SEMI-MECHANISTIC MODELLING IN NONLINEAR REGRESSION:
A CASE STUDY

KATARINA DOMIJAN1*, MURRAY JORGENSEN2 AND JEFF REID3

Trinity College Dublin, University of Waikato and New Zealand Institute for Crop and Food Research Ltd

Summary

This paper discusses the use of highly parameterized semi-mechanistic nonlinear models with particular reference to the PARJIB crop response model of Reid (2002) [Yield response to nutrient supply across a wide range of conditions 1. Model derivation. Field Crops Research 77, 161–171]. Compared to empirical linear approaches, such models promise improved generality of application but present considerable challenges for estimation. Some success has been achieved with a fitting approach that uses a Levenberg–Marquardt algorithm starting from initial values determined by a genetic algorithm. Attention must be paid, however, to correlations between parameter estimates and an approach is described to identify these based on large simulated datasets. This work illustrates the value for the scientist in exploring the correlation structure in mechanistic or semi-mechanistic models. Such information might be used to reappraise the structure of the model itself, especially if the experimental evidence is not strong enough to allow estimation of a parameter free of assumptions about the values of others. Thus statistical modelling and analysis can complement mechanistic studies, making more explicit what is known and what is not known about the processes being modelled and guiding further research.

Key words: crop response to fertilizers; descriptive model; genetic algorithm; Levenberg–Marquardt algorithm; nutrients; parameter correlation; simulated data.

1. Introduction

The contrast between linear and nonlinear regression modelling is deeper than the formal definitions might suggest. Linear models are often purely descriptive in that they seek to describe relationships between a response variable and predictor variables as economically as possible for a particular dataset. Nonlinear models, while they may be purely descriptive, often arise through subject matter considerations about the situation being modelled. In an ideal world, fitting should be part of an ongoing process of model development and testing, rather than an endpoint. In a statistical consulting situation, tension can arise between the statistician and the client if they are at cross purposes about the type of modelling being undertaken. Often the statistician tends to seek the simplest empirical model fitting the data but the client tries to build a model corresponding to his/her assumptions about the process.

*Author to whom correspondence should be addressed.

1Department of Statistics, Trinity College, University of Dublin, Dublin 2, Ireland.
e-mail: domijank@tcd.ie

2Department of Statistics, University of Waikato, Private Bag 3105, Hamilton, New Zealand.

3New Zealand Institute for Crop and Food Research Ltd, RD2 Hastings, New Zealand.

© 2006 Australian Statistical Publishing Association Inc. Published by Blackwell Publishing Asia Pty Ltd.
Models generally divide into two types, mechanistic and empirical. Empirical models are essentially descriptions of the observational data, most often associated with curve fitting and regression. Empirical modelling is not constrained by biological principles and often does not require detailed knowledge of the mechanism.

In the context of crop performance modelling, one is often interested in modelling system behaviour across a wide range of field conditions. Under such circumstances, empirical models can be of limited value, as they will rarely account for changes in covariates such as weather or plant density.

Mechanistic models, on the other hand, are reductionist in approach and are concerned with the mechanism since they aim to contribute to understanding of the processes being modelled. In general, mechanistic modelling involves breaking the system down into components and assigning properties and processes to these components, usually introducing many extra variables compared to the empirical approach.

Typically, mechanistic models are very rich in content; they may apply to a wide range of phenomena and relate them to each other (Thornley & Johnson, 1990). However, in the context of fertilizer response modelling, there are some drawbacks, as Reid (2002) observes:

...detailed mechanistic simulation models are often ill suited for calculations that span a wide range of field conditions. Such models may need large amounts of site-specific soil and crop details, and can be difficult to validate at the level at which they ostensibly simulate the processes involved.

In agricultural science, it is rare that fully mechanistic models based on good science are available. Often it is necessary to replace parts of a model which would ideally be mechanistic with assumed or empirically estimated functional forms. The resultant models are neither fully mechanistic nor fully descriptive and are here called semi-mechanistic models. These types of models can also be rich in content. Typically they are highly parameterized and fitting them requires strong statistical expertise.

Apart from a professional tendency toward scepticism, statisticians are averse to highly parameterized nonlinear models because their parameters are poorly identified and they are very difficult to fit by common statistical methods, such as the Gauss–Newton algorithm. Raw computing power and computer intensive algorithms, such as Nelder–Mead or Genetic Algorithms, offer ways of finding good least-squares solutions for complicated nonlinear models. Nevertheless, the quality of such solutions needs to be evaluated by calculation of an information matrix and exploration of the likelihood surface in the neighbourhood of the solution.

An obvious way to proceed would be to start from parameter estimates provided by a genetic algorithm and use Gauss–Newton to improve these estimates. However, when a nonlinear model has a large number of parameters, the pseudo-design matrix often fails to be of full column rank. In turn this will cause the Gauss–Newton update to fail while attempting to invert a singular or near-singular matrix.

The failure of the Gauss–Newton algorithm need not bring fitting efforts to a close. It is sometimes possible to reduce the number of parameters being estimated by specifying some of them as constants. A difficulty with proceeding in this way is that it is not easy to see which parameters are responsible for the failure, and hence which need to be specified as constants.

This paper presents a case study of fitting a semi-mechanistic biological model. It is not intended to argue for or against the validity and usefulness of that model. Some of the
practical and philosophical issues that confront the statistician in fitting such models and the biologist in using the statistician’s findings to appraise the soundness of the original model are discussed. In particular, it is shown how simulated data may be used to predict which parameters may be difficult to estimate with the actual data available.

2. Case study outline

2.1. Background

Reid (2002) developed a model (named PARJIB) to describe how crop yield varied in response to nutrient supply. Reid et al. (2002) carried out an initial fitting of that model using a genetic algorithm technique. That technique yielded parameter values that made biological sense, but any detailed interpretation of those values was limited by the fact that the genetic algorithm technique gave no indication of reliability for individual parameter values.

