Eliciting Sources of Uncertainty in Ecological Simulation Models

V. Brilhante [a], J. L. Campos dos Santos [b]

[a]Computing Science Department, Federal University of Amazonas – UFAM, Manaus, Brazil
virginia@dcc.ufam.edu.br

[b]National Institute for Research of the Amazon – INPA, Manaus, Brazil
lcampos@inpa.gov.br

Abstract:

Uncertainty is an intrinsic feature of complex ecological models. Given that it is not possible to rid the models from uncertainty, we are left with taking notice of it for consideration in model-based decision making. Traditional ecological modelling methods and tools do not support explicit accounts of model uncertainty. This work gives a contribution towards making known, or bringing to the surface, sources of uncertainty that are embedded in ecological models. The sources of uncertainty are related to the models' supporting data and equations. A metadata standard is used to specify data-related sources of uncertainty, such as creator and coverage. In the technique developed, models are described and simulated using logic, which allows the sources of uncertainty to be easily represented, and later propagated and combined during simulation. The combined sources of uncertainty can then be presented to the user who can assess their impact on model outputs and tune up his confidence in the model for decision making.

Keywords: Uncertainty elicitation; logic-based ecological modelling; metadata.

1 INTRODUCTION

Artificial Intelligence has long been pursuing the successful handling of uncertainty by its systems [Cohen, 1985], largely motivated by the many intrinsically uncertain tasks we perform in everyday life – decision-making, argumentation, learning – and by how well we manage.

The domain of ecological modelling offers a challenging context for uncertainty handling, where well-defined mathematical and simulation techniques are applied to problems that are not totally understood. A significant part of existing functional relationships in an ecosystem can be ignored in a model, for being too complex, not well understood or simply unknown. Moreover, supporting field or experimental data is usually incomplete, partly due to logistic problems and inherent difficulties in data collection and accurate measurement and sensing in the natural environment. In the light of that, it does not seem promising to try and build ecological models that are uncertainty free.

We can, nevertheless, live in better terms with it by getting *to know* the uncertainty that is embedded in the models, which may lead to better informed model-based decision making. Conventional modelling techniques and tools do not provide for representation or reporting of model uncertainty. This work presents a computational technique where models' sources of uncertainty are explicitly represented, propagated throughout the model during simulation, and shown to users for assessment of their impact on simulation results and decision making that may follow.

2 SYMBOLIC REPRESENTATION OF UNCERTAINTY

The approach followed for uncertainty representation in this work is symbolic. This has been little explored in Artificial Intelligence as opposed to the more popular numerical approaches, such as degrees of belief handled by Bayesian methods [Parsons and Hunter, 1998]. Numerical approaches seem to perform well on domains such as the stock

market, where quantitative data is available to characterise problems, and degrees of belief are quantifiable with clearly defined semantics. On the other hand, symbolic approaches seem to fit better in large complex domains such as environmental applications or medicine, where the nature of data and its interpretations cannot be purely quantitative. A number representing a probability, a degree of belief, an evidence or any other so called uncertainty measurement may be an overly concise representation that obscures the reasons that one took into account to reach that number.

2.1 Endorsements Theory Revisited

Our non-numerical and declarative representation of uncertainty is based on Cohen's theory of endorsements [Cohen, 1985]. The theory advocates explicit representation of uncertainty related to domain information handled by a reasoning system, allowing users to reason about uncertainty directly, instead of implicitly through some numerical calculus, and assess how much to believe in the system's outcomes.

In modelling, a *source of uncertainty* is any information that can suggest to model users reasons for strengthening or weakening their belief in the model's results. Sources of uncertainty and their largely domain dependent values are represented by data structures called *endorsements*, which can be attached to data, rules, conclusions, tasks and resolution procedures. Building an endorsement-based system involves: 1. identification and naming of the sources of uncertainty, or endorsements, in a domain; 2. specification of how these sources interact, so that they can be combined; and 3. specification of rules for ranking combinations of sources of uncertainty so that decisions can be made.

Our identified sources of uncertainty (Section 3.2) are attached to logical clauses that define elements of system dynamics models, namely, state variables, intermediate variables, parameters, flows and links [Ford, 1999]. An automated mechanism has been implemented to combine the endorsements (Section 5). Step 3 above was not carried out in this work because of the subjective kind of analysis that the identified endorsements lend themselves to in the domain of ecological modelling (Section 6).

