Appeared as: Cox GM, Gibbons JM, Wood ATA, Craigon J, Ramsden SJ, Crout NMJ (2006). *Towards the systematic simplification of mechanistic models*. Ecological Modelling 198:240-246.

Towards the systematic simplification of mechanistic models

[1]G. M. Cox, [1]J. M. Gibbons, [2]A. T. A. Wood, [1]J. Craigon, [1]S. J. Ramsden and [1]N. M. J. Crout.

[1]School of Biosciences, University of Nottingham, Sutton Bonington, LE12 5RD.

[2]School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD.

Abstract

Mechanistic models used for prediction should be parsimonious, as models which are over-parameterised may have poor predictive performance. Determining whether a model is parsimonious requires comparisons with alternative model formulations with differing levels of complexity. However, creating alternative formulations for large mechanistic models is often problematic, and usually time-consuming. Consequently, few are ever investigated. In this paper, we present an approach which rapidly generates reduced model formulations by replacing a model's variables with constants. These reduced alternatives can be compared to the original model, using data based model selection criteria, to assist in the identification of potentially unnecessary model complexity, and thereby inform reformulation of the model. To illustrate the approach, we present its application to a published radiocaesium plant-uptake model, which predicts uptake on the basis of soil characteristics (e.g. pH, organic matter content, clay content). A total of 1024 reduced model formulations were generated, and ranked according to five model selection criteria: Residual Sum of Squares (RSS), AIC$_c$, BIC, MDL and ICOMP. The lowest scores for RSS and AIC$_c$ occurred for the same reduced model in which pH dependent model components were replaced. The lowest scores for BIC, MDL and ICOMP occurred for a further reduced model in which model components related to the distinction between adsorption on clay and organic surfaces were replaced. Both these reduced models had a lower RSS for the parameterisation dataset than the original model. As a test of their predictive performance, the original model and the two reduced models outlined above were used to predict an independent dataset. The reduced models have lower prediction sums of squares than the original model, suggesting that the latter may be overfitted. The approach presented has the potential to inform model development by rapidly creating a class of alternative model formulations, which can be compared.

Introduction

Mechanistic, or process based, models are generally highly structured and have inter-related components whose mathematical specification is informed by scientific knowledge of relevant processes. Models of this type are widely used. Mechanistic models are usually developed using expert knowledge of the processes involved in the system under consideration. This development may include the amalgamation of previously established relationships (e.g. Gibbons et *al*., 2005), the development of new relationships (e.g. Crout *et al*., 1998), or, more commonly, a combination of both (e.g. Jamieson *et al*., 1998). If an appropriate dataset is available, the model parameters may be chosen to achieve the best "fit", in which case the model may be described as being semi-mechanistic. If parameter values are chosen using a numerical procedure (e.g. least squares), we term this "formal parameterisation". Often, if the goodness-of-fit (GOF) is considered inadequate, the model may be modified by the addition of new parameters or relationships. Throughout this development process, judgements (which are often implicit) are made about the appropriate level of complexity in the model. However, it is well known that a model's fit to a particular dataset can always be improved by the addition of new parameters, and that this may lead to over-fitting and poor predictive performance when the model is applied to a new situation (e.g. Myung and Pitt, 2002). To avoid these difficulties model developers may adhere to the parsimony principle, which states that "models should be as simple as possible, but no simpler". Unfortunately, determining the point of optimal model simplicity is often difficult in practice, as this requires the generation and comparison of alternative model formulations. Generating alternative formulations of large mechanistic or semi-mechanistic models may not be straightforward, and can be very time-consuming. Consequently, although there are often many plausible representations of a given system, simpler alternatives are rarely assessed. This is in sharp contrast with, for example, linear statistical models for which coefficients can be readily set to zero to investigate reduced models.

One approach to creating a set of alternative models is "model generation". For example, Atanasova *et al*. (2006) describe an automated modelling tool where experts define a "knowledge library" containing context free grammar statements that characterise the general processes involved in the system under study. Different models are generated by combining the various expressions specified for each general process. The models are then parameterised by the fitting of constants, and the best performing models identified.

A limitation of such approaches is that for complex systems, where there may exist many alternative explanations of the underlying processes, the number of possible models can be very large, rendering parameterisation of the candidate models infeasible.

