SEMI-MECHANISTIC MODELLING IN NONLINEAR REGRESSION: A CASE STUDY

KATARINA DOMIJAN[1*], MURRAY JORGENSEN[2] and JEFF REID[3]

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

Summary

We discuss the use of highly parameterized semi-mechanistic nonlinear models with particular reference to the PARJIB crop response model of Reid (2002). Compared to empirical linear approaches, such models promise improved generality of application but present considerable challenges for estimation. We have had some success 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 we describe an approach to identifying 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.

---

*Author to whom correspondence should be addressed.

[1] Department of Statistics, Trinity College, University of Dublin, Dublin 2, Ireland.
e-mail: domijank@tcd.ie
[2] Department of Statistics, University of Waikato, Private Bag 3105, Hamilton, New Zealand
[3] New Zealand Institute for Crop and Food Research Ltd, RD2 Hastings, New Zealand

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 theory 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.

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 we call them here *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 pseudodesign 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 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. We discuss 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. In particular, we show 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_{max}, is the yield that would be achievable in the absence of mineral nutrient stresses (Reid, 2002). Modelled yield (Y_{model}) is obtained from Y_{max} and what Reid called the scaled yield (Y^*):

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

The maximum yield variable (Y_{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,

$$Y_{max} = \frac{\tilde{Y} H h W}{1000}, \qquad (1)$$

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_{nut})

$$Y^* = q_{nut} \,. \qquad (2)$$

In this paper we consider two ways of calculating the q_{nut} term; the standard way outlined by Reid (2002), and a 'simple model' that involves fewer nutrients. In this particular case, the input variable $\widetilde{Y}$ was calculated using Wilson, Muchow & Murgatroyd's (1995) modification of the potential yield model presented by Muchow, Sinclair & Bennett (1990). The quantities Y_{max} and Y are by definition constrained to be positive and less than $\widetilde{Y}$. Variables and parameters used to calculate $\widetilde{Y}$ 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 $\widetilde{Y}$ and the input D_{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. (We will often speak of nutrient X, 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

$$X_{supply} = X_{soil} + X_{broad}\, \xi_{X1} + X_{band}\, \xi_{X2}.$$

Here X_{soil} is the amount of X present in the soil before fertilizer application, while X_{broad} and X_{band} refer to the amount of fertiliser X applied to the soil in kg/ha in broadcast and banded applications respectively. The parameters ξ_{X1} and ξ_{X2} respectively denote the efficiency of supplying X in broadcast or banded fertilizer form compared to X_{soil}.

All nutrient response curves rise from 0 when $X_{supply} \le X_{\min}$ to a maximum of 1 when $X_{\text{supply}} \ge X_{opt}$. 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 $X_{\min}$ is supplied and 1 when X_{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

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

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

In PARJIB then, the three parameters $X_{\min}$, X_{opt} and γ_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. We will not be considering pH stress 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

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

The plant population multiplier H is defined by

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

where $\eta = \eta_1$ when $h \le h_{ref}$ and $\eta = \eta_2$ when $h > h_{ref}$. The quantity h is the observed plant population density in plants/hectare and h_{ref} is a fixed reference population density taken, for maize, to be 90468 (the industry average). Parameters η_1 and η_2 are plant population coefficients for populations less than or greater than the standard population, respectively.

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

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

and W=1 otherwise. Here C is the soil's available water capacity, D_{max} is the maximum soil water deficit that a crop experiences during growth, D_{lim} is the soil water deficit beyond which the crop experiences water stress, β 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_{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.

TABLE 1

Variables used as inputs in the model.

Description	Units	Variable
Soil available N	kg/ha	N_{lab}
Fertiliser N (broadcast applications)	kg/ha	N_{broad}
Fertiliser N (banded applications)	kg/ha	N_{band}
Soil P (extractable)	µg/ml	P
Fertiliser P (broadcast applications)	kg /ha	P_{broad}
Fertiliser P (banded applications)	kg /ha	P_{band}
Soil exchangeable K	meq/100g	K
Fertiliser K (broadcast applications)	kg /ha	K_{broad}
Fertiliser K (banded applications)	kg/ha	K_{band}
Soil exchangeable Mg	meq/100g	M
Fertiliser Mg (broadcast applications)	kg /ha	M_{broad}
Fertiliser Mg (banded applications)	kg /ha	M_{band}
Soil pH	no units	pH
Soil density in the field	g/ml	ρ_{field}
Soil density in laboratory chemical tests	g/ml	ρ_{lab}
Available water capacity of the soil	mm	C
Maximum soil water deficit	mm	D_{max}
Total evapotranspiration	mm	E
Plant population	plants/ha	H
Potential yield at standard population	kg/plant	$\widetilde{Y}$
Actual yield in t/ha	t/ha	Y

TABLE 2

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

Description	Units	Parameter	Estimate
Min. N supply per unit Y_{max} to achieve a positive yield	kg N/t D.M.	N_{min}	0.885
N supply per unit Y_{max} needed to achieve max. yield	kg N/t D.M.	N_{opt}	16.78
N response coefficient	no units	γ_N	0.551
Efficiency of broadcast N fertilizer compared to N_{soil}	no units	ξ_{N1}	0.327
Efficiency of banded N fertilizer compared to N_{soil}	no units	ξ_{N2}	0.613
Min. P supply per unit Y_{max} to achieve a positive yield	kg P/t D.M.	P_{min}	0.709
P supply per unit Y_{max} needed to achieve max. yield	kg P/t D.M.	P_{opt}	1.068
P response coefficient	no units	γ_P	0.217
Efficiency of broadcast P fertilizer compared to P_{soil}	no units	ξ_{P1}	1[a]
Efficiency of banded P fertilizer compared to P_{soil}	no units	ξ_{P2}	1[a]
Min. K supply per unit Y_{max} to achieve a positive yield	kg K/t D.M.	K_{min}	2.93
K supply per unit Y_{max} needed to achieve max. yield	kg K/t D.M.	K_{opt}	81.79
K response coefficient	no units	γ_K	0.272
Efficiency of broadcast K fertilizer compared to K_{soil}	no units	ξ_{K1}	1
Efficiency of banded K fertilizer compared to K_{soil}	no units	ξ_{K2}	0.01[b]
Min. Mg supply per unit Y_{max} to achieve a positive yield	kg Mg/t D.M.	M_{min}	0.193[c]
Mg supply per unit Y_{max} needed to achieve max. yield	kg Mg/t D.M.	M_{opt}	0.607[c]
Mg response coefficient	no units	γ_M	0.123[c]
Efficiency of broadcast Mg fertilizer compared to Mg_{soil}	no units	ξ_{M1}	1[c]
Efficiency of banded Mg fertilizer compared to Mg_{soil}	no units	ξ_{M2}	1[c]
Critical value of soil pH	no units	pH_{crit}	5[c]
Slope of scaled yield on soil *pH*	no units	λ_{pH}	0.018
Fractional reduction in daily growth under water stress	no units	β	0.89
Scaled soil water deficit at which water stress begins	no units	D_{lim}	0.538
Plant population coefficient for population densities ≤ standard value	no units	η_1	0.379
Plant population coefficient for population densities > standard value	no units	η_2	0.633

[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.