2.2 Logic as the Modelling Language

Logic-based approaches for ecological modelling have been proposed in [Robertson et al., 1991] on the grounds of language accessibility for modellers and representational power. Logic is adequate to express declaratively domain knowledge and model assumptions, which in turn enable some forms of modelling automation [Brilhante, 2003] and more informed model analysis. In this work we use Horn clauses with negation under the closed world assumption [Apt, 1997] to specify models and their sources of uncertainty. Adopting this well-known representational formalism has the advantage that we can straightforwardly build systems that reason upon the represented knowledge using an off-the-shelf implementation language which is Prolog.

Particular forms of clauses are used for each kind of model element. These clauses describe a model's static structure that is equivalent to its structure denoted by the system dynamics diagrams, as illustrated by the fragment from a forest Carbon cycle model in Figure 1. The figure shows just one state variable representing the stock of Carbon in the woody litter lying on the forest floor, regulated by the incoming flow of woody litter production and the outgoing flows of Carbon to soil organic matter and the atmosphere. Intermediate variables (denoted by circles) and parameters (denoted by squares) appear with fictitious names.

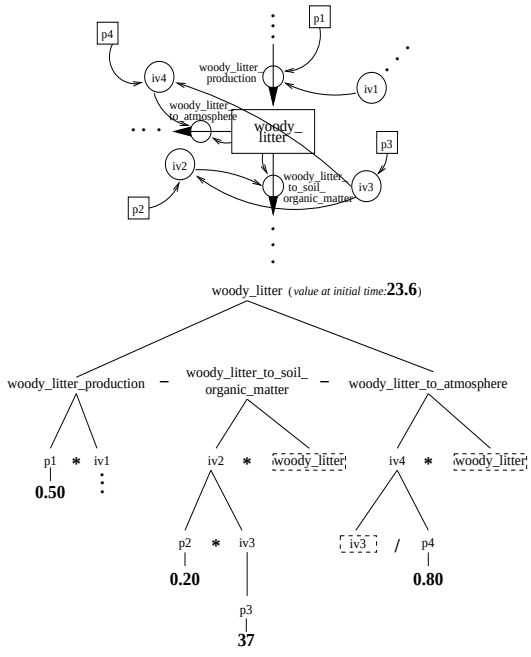

Figure 1: Fragment of a model in system dynamics notation and its visualisation as a tree structure.

Running a model consists of solving differential or difference equations which regulate changes in the values of the model variables as simulation time progresses. We can visualise a system dynamics model simulation at one tick of the simulation clock

as a set of tree-like structures, one to each state variable, having nodes representing variables which are interrelated by mathematical operations. Their roots are state variables to be solved, intermediate nodes are flows and intermediate variables, and leaves are the model parameters. The idea is depicted in Figure 1.

We have implemented the sources of uncertainty elicitation technique in a working logic-based system which includes an interpreter to run simulations of system dynamics models and a mechanism to combine instances of sources of uncertainty associated with model elements. The interpreter operates over the static structure of the model specified as a Prolog program, and is able to calculate the value of any model element at any simulation time. It goes into action when given goals of the form $goal(E, V, T)$ where E is a model element and V is the value it holds at time T. The interpreter works recursively backwards in time. Time T given in the top-level goal is successively decremented by 1 until the simulation gets to the initial time 0. A subgoal in the proof of $goal(E, V, T)$ is $goal(E, V_{\mathrm{p}}, T_{\mathrm{p}})$, where V_{p} and T_{p} stand for the previous value of the model element at the previous simulation time point.

Using logic programming has allowed us to represent model elements together with their endorsements under the homogeneous representational framework of logical clauses. The simulation interpreter, in turn, applies Prolog's built-in proof procedure to solve goals. This is a system development approach that contrasts with devising model-element specific data structures and a how-to-simulate procedural program.

3 METADATA IN MODEL UNCERTAINTY ELICITATION

As we have discussed, sources of uncertainty buried in ecological models can be many, related to procedures of data collection, to data integrity and to the modelling process itself. The approach we have taken for bringing some of this uncertainty to the surface is to "see the forest for the trees": to identify uncertainty related to model components and later combine it to provide an overall reading of uncertainty in the model.

Within this approach, the prime model components for sources of uncertainty identification are parameters and state variables. Model simulation starts from initial values assigned to these model elements. As shown in Figure 1, if we visualise the model structure as a tree of influences, the initial values are at the tree leaves (influenced by no other element in the model). The parameters' values remain constant, while the state variables' values can (and are usually expected to) change during simulation. These initial values are measures, rates, averages, constants, coefficients, percentages etc. that in the best case come from a modelling dataset built from field experiments and statistical data treatment. Sometimes, the initial values come from the literature or are mere guess estimations.