2.2. Outline of the PARJIB model

The cornerstone of PARJIB is the idea that crop responsiveness to nutrient supply is very strongly influenced by the maximum yield. Maximum yield, denoted by $Y_{\text{max}}$, is the yield that would be achievable in the absence of mineral nutrient stresses (Reid, 2002). Modelled yield ($Y_{\text{model}}$) is obtained from $Y_{\text{max}}$ and what Reid called the scaled yield ($Y^*$):

$$Y_{\text{model}} = Y^* Y_{\text{max}}.$$

The maximum yield variable ($Y_{\text{max}}$ in t/ha) itself is derived from estimates of potential yield (the yield achievable in the absence of water and nutrient stresses, dictated by weather and cultivar characteristics) adjusted for plant density and water stress,

$$Y_{\text{max}} = \tilde{Y} H h W \frac{1000}{1000},$$

where $\tilde{Y}$ is the potential yield in kg/plant calculated at a standard population density (Reid, 2002), $H$ is the plant population multiplier (no units), $h$ is the plant population (plants/hectare), and $W$ is the water stress multiplier (no units). Potential yield is calculated by a separate model which takes into account cultivar characteristics and the weather conditions experienced. The variables $H$ and $W$ will be defined below. PARJIB is strongly concerned with scaling, particularly by relating scaled yield ($Y^*$) to an integrated nutrient multiplier ($q_{\text{nut}}$)

$$Y^* = q_{\text{nut}}.$$

In this paper, two ways are considered of calculating the $q_{\text{nut}}$ term; the standard way outlined by Reid (2002), and a ‘simple model’ that involves fewer nutrients. In this particular case, the input variable $\tilde{Y}$ was calculated using Wilson, Muchow & Murgatroyd’s (1995) modification of the potential yield model presented by Muchow, Sinclair & Bennett (1990). The quantities $Y_{\text{max}}$ and $Y$ are by definition constrained to be positive and less than $\tilde{Y}$. Variables and parameters used to calculate $\tilde{Y}$ do not appear elsewhere in PARJIB, although the input $h$ is used to calculate $H$ (see below) and solar radiation and air temperature appear in the calculation of both $\tilde{Y}$ and the input $D_{\text{max}}$ (see below).
The PARJIB model predicts crop yield in kg/ha as a hypothetical ideal yield attenuated by factors expressive of stresses due to plant density, water supply and nutrient supply and soil pH. The nutrients considered are nitrogen, phosphorus, potassium and magnesium. (In this paper nutrient $X$ is often used, where $X$ is understood to range over N, P, K and Mg.) For each nutrient there is a response curve relating crop yield to the supplied amount of that nutrient, when all other nutrients are at optimal levels.

Soil nitrogen is measured in the laboratory and corrected for differences in bulk soil density between the laboratory and the field. The soil concentrations of other nutrients are estimated in a similar fashion. The amount of nutrient supplied to the crop is calculated from both soil and fertilizer forms of the nutrient. So, for a nutrient $X$ we have

$$X_{\text{supply}} = X_{\text{soil}} + X_{\text{broad}}\xi_1 + X_{\text{band}}\xi_2.$$ 

Here $X_{\text{soil}}$ is the amount of $X$ present in the soil before fertilizer application, while $X_{\text{broad}}$ and $X_{\text{band}}$ refer to the amount of fertiliser $X$ applied to the soil in kg/ha in broadcast and banded applications respectively. The parameters $\xi_1$ and $\xi_2$ respectively denote the efficiency of supplying $X$ in broadcast or banded fertilizer form compared to $X_{\text{soil}}$.

All nutrient response curves rise from 0 when $X_{\text{supply}} \leq X_{\text{min}}$ to a maximum of 1 when $X_{\text{supply}} \geq X_{\text{opt}}$. It is convenient to plot the non-constant part of the supply curves on the unit square, so we introduce the dimensionless nutrient supply index, which is 0 when $X_{\text{min}}$ is supplied and 1 when $X_{\text{opt}}$ is supplied. Considering one nutrient at a time, the effect of nutrient supply index $x$ on scaled yield (i.e. yield as a fraction of the maximum yield) is modelled using the family of curves

$$q = g_\gamma(x) = (1 + \gamma)x^\gamma - \gamma x^{1+\gamma}.$$ 

Note that $g_\gamma(0) = 0$, $g_\gamma(1) = 1$ and that $g_\gamma$ increases smoothly over the unit interval. The positive shape parameter $\gamma$ governs where most of the growth takes place, near 0 for small $\gamma$ and near 1 for large $\gamma$.

In PARJIB then, the three parameters $X_{\text{min}}$, $X_{\text{opt}}$ and $\gamma_X$ define the nutrient response curve for nutrient $X$ (where $X = N, P, K$ or Mg). Soil pH stress is treated in a similar fashion to nutrient response but using a different curve. Soil pH stress will not be considered in this paper.

A nutrient response curve applies directly only when all other nutrients are not limiting yield. In other situations scaled yield is calculated by the ‘copula-like’ function

$$q_{\text{nut}} = \max \left(0, 1 - \sqrt{(1 - q_N)^2 + (1 - q_P)^2 + (1 - q_K)^2 + (1 - q_M)^2}\right).$$ 

The plant population multiplier $H$ is defined by

$$H = 1 - \eta \log \left(\frac{h}{h_{\text{ref}}}\right)$$

where $\eta = \eta_1$ when $h \leq h_{\text{ref}}$ and $\eta = \eta_2$ when $h > h_{\text{ref}}$. The quantity $h$ is the observed plant population density in plants/hectare and $h_{\text{ref}}$ is a fixed reference population density taken, for maize, to be 90 468 (the industry average). Parameters $\eta_1$ and $\eta_2$ are plant population coefficients for populations less than or greater than the standard population, respectively.
TABLE 1

Variables used as inputs in the model

<table>
<thead>
<tr>
<th>Description</th>
<th>Units</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil available N</td>
<td>kg/ha</td>
<td>Nlab</td>
</tr>
<tr>
<td>Fertiliser N (broadcast applications)</td>
<td>kg/ha</td>
<td>Nbroad</td>
</tr>
<tr>
<td>Fertiliser N (banded applications)</td>
<td>kg/ha</td>
<td>Nband</td>
</tr>
<tr>
<td>Soil P (extractable)</td>
<td>µg/ml</td>
<td>P</td>
</tr>
<tr>
<td>Fertiliser P (broadcast applications)</td>
<td>kg/ha</td>
<td>Pbroad</td>
</tr>
<tr>
<td>Fertiliser P (banded applications)</td>
<td>kg/ha</td>
<td>Pband</td>
</tr>
<tr>
<td>Soil exchangeable K</td>
<td>meq/100g</td>
<td>K</td>
</tr>
<tr>
<td>Fertiliser K (broadcast applications)</td>
<td>kg/ha</td>
<td>Kbroad</td>
</tr>
<tr>
<td>Fertiliser K (banded applications)</td>
<td>kg/ha</td>
<td>Kband</td>
</tr>
<tr>
<td>Soil exchangeable Mg</td>
<td>meq/100g</td>
<td>M</td>
</tr>
<tr>
<td>Fertiliser Mg (broadcast applications)</td>
<td>kg/ha</td>
<td>Mbroad</td>
</tr>
<tr>
<td>Fertiliser Mg (banded applications)</td>
<td>kg/ha</td>
<td>Mband</td>
</tr>
<tr>
<td>Soil pH</td>
<td>no units</td>
<td>pH</td>
</tr>
<tr>
<td>Soil density in the field</td>
<td>g/ml</td>
<td>ρfield</td>
</tr>
<tr>
<td>Soil density in laboratory chemical tests</td>
<td>g/ml</td>
<td>ρlab</td>
</tr>
<tr>
<td>Available water capacity of the soil</td>
<td>mm</td>
<td>C</td>
</tr>
<tr>
<td>Maximum soil water deficit</td>
<td>mm</td>
<td>Dmax</td>
</tr>
<tr>
<td>Total evapotranspiration</td>
<td>mm</td>
<td>E</td>
</tr>
<tr>
<td>Plant population</td>
<td>plants/ha</td>
<td>H</td>
</tr>
<tr>
<td>Potential yield at standard population</td>
<td>kg/plant</td>
<td>Ỹ</td>
</tr>
<tr>
<td>Actual yield in t/ha</td>
<td>t/ha</td>
<td>Y</td>
</tr>
</tbody>
</table>

