Uncertainty management in complex models: the NUSAP method

Jeroen van der Sluijs[a], **James Risbey**[a], **Serafin Corral Quintana**[b], **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] *EC Joint Research Centre, Institute for Systems, Informatics and Safety, Ispra (VA), Italy*
[c] *Research Method Consultancy (RMC), London*

Abstract: A novel approach to uncertainty assessment, known as the NUSAP method (Numeral Unit Spread Assessment Pedigree) was applied to assess qualitative and quantitative uncertainties in the TIMER energy model, part of RIVMs IMAGE Model. The TIMER model is a system dynamics energy model that has been used, for instance, in the development of the new IPCC baseline scenarios (SRES). For our analysis we have used the IMAGE B1 scenario as case study. We used two complementary tools to assess uncertainty: (1) The Morris algorithm for global sensitivity analysis and (2) a NUSAP expert elicitation workshop, which assessed different aspects of the strength of the knowledgebase of the parameters. Results of (1) and (2) were combined into a diagnostic diagram putting spread and strength together to provide guidance in prioritisation of key uncertainties. The project has shown that the NUSAP method can be applied to complex models in a meaningful way. The method provides a means to focus research efforts on the potentially most problematic parameters, identifying at the same time specific weaknesses in these parameters.

Keywords: uncertainty; pedigree; NUSAP; quality; integrated assessment models

1. INTRODUCTION

This paper summarizes a project in which we assessed uncertainties in the TIMER energy model, both qualitatively and quantitatively. The TIMER energy model is part of RIVMs Integrated Model to Assess the Global Environment (IMAGE). The TIMER model (Targets IMage Energy Regional model) is a system-dynamics energy model that has, amongst others, has been used in the development of the new IPCC emission scenarios. For our analysis, we used the B1 scenario produced with IMAGE/TIMER for the IPCC Special Report on Emissions Scenarios as case study.

In the field of integrated assessment modelling, 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. This project has implemented a novel approach to uncertainty assessment, known as the NUSAP method (acronym for Numeral Unit Spread Assessment Pedigree).

2. NUSAP AND THE DIAGNOSTIC DIAGRAM

The NUSAP (Numeral Unit Spread Assessment, Pedigree) method aims to provide an analysis and diagnosis of uncertainty. It captures both quantitative dimensions and qualitative dimensions of uncertainty (Funtowicz and Ravetz 1990). The method addresses two independent properties related to uncertainty in numbers, namely spread and strength. 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. 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.

3. SENSITIVITY ANALYSIS

By means of a sensitivity analysis we explored criticality of 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, all liable to uncertainties of different orders of magnitude. A proper sensitivity analysis asks in such situation for an approach that covers the entire range of possible values for a given input variable. The Morris (1991) method facilitates such global sensitivity analysis in a minimum number of model runs.

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. The analysis differentiated clearly between sensitive and less sensitive model components. TIMER contains 300 variables that serve as input to the model. Parameters were varied over a range from 0.5 to 1.5 times the default values. The method and full results are documented in chapter 5 of Van der Sluijs *et al.* (2002).

The results show that the model outcome is sensitive to uncertainty in a substantial number of parameters (about one third). The combination of these uncertainties may hence produce substantial spread in model outcome. We also found that the sensitivity to uncertainty in a large number of parameters was contingent on the particular combinations of samplings for other parameters, reflecting the curvi-linear nature of many components of the TIMER model. The following input variables and model components (groups of input variables) were identified as most sensitive with regard to model output (CO_2 emission projections):

- Population levels and economic activity as main drivers;
- Variables related to the formulation of intra-sectoral structural change;
- 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;

4. PARAMETER STRENGTH AND PEDIGREE

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 we used are: proxy, empirical basis, theoretical understanding, methodological rigour, and validation (Table 1).

Code	*Proxy*	*Empirical*	*Theoretical basis*	*Method*	*Validation*
4	Exact measure	Large sample direct mmts	Well established theory	Best available practice	Compared with indep. mmts of same variable
3	Good fit or measure	Small sample direct mmts	Accepted theory partial in nature	Reliable method commonly accepted	Compared with indep. mmts of closely related variable
2	Well correlated	Modeled/ derived data	Partial theory limited consensus on reliability	Acceptable method limited consensus on reliability	Compared with mmts not independent
1	Weak correlation	Educated guesses / rule of thumb est	Preliminary theory	Preliminary methods unknown reliability	Weak / indirect validation
0	Not clearly related	Crude speculation	Crude speculation	No discernible rigour	No validation

Table 1. Pedigree matrix for parameter strength. Note that the columns are independent. (Ellis et al., 2000a, b; Risbey et al., 2001)

Assessment of pedigree involves qualitative expert judgement. To minimise arbitrariness and subjectivity in measuring strength a pedigree matrix is used to code qualitative expert judgements 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 project.