Our endorsements constitute information about the modelling data, and as such, can be specified as *metadata*.

3.1 Why Metadata?

Metadata is, quite simply, data about data. It describes the attributes and contents of a document, work or dataset [Duval et al., 2002]. The scope of questions it can cover in this way include: **Who** collected and who distributed the data? **What** is the subject (e.g., a dataset)? **When** was the data collected? **Why** was the data collected (its purpose)? **How** was the data collected? **How** should it be used? **How** much does it cost? Answers to these questions in the form of metadata can span over a quality spectrum – from raw to quality assured metadata. The better the quality, the more able users will be to evaluate, with less uncertainty, whether or not the data is useful to them.

Difficult and time-consuming to produce as it may be at first, metadata can help decision makers and researchers to find and use data, and can also benefit the primary creators of the data by maintaining its value and assuring its continued use over a span of years. From all its benefits, the one that is most relevant to the application of metadata in this work, in particular, is that metadata *allows for data understanding*, which is essential in data modelling and sharing. The source of uncertainty elicitation technique we present, harnesses metadata to extend such benefit to simulation models. Models annotated with metadata become more re-usable and sharable as users become less dependent on the modeller and the data provider.

3.2 Applying a Metadata Standard

The formalisation and use of metadata standards for data description is necessary to make it more precise and accessible to an audience that is as wide as possible [Campos dos Santos, 2003]. Standards provide a common set of terms naming key data attributes, which, if consistently used as recommended, fa-

cilitate data understanding to human users and the development of computer applications that handle metadata. Standard metadata terms can easily be translated to encoding formalisms such as XML, RDF, etc. In recent years we have seen an increase in the adoption of the Dublin Core Metadata Standard (DCMS) [Dublin Core Metadata Initiative, 2003]. The standard is considered to comprise a set of attributes that is simple and effective for describing a wide range of data resources.

We have identified sources of uncertainty by considering characteristics of models' supporting data that may strengthen or weaken one's understanding of or reliance on the model. The DCMS attributes lent themselves well to representation of such sources of uncertainty as metadata. A non-exhaustive list of these attributes is shown below.

Source – Where the data comes from. The range of possibilities includes: *literature*, *field experiments*, *model* (in case of a model element whose values are generated by the model itself), *modeller's definition*, and *modeller's assumption*. Every parameter and initial value has at least this source of uncertainty.

Creator – Who is responsible for making the data available, either through a scientific publication or other forms of communication.

Date – When the data was published or provided.

Identifier – An identifier of the place of publication of the data, e.g., journal, proceedings, book, URL, etc.

Coverage – Spatial location from where the data has been sampled; e.g., a research station, a geographical area like 'central Amazonia', etc.

Description – The state of the system from which the data was collected, by the time of collection. For example, a forest in equilibrium, or a forest that has suffered logging, burning or cultivation.

Description – Sampling information such as which data has been sampled (e.g., nutrients content in litter) and in which sampling campaigns (e.g. before logging, after logging), sampling design used (e.g. census, at random), number of samples, sampling frequency (annually, weekly), etc.

The 'Description' attribute is iterated, as DCMS allows, to distinguish the two different categories of descriptional information. References to and separate manipulation of the iterations' content can be resolved at the implementation level of the sources of uncertainty elicitation technique.

4 Model Equations and Sources of Uncertainty

Besides metadata, we also associate sources of uncertainty with model equations, which regulate the changing values of model elements. Parameters, the leaves of the models' tree-like structures, are constants and as such do not have regulating equations. It is through resolving the equations that the data and metadata associated with the parameters and initial values of state variables are fed into and propagated throughout model simulation. Two sources of uncertainty related to model equations are:

Equation Source – Similarly to data source, this can be *literature*, *field experiments*, *modeller's definition*, and *modeller's assumption*.

Equation Description – Explains what an equation means. E.g., woody_litter_production = 0.50 * biomass_mortality means that woody litter production is derived from the forest biomass mortality, assuming a 50% conversion of biomass to Carbon.

The identified sources of uncertainty – metadata or equation-related – are instantiated to values according to the specific model given. Hereafter, we shall use the unifying term 'endorsements' to refer to instantiated sources of uncertainty.

