UNCERTAINTY ANALYSIS OF FOOD-CHAIN MODELS

Michiel J.W. Jansen
Centre for Biometry Wageningen, CPRO-DLO,
P.O.Box 16, 6700 AA Wageningen, The Netherlands
E-mail: m.j.w.jansen@cpro.dlo.nl

Key words: uncertainty contributions; variance components; regression-based and regression-free uncertainty analysis.

Introduction

A simple food chain might be modelled by a train of sub-models, say A, B, C, D..., the output of one sub-model serving as input of the next. The whole food-chain model yields an output y.

$$\begin{array}{ccccccccc} \mathbf{A} & \rightarrow & \mathbf{B} & \rightarrow & \mathbf{C} & \rightarrow & \mathbf{D} & \rightarrow \ldots \rightarrow & \mathbf{y} \\ \uparrow & & \uparrow & & \uparrow & & \uparrow & & \end{array}$$

Each sub-model has to be specified by a set of parameters, which are imperfectly known; it may also be influenced by unpredictable exogenous input. These uncertain inputs are collectively represented by the upward arrows in the diagram. The uncertainty about these inputs is often modelled by considering them as random variables.

The random effects propagate through the chain, by which some accumulation will occur. Uncertainty analysis provides methods to study the consequences of these random effects. The goal of such analyses is to evaluate if the joint effect on the final output y is still acceptably small, and to pinpoint the sources that contribute most to the uncertainty about the final output y. The results may be used to assess the potential benefits of control measures and of more accurate estimation of model parameters.

At present, most uncertainty analyses are performed computer-experimentally, by Monte Carlo methods. The model output is calculated for a representative sample of the uncertain model inputs. If the sample is large, the computational burden may be considerable. Several statistical methods have been developed to produce efficient Monte Carlo samples in order to keep the computational efforts as small as possible and to efficiently analyse the results. For a survey, see e.g. Helton (1993).

In Section 1, uncertainty and uncertainty contributions are characterised in terms of variances and variance components. Section 2 treats standard analyses based on regression-type approximations of the model output, the uncertain inputs serving as regressors. The analysis, which yields uncertainty contributions per individual input, is valid if the regression-approximation is of good quality. The approximation may fail because of non-linearities, especially in combination with a large number of inputs. Moreover, one may be interested in contributions of groups of inputs rather than individual inputs. Section 3 describes how, in such cases, one can have recourse to methods that are not based on regression-approximations, but that are more computer-intensive. Sections 2 and 3 contain examples of uncertainty analyses of models at the beginning of the food-chain. The paper ends with a discussion section.

Mathematical specifications

The object of study is a scalar output y of a deterministic model:

$$y = f(x_1...x_k),$$

with $x_1...x_k$ distinct inputs or vectors of inputs. If there are several scalar outputs of interest, these are analysed separately. The uncertain input vectors distinguished may have different sizes. The inputs are grouped into vectors because it is usually more relevant to assess uncertainty from meaningful groups of inputs, rather than from individual scalar inputs.

A bewildering variety of measures of uncertainty has been proposed in past. Recently, however, there is a growing consensus that the variance of the output is pre-eminently suitable as uncertainty measure, because the variance can be decomposed into meaningful parts, and because it is closely related to loss due to uncertainty (Krzykacz, 1990; Sobol, 1990; Jansen et al., 1994; McKay, 1996 and 1998).

Two important variance components related to group x_i

The *top-marginal variance* of x_i, TMV[x_i], intended as a measure of the premium for perfectly knowing x_i, is defined as the expected variance reduction when x_i would become known, that is

$$TMV[x_i] = Var[y] - E[Var[y|x_i]].$$

In this expression, E[] denotes the expectation, Var[] the variance, and Var[y|x_i] the conditional variance of y at the given value of x_i. According to the decomposition Var[y] = E[Var[y|x_i]] + Var[E[y|x_i]], which can be found in almost every introductory chapter on conditional distributions, the top-marginal variance may also be written as TMV[x_i] = Var[E[y|x_i]].

The *bottom-marginal variance* of x_i, BMV[x_i], intended as a measure of the price of not knowing x_i, is defined as the expected variance remaining as long as x_i stays unknown, that is