An equivalent multiplier for stress due to inadequate rainfall or irrigation is given by

\[ W = 1 - \frac{\beta(D_{\text{max}} - D_{\text{lim}}C)}{E} \quad \text{when } D_{\text{max}} > D_{\text{lim}}C, \]

and \( W = 1 \) otherwise. Here \( C \) is the soil’s available water capacity, \( D_{\text{max}} \) is the maximum soil water deficit that a crop experiences during growth, \( D_{\text{lim}} \) is the soil water deficit beyond which the crop experiences water stress, \( \beta \) is the daily fractional loss in growth under water stress, and \( E \) is the total evaporation by the crop for the period of growth (Reid, 2002). The quantity \( D_{\text{max}} \) is calculated from weather observations (solar radiation, temperature, rainfall). Wherever possible \( C \) is taken from published independent measurements of water retention by the soil, but in this case at some sites it was estimated indirectly from assessments of soil texture. Soil chemical properties were assessed using the New Zealand standard methods given by Cornforth (1980), except for soil available N which was measured using the anaerobic incubation technique of Keeney & Bremner (1966). Note that exchangeable cation concentrations are given as milliequivalents per 100g of dry soil, which is the New Zealand standard unit.

Tables 1 and 2 give a comprehensive summary of all model parameters and input variables. In Table 2, D.M. stands for “dry matter” and the numerical estimates are for the maize data discussed in Section 4.

2.3. Simple model derivation for one nutrient

One useful approach when fitting complex nonlinear models is to work with a simplified version of the model at first, and subsequently build the model up. Given that its structure is heavily based on the concept of scaling, PARJIB is particularly suited to this approach,
TABLE 2

Model parameters and Reid’s fitted estimates (Reid et al., 2002)

<table>
<thead>
<tr>
<th>Description</th>
<th>Units</th>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min. N supply per unit $Y_{\text{max}}$ to achieve a positive yield</td>
<td>kg N/t D.M.</td>
<td>$N_{\text{min}}$</td>
<td>0.885</td>
</tr>
<tr>
<td>N supply per unit $Y_{\text{max}}$ needed to achieve max. yield</td>
<td>kg N/t D.M.</td>
<td>$N_{\text{opt}}$</td>
<td>16.78</td>
</tr>
<tr>
<td>N response coefficient</td>
<td>no units</td>
<td>$\gamma_N$</td>
<td>0.551</td>
</tr>
<tr>
<td>Efficiency of broadcast N fertilizer compared to N$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{N1}$</td>
<td>0.327</td>
</tr>
<tr>
<td>Efficiency of banded N fertilizer compared to N$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{N2}$</td>
<td>0.613</td>
</tr>
<tr>
<td>Min. P supply per unit $Y_{\text{max}}$ to achieve a positive yield</td>
<td>kg P/t D.M.</td>
<td>$P_{\text{min}}$</td>
<td>0.709</td>
</tr>
<tr>
<td>P supply per unit $Y_{\text{max}}$ needed to achieve max. yield</td>
<td>kg P/t D.M.</td>
<td>$P_{\text{opt}}$</td>
<td>1.068</td>
</tr>
<tr>
<td>P response coefficient</td>
<td>no units</td>
<td>$\gamma_P$</td>
<td>0.217</td>
</tr>
<tr>
<td>Efficiency of broadcast P fertilizer compared to P$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{P1}$</td>
<td>1$^a$</td>
</tr>
<tr>
<td>Efficiency of banded P fertilizer compared to P$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{P2}$</td>
<td>1$^a$</td>
</tr>
<tr>
<td>Min. K supply per unit $Y_{\text{max}}$ to achieve a positive yield</td>
<td>kg K/t D.M.</td>
<td>$K_{\text{min}}$</td>
<td>2.93</td>
</tr>
<tr>
<td>K supply per unit $Y_{\text{max}}$ needed to achieve max. yield</td>
<td>kg K/t D.M.</td>
<td>$K_{\text{opt}}$</td>
<td>81.79</td>
</tr>
<tr>
<td>K response coefficient</td>
<td>no units</td>
<td>$\gamma_K$</td>
<td>0.272</td>
</tr>
<tr>
<td>Efficiency of broadcast K fertilizer compared to K$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{K1}$</td>
<td>1</td>
</tr>
<tr>
<td>Efficiency of banded K fertilizer compared to K$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{K2}$</td>
<td>0.01$^b$</td>
</tr>
<tr>
<td>Min. Mg supply per unit $Y_{\text{max}}$ to achieve a positive yield</td>
<td>kg Mg/t D.M.</td>
<td>$M_{\text{min}}$</td>
<td>0.193$^c$</td>
</tr>
<tr>
<td>Mg supply per unit $Y_{\text{max}}$ needed to achieve max. yield</td>
<td>kg Mg/t D.M.</td>
<td>$M_{\text{opt}}$</td>
<td>0.607$^c$</td>
</tr>
<tr>
<td>Mg response coefficient</td>
<td>no units</td>
<td>$\gamma_M$</td>
<td>0.123$^c$</td>
</tr>
<tr>
<td>Efficiency of broadcast Mg fertilizer compared to Mg$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{M1}$</td>
<td>1$^c$</td>
</tr>
<tr>
<td>Efficiency of banded Mg fertilizer compared to Mg$_{\text{soil}}$</td>
<td>no units</td>
<td>$\xi_{M2}$</td>
<td>1$^c$</td>
</tr>
<tr>
<td>Critical value of soil pH</td>
<td>no units</td>
<td>pH$_{\text{crit}}$</td>
<td>5$^c$</td>
</tr>
<tr>
<td>Slope of scaled yield on soil pH</td>
<td>no units</td>
<td>$\lambda_{\text{pH}}$</td>
<td>0.018</td>
</tr>
<tr>
<td>Fractional reduction in daily growth under water stress</td>
<td>no units</td>
<td>$\beta$</td>
<td>0.89</td>
</tr>
<tr>
<td>Scaled soil water deficit at which water stress begins</td>
<td>no units</td>
<td>$D_{\text{lim}}$</td>
<td>0.538</td>
</tr>
<tr>
<td>Plant population coefficient for population densities $\leq$ standard value</td>
<td>no units</td>
<td>$\eta_1$</td>
<td>0.379</td>
</tr>
<tr>
<td>Plant population coefficient for population densities $&gt;$ standard value</td>
<td>no units</td>
<td>$\eta_2$</td>
<td>0.633</td>
</tr>
</tbody>
</table>