5 Combining Sources of Uncertainty

The interpreter produces the proof tree of each $goal(E, V, T)$ simulation. The proof tree is a data structure containing information about how the goal has been proved [Apt, 1997] over the model description enriched with the endorsements attached to the model components involved in the proof. It is gradually constructed as the interpreter operates recursively resulting in a nested data structure with hierarchy levels correspondent to the levels of recursion. This data structure is then processed for the endorsements to be combined.

Each run of the interpreter solves a given $goal(E, V, T)$ for a specific model element E. This model element will bear uncertainty that encompasses the uncertainty of all other model elements which directly or indirectly influence it. In system dynamics model diagrams, such influence is represented by the network of flows and links interconnecting model elements. In the numerical simula-

tion, the values of all other model elements connected to a model element are operands in calculating its value (see Figure 1). To provide for this, the proof tree of $goal(E, V, T)$ contains the endorsements of all such model elements connected to E. The endorsements are combined by means of a combination function that takes two sets of values for a source of uncertainty and finds the *union* and *intersection* of these two sets. The combination function is applied progressively, combining values of sources of uncertainty[1] attached to pairs of model elements in the goal's proof tree until it is fully parsed. We call the resulting intersection and union sets of each source of uncertainty its *lower bound* and *upper bound*, respectively.

Let us now see an example of application of the combination function given a certain model. Suppose the goal $goal$('leaf_litter_production',5,V) is given to the interpreter. Also suppose that 'leaf_litter_production' is a flow element in the model, directly influenced by the state variable 'above_ground_vegetation' and the intermediate variable 'leaf_litter_production_coefficient', which, in turn, is directly influenced by the parameters 'measured_leaf_litter_production' and 'measured_biomass'. Figure 2 depicts the example, with the endorsements being propagated and combined bottom up. The Ⓒ symbol in the figure stands for the endorsements combination function. The final lower and upper bounds of each source of uncertainty associated with model elements involved in the simulation are shown at the top, next to the 'leaf_litter_production' flow, the model element given in the simulation goal.

6 Interpreting Combined Sources of Uncertainty

The examples in Figures 1 and 2 are small excerpts from a much larger model of the cycle of Carbon in a logged forest to which we have applied the sources of uncertainty elicitation technique. The fall of leaves from the trees is one of the ways in which Carbon flows through forest systems. In the model, the rate in which leaf litter is produced (represented by the 'leaf_litter_production' flow) is calculated using a coefficient based on measured volumes of leaf litter and biomass, reported in relevant literature. The measurements have been taken from areas nearby the modelled logged forest, which is desirable, however from forests that were, at the time of sampling, in different states: leaf litter was measured in a forest in equilibrium, and biomass in a forest in post-burning state.

[1] Except values of the source of uncertainty 'equation reasoning'.

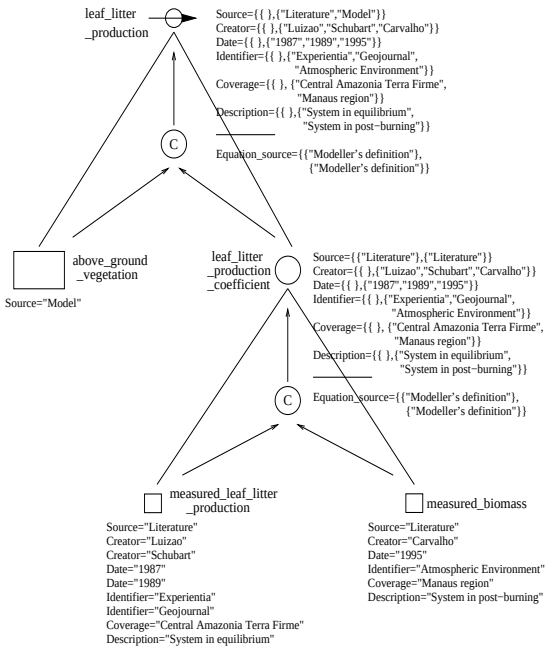

Figure 2: An example of endorsements propagation and combination.

The generic combination function explained in Section 5 calculates endorsements bounds in order to give a user access to a condensed account of information about the origins of parameters that calibrate the model, which can be relevant to the user's assessment of how adequate is that model to his purposes. The interpretation of the endorsements bounds is a subjective, non-deterministic task that is left to the user. For instance, in the 'leaf_litter_production' example, the bounds found for the Coverage source of uncertainty could weaken one's confidence in the model. The values calculated for the flow could be considered more applicable if all the data used had come from the same site as the modelled system or from sites with similar characteristics. If homogeneous endorsements are desirable, the user will be looking for coinciding upper and lower bounds. On the other hand, for the Identifier source of uncertainty in the 'leaf_litter_production' example, diverse and numerous places of publication could be preferable – the more widely published the data the better. In this case, the user would be looking for an upper bound set with many elements and an empty lower bound set associated with the source of uncertainty.