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, 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_{nut}:

$$q_{nut} = q_N .$$

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 η_1, η_2, D_{lim}, β, ξ_{N1}, ξ_{N2}, N_{min}, N_{opt} and γ_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 three 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, GAs 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 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 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 our 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, we have combined Reid's genetic algorithm results 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

Firstly 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 pseudodesign 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_{lab}, ρ_{field}, ρ_{lab}, C, δ_{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_{broad} and N_{band} were simulated in the following way; with probability 0.2 N_{band} was taken as zero and N_{broad} was drawn from Uniform (95.3, 576.5), otherwise N_{broad} was taken as zero and N_{band} was drawn from Uniform (122, 250). This mimicked the observed joint distribution of N_{broad} and N_{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, σ^2), where σ^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_{sim} = 300) and a small error standard deviation (σ = 0.1). The 'complete model' was fitted with a larger dataset (n_{sim} = 50000) and a very small residual standard deviation (σ = 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 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 us to fit 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 we found problems estimating the shape parameter γ_N, and N_{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 γ_N and N_{min} . The plot shows a steep valley, along the floor of which the surface changes little. The best values of γ_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 γ_N and N_{min}.

Similar patterns of results were obtained in fitting the 'complete model'. As there are 26 parameters fitted in this run, we do not display the resulting correlation matrix of the parameter estimates, however its general structure may be simply described. Firstly note that the parameters break into five groups, with $X_{\min}, X_{opt}, \gamma_X, \zeta_{X1}$, and ζ_{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 λ_{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 β and $D_{\lim}$, between η_1 and $\eta_{2,}$, and between pH_{crit} and λ_{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: γ_N and N_{min}, ξ_{N1} and ξ_{N2}, γ_P and P_{opt}, ξ_{P1} and P_{min}, ξ_{P2} and P_{min}, ξ_{P1} and ξ_{P2}, γ_K and K_{opt}, γ_M and Mg_{min}, and λ_{pH} and pH_{crit}.

TABLE 3

Correlation matrix of the estimates obtained from one run of fitting the 'simple model' to a simulated dataset.

	γ_N	N_{min}	N_{opt}	ξ_{N1}	ξ_{N2}	β	D_{lim}	η_1
N_{min}	-0.99							
N_{opt}	0.47	-0.46						
ξ_{N1}	0.02	-0.02	-0.02					
ξ_{N2}	0.00	-0.01	-0.02	0.70				
β	-0.02	0.02	0.00	-0.36	-0.03			
D_{lim}	0.01	-0.01	0.01	0.38	0.02	-0.25		
η_1	0.08	-0.04	0.76	-0.02	-0.02	0.02	0.00	
η_2	0.02	0.01	0.84	-0.02	-0.01	0.02	-0.01	0.83

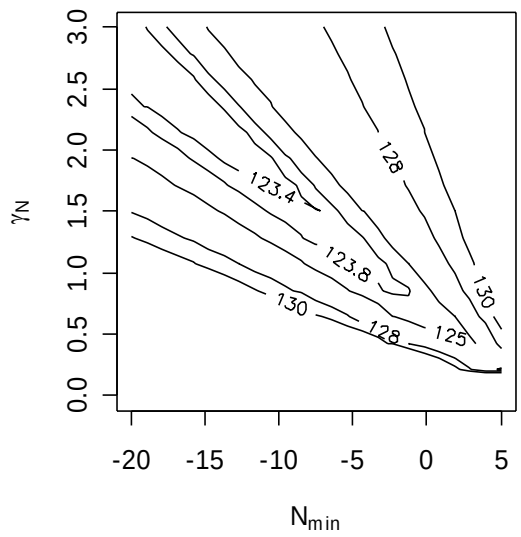

Figure 1. Contour plot of the residual sum of squares surface against γ_N and N_{min} in the 'simple model'.

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 $\boldsymbol{X}^T\boldsymbol{X}$ 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,…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 (-762.8265) 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 ($-5.51e^{-06}$). 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_{min} held constant at three selected values. The regression is quite noisy indicating that yield was influenced by other variables besides N_{supply}. The sub-

model ignores the effect of other important nutrients, such as potassium, on the yield. Parameter N_{min} was set constant at values 0, 0.8852 and 5.5. The 'simple model' was fitted at each fixed value of N_{min} and response curves of scaled yield on the supply of nitrogen were plotted along with residual plots for each fit.

The plots in Figure 2 show how the response curve changes shape for different fixed values of N_{min}; as N_{min} gets smaller, the curve gets flatter. The fit improves for smaller values of N_{min}; for N_{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_{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_{min}. Consequently, we fixed the parameter N_{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_{min}, in a highly correlated pair greatly reduced the estimated variance of the other parameter γ_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 $\boldsymbol{X(\theta)^T X(\theta)}$ 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 $\boldsymbol{X(\theta)^T X(\theta)}$ 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.

