W5: Systematic Simplification of Mechanistic Models

Neil Crout, Glen Cox, James Gibbons (Environmental Science, Nottingham)

Abstract

Models of environmental systems are often complex, reflecting the complexity of the systems they try and describe. This complexity is difficult to manage and, especially in the face of limited observed data, there is a risk that models become over-parameterised with the result that predictions are less reliable than they need be.

An approach to investigating the influence of model complexity on prediction accuracy is to compare the performance of alternative (simpler) model formulations. However this is difficult to achieve in practice as the process of simplification can be time-consuming with the consequence that only a few alternative formulations can be investigated.

Automatic, or perhaps semi-automatic, methods of model simplification are potentially useful in addressing this problem. This papers makes the case for such methods and discusses some of the issues arising from their use, with a practical example for a mechanistic model of plant uptake of radiocaesium.

Rationale

The use of predictive models is widespread in almost all areas of science and model outputs are often used for decision making. In applications, the predictive reliability of a model is of central importance.

Within the environmental sciences, substantial progress has been made in the development of models of complex systems. The dominant paradigm is the 'mechanistic' modelling of processes. Typically models are developed by scientists with an interest in the various processes at work within a system. Consequently more and more component 'processes' are added, increasing the level of detail represented. While more detailed models may be scientifically credible they do not necessarily give more reliable predictions than simpler models. Indeed, given the limitations of available data, complex mechanistic models can easily become over-parameterised, with the consequence that they make less reliable predictions.

An alternative approach is to develop simpler models, which utilise less process based information. However, such models are open to the criticism that they are less generalisable and the exclusion of detailed processes often makes them less credible to specialist scientists.

These considerations raise the question of how we can best apply our scientific understanding of a system within predictive environmental models? We are implicitly assuming that we are not able to specify the 'true' model of the system. The challenge is to find ways to combine the rigour of statistical model development with process based credibility. While mechanistic models tend to be detailed, they are less detailed than the real systems they seek to describe, so implicit judgements *are* being made about the appropriate level of detail within the process of model development. However, this judgement is often subjective

and little use is being made of the tools developed by statisticians for model selection. This is not through ignorance; there are barriers to the application of these methods.

Statistical methods for model selection *compare* a *set* of candidate models. The way in which the comparison is made varies between the different methods, but the key words are 'compare' and 'set'. The performance (however one defines it) of different models are *compared*. To do this one needs a *set* of different models. The development of a typical model of a typical environmental system is normally time consuming and expensive. It is not straightforward to readily develop a set of alternative model formulations. One solution to this problem is to take an existing detailed model and simplify it (Ziegler *et al*, 2000), which is easier to say than to accomplish.

One Possible Approach

We propose that it is useful to include scientific understanding of a system in detailed process based models, but if such models are to be used predictively they need to be tested against alternative model formulations. As suggested above this can (to some extent) be accomplished by taking a detailed process based model and systematically simplifying it to create a set of inter-related models. These models can then be compared, testing whether the inclusion of a given process description contributes to the model's performance. If the ultimate purpose of the model is prediction, then model performance will be based on comparisons with observed data. As we shall discuss later this type of approach may provide useful diagnostics to support model development, the output will always require interpretation in the context of the scientific understanding of the system under study

How can this be done? Below we give an example of the application of one possible method.

Simplification by Replacement

The typical structure of a mechanistic model does not generally allow a model relationship or parameter to be simply omitted. Therefore a strategy is required in which the omitted variable is either replaced or aggregated with other model terms. The procedure is analogous to classic regression methods such as backwards elimination or stepwise, although the implementation is more complex due to the structured and inter-connected nature of typical mechanistic models. In this example we consider the case where model variables are replaced by a constant.

Terminology

Before describing the approach, we define some terminology. Constant values within a model are *parameters*. For the purposes of the model development, they may be fixed, in which case their value is set before the model was developed, or they may be adjustable in which case their value is estimated as part of the model development process, usually through the use of data. *Input variables* are values obtained directly from data, and are independent of a model's calculations. *Model variables* are internal quantities calculated using an assumed relationship expressed in terms of the model's parameters, input variables and other model variables. The definition of model variables is partially subjective because intermediate steps in a model calculation could be defined as individual model variables, or combined into a larger relationship as a single model variable. Such choices will often depend upon the requirements of specific computer implementation. However, for our purposes, we shall regard each model variable as having a specific mechanistic interpretation. This is illustrated

later in the example application. Throughout we use M to denote the number of model variables, p to denote the number of parameters and n to denote the number of data.