$^a$ Values imprecise due to weak response of yield to P fertilizer in the range of data available for fitting.

$^b$ Value imprecise due to small number of experimental sites where K was banded.

$^c$ Values could not be estimated with accuracy.

D.M. = dry matter.

as large components of the model can be left out or simplified. Here we used a simplified version of the model, achieved by reducing the terms used to define $q_{\text{nut}}$:

$$q_{\text{nut}} = q_N.$$  

Hence, the PARJIB model was reduced to one nutrient (nitrogen) and nine parameters. Note that this is equivalent to assuming that the rest of the nutrients are at optimal supply, and the soil pH is above the threshold. The parameters remaining in the model are $\eta_1$, $\eta_2$, $D_{\text{lim}}$, $\beta$, $\xi_{N1}$, $\xi_{N2}$, $N_{\text{min}}$, $N_{\text{opt}}$ and $\gamma_N$. It is important to note that there are some biological restrictions on the parameters which are not built into the model; in particular all of them need to be positive to be interpretable.

More explicitly, the structure of this simple model may be gleaned from the model function in the R code listed in the Appendix.

3. Adequacy of the original fitting method

The PARJIB model of Reid (2002) was fitted by Reid et al. (2002) to a dataset that collated information from three separate studies of maize crops grown in New Zealand in the
3-year span from 1996 to 1999. Reid et al. (2002) used a genetic algorithm for this fitting. The resulting parameter estimates are given in Table 2.

3.1. Genetic algorithms – background

Genetic algorithms (GAs) are search algorithms that are based on concepts of natural selection and genetics (Holland, 1975). Their transition schemes are probabilistic, and they do not require the model function to be differentiable in the parameters, nor do they need any prior information about the model parameters. This makes them a very convenient nonlinear optimisation tool for fitting models such as PARJIB.

In practice, genetic algorithms have been demonstrated to outperform derivative-based methods in applications with non-differentiable or multi-modal objective functions (Goldberg, 1989). When a function to be optimised is not globally concave, it may have multiple local optima, saddle points, boundary solutions or discontinuous jumps. In these cases, methods of optimisation that depend on derivative information will encounter difficulties unless starting from near-optimal initial values — if the method is able to find any optimum at all, it is unlikely that it will converge to a global optimum. Genetic algorithms are generally more robust to these difficulties. Provided the parameter set that defines the global optimum is within the domains over which the GA is allowed to search, GAs can be more effective at finding the neighbourhood of a global optimum than gradient methods. On the downside, however, the GA can be quite slow to move from a near-optimal point to the exact optimum point (Sekhon & Mebane, 1998).

3.1.1. How good was the fit obtained by the GA?

The genetic algorithm implemented by Reid et al. (2002) was able to obtain parameter estimates that were in accordance with their judgement based on prior knowledge.

However, a drawback of the GAs is that they do not provide any measure of confidence for individual parameter estimates. Chatterjee, Laudato & Lynch (1996) suggested using the bootstrap to estimate standard errors once the genetic algorithm has obtained the parameter estimates of a model. This approach is not pursued here, as it would have been too time-consuming in this application. Instead, the focus was placed on a more detailed exploration of the residual sum of squares surface near the parameter estimates; the measures of confidence for individual parameters in PARJIB are obtained through gradient information and likelihood methods.

Analytic derivatives of the sum of squares function were derived in order to identify if Reid’s parameter estimates are located at an exact optimum value. The derivatives were not equal to zero with respect to all the parameters, indicating that a local optimum has not been reached. Note that the only assurance of the global nature of this optimum comes from the GA itself, as the algorithm has carried out an exhaustive search for such an optimum. In order to improve on these estimates, Reid’s genetic algorithm results are combined with a derivative based method. Reid’s results were set as initial parameter estimates, and the model was then fitted using a standard derivative based method. Hence, while the genetic algorithm may have obtained estimates in the neighbourhood of an optimum, the gradient methods are used to expedite the final convergence, closing in on the optimum point itself. Once the parameter estimates at an optimum are found, standard errors for these parameters are estimated using local approximation and profile likelihood methods.
4. Alternative fitting methods: using simulated data

4.1. Approach

First an attempt was made to fit the model using conventional nonlinear regression in the statistical package R. By default, R uses the Gauss–Newton algorithm which does not work for datasets where the columns of the pseudo-design matrix are almost linearly dependent, as is the case here. Even if the model is successfully fitted, highly correlated parameters will be very poorly estimated, so it is a good idea to keep one member of a highly correlated pair constant.

The structure of the correlation matrix of the parameter estimates can be investigated by fitting the model to a large simulated dataset. It is much easier to fit a model to simulated data, since it is possible to choose the sample size and make the simulated dataset more balanced and less multicollinear than the original data. In addition, because the ‘true values’ of all the parameters are known, it is possible to leave sets of parameters out of the estimation by starting with them constant at their true values and progressively re-introducing them to the model. Furthermore, it is numerically easier to work with a model that actually generated the data to which it is being fitted.

Reid fitted the model to a maize dataset, which was composed of three different sources of measurements of experimental and commercial crops of maize grown in the North Island of New Zealand between 1996 and 1999 (Reid et al., 2002). The dataset contains observations from twelve sites, which differ substantially in observed yield and associated regressor variables related to weather, cultivar, soil properties, plant density and fertilizer applications. Altogether there are 84 observations in the dataset, which is rather few for fitting such a highly parameterized nonlinear model as PARJIB.