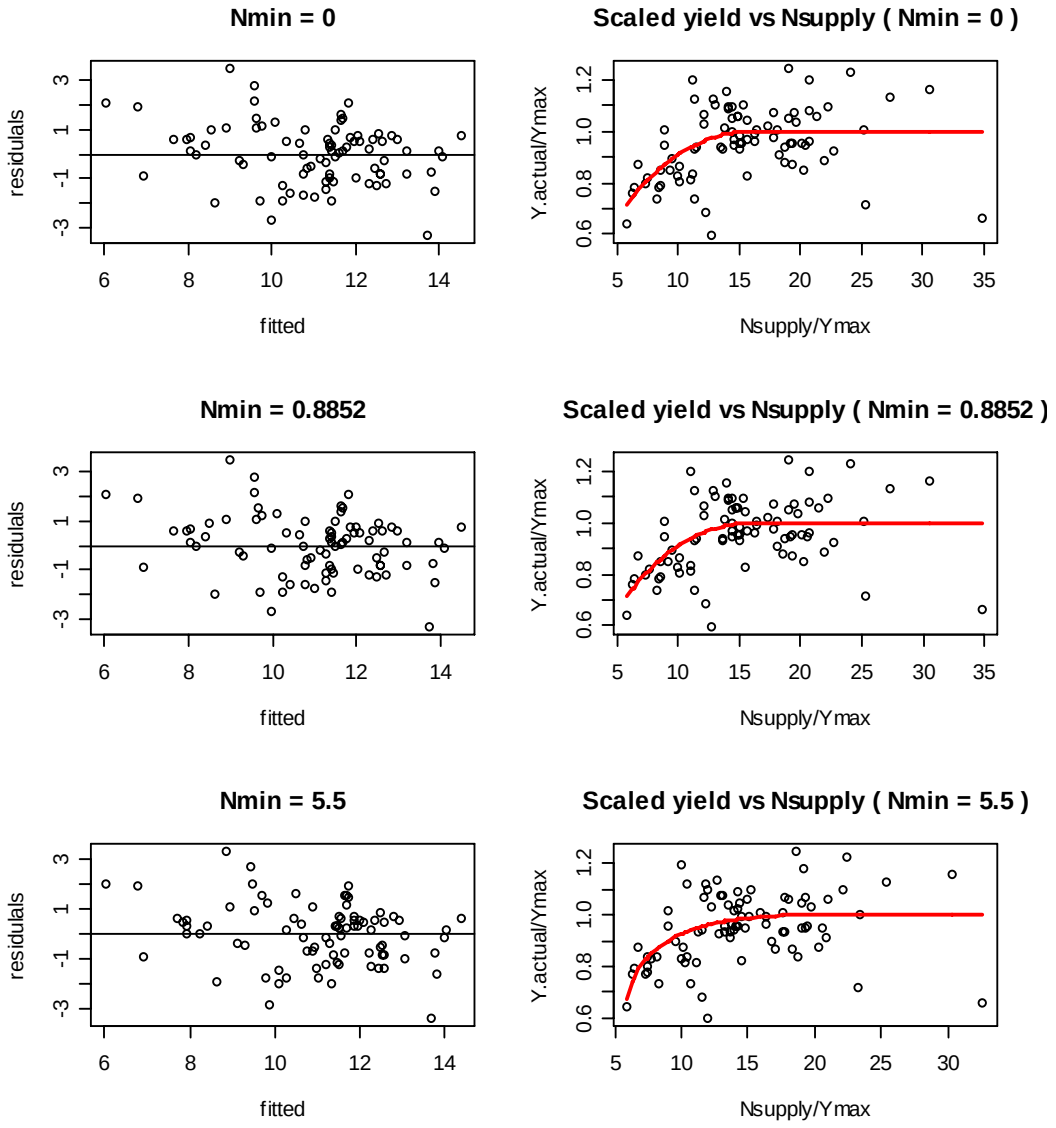

Figure 2. Illustration of the effect that different ($\hat{\gamma}_N$, fixed N_{min}) pairs had on the response curve of scaled yield on nitrogen supply (simple model).

TABLE 4

Estimates of the eight parameters in the 'simple model'.

Parameter	Estimate ($\hat{\theta}_j$)	Std error	$\partial S/\partial\hat{\theta}_j$	95% Wald CI	95% Likelihood CI
N_{opt}	16.248	3.95	2E-06	(8.391, 24.106)	(10.411, 27.767)
γ_N	0.679	0.26	1E-05	(0.157, 1.201)	(0.315, 2.157)
ξ_{N1}	0.218	0.10	-3E-05	(0.012, 0.423)	(0.076, ∞)
ξ_{N2}	0.255	0.10	-2E-05	(0.058, 0.452)	(0.111, 0.543)
β	0.600	0.11	1E-04	(0.377, 0.823)	(0.377, 0.826)
D_{lim}	0.222	0.12	-1E-04	(-0.012, 0.455)	(-0.125, 0.413)
η_1	0.612	0.12	-2E-05	(0.371, 0.853)	(0.377, 0.859)
η_2	0.650	0.08	5E-05	(0.485, 0.815)	(0.462, 0.798)

TABLE 5

Parameter correlation matrix of the 'simple model' fitted to the maize data.

	N_{opt}	γ_N	ξ_{N1}	ξ_{N2}	β	D_{lim}	η_1
γ_N	-0.9						
ξ_{N1}	0.37	-0.27					
ξ_{N2}	0.62	-0.44	0.32				
β	0.36	-0.19	-0.13	0.39			
D_{lim}	0.52	-0.32	-0.13	0.39	0.86		
η_1	-0.12	0.10	-0.02	-0.02	0.23	-0.09	
η_2	-0.24	0.18	-0.01	-0.21	-0.63	-0.35	-0.39

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, we employed profile likelihood methods. 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 θ_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 θ_j, $S(\theta_j)$ is the residual sum of squares based on optimising all parameters except the fixed θ_j, and $S(\hat{\boldsymbol{\theta}})$ is the residual sum of squares at $\hat{\boldsymbol{\theta}}$. A profile plot of the parameter θ_j is a plot of $\tau(\theta_j)$ against a range of values for the parameter. A (100-α)% marginal likelihood interval for θ_j is defined as the set of all θ_j

for which $-t(n\text{-}p;\ \alpha/2) \leq\ \tau(\theta_j)\ \leq t(n\text{-}p;\ \alpha/2)$, 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 θ_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 γ_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.

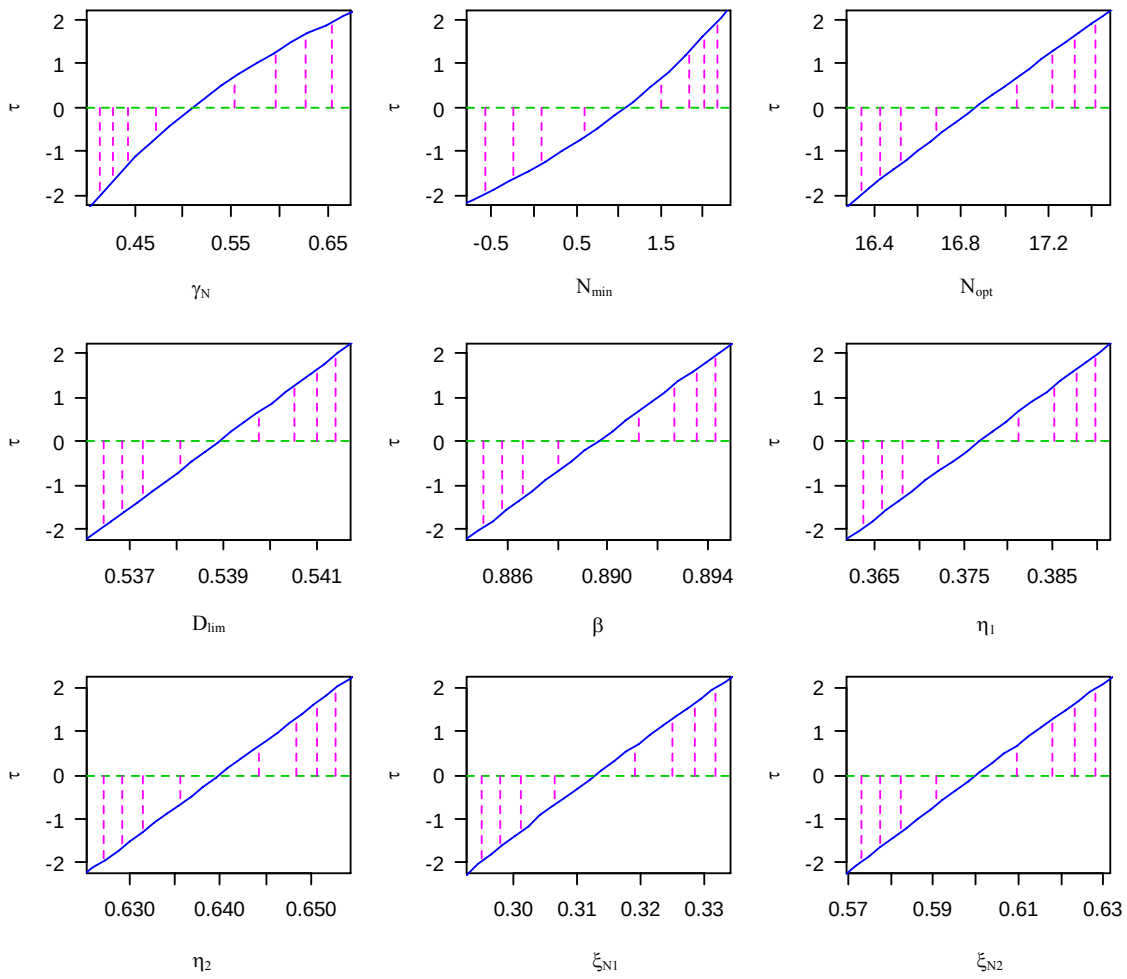

Figure 3. Profile plots for the 'simple model' fitted to a simulated dataset. Marked are 80%, 90%, 95%, and 99% confidence intervals.