Traditional statistical approaches to model selection have focussed on the number of adjustable parameters as a measure of model complexity (either explicitly or implicitly). Here we are also considering the number of model variables and inputs as a further measure of model complexity in order to reflect the structured and inter-related nature of typical mechanistic models. This distinction is further illustrated with reference to the example we present later.

The approach investigated involves the systematic replacement of model variables by constant values to produce a class of reduced models. The performance of these reduced models can then be compared using various criteria to assist the identification of model variables whose inclusion are not justified by the data, and which may, therefore, be unnecessarily increasing the complexity of the model. The procedure is not intended to generate the *best* model, rather, it is hoped that it may be used as an iterative diagnostic to inform model development.

Consider a model comprised of M model variables, V_i, each of which is defined by a relationship in terms of parameters, input variables or other model variables. If all of the possible combinations of variable replacements, R_i, are considered (i.e. an exhaustive search), 2^M simplified models will be generated and require assessment.

Comparing Model Performance

The ideal measure of a model's predictive performance is how well it can predict observed values of interest for a new situation. When a suitable dataset, which has not been used for model development, is available its predictive performance can be assessed by a measure such as the prediction residual sum of squares (PSS), defined as the sum of squared differences between the observed and predicted values.

If independent data are not available, an alternative approach is to rely on RSS (or other GOF statistics) derived using the data employed during model development. However, this does not take into account the possibility that the model is over-fitted. In these cases model selection criteria are a useful alternative, although it should be noted that they are only applicable if the model has been formally parameterised.

Several model selection criteria have been developed in the fields of information science and statistics, some of which are summarised in Table 2 (see Myung (2000) for a general review). Each comprises a term based on the model's GOF and a term which estimates the influence of the model's complexity on its predictive capability.

Choice of replacement value

How should the replacement values be selected? In principle, our objectives could be met by setting R_i to arbitrary values. However, R_i needs to be chosen in such a way that the rest of the model calculations can proceed successfully. A feature of many mechanistic models is the high degree of inter-connection between model variables, where one variable may depend upon another and so on. Consequently, an inappropriate choice of R_i may lead to poor model performance and/or numerical problems (e.g. if the value of the replacement constant results in taking the logarithm of a negative number). For this reason the standard approach for linear models, in which coefficients are set to zero, is not appropriate. One practical method is to set R_i equal to the mean value V_i

attains over the course of a simulation with no replacements (i.e. using the original model). The rationale for this method is that the replacement value is broadly appropriate, and a comparison between models becomes a test of whether the variation of a model variable about its mean is worth including in the model.

An appealing alternative approach when selecting a replacement value is to regard the R_i as adjustable parameters and to estimate them by fitting the reduced model to the observed data. This should improve the fit of the reduced models, relative to the models with mean value replacements. However, because these parameters will have been fitted to the data, this will be at the expense of additional adjustable parameters. Moreover this approach is computationally more intensive than simply using mean values, due to the fitting of the replacements. This may be significant when performing exhaustive searches with many replacement candidates, especially for large models.

Example Application: Model description

The model developed by Absalom *et al*. (2001) predicts the plant uptake of radiocaesium from contaminated soils. It is a semi-mechanistic model which estimates the partitioning of radiocaesium between the clay and humic fractions of soils; the time-dependent fixation of radiocaesium to clay particles; and competition between radiocaesium and potassium ions for plant uptake. The input variables for the model are the physical and chemical characteristics of the contaminated soils, namely: pH, fractional clay content, fractional organic matter content, the radiocaesium activity concentration and the concentrations of exchangeable potassium and ammonium in the soil. The model is schematically presented in Figure 1, which shows the extensive inter-connection between the model's variables, each of which has a specific mechanistic interpretation.

The model was parameterised using data from two comparable experiments in which radiocaesium uptake by grass was measured for a wide range of soil types (Smolders *et al.* (1997); Sanchez *et al.* (1999)). Employing the definitions given above, the model comprises 6 input variables, 16 model variables, 8 fixed parameters and 7 adjustable parameters. The adjustable parameters were estimated by fitting the model to the combined data set using the Marquardt non-linear regression method (Press *et al*., 1989). An additional data set, derived from the work of Nisbet *et al.* (1999), provided an independent test of the model's predictive performance.