More recently, Asgharbeygi *et al*. (2006) have developed an algorithm which generates a set of alternative models based on an initial model, i.e. "model revision". Users specify which parts of the initial model are "fixed" and which parts can be removed or have their parameters changed, reflecting the areas of uncertainty within the model. The algorithm generates all models that are consistent with the constraints specified, and each model structure is parameterised using observed data. The method we describe here is similar, although simpler, and we are focussed upon the systematic removal of variables from a model, rather than the insertion or alteration of processes.

We illustrate our approach through its application to a published model, and discuss the results both in the context of the example model and more general application.

Approach

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. If the model considered contains parameters which have been estimated using data then it may be appropriate to re-estimate these values for each reduced model.

Choice of replacement value

An important question when simplifying mechanistic models by replacing model variables with constants is: how should the replacement values be selected? In principle, our objectives could be met by setting R_i to arbitrary values. However, the R_i need 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 in which there are no replacements (i.e. using the original model). The rationale for this method is that the replacement value is broadly appropriate, and our comparison between models becomes a test of whether the variation of a model variable about its mean is worth including in the model.

An obvious temptation here would be to select values for the $R_{i,}$ via formal parameterisation, which maximised the likelihood function. However, whilst this would improve the fit of the reduced models, it would effectively be introducing new adjustable parameters and consequently increase model complexity. This would conflict with our objective of identifying parsimonious models.

A further problem with using fitted replacement constants is that they may make interpretation of the results more difficult if the optimised values obtained are not mechanistically feasible. This can be avoided if the parameters' values are constrained in some way, although, care must taken when defining parameter boundaries, as limits which are too restrictive may affect the predictive performance of any reduced models generated. A further limitation to this approach is that it 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.

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, as discussed earlier, 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 1. 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.

The models we consider are all of the following general form:

$$y_j = f(I_j, \theta) + \in_j, \quad j = 1,\ldots,n,$$

where *n* is the sample size, y_j is the response for observational unit *j*, I_j is the corresponding vector of values of the input variables, *θ* is the parameter vector for the model under consideration, f is a known function of I_j and θ, and $\in_1,\ldots,\in_n$ are independent random error terms which are normally distributed with mean zero and variance σ^2. Each model determines an f. For the models under consideration, f is too complicated to specify explicitly here; an idea of the structure of a typical f is given by Figure 1. In practice, each f is specified through a computer program.

The log-likelihood for a model is given by

$$l(\theta,\sigma^2) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - f(I_j,\theta))^2$$

The maximised log-likelihood is given by

$$\ln(ML) = l(\hat{\theta},\hat{\sigma}^2),$$

where $\hat{\theta}$ is found by numerically by using the Marquardt parameterisation procedure (Press *et al.*, 1989) which minimises the residual sum of squares

$$\sum_{j=1}^{n}(y_i - f(I_j,\theta))^2,$$

and $\hat{\sigma}^2$ is set equal to the minimised mean residual sum of squares, i.e.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(y_i - f(I_j,\hat{\theta}))^2.$$

The principal difference between the model selection criteria is the approach used to estimate model complexity. In AIC (Akaike’s Information Criterion), the complexity term is simply twice the number of adjustable parameters in the model. However, where sample sizes are small, Burnham and Anderson (2002) recommend using AIC$_c$, a corrected version of AIC, when n/p<40. In BIC (Bayesian Information Criterion), the number of data points used to calculate the maximum likelihood is introduced, and consequently BIC penalises parameters more than AIC when n>8. However, complexity may not be related simply to the *number* of parameters in a model, but also the model's functional form. The MDL (Minimum Description Length) and ICOMP (Information Complexity) criteria attempt to take this into account through the Hessian matrix (which is the matrix of second derivatives, with respect to θ and σ^2, of the log-likelihood $l(\theta,\sigma^2)$, evaluated at $\theta=\hat{\theta}$ and $\sigma^2=\hat{\sigma}^2$) and the asymptotic covariance matrix of the parameter estimates respectively. These matrices are estimated during the Marquardt parameterisation procedure.

In the context of the model selection criteria, only adjustable parameters that are estimated using data are considered when determining the level of model complexity. However, determining the number of parameters to be included within the criteria may not be straightforward, as frequently some "fixed" parameters (which are not included in formal parameterisation procedures) are "tweaked" (i.e. adjusted manually by model developers) during model development to obtain a better fit, which amounts to *ad hoc* parameterisation. If that is the case, those parameters should be considered by the selection criteria.