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.

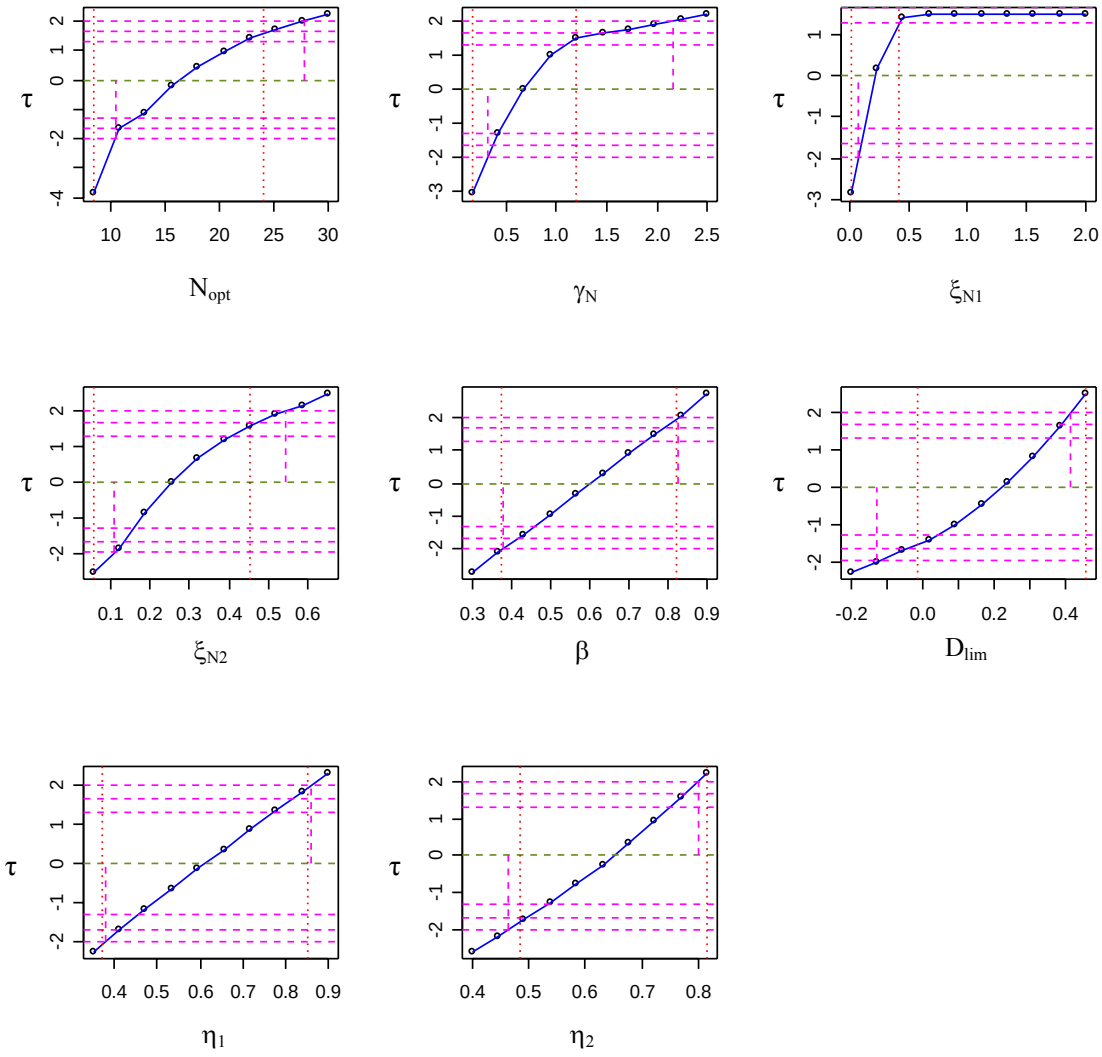

Figure 4. Profile plots for the 'simple model' fitted to the maize data. The 95%, 90% and 80% likelihood confidence intervals are marked with dashed lines. Vertical dotted lines mark 95% linear approximation confidence intervals.

Since n = 84 and p = 8, the numbers ± 1.9917, ±1.6652 and ±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 θ_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.

The profile plot of ξ_{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}, γ_N and ξ_{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 β, η_1 and η_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 β 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 (γ_N), we removed the N_{min} term 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 experimentation targeted at understanding a particular sub-process governed by a poorly understood parameter.