Implementation

The original model was run using the full range of soil input variables within Absalom *et al*.'s (2001) parameterisation data, to allow the mean values of the model variables to be calculated.

As a preliminary screening procedure all the model variables were individually replaced (i.e. with all other variables retaining their original formulation) to identify potential replacement candidates. Any model variable whose replacement did not more than double the RSS with respect to the parameterisation dataset was deemed a replacement candidate. This procedure identified 10 model variables: ph, M_{CaMg}, CEC_h, CEC_c, θ_h, Kx_s, NH_4, Kd_h, θ_c and RIP_c. An exhaustive simplification was then performed, whereby a model formulation was generated for every possible combination of replacement of these model and input variables (2^{10}=1024 in total).

For each reduced model the adjustable parameters were re-estimated using the Marquardt procedure (Press *et al*., 1989) originally employed by Absalom *et al*. (2001). In each case, the parameterisation data were used to calculate RSS,

AIC_c, BIC, MDL and ICOMP. The independent data derived from Nisbet *et al*. (1999) were used to calculate the prediction sum of squares (PSS), which was used as an indicator of the model's general predictive capability.

Results

The models with the best performance measures for each criterion are summarised in Table 2. Two measures of model complexity are shown: the number of adjustable parameters (p), which is the conventional measure of complexity of statistical models, and the number of model and input variables (M), which is arguably a more relevant measure of complexity for mechanistic models although not normally considered in statistical model selection.

The lowest values of RSS and AIC_c occurred for the same model, in which M_{CaMg}, CEC_h, and pH were replaced. As can be seen in Figure 1, these three variables are directly related, and replacing pH has the effect of also replacing CEC_h and M_{CaMg} with constants. Similarly, if both CEC_h and M_{CaMg} are replaced, pH can effectively be considered a constant. In this case the number of adjustable parameters is the same as in the original model (7), although the number of model and input variables is reduced from 22 to 19. This arises because the replaced variables (M_{CaMg}, CEC_h, and pH) do not utilise any adjustable parameters (the use of adjustable parameters is indicated in Figure 1).

The lowest values of BIC, MDL and ICOMP were all associated with a further reduced model in which Kd_h and RIP_C were replaced, in addition to M_{CaMg}, CEC_h, and pH. This model had a higher RSS than the original model. However, p is reduced to 5 due to the replacement of the model variable RIP_c, which more than compensates for the loss of fit in the calculation of BIC, MDL and ICOMP.

Both reduced models resulted in lower values of PSS than the full model, with the RSS-AIC_c selected model slightly outperforming the BIC-MDL-ICOMP selected model; although this difference appears trivial.

For each of the criteria, there was little difference between the best performing models and those models with second lowest criteria scores, although further simplification resulted in significant increases in the respective criteria scores.

In the best performing reduced-models the pH input variable was replaced, together with the model variables solely dependent upon it. This does not imply that pH does not play a role in the uptake of radiocaesium, merely that the description of pH in this model does not contribute to its predictive capability. Pragmatically, the removal of pH increases the utility of the model, as it reduces the model's input requirements. This is especially important in the case of the Absalom model as it has been applied spatially (Gillett *et al.* (2001)), and pH is a difficult soil parameter to obtain from spatial data sets.

Extension to Model Averaging

The thrust of the above example is the selection of a model representing the optimal level of complexity. An attractive alternative approach is to average predictions over a class of possible models, weighted by their performance (e.g. Hoeting *et al*. (1999). This type of method is equally applicable to alternative mechanistic model formulations and the set of models produced by systematic simplification may provide a means of creating an appropriate model set. Averaging may have advantages in cases where there are a number of models with a similar performance and selecting a single model is arbitrary. Although not presented here, we have investigated this approach in the context of the Absalom example with useful results.

Pragmatic Considerations

Replacing model variables with their mean value, as outlined above, is conceptually simple. However, the procedure is not simple to implement for most complex environmental models which are often implemented within a procedural computer language (e.g. C, Fortran). Significant code modification would be required, especially if this kind of simplification procedure was not anticipated in the code design. This difficulty is further exacerbated if we wish to try a range of different replacement strategies other than simply using a mean value.