Simulated datasets were generated so that they mimicked the real data as much as possible. The histograms of the variables in the maize dataset indicated that \( N_{\text{lab}} \), \( \rho_{\text{field}} \), \( \rho_{\text{lab}} \), \( C \), \( \delta_{\text{max}} \), \( E \), \( h \) and \( \tilde{Y} \) are roughly normally distributed. These variables were thus simulated by drawing random samples from normal distributions. Means and standard deviations for each variable were set equal to the estimates of these parameters obtained from the maize dataset. The sample size could be set at any value in the simulations.

The variables \( N_{\text{broad}} \) and \( N_{\text{band}} \) were simulated in the following way; with probability 0.2 \( N_{\text{band}} \) was taken as zero and \( N_{\text{broad}} \) was drawn from Uniform (95.3, 576.5), otherwise \( N_{\text{broad}} \) was taken as zero and \( N_{\text{band}} \) was drawn from Uniform (122, 250). This mimicked the observed joint distribution of \( N_{\text{broad}} \) and \( N_{\text{band}} \) in the maize data.

The observed yield was generated by the model with the parameters set equal to the estimates obtained by Reid from the genetic algorithm fitting (Reid et al., 2002). To this observed yield we added independent error terms distributed as Normal \((0, \sigma^2)\), where \( \sigma^2 \) was at the disposal of the simulator.

Using this method, the simple model was fitted to a series of simulated datasets with varying sample sizes and error standard deviations. Using the Gauss–Newton algorithm, the smallest dataset to which the model was successfully fitted had 300 observations \((n_{\text{sim}} = 300)\) and a small error standard deviation \((\sigma = 0.1)\). The complete model was fitted with a larger dataset \((n_{\text{sim}} = 50000)\) and a very small residual standard deviation \((\sigma = 0.001)\).

The role of these simulated datasets in this application is to explore a surrogate for the actual log-likelihood surface in order to identify highly correlated parameters. The simulations
Table 3
Correlation matrix of the estimates obtained from one run of fitting the simple model to a simulated dataset

<table>
<thead>
<tr>
<th></th>
<th>$\gamma_N$</th>
<th>$N_{\text{min}}$</th>
<th>$N_{\text{opt}}$</th>
<th>$\xi_{N1}$</th>
<th>$\xi_{N2}$</th>
<th>$\beta$</th>
<th>$D_{\text{lim}}$</th>
<th>$\eta_1$</th>
<th>$\eta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\text{min}}$</td>
<td>-0.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N_{\text{opt}}$</td>
<td>0.47</td>
<td>-0.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\xi_{N1}$</td>
<td>0.02</td>
<td>-0.02</td>
<td>-0.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\xi_{N2}$</td>
<td>0.00</td>
<td>-0.01</td>
<td>-0.02</td>
<td>0.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>-0.36</td>
<td>-0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{\text{lim}}$</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td>0.38</td>
<td>0.02</td>
<td>-0.25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>0.08</td>
<td>-0.04</td>
<td>0.76</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.02</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>0.02</td>
<td>0.01</td>
<td>0.84</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.83</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1. Contour plot of the residual sum of squares surface against $\gamma_N$ and $N_{\text{min}}$ in the simple model.

are based on marginal distributions and, as such, they underestimate the collinearity of the real data, hence fitting the model to them may not yield enough correlated parameters to enable the fit of reduced parameter models to the real data. An alternative way of obtaining simulated datasets would be to take a large sample of observations from the maize dataset and ‘jitter’ the components of the resampled observations, hence retaining the multivariate structure of the data.

4.2. Results

When fitting the simple model, problems were encountered estimating $\gamma_N$, the shape parameter, and $N_{\text{min}}$, the minimum value of nitrogen supply required to produce a positive yield. The correlation matrix in Table 3 indicates that these parameters have estimates which are highly correlated with each other and are thus very poorly estimated. This relationship is illustrated in more detail in Figure 1. Here we plot the contours for the minimized concentrated sum of squares surface in the space of the parameters $\gamma_N$ and $N_{\text{min}}$. The plot shows a steep
valley, along the floor of which the surface changes little. The best values of $\gamma_N$ and $N_{min}$ are found along the line of the bottom of the valley, showing the strong negative linear relationship between the estimates of $\gamma_N$ and $N_{min}$.

Similar patterns of results were obtained in fitting the complete model. As there are 26 parameters fitted in this run, the resulting correlation matrix of the parameter estimates is not displayed, however its general structure may be simply described. First note that the parameters break into five groups, with $X_{min}$, $X_{opt}$, $\gamma_X$, $\xi_X1$ and $\xi_X2$ being associated with nutrient $X$ ($X = N$, P, K and Mg), and the remaining 6 parameters $\beta$, $D_{lim}$, $\eta_1$, $\eta_2$, $pH_{crit}$ and $\lambda_{pH}$ forming another group. The matrix is approximately block-diagonal with no correlations between parameter estimates for parameters from different groups exceeding 0.3. Within the remaining parameters only the correlations between $\beta$ and $D_{lim}$, between $\eta_1$ and $\eta_2$, and between $pH_{crit}$ and $\lambda_{pH}$ exceeded 0.3 and only the last of these exceeded 0.8. In the whole correlation matrix the only pairs of parameters with correlation exceeding 0.8 were: $\gamma_N$ and $N_{min}$, $\xi_{N1}$ and $\xi_{N2}$, $\gamma_P$ and $P_{opt}$, $\xi_{P1}$ and $P_{min}$, $\xi_{P2}$ and $P_{min}$, $\gamma_K$ and $K_{opt}$, $\gamma_M$ and $Mg_{min}$, and $\lambda_{pH}$ and $pH_{crit}$.

5. Alternative fitting methods: using field data

5.1. Fitting the model with the Levenberg–Marquardt algorithm

The simple model was fitted using the Levenberg–Marquardt algorithm, a modification to the Gauss–Newton increment that involves inflating the diagonal of the $X^TX$ matrix in order to transform it to a better-conditioned full rank matrix (Bates & Watts, 1988, p. 81). At the time this was done the Levenberg–Marquardt method was not implemented in R, so code was written for it. The R function written requires the analytically derived derivatives of the fitted values function with respect to all the parameters in the model to be supplied. (The Levenberg–Marquardt algorithm is now available in R through the package minpack, which provides an interface to the MINPACK library of Fortran subroutines for nonlinear optimization.)