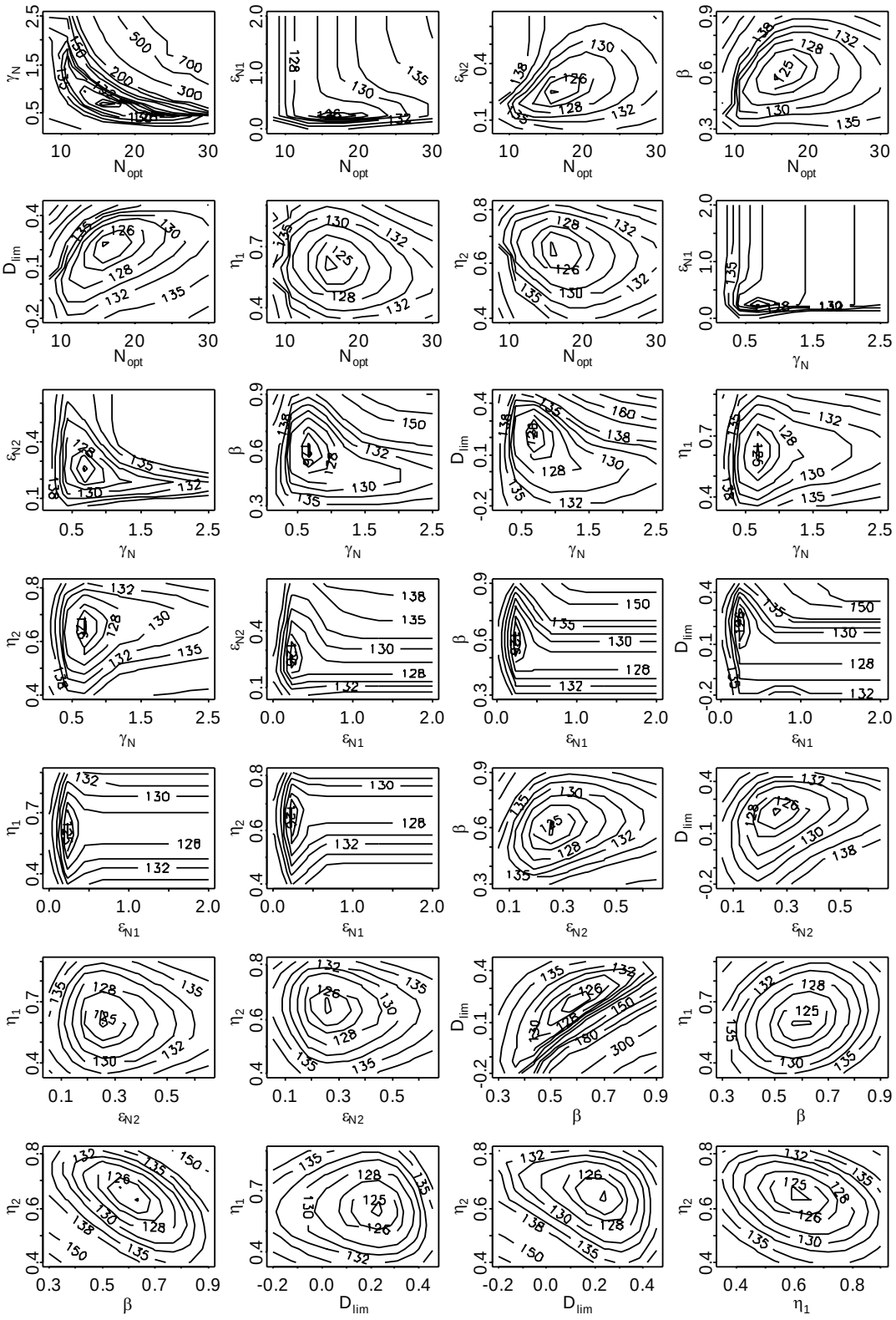

Figure 5. Contour plots of profile log-likelihoods (residual sum of squares) of pairs of parameters in the 'simple model'.