$$BMV[x_i] = E[Var[y|x_{(i)}]],$$

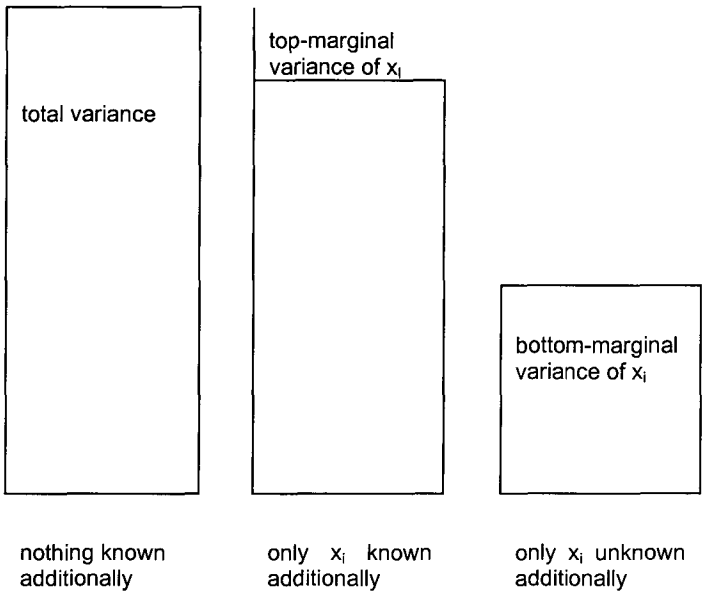
in which $x_{(i)}$ denotes the set of all sources except x_i. According to the decomposition just mentioned, the bottom-marginal variance may also be written as BMV[x_i] = Var[y] - Var[E[y|$x_{(i)}$]].

It follows directly from the definitions that the bottom-marginal variance of x_i is complementary to the top-marginal variance of the group all other sources $x_{(i)}$, that is: BMV[x_i] + TMV[$x_{(i)}$] = Var[y].

Unfortunately, the two variance components need not be equal. It can be shown, however, that BMV[x_i] = TMV[x_i] if x_i is stochastically independent of the others x-es and does not interact with the other x-es in making y. (Non-interaction means that y can be written as a sum of functions of the separate x-es.) Thus, inequality of BMV and TMV is a consequence of dependence and/or interaction. Moreover, if x_i is stochastically independent of the others, which often happens in practice, one has BMV[x_i] $\geq$ TMV[x_i], and a difference between the two is a manifestation of interaction.

The top- and bottom-marginal variances are often expressed as percentage of the full variance Var[y]. A graphical representation of the three variances is given in table 1.

Table 1 Graphical representation of total variance, top-marginal variance, and bottom-marginal variance



Regression-based uncertainty analysis

Regression-based uncertainty analysis is mainly applied for scalar x-es, which need not be independent. The method is based on least-squares regression of the model output y on several subsets of the inputs $x_1...x_k$. It is only valid if the full model, the one in which all inputs are represented, fits well (as can only be checked afterwards). Thus, the analysis is hardly ever perfectly valid, which is a draw-back. The regression is performed on a representative sample from the distribution of the inputs, with added output y. The computational effort is proportional to the size of the sample, i.e. the required number of model runs. In general, the sample size required for a regression-based analysis is small compared to that required for a regression-free analysis, which is an asset. There are several software-packages for this type of uncertainty analysis.

The relative uncertainty contributions of x_i are estimated as the increase, due to adding x_i, of the fraction of variance accounted for (adjusted R^2): TMV[x_i] by the increase when x_i is first term to be fitted, and BMV[x_i] by the increase when x_i is last term to be added. The most common analysis is based on linear regression in which y is approximated by an expression of the type $y \approx a + \Sigma_j b_j x_j$. Often, a worthwhile increase in the fraction of variance accounted for by the full model can be obtained without much extra effort by means of an additive spline approximation of the form $y \approx a + \Sigma_j s_j(x_j)$, in which each $s_j()$ is a smoothing spline (e.g. Jansen & Withagen, 1997).

Example: uncertainty contributions of parameters of a crop-growth model

The model SUCROS87-maize can be used to simulate grain dry-weight per hectare, depending on weather circumstances. Metselaar & Jansen (1994) study the uncertainty in the mentioned model output due to (only) 8 selected parameters: rglv, bsla, eff, q10, mnrt, mnlv, arso, arrt, which pertain to leave growth, photosynthesis, maintenance respiration and dry-matter conversion. The uncertainty as to the value of these 8 parameters had been assessed by world-wide literature search, and subsequent meta-analysis. With these sources of uncertainty, the model simulated grain dry-weight under the weather circumstances of Wageningen in 1975. The goal of the analysis was to assist in the selection of parameters that needed to be more accurately estimated. Metselaar (1998) contains a more elaborate analysis for a much larger set of parameters.