7 An Application Scenario

The technique has been applied to a large system dynamics model, originally implemented in Stella

II[2], of a tropical forest ecosystem in *terra firme*[3] areas of Central Amazonia, with emphasis on nutrient cycling and DBH[4] growth of trees of commercial and non-commercial species. The model was built at INPA (Instituto Nacional de Pesquisas da Amazônia) in Manaus, Brazil, to simulate logging strategies and predict their effect on the forest's sustainability, supporting the design of guidelines for sustainable timber exploitation in the region.

The model was taken as a representative example of complex models which require a varied and wide range of supporting data. Its dataset contained roughly 250.000 data records, regarding vegetation, litter, mesofauna, micro-biology, topsoil chemistry, hydrochemistry, soil physics, and other ecological entities and phenomena. Given such a diverse dataset, in both content and methods, in most simulations, for various model elements in the simulation goal, we obtained empty lower bounds and broad upper bound sets for the model's sources of uncertainty, as it was intuitively expected. This is particularly apparent if in the goal we have a model element that is influenced by many others, which causes the combination mechanism to try and pull together the diversity of the model's endorsements.

8 Results and Conclusions

We have presented a computational technique by way of which sources of uncertainty that would otherwise remain implicit in ecological simulation models are identified, propagated by simulation, combined and made available to a model user. This was achieved through a novel application of metadata, namely, its integration with executable simulation models. A modelling feature such as this can enlarge the understanding of an ecological model, as well as enhance its usage as a decision-making or research tool. The feature can also assist modellers in incremental development of models, not only helping to identify gaps in knowledge (as any modelling activity does), but also in pointing out *uncertain* knowledge.

9 Future Work

The range of models' sources of uncertainty can certainly be widened. We have identified only a sample of them to which applying simple Dublin Core has sufficed. Widening the range of sources of uncertainty will lead to consideration of using qualified or more specific Dublin Core encoding schemes. Moreover, combination heuristics could be tailor-made having in mind specific sources of uncertainty. For the coverage source of uncertainty, for example, combination heuristics could be based on spatial relations between locations, such as distance, subsumption, similarities and differences of environmental conditions in the locations, etc.

We now have the opportunity of exploring the technique within the LBA – Large Scale Biosphere-Atmosphere Experiment in Amazonia – project. We would also like to make the research-prototype system we have into a tool for others to use, possibly with an interface to main-stream ecological modelling graphical tools.

[2]High Performance Systems, Inc.

[3]Area that is not fboded when a river's water level rises.

[4]Diameter at Breast Hight of tree trunks.

Acknowledgements

The authors wish to thank FAPEAM (Fundação de Amparo a Pesquisa do Estado do Amazonas) for its sponsorship through the research project Metadata, Ontologies and Sustainability Indicators integrated to Environmental Modelling.

References

Apt, K. R. *From logic programming to Prolog*. Prenctice Hall, 1997.

Brilhante, V. *Ontology and Reuse in Model Synthesis*. PhD thesis, School of Informatics, University of Edinburgh, 2003.

Campos dos Santos, J. L. *A Biodiversity Information System in an Open Data/Metadatabase Architecture*. ITC Printing Departament, 2003.

Cohen, P. *Heuristic Reasoning About Uncertainty: an Artificial Intelligence Approach*. Pitman, London, 1985.

Dublin Core Metadata Initiative. Dublin Core Metadata Standard. http://dublincore.org/, 2003. Accessed on 20 Feb. 2004.

Duval, E., W. Hodgins, S. Sutton, and S. Weibel. Metadata principles and practicalities. *D-Lib Magazine*, 8(4), 2002.

Ford, A. *Modeling the environment: an introduction to system dynamics models of environmental systems*. Island Press, 1999.

Parsons, S. and A. Hunter. A review of uncertainty handling formalisms. In *Applications of Uncertainty Formalisms*, pages 8–37, 1998.

Robertson, D., A. Bundy, R. Muetzelfeldt, M. Haggith, and M. Uschold. *Eco-logic: logic-based approaches to ecological modelling*. MIT Press, Cambridge, Massachusetts, 1991.