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.

```
library(nls)
# 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
      }
# 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('gN', '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

BATES, D. M. & WATTS, D. G. (1988). *Nonlinear Regression Analysis and its Applications*. New York: Wiley.

CHATTERJEE, S., LAUDATO, M. & LYNCH, L.A. (1996). Genetic Algorithms and their Statistical Applications: an Introduction. *Computational Statistics and Data Analysis* **22**, 633-651.

CORNFORTH, I.S., (1980). *Soils and fertilizers - soil analysis and interpretation*. Aglink AST 8. Media Services, NZ Ministry of Agriculture and Fisheries, Wellington.

GOLDBERG, D.E. (1989). *Genetic Algorithms in Search, Optimisation and Machine Learning*. Reading, Mass.: Addison-Wesley.

HOLLAND, J. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.

KEENEY, R.R., & BREMNER, J.M., (1966). Comparison and evaluation of laboratory methods of obtaining an index of soil nitrogen availability. *Agronomy Journal* **58**, 493-503.

MUCHOW, R.C., SINCLAIR, T.R. & BENNETT, J.M. (1990). Temperature and solar radiation effects on potential maize yield across locations. *Agronomy Journal* **82**, 338-343.

RATKOWSKY, D.A. (1983). *Nonlinear Regression Modelling*. New York: Dekker.

REID, J.B. (2002). Yield Response to Nutrient Supply Across a Wide Range of Conditions 1. Model Derivation. *Field Crops Research* **77***,* 161-171.

REID, J. B., STONE, P. J., PEARSON, A. J. & WILSON, D. R. (2002). Yield Response to Nutrient Supply Across a Wide Range of Conditions 2. Analysis of Maize Yields. *Field Crops Research* **77**, 173-189.

SEBER, G.A.F. & WILD, C.J. (1989). *Nonlinear Regression*. New York: Wiley.

SEKHON, J. S. & MEBANE, W. R. Jr. (1998). Genetic Optimisation Using Derivatives: Theory and Application to Nonlinear Models. *Political Analysis* **7**, 189-203.

THORNLEY, J. H. M & JOHNSON, I. R. (1990). *Plant and Crop Modelling: a Mathematical Approach to Plant and Crop Physiology*. Oxford : Clarendon Press.

WILSON, D.R., MUCHOW, R.C. & MURGATROYD, C.J. (1995). Model analysis of temperature and solar radiation limitations to maize potential productivity in a cool climate. *Field Crops Research* **43**, 1-18.

Comment on
Semi-Mechanistic Modelling in Nonlinear Regression: A Case Study

by A.H. Welsh
Centre for Mathematics and its Applications, The Australian National University, Canberra ACT 0200

Mechanistic modelling is very widely used in science, often without reference to Statistics or even as an alternative in opposition to it. This paper combines both approaches by developing a particular mechanistic model and then treating it statistically. It is timely because there is increasing interest in applying statistical methods to fit and use mechanistic models; it is stimulating because it highlights issues which arise in bringing Mechanistic Modelling and Statistics together and provides a good focus for thought and discussion on how these approaches might interact

The main statistical contribution in this paper is the explicit recommendation that it is useful to identify singularities and near singularities in the $X(\theta)^T X(\theta)$ matrix in a nonlinear model (i.e. the nonlinear model equivalent of mulitcollinearity) and the proposal of a method for doing so through detecting "highly correlated parameters". The simulation method can be viewed as a parametric bootstrap where some features may be determined by data but others are not. For example, in the paper, the values of explanatory variables and the model parameters used are informed by data but the variety of sample sizes and the intentionally small standard deviations (say, $\sigma = 0.1$, 0.01 or 0.001) of the independent and identically distributed Gaussian errors are not. Simulating very large samples with small errors is intended to allow the definitive identification of problems which persist even under ideal conditions: varying the settings should show how rapidly things become more difficult in realistic settings.

In the paper, the p explanatory variables are generated independently from marginal distributions $F_1, \ldots, F_p$ suggested by a set of data and the value of θ used is the parameter estimate obtained by fitting the model to these data by a genetic algorithm. The first choice is discussed briefly in the paper but the second may also be controversial: If the Gauss-Newton algorithm fails because of lack of identifiability, then other algorithms may produce numbers which are not meaningful and exploring the model at or near this point in the parameter space (and indeed other analyses based on this point) will also be devoid of meaning. It may well be better to try to move around the parameter space in a sensible way such as by an appropriate Markov Chain Monte Carlo algorithm than to use a fixed point. Importantly, both of these choices can be varied to taste so should not detract from the basic idea which may prove very useful.

The authors point out that singularities and near singularities in the $X(\theta)^T X(\theta)$ matrix may be due to unfortunate sets of data or to structural relationships between variables. They recommend treating near singularities by collecting more

information either by theoretical argument or additional experimentation. In either case this seems to come down to fixing the value of a parameter so that it is no longer estimated from the current data. This is a pragmatic approach but it can introduce biases, essentially because of a generalized version of the fallacy of averages (nonlinear functions of means are not the same as means of nonlinear functions). Whether this effect matters or not can be evaluated by assessing fit.

It is unfortunate that the paper does not emphasise the assessment of fit, predictive ability etc. as a primary concept. The authors mention a need to "adequately test the model at the mechanistic level" and imply that (statistically) fit is measured by an information matrix or standard errors. I don't understand the first statement and feel that the assessment of fit (like model-fitting or estimation) is intrinsically statistical. In addition to evaluating the proposed mean function, an absolutely critical aspect of the model is the assumption of independent, additive, homoscedastic errors which are tacked on to the model more or less as an afterthought. A more complete discussion should consider the use of diagnostics and other sophisticated (statistical) evaluations of fit, predictive ability etc. Promoting these assessments makes explicit the scientifically vital connection between models and data and puts into context the advantages of mechanistic modelling: What is the value of enhanced interpretability if the model does not fit? The fact that mechanistic models can be used for extrapolation into extreme circumstances does not mean that they should be so used, particularly without any assessment of fit, and, even when some assessment of fit has been made, extrapolation beyond regions where fit has been assessed remains just as problematic as in empirical modelling.

This paper describes the current distinction between empirical and mechanistic models eloquently and accurately. It reduces the distinction by introducing statistical methods to fit the model and we can reduce it further in our evaluations of the model. However, both of these tend to come after the modelling exercise is complete, at a stage when the model seems to be regarded as subject to only minor adjustment. Mechanistic modelling could benefit from far greater attention to data throughout the modelling process. This would encourage focus on fundamental issues such as whether the explanatory variables can be observed, the identification of sources of error and variability in the observation process (which should be incorporated into the model), whether any data are available (particularly important if the model is dynamic and needs to be fitted to historical data), whether there are relationships between explanatory variables, and consequent lack of identifiability of parameters and so on. It would lead to a clearer understanding of what are variables (observables), errors or latent variables and parameters (unobservables) and make possible the adoption of the sensible statistical convention of distinguishing in notation between these quantities. It would place the model on a firmer basis and make it more likely to fit, it would improve the process and the outcome. And it might help to reinforce the statistical insight that even after doing all we can, the model may still be wrong.

Discussion—Semi-Mechanistic Modelling in Nonlinear Regression: A Case Study

K.B. Newman

Domijan, Jorgensen, and Reid (DJR) present approaches to detecting and remedying parameter identifiability issues in realistic but very complex nonlinear models. The problem of estimating parameters for even the "simple" model based on a single nutrient can be daunting. The model structure is described in pieces in the article, and the full complexity of the model can be better appreciated by seeing it written as a single function. Let $\theta = (\gamma_N, N_{min}, N_{opt}, D_{lim}, \beta, \eta_1, \eta_2, \xi_{N1}, \xi_{N2})$ denote the unknown parameters, $\psi \equiv h_{ref}$ be a fixed value, and $\mathbf{x}$= (h, D_{max},C, E, $\tilde{Y}$,Bd_{field},Bd_{lab},N_{soil}, N_{broad},N_{band}) denote the covariates. The simple nitrogen-based model for yield, y, is as follows.

$$f(\theta, \psi, \mathbf{x}) = Y_{max} q_{nut} \equiv \left(\tilde{Y} \frac{h}{1000} HW\right) q_{nut}$$

where

$$\begin{aligned} H &= 1 - [I_{h \le h_{ref}} \eta_1 + I_{h > h_{ref}} \eta_2] \log\left(\frac{h}{h_{ref}}\right) \\ W &= \min\left(1, 1 - \frac{\beta(D_{max} - D_{lim} C)}{E}\right) \\ q_{nut} &= (1 + \gamma_N) N^{\gamma_N} - \gamma_N N^{1+\gamma_N} \end{aligned}$$

with

$$N = \min\left[1, \max\left(0, \frac{N_{soil} \frac{Bd_{field}}{Bd_{lab}} + \xi_{N1} N_{broad} + \xi_{N2} N_{band} - N_{min} Y_{max}}{N_{opt} Y_{max} - N_{min} Y_{max}}\right)\right]$$

And this is the simple model!

Several questions come to mind upon seeing such a complicated model: (1) Why such a complex model? (2) Are all the parameters identifiable, and if so, which ones are well identified?; (3) If some parameters are unidentifable or weakly identifiable, what is a remedy?

1 Why such a complex model?

Scientists want to understand the mechanicisms underlying natural phenomena, to identify the causal factors, and to explain how particular inputs, say $\mathbf{x}$, combine and interact to produce a particular output, y. Mathematical models of the general form $y = f(\mathbf{x}, \theta)$ are a means of quantifying the relationship between the imputs and the output. There are two problems, one to determine the