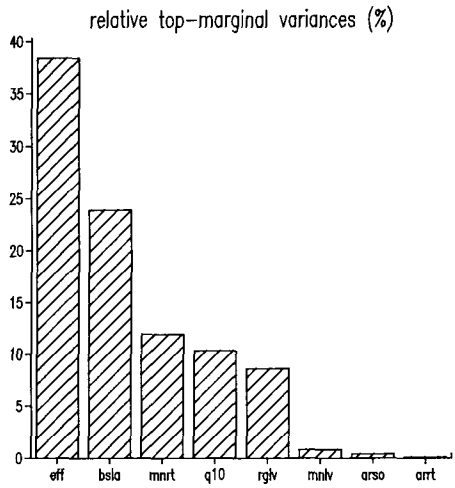


Figure 1 the relative top-marginal variances in order of decreasing importance

The analysis started with the construction of an ordinary random sample of 1000 parameter vectors. Uncertainty analysis via linear regression gave a coefficient of determination R^2 = 87.5% for the full model in which all parameters were present. Which is not too bad: only 12.5% of the output remained out-of-sight of the linear approximation. In figure 1, the relative top-marginal variances are sketched in order of decreasing importance.

Regression-free uncertainty analysis

The following so-called *winding stairs* sample for three independent input groups u, v, w, allows efficient regression-free estimation of TMV and BMV of the three input groups considered (Jansen et al, 1994; Jansen, 1996; Chan, 1998). To avoid multiple indices the groups are called u, v, w instead of x_1, x_2, x_3.

$f(u_1,v_1,w_1)$ $f(u_1,v_2,w_1)$ $f(u_1,v_2,w_2)$
$f(u_2,v_2,w_2)$ $f(u_2,v_3,w_2)$ $f(u_2,v_3,w_3)$
$f(u_3,v_3,w_3)$ $f(u_3,v_4,w_3)$ $f(u_3,v_4,w_4)$
$f(u_4,v_4,w_4)$ $f(u_4,v_5,w_4)$ $f(u_4,v_5,w_5)$
...

In this scheme $u_1,u_2...,v_1,v_2...,w_1,w_2...$ denote independent random draws of u, v and w; the function f() denotes the model output studied. When going through the sample matrix in reading order, one encounters in turn new random draws of the sources u, v, and w. It is easily seen that each column consists of independent realizations of f(u,v,w). The sample goes stepwise from one independent realization to next. The sampling method can be generalized in an obvious way to any number (at least 2) of independent input groups.

In the estimation of the uncertainty contributions, the values of the inputs u, v, w are ignored. The analysis is just based on the degree of resemblance between the columns of

the sample of outputs. For instance, if the first two columns, col_1 and col_2, are nearly the same, it is clear that the output is already accurately known when u and w are known, since the two columns differ only with respect to the value of v. In line with this, the bottom marginal variance of v can be estimated by

$$B\hat{},MV[v] = \tfrac{1}{2}Var[col_1-col_2].$$

Accordingly, the top-marginal variance of the pooled sources u and w can be estimated by

$$T\hat{},MV[u+w] = Covar[col_1, col_2],$$

since $Covar[col_1, col_2]$ and $\tfrac{1}{2}Var[col_1-col_2]$ sum up to an estimate of the total variance, and because of the complementarity rule mentioned in Section 1. For some of the estimations, a shift of a column is required: the top-marginal variance of w, for instance, can be estimated by

$$T\hat{},MV[w] = Covar[col_2, shift(col_3)],$$

where $shift(col_3)$ denotes col_3 shifted one step downwards. In general, a winding stairs sample enables the estimation of TMV and BMV of all sources distinguished, and of pools of sources that are adjacent in the cycle of renewal of the sources. For more details, see the references cited.

Example: brown rust in winter wheat

The next example involves a computer-based advisory system, EPIPRE, that supplies farmers with field-specific recommendations on chemical control of diseases and aphids. The goals of the uncertainty analysis are: to find optimal decisions under uncertainty; to determine the price of uncertainty; and to suggest research that might reduce uncertainty. For a more complete discussion, see Rossing et al. (1994-1 and 1994-2) and Jansen et al. (1994).

The model will be sketched summarily. The incidence of brown rust is measured as the number of leaves with spots from a sample of leaves. Rust incidence is translated into rust density by an empirical relation; a random effect is added to account for lack of fit (LOF). The density grows exponentially, with a constant plus a white noise as relative growth rate. The growth ends when the crop attains the mature stage. The stage is function of the cumulative temperature, with an added random effect for lack of fit. Finally, loss is calculated as a function of the time-integral of the number of spots.

The various sources of uncertainty were quantified by analysis of experimental data. Temperature uncertainty was represented by 30 years of historical data. This led to the following inventory of independent multivariate sources of uncertainty: -estimation error in parameters; -measurement error in initial value; -noise on relative growth rate; -lack of fit in crop development function; -lack of fit in density function; -temperature. The sampling was done in the order given, which enables grouping of the input vectors into methodologically meaningful classes. The first two input vectors might become known more accurately, via accurate measurements of the initial conditions, and additional experiments. On the other hand, subprocess lack of fit and noises will remain in force until model structure is modified. And finally, long-range temperature prediction is notorious for its bad quality.