Finally, it should be noted that the derivations of these selection criteria include a series of simplifying assumptions, which may not be satisfied in all cases. Consequently, some caution is required in the application of these measures. See Burnham and Andersen (2002) and Raftery (1995) for relevant discussion.

Example Application

<u>Model description</u>

The model developed by Absalom *et al*. (2001) predicts the plant uptake of radiocaesium from contaminated soils. It is a semi-mechanistic model which considers 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 (Table 2).

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. The study by Smolders *et al.* (1997) focussed on mineral soils (with relatively low radiocaesium uptake), whereas the study by Sanchez *et al.* (1999) considered organic soils (with relatively high radiocaesium uptake). Employing the definitions given above, the model comprises 6 input variables, 17 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. This data provided sufficient information for the application of the model, although it considers a range of graminaceous cereals rather than grass specifically. Consequently, it might be expected to show a higher degree of variability than the data set used to fit the models (the parameterisation data set).

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 3. Two measures of model complexity are shown: the number of adjustable parameters (p), which is the traditional 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. This model had a lower RSS than the full model (36.84 c.f. 39.15). In this case the number of adjustable parameters is the same as in the original model (i.e. 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. In all cases, the only difference was the inclusion or exclusion of Kd_h (depending on whether it was present in the best model). Furthermore, this replacement had a relatively small effect on the criteria scores. For example, RSS_p increased from 36.84 to 37.63, BIC increased from 69.03 to 69.38, MDL increased from 23.98 to 24.07 and ICOMP increased from 25.73 to 25.97 for the best and second-best models respectively.

The models with the third lowest criteria scores all involved the replacement of CEC_c. This resulted in more significant increases in the respective criteria scores.

Discussion of example application

The two reduced models selected both had the pH input variable replaced, together with the model variables solely dependent upon it. Although this is a very clear finding across all of the performance criteria it is mechanistically surprising. Many subject specialists would expect pH to be related to plant uptake of radiocaesium. However, these results suggest that the pH input variable is introducing additional variation into the model predictions, which is not accounted for by the relationships that predict the soil solution concentration of Ca and Mg (M_{CaMg}) and the cation exchange capacity of the humic fraction (CEC_h). This does not imply that pH does not play a role in the uptake of radiocaesium, merely that the pH input variable 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.

The further replacement of RIP_c and Kd_h is recommended by BIC-MDL-ICOMP, notwithstanding the increase in RSS_P, as this reduces the number of adjustable parameters. These model variables seek to refine the model's description of Cs adsorption in soils, accounting for the differences between adsorption on mineral and clay surfaces. While these may well be real processes the implication of the BIC-MDL-ICOMP result is that these refinements are over-fitting the model to the parameterisation data. Although, the results of the independent test of the model's predictive performances do not support this conclusion, they do suggest there is very little benefit from the inclusion of these variables.

General Discussion

The widely used approach of comparing the predictions of a model to corresponding 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.

The approach described here provides a method for rapidly generating many alternative model formulations, which may then be compared using various performance measures. Of course, all of the model formulations that are generated are based on the structure of the original model. Clearly, we are not investigating all possible models for a system but a related sub-set. For this reason, we regard the 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 with the full model.

The example we have presented here included a formal parameterisation step. The application of AIC, BIC, MDL and ICOMP is dependent on this as they are based on the concept of formally fitted parameters and, in the case of MDL and ICOMP, information about the variances and co-variances of parameter estimates. However, this is not a requirement for the application of the simplification approach. 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 valuable alternative.

A limitation to this approach is that an exhaustive search of all possible combinations of model variable replacements may become computationally prohibitive in situations where there are large numbers of candidate variables for replacement. This would be especially true for models that were computationally intensive in their original form. In such cases, it may be that some form of successive search, analogous to stepwise regression procedures, could be developed.

An alternative approach to selecting a best model, which is now commonly used in the case of statistical models, is to average predictions over a class of possible models, weighted in some way by their performance (e.g. Hoeting *et al*. (1999)). This type of method is also applicable to alternative mechanistic model formulations and the proposed approach may provide a means of creating appropriate alternative models.

Acknowledgements

We would like to thank the 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., 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., Csaki, F. (Editors) Second International Symposium on Information Theory. Akademiai Kiado, Budapest. 267-281.

Asgharbeygi, N., Langley, P., Bay, S., Arrigo, 2006. Inductive revision of quantitative process models. Ecol. Mod., 194:70-79.