nature of the function f and the other to specify, or estimate, the parameters θ. Statistics can play an important part in the attempt to solve both problems and efforts like DJR's are important.

One can imagine a continuum where at one end are black box models and at the other are fully mechanistic models. Black box models have been defined as models "which try to give an efficient description of the input-output behavior of a system without relying on hypotheses about how the system works internally" (Brun, Reichert, and Künsch 2001; Beck 1987), while fully mechanistic models attempt to explain exactly, or with a relatively high degree of realism, how the inputs interact and combine to yield the output. Empirical linear models, such as multiple linear regression and generalized linear models, often lie somewhat nearer to the black box end of the continuum in that they are often models of convenience, whereby parameter estimation and fitting can be relatively mechanical and easy. At the same time, transformations of y and $\mathbf{x}$ (e.g., Box-Cox and Box-Tidwell transformations (Myers 1990, Chapter 7)) and other remedies can be used to "shoehorn" somewhat realistic nonlinear models into a linear model structure. The PARJIB crop response model, labeled a semi-mechanistic model, is of course closer to the fully mechanistic end of the continuum, having a combination of mathematically convenient components (as in the linear model for total nitrogen supply, N_{supply}) and more realistic, mechanistic parts (as in calculation of the water stress multiplier, W).

If the parameters of a semi- or fully-mechanistic model can be estimated and *if* the resulting model more accurately and precisely predicts y for a given $\mathbf{x}$ than an empirical linear model, then such mechanistic models are generally preferable because of their greater explanatory and predictive power. In this paper, DJR focus on parameter estimation, e.g., using a genetic algorithm to come up with an initial estimate of θ, which is then used as an initial value for a more stringent optimization routine, and determining which parameters can be well estimated. The second *if*, the predictive ability of the resulting model, is not explored in this paper, however previous work (Reid, et al. 2002) reports that 66-73% of observed yield variation was accounted for in a cross-validation analysis of real crop data. Incidentally, it would be interesting to repeat this analysis using the the Levenberg-Marquardt estimates.

While mechanistic models can be scientifically attractive, black box models, in particular algorithmic models (Breiman 2001) or statistical learning methods (Hastie, Tibshirani, and Friedman 2001), do have potential explanatory and predictive capabilities, and a direct comparison with such complex semi-mechanistic models could be informative. Algorithmic models include decision

trees (e.g., CART), neural nets, generalized additive models, and multivariate adaptive regression splines (MARS). With algorithmic modeling algorithms are used to "find" the functional form, f, and model validation is measured by predictive accuracy. The resulting functional forms are not necessarily simple to interpret. However, as argued by Breiman (2001), information can be gained about the relative importance of input variables by using a combination of cross-validation and random permuting of a single covariate in the test sets. Permuted covariates which have the largest prediction errors in the test set can be interpreted as being the more important or influential covariates.

2 Parameter identifiability?

There are at least three kinds of identifiability, or conversely unidentifiability: theoretical, practical, and weak. Parameters are theoretically unidentifiable if they cannot be estimated for any data set. For example, if $x_i \overset{iid}{\sim} N(\xi + \theta, \sigma^2)$, $i = 1, \ldots, n$, $\xi + \theta$ is identifiable, but not ξ and θ separately (Lehman 1983, p 335). Parameters are practically unidentifiable if they cannot be estimated for a given data set, but could be estimated with another data set. For example, in multiple linear regression with two predictors, where $E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, β_1 and β_2 are practically unidentifiable when $|r_{x_1,x_2}|$=1, but they are theoretically unidentifiable only if this perfect correlation exists for any possible sample. Weak identifiability is perhaps less well-defined, but one could say it refers to a parameter that is theoretically identifiable, and practically identifiable for a particular data set, but relatively minor changes in the data can lead to considerable changes in the parameter estimate and thus the variance of the estimate is large.

2.1 Covariate structure.

It is worth emphasizing the effect of the structure of the inputs or the experimental layout, $\mathbf{x}$, on identifiability. In the case of linear regression models, orthogonal experiment designs yield uncorrelated estimated regression coefficients. In the case of nonlinear regression models, experiment designs that are optimal in some sense, e.g., D-optimal, where the generalized variance of the parameter estimates is minimized, depend upon the covariate structure and the value of θ (Atkinson 1996), thus *a priori* estimates of the parameters are needed. The complexity of the PARJIB model appears to be far beyond those considered for optimal designs in most papers. An additional direc-

tion of research would be to explore this literature with regard to the effect of covariate structure on parameter estimation for the PARJIB model and other complex crop response models. The design problem is also interesting in that some of the covariates are beyond the control of the experimenter (e.g., the covariates involved in the water stress expression, W), and design considerations will thus need to focus on those that are manipulable, such as the choice of nutrient levels. Design considerations could also guide simulations used to assess identifiability problems.

DJR's simulated covariate data were generated assuming independence between the covariates (excepting N_{broad} and N_{band}). They recognize that the removal of dependence between covariates will have some effect on the degree of dependence amongst parameter estimates, likely reducing the dependency, and they mention jittering a resampling of the observations as a remedy. Using a multivariate normal approximation is another alternative with continuous input variables. Depending on the particular model and data, it could indeed be important to work with simulated data that nearly matches the observed data in terms of means, variances, and covariances.

2.2 Detecting and quantifying identifiability problems

For simple nonlinear models or those with few parameters, determining whether or not all the parameters are theoretically identifiable can sometimes be done relatively easily. With more complex models with many parameters, such as PARJIB, it is generally not easy, and the failure of an optimization routine to converge could be practical unidentifiability not theoretical.

With the PARJIB model and the observed data set, there is neither theoretical nor practical unidentifiability given that least squares estimates are, after some work, calculated. But there are weak identifiability problems which DJR clearly identify by means of standard errors and contour plots of the profile log-likelihoods of pairs of parameters. As a check on the results, I used their `R` code for the simple model to repeatedly generate samples of observations and responses and calculate least squares estimates of the parameters. Extreme between sample variation in some of the parameter estimates was indeed borne out, but the estimates were on average accurate. My simulations did reveal some slight differences with those shown in Table 2, however, in that the correlations r_{N_{opt},γ_N} and $r_{N_{opt},N_{min}}$ had the same magnitude as those shown but opposite signs, e.g., r_{N_{opt},γ_N}=-0.61 and $r_{N_{opt},N_{min}}$=0.59. Also I consistently got a moderate positive $r_{\beta,D_{lim}}$, e.g., 0.73, in contrast to the value of -0.25.

As an aside, the program AD Model Builder (Otter Research Ltd., Sidney, BC, Canada), which uses automatic differentiation for optimization calculations, proved to be considerably more stable than R's `nls` function for finding the least squares estimates. For roughly 7% of the simulations, `nls`, failed to converge even when initial values equalled the true values; this never happened for AD Model Builder. Also, for a large percentage of simulations, AD Model Builder was able to converge using initial values not equal to the true values (the initial value for all parameters was 0.5 except for N_{opt} which started at 15), and R's `nls` never did.

While such simulations, calculations of variance-covariance matrices, and profile likelihood contour plots are useful, a potential shortcoming of is that attention is largely focused on studying pairwise relationships between parameter estimates while it could be the case that higher order dependencies exist. A method described by Brun, Reichert, and Künsch (2001) overcomes this shortcoming. There are similarities with diagnostics for multicollinearity in linear models and the method is sketched briefly here. The estimated covariance matrix is

$$\widehat{Var}\left[\hat{\theta}\right] = \frac{SSE}{n-p}\mathbf{V}'\mathbf{V}^{-1} \tag{1}$$

where

$$\mathbf{V} = \frac{df(\theta)}{d\theta'}\mid_{\theta=\hat{\theta}} \tag{2}$$

is an n by p matrix of derivatives evaluated at $\hat{\theta}$, the least squares estimate of θ. In the case of linear regression, where $\mathbf{X}$ is the design matrix, the (estimated) covariance matrix of the least squares estimates is proportional to $(\mathbf{X}'\mathbf{X})^{-1}$. Near linear dependencies between the columns of $\mathbf{X}$ lead to one or more large values on the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$, i.e., large variance inflation factors (Myers 1990, pp 126-127), hence instability in the estimates. Similarly, near linear dependencies in $\mathbf{V}$ translate into large variances for some of the parameter estimates.

The matrix $\mathbf{V}$, labeled the sensitivity matrix by Brun, et al. (2001), measures the sensitivity of y ($\approx f(\theta, \mathbf{x})$) to small changes in θ near $\hat{\theta}$. A dimension-free measure of the sensitivity is the scaled sensitivity matrix, $\mathbf{S}$, where the i,jth element is