5. SET UP OF THE NUSAP WORKSHOP

We assessed parameter pedigree by means of a NUSAP expert elicitation workshop on June 12 and 13 2001, in Loosdrecht, The Netherlands. The workshop was attended by 19 experts on the fields of energy economy and energy systems analysis and uncertainty assessment. The primary goal of the workshop was to assess the strength of the input values for key variables. For a full description of the methodological details of the elicitation we refer to Van der Sluijs *et al.*, 2002.

We limited the elicitation to those parameters identified either as sensitive by the Morris analysis or as a key uncertain parameter by expert elicitation in a interview with one of the TIMER modellers. Our selection of variables to address in the NUSAP workshop counted 39 parameters. To further simplify the task of scoring pedigree criteria for each parameter at the NUSAP workshop, we grouped together similar parameters, either because they related to the same concept and because the pedigree scores might be to some extent similar for a group of similar parameters. We were able to group the selected parameters into 18 clusters. For each cluster a information and 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.

The workshop was set up in three phases:

- a plenary session with introductory lectures
- a expert elicitation session in 3 parallel groups
- a concluding plenary session

For the expert elicitation session, we divided the participants into 3 parallel groups of 6 person. The groups were made balanced with regard to *inter alia* expertise. Each participant received a set with all 18 cards listing the parameters to be reviewed. 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, focussing on, but not restricted to, the 5 pedigree criteria and eliciting the scores of the parameters for each of these pedigree criteria. Further we asked to provide a characterisation of the degree to which that parameter was considered to be value-laden. 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 judgement, informed by the group discussion.

We concluded the workshop with a plenary session, reflecting on our experiences with the method during the workshop. Overall there was a shared feeling amongst participants that the NUSAP method and the elicitation procedure with the cards facilitates and structures a creative process and in depth discussion on and assessment of uncertainty. The task of quality control in complex models is a complicated one and the NUSAP method disciplines and supports this process.

6. WORKSHOP RESULTS

After the workshop all the cards were collected and coded. In the analysis we treated results from the three groups separately so that we could check for intergroup differences in results.

We used two different diagrams to graphically represent the results: radar diagrams, and kite diagrams (Risbey, Van der Sluijs and Ravetz, 2001). An example of both representations is given in Figure 1.

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. Note that we inverted the scores for value-ladeness, so high value-ladeness is in the center and negligible value-ladeness in on the corner point for that axis. This is to keep for all the criteria the danger zone at the zero end of the scale (center) and the safe zone at the high end of the scale (corner point).

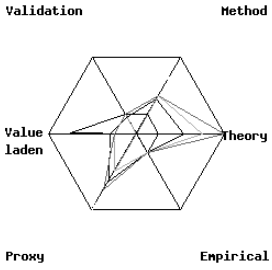

Figure 1a. Example of radar diagram of the gas depletion multiplier assessed by 6 experts

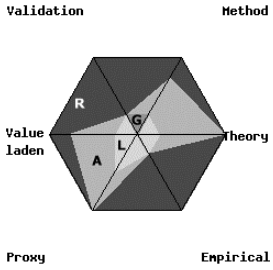

Figure 1b. same, but represented as kite diagram. G=green, L=light green, A=amber, R=red

In the radar diagrams the scoring of each expert is represented by a line, using a different color for each expert in the group, whereas also a line

connecting the average scores for each criterion is given (black line).

The kite diagrams follow a traffic light analogy. The green kite is spanned up by the minimum scores in each group for each pedigree criterion; the orange kite (amber in the traffic light) is spanned up by the maximum scores. The width of the orange band between the green kite and the red area represents expert disagreement on the pedigree scores for that variable. In some cases the size of the green area was strongly influenced by a single deviating low score given by one of the six 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.

The kite diagrams can be interpreted as follows: the size of the green colored area reflects the (apparent minimal consensus) strength of the underpinning of each parameter. The orange colored zone shows the range of expert disagreement on that underpinning. The remaining area is red. The more red, the weaker the underpinning is (all according to the assessment by the group of experts represented in the diagram). The methodological advantage of representing the group results by a kite diagram is that you can capture the information from all experts in the group without the need to average expert opinion. A second advantage is that is provides a fast and intuitive overview of parameter strength, preserving the underlying information.

