. Kluwer, 229 pp., Dordrecht, 1990.

Kloprogge, P, J.P. van der Sluijs, and A. Petersen, *A method for the analysis of assumptions in assessments applied to two indicators in the fifth Dutch Environmental Outlook*, Department of Science Technology and Society, Utrecht University, 2004.

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

Risbey, J.S., J.P. van der Sluijs and J. Ravetz, *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, Utrecht, 2001 (www.nusap.net).

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

Van der Sluijs, J.P., J. Risbey, and J. Ravetz, *Uncertainty Assessment of VOC emissions from Paint in the Netherlands*, Department of Science Technology and Society, Utrecht University, 2002a, 90 pp (www.nusap.net).

Van der Sluijs, J.P., J. Potting, J. Risbey, D. van Vuuren, B. de Vries, A. Beusen, P. Heuberger, S. Corral Quintana, S. Funtowicz, P. Kloprogge, D. Nuijten, A. Petersen, J. Ravetz., *Uncertainty assessment of the IMAGE/TIMER B1 CO_2 emissions scenario, using the NUSAP method* Dutch National Research Program on Climate Change, Report no: 410 200 104, 227 pp, Bilthoven, 2002b, 237 pp (www.nusap.net).