For example, the radiocaesium uptake model was implemented within a (locally developed) modelling package within which the model equations are specified as strings and the model compiled and solved. It was relatively straightforward to include within this system the capability to generically replace model variables with constants under program control. This would not have been the case had the model been implemented in its own computer program. Our experience is that these difficulties rapidly increase as the complexity of the models under study increase. Muetzelfeldt and Yanai (1996) have discussed this question and concluded that for systematic model manipulation to be possible, models need to be symbolically represented. Once this is the case it becomes possible, conceptually at least, to devise 'transformation rules' which would allow the model specification to be modified according to some programmatically controlled scheme. Potentially such rules could be more sophisticated than simple variable replacement, perhaps, for example, considering the aggregation of model variables, or changing the form of equations. Muetzelfeldt and Yanai (1996) suggest that the logic language Prolog might be a suitable basis for specifying both the model and its transformation rules. Unfortunately we have not been able to find any literature evidence of further development along these directions. There does seem to be increasing interest in the question of model complexity, but approaches to its investigation remain essentially manual. For example, Van Nes and Scheffer (2005) describe a general approach to simplifying large mechanistic models using a three-staged procedure, which involves "scrutinizing", "simplifying" and "synthesizing". The first stage involves performing sensitivity analyses on a model; this is used as a screening procedure to identify parts of a model that may be simplified. Simplification then proceeds either by removing state variables, replacing variables with constants, or by creating a "minimal" model which describes the dominant mechanisms of the full model. Simpler models are then compared to the full model to ascertain whether the hypotheses contained in the full model are credible. If not, then van Ness and Scheffer suggest that the model should be reformulated. Although this method utilises the information contained in a model's structure to guide the simplification process, the number of alternative model formulations that are eventually investigated may still be limited due to the level of user intervention required at the simplification stage. A practical example of this type of approach, in the context of crop modelling, is given by Brooks *et al*. (2001).

Discussion

The widely used approach of comparing the predictions of a model to observed values provides a basis for assessing the performance of the model. However, this is a test without a 'scale' unless there is a comparison *between* different models of the same system. This requires a set of models to compare.

Systematic model simplification potentially provides a means for rapidly generating many alternative model formulations, which may then be compared using appropriate performance measures. In the case of the presented replacement method all the model formulations that are generated are based on

the structure of the original model. This would probably be a characteristic of any comparable practical method. For this reason, we regard the general approach as a potentially useful diagnostic, which can be used to inform model formulation, rather than as a method for definitively identifying the best model. For example, in the case of the Absalom model the results suggest specific aspects of the model's formulation that could be re-visited.

The importance of expert scientific knowledge when designing mechanistic models remains paramount. However, if models are to be used for predictive purposes it is also important that they have empirical support and are not over-fitted. The proposed approach is potentially valuable in this regard, as useful information can be obtained about the empirical justification of hypotheses contained in a model by comparing the numerous simpler models generated from the full model.

Assessing model performance with reference to observed data can be criticised as being data driven. Certainly it is the case that if the data do not encompass the intended range of operation for the model, then results would need to be carefully interpreted. However, a model that is to be used for prediction should have empirical support, which requires observed values.

The example we have presented included a formal parameterisation step. The application of the model selection criteria is normally dependent on this. However, this is not a requirement for the application of the simplification approach itself. The simple comparisons to observed data could be applied to any model, and the use of a data set truly independent of model development is probably a more valuable alternative in any case.

If automatic, or even semi-automatic, simplification of complex models is to be widely applied it will require models to be developed in a way that enables the model specification itself to be manipulated within the system used for simplification. Moreover, if reliable and practical methods for model simplification were available, the process of primary model development could focus on the description of current scientific knowledge within the 'full' model, which could then be 'shrunk' as appropriate for a particular application.

Acknowledgements

We would like to thank the UK Biotechnology and Biological Sciences Research Council for financially supporting this work (grant reference BBS/B/05672).

References

Absalom, J. P., Young, S. D., Crout, N. M. J., Sanchez, A., Wright, S. M., Smolders, E., Nisbet, A. F. and Gillett, A. G., 2001. Predicting the transfer of radiocaesium to plants using soil characteristics. J. Environ. Radioactiv., 52:31-43.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., and Csaki, F. (Editors) Second International Symposium on Information Theory. Akademiai Kiado, Budapest. 267-281.

Bozdogan, H., 2000. Akaike's information criterion and recent developments in information complexity. J. Math. Psych., 44:62-91.

Brooks, R. J., Semenov, M. A., and Jamieson, P. D., 2001. Simplifying Sirius: sensitivity analysis and development of a meta-model for wheat yield prediction. Eur. J. Agron., 14:43-60.