We also calculated an overall average number for parameter strength, attributing equal weight to each of the 5 pedigree criteria we used. Further, we calculated the standard deviations in parameter strength to reflect the level of (dis)agreement amongst the experts.

Results indicate a range of attributes for the key TIMER parameters. For some parameters, there is reasonable consistency across the group results, indicating a convergence in view of the underpinnings of these parameters. We found that the convergence within groups tended to be larger, and sometimes much larger, than across the groups. This reflects the influence of discussion among the group members in the evaluation of strength. For other parameters there is considerable disagreement within and across groups. For instance, large intra-group disagreement was found on the value ladeness of the initial gas resource base. We found a similar pattern of disagreement on this score in all three groups. This could reflect a lack of expertise within some groups, with some participants simply not able to make appropriate judgements. However we have indications that this is not necessarily the case. That is because participants working in the core of the field of energy systems analysis and experts who gave a high score for their self assessment of competence on those parameters also diverge from each other on their pedigree scores for these parameters.

7. DIAGNOSTIC DIAGRAM

Results from the sensitivity analysis and strength assessments were combined in figure 2 to produce a diagnostic diagram. The diagram shows each of the selected parameters plotted according to sensitivity and strength. The sensitivity axis measures criticality of quantitative parameter uncertainty, using the contribution to change in CO_2 emissions from the Morris sensitivity runs. Results have been normalised for display.

The strength axis displays the pedigree scores for each variable averaged over the five pedigree criteria and the experts who ranked the variable. The error bars about these values indicate one standard deviation about the average expert value, to reflect the associated degree of expert disagreement on pedigree scores. Results have been plotted on a scale from 1 at the origin to zero on the right. With this convention the more "dangerous" variables are in the top right quadrant of the plot where sensitivity is high and strength is low.

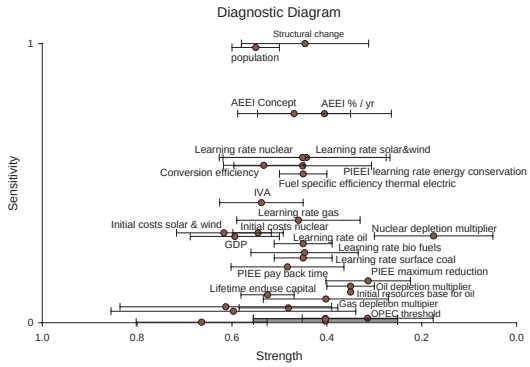

Figure 2. Diagnostic diagram for key uncertainties in TIMER model parameters.

We refer to the top right quadrant region of the diagnostic diagram in particular as the "danger zone". However, note that it does not have sharp boundaries and is not confined simply to the top right quadrant. The lower the contribution to sensitivity and the higher the strength, the further out of the danger zone, but there is no strictly "safe zone" as such. The term "danger zone" is simply

shorthand notation for higher sensitivity, lower strength combinations.

Since the majority of averages is between a relatively small band (values between 0.3 and 0.6), the resolving power on strength is relative weak in this case. Reasons for this might include the convergence process within groups, the process of calculating averages for pedigree scores and groups (divergence cancels out in this way), and the extent to which theory in the field of energy system analysis and modelling is crystallised according to the participants.

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. This makes intuitive sense in each case. Structural change in the economy has a large bearing on energy demand and use and heavily conditions CO_2 emissions. Further, structural change is represented in a highly idealised way in energy models and has limited grounding in theory or data and has not well been validated. Thus, its strength is weak to moderate (with the range here dependent on the particular experts assessment). Similarly, the population scenario has a large bearing on emissions via population loading on energy demand. The underpinnings of population scenarios in terms of theory, data, and method are judged to be slightly stronger (on average) than for the structural change variable, though still of only moderate strength. Autonomous energy efficiency improvements affect emissions strongly because of the role they play in translating demand for energy services into actual consumption. The theory behind this concept and its rate of change is fairly weak and there is little data to validate it. However there is some disagreement about the actual underpinnings of the AEEI variables and its strength spans from weak to moderate depending on the expert.

In interpreting the diagram, we must keep in mind that in calculating strength figures, we weighted all pedigree criteria equally, whereas it may depend on the specific nature of the parameter at hand what pedigree criteria are critical. Therefore, when a variable is identified as important from the diagnostic diagram, one can get further diagnostic aid by considering its underlying pedigree elements. If it has low strength, the pedigree scores will reveal in particular why the average pedigree (strength) is low. Reflection on the relative importance of these weaknesses in parameter underpinning in view of the nature and characteristics of that parameter may be needed for further meaningful interpretation and prioritisation of uncertainties. Knowing which parts of the pedigree are weakest guidance on where to possibly improve it.