The model was fitted to the maize yield dataset with the initial values set at Reid’s estimates from the genetic algorithm (Reid et al., 2002). The residual sum of squares (RSS) obtained from this fitting was 122.7847 (on 9 degrees of freedom), which was a reduction from $RSS = 206.9336$ obtained from Reid’s estimates. The derivatives of the residual sum of squares function with respect to all the model parameters were evaluated for the maize dataset and the parameter estimates ($\hat{\theta}$). These gradient calculations indicated that the algorithm had reached a local optimum, since $\partial S/\partial \hat{\theta}_j$ was very close to zero for all nine parameters ($j = 1, 2, \ldots, 9$).

As with the simulated data, we found difficulties due to the $N_{min}$ parameter. The estimate obtained for $N_{min}$ was a large negative value ($-762.8$) which, given the biological interpretation of the parameter, did not make sense. For the Levenberg–Marquardt algorithm to take into account biological constraints on the parameters when attempting to optimise the fit, the constraints need to be built into the model. To enforce a constraint to positive values the model was re-parameterised; instead of $N_{min}$, a new parameter ($N_{smin}$) was introduced such that $N_{smin}^2 = N_{min}$. The model was refitted and the estimate for $N_{smin}$ was approximately equal to zero ($-5.51 \times 10^{-6}$). The results of this fit imply that any supplied nitrogen will cause a response in yield in the conditions of the experiment, which biologically, may not be
very likely. In fact, parameters being estimated at biologically implausible values are likely to be a symptom of the lack of fit to data of some of the functional forms assumed in the model. In this case, the estimation involves extrapolation beyond the conditions measured in the experiment, as the maize dataset does not contain observations with the supply of nitrogen so low to make the observed yield equal to zero.

In Figure 2 the fitted response curve of scaled yield on nitrogen supply is plotted for the simple model with $N_{\text{min}}$ held constant at three selected values. The regression is quite noisy indicating that yield was influenced by other variables besides $N_{\text{supply}}$. The sub-model ignores the effect of other important nutrients, such as potassium, on the yield. Parameter $N_{\text{min}}$ was set constant at values 0, 0.8852 and 5.5. The simple model was fitted at each fixed value of $N_{\text{min}}$ and response curves of scaled yield on the supply of nitrogen were plotted along with residual plots for each fit.

Figure 2. Illustration of the effect that different ($\gamma_N$, fixed $N_{\text{min}}$) pairs had on the response curve of scaled yield on nitrogen supply (simple model).
TABLE 4
Estimates of the eight parameters in the simple model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate ($\hat{\theta}_j$)</th>
<th>Std error</th>
<th>$\partial \bar{y}/\partial \hat{\theta}_j$</th>
<th>95% Wald CI</th>
<th>95% Likelihood CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N_{\text{opt}}$</td>
<td>16.248</td>
<td>3.95</td>
<td>2E−06</td>
<td>(8.391, 24.106)</td>
<td>(10.411, 27.767)</td>
</tr>
<tr>
<td>$\gamma_N$</td>
<td>0.679</td>
<td>0.26</td>
<td>1E−05</td>
<td>(0.157, 1.201)</td>
<td>(0.315, 2.157)</td>
</tr>
<tr>
<td>$\xi_{N1}$</td>
<td>0.218</td>
<td>0.10</td>
<td>−3E−05</td>
<td>(0.012, 0.423)</td>
<td>(0.076, ∞)</td>
</tr>
<tr>
<td>$\xi_{N2}$</td>
<td>0.255</td>
<td>0.10</td>
<td>−2E−05</td>
<td>(0.058, 0.452)</td>
<td>(0.111, 0.543)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.600</td>
<td>0.11</td>
<td>1E−04</td>
<td>(0.377, 0.823)</td>
<td>(0.377, 0.826)</td>
</tr>
<tr>
<td>$D_{\text{lim}}$</td>
<td>0.222</td>
<td>0.12</td>
<td>−1E−04</td>
<td>(−0.012, 0.455)</td>
<td>(−0.125, 0.413)</td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>0.612</td>
<td>0.12</td>
<td>−2E−05</td>
<td>(0.371, 0.853)</td>
<td>(0.377, 0.859)</td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>0.650</td>
<td>0.08</td>
<td>5E−05</td>
<td>(0.485, 0.815)</td>
<td>(0.462, 0.798)</td>
</tr>
</tbody>
</table>

TABLE 5
Parameter correlation matrix of the simple model fitted to the maize data

<table>
<thead>
<tr>
<th></th>
<th>$N_{\text{opt}}$</th>
<th>$\gamma_N$</th>
<th>$\xi_{N1}$</th>
<th>$\xi_{N2}$</th>
<th>$\beta$</th>
<th>$D_{\text{lim}}$</th>
<th>$\eta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_N$</td>
<td>-0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\xi_{N1}$</td>
<td>0.37</td>
<td>-0.27</td>
<td></td>
<td></td>
<td>0.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\xi_{N2}$</td>
<td>0.62</td>
<td>-0.44</td>
<td>0.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.36</td>
<td>-0.19</td>
<td>-0.13</td>
<td>0.39</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_{\text{lim}}$</td>
<td>0.52</td>
<td>-0.32</td>
<td>-0.13</td>
<td>0.39</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\eta_1$</td>
<td>-0.12</td>
<td>0.10</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.23</td>
<td>-0.09</td>
<td>-0.39</td>
</tr>
<tr>
<td>$\eta_2$</td>
<td>-0.24</td>
<td>0.18</td>
<td>-0.01</td>
<td>-0.21</td>
<td>-0.63</td>
<td>-0.35</td>
<td>-0.39</td>
</tr>
</tbody>
</table>

The plots in Figure 2 show how the response curve changes shape for different fixed values of $N_{\text{min}}$; as $N_{\text{min}}$ gets smaller, the curve gets flatter. The fit improves for smaller values of $N_{\text{min}}$; for $N_{\text{min}}$ fixed at 0, 0.8852 and 5.5, the RSS is calculated as 123.76, 123.85 and 124.92 respectively. However, the changes in the response curves are only very slight for this range of $N_{\text{min}}$ values, and the plots of residuals on fitted values are almost the same for all three fits. This would indicate that the maize dataset does not contain a lot of information on $N_{\text{min}}$. Consequently, we fixed the parameter $N_{\text{min}}$ at zero and treated it as a constant. The parameter estimates and their standard errors, obtained from the expected information matrix, are given in Table 4. The correlation matrix of this fit is given in Table 5. As in previous fits, the Levenberg–Marquardt algorithm was used.