Gillett, A. G., Crout, N. M. J., Absalom, J. P., Wright, S. M., Young, S. D., Howard, B. J., Barnett, C. L., McGrath, S. P., Beresford, N. A. and Voigt, G., 2001. Temporal and spatial prediction of radiocaesium transfer to food products. Radiat. Environ. Biophys., 40:227-235.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T., 1999. Bayesian model averaging: A tutorial. Stat. Sci., 14:382-401.

Hurvich, C. M., and Tsai, C-L., 1989. Regression and time series model selection in small samples. Biometrika, 76:297-307.

Jamieson, P. D., Semenov, M. A., Brooking, I. R. and Francis, G. S., 1998. Sirius: a mechanistic model of wheat response to environmental variation. Eur. J. Agron., 8:161-179.

Muetzelfeldt, R.I., Yanai, R.D. (1996). Model transformation rules and model disaggregation. Science Total Environment 183:25-31.

Myung, J., 2000. The importance of complexity in model selection. J. Math. Psych., 44:190-204.

Nisbet, A. F., Woodman, R. F. M., and Haylock, R. G. E., 1999. Recommended soil-to-plant transfer factors for radiocaesium for use in arable systems. NRPB-R304. National Radiological Protection Board, Chilton, Didcot, UK.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. 1989. Numerical recipes in Pascal. Cambridge University Press, Cambridge, UK.

Rissanen, J., 1987. Stochastic complexity and the MDL principle. Econometric Reviews, 6:85-102.

Sanchez, A. L., Wright, S. M. Smolders, E., Naylor, C. Stevens, P. A., Kennedy, V. H., Dodd, B. A., Singleton, D. L. and Barnett, C. L., 1999. High plant uptake of radiocaesium from organic soils due to Cs mobility and low soil K content. Environ. Sci. Technol., 33:2752-2757.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist., 6: 461-464.

Smolders, E., Van Den Brande, K., and Merckx, R., 1997. The concentrations of ^{137}Cs and K in soil solution predict the plant availability of ^{137}Cs in soils. Environ. Sci. Technol., 31:3432-3438.

Van Nes, E. H., and Scheffer, M., 2005. A strategy to improve the contribution of complex simulation models to ecological theory. Ecol. Mod., 185:153-164.

Ziegler, BP, Praehofer, H, Kim, TG (2000). Theory of modeling and simulation (2nd Edn). Academic Press.

Tables

Table 1. Commonly used model selection criteria. Where: ML is the maximised likelihood; p is the number of parameters estimated using data; n is the number of data points used to determine the maximum likelihood; H is the Hessian matrix; tr(θ) is the trace of the parameter covariance matrix.

Criterion	Calculation			Reference
	GOF term		Complexity term	
AIC	-2ln(ML)	+	2p	Akaike (1973)
AIC_c	-2ln(ML)	+	2p+ 2p(p+1)/(n-p-1)	Hurvich and Tsai (1989)
BIC	-2ln(ML)	+	p*ln(n)	Schwarz (1978)
MDL	-ln(ML)	+	½ln(|H|)	Rissanen (1987)
ICOMP	-ln(ML)	+	(p/2)ln(tr(θ)/p) – ½ ln|θ|)	Bozdogan (2000)

Table 2. Summary of the original model and the best performing reduced models selected by RSS, AIC_c, BIC, MDL and ICOMP. ✓ indicates that the variable remains in the model in its original form and × denotes that the variable is replaced by a constant. RSS is the residual sum of squares for the parameterisation dataset; PSS is the prediction sum of squares for the independent dataset; *p* indicates the number of adjustable parameters present in the model; M indicates the number of model and input variables in the model.

Selection criterion	Model variable													
	M_{camg}	CEC_h	NH_4	CEC_c	pH	θ_h	Kx_s	θ_c	Kd_h	RIP_c	p	M	RSS	PSS
None (full model)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	7	22	39.15	20.69
RSS, AIC_c	×	×	✓	✓	×	✓	✓	✓	✓	✓	7	19	36.84	16.59
BIC, MDL, ICOMP	×	×	✓	✓	×	✓	✓	✓	×	×	5	17	43.69	16.68

.

Figures

Figure 1. Schematic diagram of the radiocaesium plant uptake model. Shaded boxes indicate input variables. Open boxes represent model variables. Figures shown in parentheses on some of the model variables indicate the number of associated adjustable parameters (Note: One of the adjustable parameters is associated with two model variables).