Attempts to increase the robustness of energy related CO_2 emissions projections from the model would naturally focus first in improving the underpinnings of the variables closest to the danger zone discussed above. The next cluster of variables apparent in the diagnostic diagram is the group with moderate sensitivity contributions and weak to moderate strength in the centre of the diagram. This includes nuclear, learning solar wind, price induced energy efficiency improvement, and the fuel specific efficiency in thermal electric. As one descends the sensitivity axis in the diagnostic diagram to cover variables with increasingly lower sensitivity contributions it is important to pay particular attention to variables low in strength. 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. Should more knowledge come to light on the factors underpinning this variable, an alternative representation may move it higher or lower in the diagram. A move higher would likely bring it more squarely into the danger zone.

8. CONCLUSION AND DISCUSSION

We applied the NUSAP method to assess quantitative and qualitative dimensions of uncertainty in the TIMER model. We assessed parameter sensitivity using the Morris method and parameter pedigree and parameter value loading by means of a NUSAP expert elicitation workshop. We focussed the elicitation on those parameters identified either as sensitive by the Morris analysis or as key uncertain parameter by expert elicitation. The pedigree of these parameters was assessed by systematically evaluating the underpinning of the numerals and the status and nature of the knowledge from which they stem. We looked particularly at the following dimensions of parameter pedigree: proxy, empirical basis, theoretical understanding, methodological rigour and validation.

Results indicate a range of attributes for the key TIMER parameters. For some parameters there is reasonable consistency across the group results, indicating a common view of the underpinnings of these parameters and that the pedigree scores are meaningful. For other parameters there is considerable disagreement within and across groups. We interpret these diverging scores to reflect a higher degree of ignorance on the underpinning of those key uncertain parameters.

Pedigree results show slightly higher average score for theoretical understanding compared to empirical basis combined with the consistently low scores for validation nicely reflect the inherent theory ladeness of scenario studies of future developments, but in this case based on not so well crystallised theory. The latter may reflect that the scientific discipline of energy modelling and energy systems analysis is, seen from an epistemological perspective, in a relatively early stage of its development. One implication is that it seems more expectable that quantitative energy related CO_2 emission projections will remain in flux over the coming years than that they will show to have converged in the coming decade. It could also imply that the discipline of energy modelling is in a phase of development where more research may initially increase uncertainties by revealing new complexities not accounted for earlier. Consequently, the level of uncertainty is not a suitable indicator for the quality and progress in this complex field.

A diagnostic diagram combines results for parameter sensitivity and parameter strength. It provides a convenient way in which to view each of the key parameters in terms of their relative contribution to sensitivity in output and relative strength underlying their determination and representation. It is clear from the diagram which variables have relatively lower priority when one aims to increase the insightfulness and reliability of model projections, and which ones are more substantial contributors to overall uncertainty within the TIMER B1 scenario. Variables with relatively weaker or stronger pedigrees can be reasonably identified using the expert elicitation methods applied here. As a result, the NUSAP method provides a useful means to focus research efforts on the potentially most problematic parameters while it at the same time pinpoints specific weaknesses in these parameters.

This has been the first test of the use of NUSAP on a model of such complexity as TIMER. The results give support to the thought that the method can usefully be adapted and used for other complex model applications as well. This interpretation is supported by an evaluative survey held after the workshop: the responding participants unanimously answered the question whether they would like to see this type of NUSAP workshop further applied, with Yes. The overall judgement of the usefulness of the NUSAP workshop by the respondents to the survey was useful (62%) to very useful (38%) on a five point scale from not useful at all to very useful.

9. ACKNOWLEDGEMENTS

This project was carried out in the framework of the Dutch National Research Programme on Global Air Pollution and Climate Change, registered under no. 954267, entitled "Uncertainty Assessment IMAGE 2". We thank José Potting, Detlef van Vuuren, Penny Kloprogge and Bruce Beck for useful comments and/or valuable inputs used in this paper.

10. REFERENCES

Ellis, E.C., R.G. Li, L.Z. Yang, X. Cheng, Long-term change in village-scale ecosystems in China using landscape and statistical methods, *Ecological Applications*, 10 (4), 2000a, pp. 1057-1073.

Ellis, E.C., R.G. Li, L.Z. Yang, X. Cheng, Changes in Village-scale nitrogen storage in China's Tai Lake region. *Ecological Applications*, 10 (4), 2000b, pp. 1074-1096.

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

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

Risby, 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.

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, 2002 (available from www.nusap.net).