Removing from the model one of the parameters, $N_{\text{min}}$, in a highly correlated pair greatly reduced the estimated variance of the other parameter $\gamma_N$. In this application we have a relatively small amount of trial data compared to the complexity of the model and, as a result, the estimated standard errors are fairly large. This problem is further exacerbated by the fact that not all the observations are independent; the maize dataset is composed of measurements from twelve different sites, and any correlation of observations within sites would also have had an effect on standard errors. Three options suggest themselves for dealing with this problem: do more experimentation and collect more information about the parameters; add more knowledge external to the trials about any of the parameters; or further reduce the size of the model being fitted.

In general, the problems of correlated estimates and poor precision of estimation in certain directions are common for nonlinear models. The problems are usually caused by the
\( X(\theta)^\top X(\theta) \) matrix being singular, or nearly so. Another problem that often occurs in a nonlinear setting is that of parameter unidentifiability, which is also signalled by ill-conditioning of the \( X(\theta)^\top X(\theta) \) matrix. However, the difficulties associated with unidentifiability may stem from the structure of the model and the method of parameterization rather than unfortunate experimental design (Seber & Wild, 1989, p. 126). The so-called structural relationships in the model that cause the identifiability problems appear often in PARJIB. One common way of dealing with this problem is to impose identifiability constraints that identify which solution is required (Seber & Wild, 1989, p. 102). In this application it is recommended to collect more information about the approximately unidentifiable parameters in the model, either by theoretical argument or additional experimentation. For example, further study could be specifically directed at measuring efficiencies of fertilizer forms.

5.2. Profile likelihood methods

Linear approximation intervals obtained from the information matrix are likely to be misleading for models with either high intrinsic or parameter-effects nonlinearity. An additional disadvantage of the linear approximation methods is that the validity of the approximation over the region of interest is not known (Bates & Watts, 1988). In general, the estimation situation will be highly nonlinear for most nonlinear models with many parameters and relatively few observations. These models may exhibit near-linear behaviour for very large sample sizes and small residual variance (Ratkowsky, 1983, p. 183), but in practice such datasets are often beyond the resources of the experimenter. In order to obtain more accurate summaries of inferential results for the parameter estimates in PARJIB, profile likelihood methods were employed. Such methods are described in more detail by Bates & Watts (1988, section 6.1.2).

The approach for multi-parameter models is to evaluate the (one-dimensional) profile likelihood function by varying a parameter of interest (say \( \theta_j \)) over fixed values, optimising the objective function over the other parameters. The profile \( t \) function is given by

\[
\tau(\theta_j) = \text{sign}(\theta_j - \hat{\theta}_j) \frac{\sqrt{S(\theta_j) - S(\hat{\theta})}}{s},
\]

where \( \hat{\theta}_j \) is the model estimate of \( \theta_j \), \( S(\theta_j) \) is the residual sum of squares based on optimising all parameters except the fixed \( \theta_j \), and \( S(\hat{\theta}) \) is the residual sum of squares at \( \hat{\theta} \). A profile plot of the parameter \( \theta_j \) is a plot of \( \tau(\theta_j) \) against a range of values for the parameter. A \((100 - \alpha)\%\) profile likelihood interval for \( \theta_j \) is defined as the set of all \( \theta_j \) for which

\[-t(n - p; \alpha/2) \leq \tau(\theta_j) \leq t(n - p; \alpha/2),\]

where \( n \) is the number of observations and \( p \) is the number of parameters.

In addition to providing likelihood intervals for individual parameters, profile plots can serve as a general way of assessing validity of the linear approximation over the domain of interest; if the estimation situation is linear, the plot of \( \tau(\theta_j) \) on \( \theta_j \) will be a straight line, but any deviations from straightness indicate that the linear approximation might be misleading in that direction.

The R function `profile` can evaluate the profile \( t \) function for parameters in a nonlinear model fitted using the `nls` package. When the simple model was fitted to a simulated dataset (with sample size = 300 and \( \sigma = 0.1 \)), the resultant profile functions were almost linear. The
resulting profile plots are given in Figure 3. The plots for parameters $\gamma_N$ and $N_{\min}$ look slightly skewed and the estimated confidence intervals are somewhat asymmetric. For the rest of the parameters the surface seems relatively linear, so the linear approximation confidence intervals are adequate. Hence, for a large sample with small residual variance, the PARJIB model has near-linear behaviour.

A new profile function was written in order to work out profiles for the simple model fitted to the maize dataset using the Levenberg–Marquardt algorithm. The new function works out the t-statistic equivalent $\tau(\theta_j)$ on a user-selected grid of values around the least square estimates. The profile plots for the simple model fitted to the maize dataset are given in Figure 4 (in this fit $N_{\min}$ is set equal to 0 kg N/ha.) The $95\%$, $90\%$ and $80\%$ likelihood confidence intervals are marked on these profile plots. Note that for some parameters the curves do not reach $95\%$ or $90\%$ confidence interval horizontal lines, meaning that the corresponding confidence interval is one sided.

Since $n = 84$ and $p = 8$, the numbers $\pm 1.9917$, $\pm 1.6652$ and $\pm 1.2928$ are the critical values for $\tau(\theta_j)$ bounding the $95\%$, $90\%$ and $80\%$ likelihood confidence intervals respectively. The linear interpolation method was used to obtain the parameter values $\theta_j$ for which $-t(n - p;\alpha/2) \leq \tau(\theta_j) \leq t(n - p;\alpha/2)$. The results are given in Table 4.

© 2006 Australian Statistical Publishing Association Inc.
Figure 4. Profile plots for the simple model fitted to the maize data. The 95%, 90% and 80% likelihood confidence intervals are marked with dashed lines. Vertical dotted lines mark 95% linear approximation confidence intervals.

The profile plot of $\xi_{N1}$ is strongly curved and tends to an asymptote, indicating nonlinearity. The likelihood interval is skewed and does not close for 90%. Since the nonlinearity assumption is violated, it can be concluded that the standard errors given in Table 4 do not accurately summarise the uncertainty in this parameter. For parameter $D_{lim}$, the curve will reach the lower 90% confidence interval horizontal line only when $D_{lim}$ is negative, which is outside the biological constraints for this parameter. The profile plots of parameters $D_{lim}$, $N_{opt}$, $\gamma_N$ and $\xi_{N2}$ are somewhat nonlinear and the corresponding likelihood confidence intervals are skewed, hence they differ from the linear approximation confidence intervals (see Table 4). In the region of the 95% likelihood confidence intervals, the surface seems relatively linear with respect to the parameters $\beta$, $\eta_1$ and $\eta_2$. For these parameters the likelihood intervals are almost identical to the linear approximation confidence intervals and the standard errors are adequate as a summary of the uncertainty of the parameter estimates.
Profile plots provide likelihood intervals for each parameter and reveal how nonlinear each parameter is, but they do not offer any information on how the parameters interact. This information can be extracted from the contour plots of profile log-likelihoods of pairs of parameters presented in Figure 5 and the correlation matrix in Table 5. For example, the third contour plot in the sixth row of Figure 5 shows that the parameters $D_{lim}$ and $\beta$ are strongly related in the sense that specifying the value of either shrinks the plausible range of values of the other.