Table 2 Estimated top-marginal variances as percentage of the full variance.

estimation error in parameters	80	11	11
measurement error in initial value			
noise on relative growth rate		63	45
LOF in crop development function			7
LOF in density function			22
temperature	20	20	20

The output analyzed was the base-10 logarithm of the loss, which appears to be approximately normally distributed. It has an estimated mean of 1.68, and variance 0.544. The analysis of a 250-cycle winding stairs sample yielded table 2 of estimates of top-marginal variances of several groups of sources.

Discussion

In the uncertainty analyses discussed, the model structure is assumed to be correct, which it never is. Thus the results merely give a preview of prediction error.

The question whether or not to base the analysis on regression has several aspects. First of all, standard regression-based uncertainty analysis estimates uncertainty contributions of individual scalar sources, which is not always adequate in view of the research questions. A regression-based analysis uses the values of the inputs, so that it is potentially less variable from sample to sample, implying that smaller samples and less model runs might be required. On the other hand, the regression fit will never be perfect, except for trivial models, so that a regression-based analysis will almost always entail some fixed error from sample to sample. Regression-based uncertaity analysis will produce, as spin-off, an approximation of the model as function of the inputs considered; this may give a wellcome insight in the functioning of the model. In summary, it seems advisable to start with trying a regression-based analysis if that can yield the desired kind of results and if it is feasible in view of the dimensions of the inputs. One may still have recourse to regression-free methods if the quality of the fit is unsatisfactory.

The examples given involve two separate processes at the very beginning of the food chain. But uncertainty and sensitivity analysis are eminently suitable for the study of a chain of models as a whole.

References

Chan K. (1998): Winding stairs as a tool to generate samples for sensitivity analysis. P 49-52 in: Chan K. & Tarantola, S. & Campolongo, F., *SAMO'98: Second International Symposium on Sensitivity Analysis of Model Output.* Luxembourg: Office for Official Publications of the European Communities.

Helton J.C. (1993): Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliability Engineering and Systems Safety* **42**, 327-367

Jansen M.J.W. & Rossing W.A.H. & Daamen R.A. (1994): Monte Carlo estimation of uncertainty contributions from several independent multivariate sources. P 334-244 in: Grasman, J. & Van

Straten, G. (eds.), *Predictability and Nonlinear Modelling in Natural Sciences and Economics*. Kluwer, Dordrecht.

Jansen M.J.W. (1996): *Winding stairs sample analysis program WINDINGS 2.0*. Internal Note MJA-96-2, Wageningen: DLO-Agricultural Mathematics Group.

Jansen M.J.W. & Withagen J.C.M. (1997): *USAGE: uncertainty and sensitivity analysis in a Genstat environment. Manual. Version 1.0*, in prep.

Jansen M.J.W. (1998): Efficiency of designs for the analysis of variance of model output. P 143-145 in: Chan, K. & Tarantola, S. & Campolongo, F., *SAMO'98: Second International Symposium on Sensitivity Analysis of Model Output*. Luxembourg: Office for Official Publications of the European Communities.

Krzykacz B. (1990): *SAMOS: a computer program for the derivation of empirical sensitivity measures of results from large computer models*. Gesellschaft für Reaktorsicherheit, GRS-A-1700.

McKay M.D. (1996): *Variance-based methods for assessing uncertainty importance in NUREG-1150 analyses*. Los Alamos National Laboratory, LA-UR-96-2695.

McKay M.D. (1998): Evaluating uncertainty in stochastic simulation models. P 171-175 in: Chan K. & Tarantola S. & Campolongo F., *SAMO'98: Second International Symposium on Sensitivity Analysis of Model Output*, Luxembourg: Office for Official Publications of the European Communities.

Metselaar K. & Jansen M.J.W. (1995): Evaluating parameter uncertainty in crop growth models. *IMACS/IFAC First International Symposium on Mathematical Modelling and Simulation in Agriculture and Bio-industries*, Brussels, May 1995.

Metselaar K. (1998): *Auditing predictive models: a case study in crop growth*. PhD thesis, Wageningen Agricultural University, to appear.

Rossing W.A.H. & Daamen R.A. & Jansen M.J.W. (1994-1): Uncertainty analysis applied to supervised control of aphids and brown rust in winter wheat. Part 1. Quantification of uncertainty in cost-benefit calculations. *Agricultural systems* **44**, 419-448.

Rossing W.A.H. & Daamen R.A. & Jansen M.J.W. (1994-2): Uncertainty analysis applied to supervised control of brown rust in winter wheat. Part 2. Relative importance of different components of uncertainty. *Agricultural systems* **44**, 449-460.

Sobol I.M. (1990): Sensitivity estimates for nonlinear mathematical models. *Matematicheskoe Modelirovanie* **2**, 112-118 (in Russian). English translation in: Mathematical Modelling and Computational Experiments 1, 1993, 407-414.