Atanasova, N., Todorovski, L., Džeroski, S. and Kompare, B., 2006. Constructing a library of domain knowledge for automated modelling of aquatic systems. Ecol. Mod., 194:14-36.

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

Burnham, K. P., Anderson, D. R., 2002 (Second edition). Model selection and multimodel inference. Springer, New York.

Crout, N. M. J., Beresford, N. A., Howard, B. J., Mayes, R. W., Hansen, H. S., 1998. A model of radiostrontium transfer in dairy goats based on calcium metabolism. J. Dairy Sci., 81:92-99

Gibbons, J. M., Sparkes, D. L., Wilson, P., Ramsden, S.J., 2005. Modelling optimal strategies for decreasing nitrate loss with variation in weather – a farm-level approach. Agr. Syst., 83:113-134.

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., 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., Volinsky, C. T., 1999. Bayesian model averaging: A tutorial. Stat. Sci., 14:382-401.

Hurvich, C. M., 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. Francis, G. S., 1998. Sirius: a mechanistic model of wheat response to environmental variation. Eur. J. Agron., 8:161-179.

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

Myung, J., Pitt, M. A., 2002. When a good fit can be bad. Trends. Cogn. Sci., 6:421-425.

Nisbet, A. F., Woodman, R. F. M., 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., Vetterling, W. T. 1989. Numerical recipes in Pascal. Cambridge University Press, Cambridge, UK.

Raftery AE. 1995. Bayesian model selection in social research. Sociological Methodology 25:111-163.

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., 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., 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.

Tables

Table 1. Commonly used model selection criteria.

Criterion	Calculation			Reference
	GOF term		Complexity term	
AIC	$-2\ln(ML)$	+	$2p$	Akaike (1973)
AIC_c	$-2\ln(ML)$	+	$2p+ 2p(p+1)/(n-p-1)$	Hurvich and Tsai (1989)
BIC	$-2\ln(ML)$	+	$p*\ln(n)$	Schwarz (1978)
MDL	$-\ln(ML)$	+	$\frac{1}{2}\ln(\lvert H\rvert)$	Rissanen (1987)
ICOMP	$-\ln(ML)$	+	$(p/2)\ln(tr(\theta)/p) - \frac{1}{2} \ln\lvert\theta\rvert)$	Bozdogan (2000)

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.

Table 2. Mechanistic descriptions of variables in the Absalom model.

Model variable	Mechanistic interpretation	Units/scale
% clay	Fraction of clay matter in soil	%
% C	Fraction of organic matter in soil	%
K^+	Exchangeable potassium in soil	Meq $100g^{-1}$
pH	Soil pH	0-14
NH_4	Ammonium concentration in soil	Mol dm^{-3}
θ_c	Gravimetric clay content	g g^{-1}
θ_c	Gravimetric clay content	g g^{-1}
RIP_c	Radiocaesium interception potential	mmol kg^{-1}
Kx_{soil}	Exchangeable potassium in soil	$Cmol_c$ kg^{-1}
CEC_h	Cation exchange capacity on the humic soil fraction	$Cmol_c$ kg^{-1}
M_camg	Concentration of Calcium and Magnesium ions in the soil solution	Mol dm^{-3}
CEC_c	Cation exchange capacity on the clay soil fraction	$Cmol_c$ kg^{-1}
Kx_h	Exchangeable potassium on the humic soil fraction	$Cmol_c$ kg^{-1}
Kd_h	Radiocaesium distribution coefficient for the humic soil fraction	mol kg^{-1}
Kd_c	Radiocaesium distribution coefficient for the clay soil fraction	mol kg^{-1}
mk	Concentration of K^+ in the soil solution	mol dm^{-3}
Kdr	Proportion of labile Cs^+ adsorbed on the clay fraction	0-1
Kdl	Total labile radiocaesium	mol kg^{-1}
CF	Concentration factor	dm^3 kg^{-1}
D factor	Dynamic factor which describes the change in labile Cs^+ with time.	0-1
Cs_{sol}	Radiocaesium activity concentration in soil solution	Bq dm^{-3}
Cs_p	Radiocaesium activity concentration in plants	Bq kg^{-1}
Cs_{soil}	Total radiocaesium activity concentration in soil	Bq kg^{-1}

Table 3. Summary of the original model and the best performing reduced models selected by RSS, AIC_c, BIC, MDL and ICOMP.

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

✓ 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.