6. Conclusions

Nonlinear semi-mechanistic models have the advantage of enhanced interpretability when compared to purely descriptive multiple regression models. However, when they incorporate large numbers of parameters, they may prove very difficult to fit with standard nonlinear regression software. Genetic algorithms may be able to find solutions that appear to be reasonable, but the resultant parameter estimates come without standard errors. Attempts to use nonlinear regression with starting values supplied by a genetic algorithm may fail because the pseudo-design matrix is numerically less than full rank when evaluated at those starting values. In practice, it can be too costly to obtain enough experimental data to adequately test the model at the mechanistic level, and this kind of failure may be common when the size of the dataset is not large.

This case study suggests a method for using standard nonlinear regression software to fit these models despite such obstacles. The method involves constructing a large artificial dataset whose predictor variables mimic those of the original dataset in terms of their marginal distributions and whose response variable values are generated from the model with only a small amount of random error. The true parameter values used to generate the simulated dataset may be taken as the values found using the genetic algorithm applied to the original data. The model may then be fitted to the simulated data starting from the true parameter values. If the artificial dataset is large enough and the random error in the response variable is small enough, the fitting of the model is unlikely to be troublesome.

Standard nonlinear regression output includes a matrix of correlations between parameter estimates. From this matrix we may identify pairs of parameters that are highly correlated and which may lead to difficulties in fitting the model to the original data. We suggest fixing one variable from each of the most highly correlated pairs and estimating only the remaining parameters using the original data. If this fit is obtained, attempts may be made to incorporate further parameters into the optimization. In this application, due to a high correlation with the shape parameter ($\gamma_N$), the $N_{min}$ term was removed from the model. This is a difficult parameter to measure experimentally, and the model fitting process strongly suggests further experimental study could be dedicated to determining the correct level at which the parameter could be set (see below). Alternatively, effort could be directed to identifying a mechanistically satisfying formulation of the model that did not require such minimum terms for nutrients to be separated from the curve shape parameters.

This work illustrates the value for the scientist in exploring the correlation structure in mechanistic or semi-mechanistic models. The process can show what parts of the model are poorly determined or validated by the data. This might then lead to various solutions. As a first resort, parameter values might be fixed at values determined from outside the data, that is, from prior subject-area knowledge. Alternatively, it may be possible to conduct further
Figure 5. Contour plots of profile log-likelihoods (residual sum of squares) of pairs of parameters in the ‘simple model’.
experimentation targeted at understanding a particular sub-process governed by a poorly understood parameter.

For example the minimum N, P, K and Mg levels for successful maize growth could be investigated in a further series of experiments measuring the presence or absence of growth when one of these four nutrients is at or near a stressfully low level, the other nutrients being in adequate supply.

Finally the information on correlation structure might be used to reappraise the structure of the model itself, especially if the experimental evidence is not strong enough to allow estimation of a parameter free of assumptions about the value of others. Thus statistical modelling and analysis can complement mechanistic studies, making more explicit what is known and what is not known about the processes being modelled and thereby guiding further research.

Appendix

The R code for fitting the simple model to simulated data follows.

```r
# initial parameter assignments
Nmin <- 0.8852
Nopt <- 16.78
gN <- 0.5511
E.n1 <- 0.3271
E.n2 <- 0.6132
beta <- 0.8902
delta <- 0.5378
eta1 <- 0.3791
eta2 <- 0.6332
PopStd <- 90468

# define model function
Y.model <- function(gN, Nmin, Nopt, delta, beta, eta1, eta2, E.n1, E.n2) {
    Ymax <- 1 - ifelse(Popn <= PopStd, eta1, eta2) * log(Popn / PopStd)
    Ymax <- Ymax * PotYield3 * Popn / 1000
    diff <- Dmax - delta * AWC
    Ymax <- Ymax * ifelse(Dmax <= delta * AWC, 1, 1 - beta * diff / SumEp)
    Nsupply <- Nsoil * Bdfield / Bdlab + Nfert.broad * E.n1 + Nfert.band * E.n2
    Nstar <- (Nsupply - Nmin * Ymax) / (Nopt * Ymax - Nmin * Ymax)
    Nstar <- pmax(0, Nstar)
    Ystar <- ifelse(Nstar < 1, (1 + gN * (1 - Nstar)) * Nstar * (1 + gN), 1)
    Ystar <- pmax(0, Ystar)
    Y.model <- Ystar * Ymax
    Y.model
}
```

© 2006 Australian Statistical Publishing Association Inc.
# simulate experimental data for predictors
#set.seed(71201)
nsim <- 300
Popn <- rnorm(nsim,PopStd,0.1*PopStd)
Dmax <- rnorm(nsim,140.68,47.45)
AWC <- rnorm(nsim,186.86,47.41)
SumEp <- rnorm(nsim,318.54,32.53)
PotYield3 <- rnorm(nsim,0.16180,0.01167)
Nsoil <- rnorm(nsim,94.07,34.06)
Bdfield <- rnorm(nsim,1.0590,0.1420)
Bdlab <- rnorm(nsim,0.7876,0.1169)
Nfert.broad <- runif(nsim,95.3,576.5)
Nfert.band <- runif(nsim,122,250)
broad <- rbinom(nsim,1,0.2)
Nfert.broad <- Nfert.broad * broad
Nfert.band <- Nfert.band * (1 - broad)
error <- rnorm(nsim,0,1)
scale <- 0.1

# generate response variable from model
Y <- Y.model(gN, Nmin, Nopt, delta, beta, eta1, eta2, E.n1, E.n2)
Y <- Y + scale * error

# attempt to fit starting from true parameters
simparj.st <- c(gN, Nmin, Nopt, delta, beta, eta1, eta2, E.n1, E.n2)
names(simparj.st) <- c('N', 'Nmin', 'Nopt', 'delta', 'beta', 'eta1', 'eta2', 'E.n1', 'E.n2')
simparj.fm <- nls(Y ~ Y.model(gN, Nmin, Nopt, delta, beta, eta1, eta2, E.n1, E.n2), start = simparj.st, trace = T)
summary(simparj.fm)

References


© 2006 Australian Statistical Publishing Association Inc.