$$s_{ij} = v_{ij}\frac{\Delta\theta_j}{sc_i}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p. \tag{3}$$

The term $\Delta\theta_j$ is an a priori measure of the range of θ_j and sc_i is a scale factor proportional to the physical dimension of the ith observation. The degree of linear independence between the parameter estimates is then assessed via the smallest eigenvalue for the normalized scaled sensitivity matrix,

$\tilde{\mathbf{S}}$, where the jth column is

$$\tilde{\mathbf{s}}_j = \frac{\mathbf{s}_j}{||\mathbf{s}_j||}. \quad (4)$$

In particular the collinearity index for any subset K of k parameters, denoted γ_K, is defined by

$$\gamma_K = \frac{1}{\sqrt{\lambda_k}} \quad (5)$$

where λ_k is the smallest eigenvalue in the submatrix $\tilde{\mathbf{S}}_K$. Critical values of γ_K indicative of linear dependence are between 5 and 20; i.e., if $\gamma_K < 5$, then linear dependence is not a serious problem.

The normalized scaled sensitivity matrix can be calculated for various subsets of the parameters and when a subset leads to a linearly dependent, unidentifiability is present. Weak identifiability is characterized by the degree of linear dependence. Gadkar, Gunawan, and Doyle (2005), using a very similar approach, identify practically unidentifiable parameters as those with infinitely large confidence intervals, i.e., infinitely large variances based on equation (1).

Brun et al. (2001) provide additional measures based on $\mathbf{S}$ to screen important parameters, but here I will just use their collinearity index for the simple model. To calculate the sensitivity matrix, $\mathbf{V}$, a finite difference approximation (Ridder 1982), as implemented by Press et al. (1992), was used to estimate the derivatives. It was evaluated at the least squares estimates of θ based on a simulated data set (with a sample size of 300). The normalized scaled sensitivity matrix (equation 4) was next calculated; due to cancellations the terms $\Delta\theta_j$ need not be calculated and because the simulated values were of essentially the same magnitude the scale factors sc_i were set equal to 1. Brun et al. (2001) advise that γ_K be calculated for subsets K out of the total number of parameters and plot γ_K against subset size. Since the subset should contain at least two parameters, that means calculating $\binom{9}{2} + \binom{9}{3} + \ldots + \binom{9}{9} = 502$ γ_K's. Knowing *a posteriori* that N_{min} was the most likely problem, just a few cases were examined. To begin, γ_K for all 9 variables was calculated; it equalled 10.5, indicative of dependencies (using the 5-20 scale for critical values). Excluding N_{min}, the collinearity index based on the remaining eight parameters was γ_K=4.4, which is below the threshold value of 5.

This method can also be used to assess identifiability issues prior to trying to fit a nonlinear model by substituting an *a priori* estimate of θ for $\hat{\theta}$ into the sensitivity matrix calculation.

3 Remedies

Once identifiability problems have been found, DJR discuss several solutions, including (1) fixing the weakly identifiable parameters at values based on prior subject-area knowledge; (2) conducting further experimentation aimed at better understanding the processes; (3) modifying the model structure if there are not enough data to estimate parameters without making assumptions about some of the (weakly identifiable) parameters. These are all pragmatic, achievable remedies.

For the first solution of fixing parameter values, one might also carry out simulations using fixed parameter values different from the chosen values to look for bias in the estimates of the other parameters. For example, with the simple model, 100 simulations with n=300 were carried out using N_{sim}=0.8852, but for least squares estimation N_{min} was fixed at 0. The average values of the parameter estimates for the remaining 8 parameters appeared to be unbiased. This first solution could also be tackled somewhat differently by putting the model into a Bayesian framework and using relatively narrow priors on the weakly identifiable parameters to at least allow for some uncertainty in the parameters. Related to the second solution of conducting further experimentation, one could also try to find experiment designs for the manipulable variables that are in some sense optimal for estimating parameters of primary interest. Lastly, regarding the third solution, as was mentioned previously algorithmic models, e.g., MARS, might be an alternative or a means of providing guidance for restructuring the model.

References

Atkinson, A.C. (1997). The usefulness of optimum experiment designs. *Journal of the Royal Statistical Society, Series B* **58**, 59-76.

Beck, M.B. (1987). Water quality modeling; A review of the analysis of uncertainty. *Water Resources Research* **23**, 1393-1442.

Brun, R., Reichert, P., and Künsch, H.R. (2001). Practical identifiability analysis of large environmental simulation models. *Water Resources Research* **37**, 1015–1030.

Gadkar, K.G., Gunawan, R., and Doyle, F.J. (2005). Iterative approach to model identification of biological networks. *BMC Bioinformatics* **6**: 155; doi:10.1186/1471-2105-6-155.

Lehman, E.L. (1983). *Theory of Point Estimation*. New York: John Wiley and Sons.

Press, W.H, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, B.P. (1992). *Numerical Recipes in C, 2nd Ed.* New York: Cambridge University Press.

Ridders, C.J.F. (1982). *Advances in Engineering Software*, vol 4, no. 2, 75-76.

Myers, R.H. (1990). *Classical and Modern Regression with Applications, 2nd Ed.* Belmont, California: Duxbury Press.

Commentary on Domijan, Jorgensen and Reid, Semi-mechanistic modelling in nonlinear regression: A case study

Martyn Plummer

International Agency for Research on Cancer

The contrast between the "empirical" and "mechanistic" approaches to statistical modelling has a long history. Breslow (2003) reviewed this history in his presidential address to the XXI International Biometric Conference, noting that one of the earliest statisticians to make the distinction was Ronald Ross, a pioneer of infectious disease epidemiology. Ross (1916) used the terms *a posteriori* and *a priori* to distinguish the path of reasoning from observations to causes or vice versa. The success of Ross's mechanistic or *a priori* model for the spread of Malaria has, however, not frequently been replicated in other areas of biometry. Empirical model building has become the standard approach.

In the case of the crop-response model considered by Domijan, Jorgensen and Reid (2006), the best criterion for judging a model would appear to be the quality of its predictions of crop yield. The attraction of a mechanistic, or semi-mechanistic model in this case is that its predictions might extrapolate better outside of the field testing conditions, when compared to an empirical model based on a mathematically convenient, but otherwise arbitrary, dose-response curve. The disadvantage of moving outside the realm of empirical modelling is that we lose the use of standard modelling tools. This includes not only software for fitting the models, but also the ability to fit the problem into an existing conceptual framework that allows us to think about and criticise the model.

One of the key concepts of model criticism is parsimony. Statisticians are not just averse to highly parameterised models out of "professional skepticism" but from an understanding that over-parameterised models have poor reproducibility and, in particular, give worse predictions than appropriately parsimonious models. I hope that nobody would seriously attempt to fit a linear regression model with 26 parameters to a set of only 84 observations. I would expect, instead, that a statistician would either search for a subset of the best predictors, or use a technique such as ridge regression to attenuate the impact of having highly correlated parameters in the model. No such techniques are available for a complex, non-linear model based on *a priori* reasoning, and Domijan, Jorgensen and Reid (2006) have had to develop new techniques for simplifying the model, while retaining its essentially mechanistic character.

I would like to point out some ways in which the same problem can be addressed in a Bayesian framework. A Bayesian hierarchical model can be constructed as a directed acyclic graph (DAG), in which the nodes represent the model variables and the edges represent the causal relationships between them (Pearl 2000). The popular WinBUGS software (Spiegelhalter *et al*, 2004) provides a language for describing such models and has, for example, been adapted to problems in population pharmacokinetics/pharmacodynamics (Lunn *et al* 2002) for which mechanistic reasoning plays an important role in model building.

An important difference between the Bayesian and frequentist paradigms is the need to specify a distribution that encapsulates the prior information about the parameters. In practice, most people use weakly informative or "reference" priors in order to minimize the sensitivity of the conclusions to the prior. However, for a Bayesian purist, the prior distribution is an intrinsic part of the specification of the model. Arguably, if the model discussed by Domijan, Jorgensen and Reid (2006) were analyzed in the Bayesian paradigm, it would require an informative prior distribution on each parameter: one that had been carefully elicited from expert opinion. The authors note that some of the parameters in the model must be bounded to be biologically meaningful. It is only a small step from this to asking what ranges of values are biologically plausible and using this *a priori* knowledge to improve the predictions from the model.

Most modern Bayesian analyses use Markov Chain Monte Carlo (MCMC) techniques, which generate a sequence of correlated samples from the posterior distribution. In principle, this sequence provides all of the information about the parameters, and no further simulations are required to explore the joint posterior distribution. Hence, for example, the correlation between two parameters can be examined by looking at the corresponding correlations between simulated parameter values in the MCMC output. Furthermore, the posterior distribution of a parameter can be estimated from the sampled values using density estimation, and compared with the prior to see how much has been learned from the data.

The conceptual simplicity of the Bayesian approach is no panacea, however. Numerical problems in maximizing the likelihood do not disappear when MCMC techniques are used, but generally manifest themselves as poor "mixing" of the Markov Chain. When this occurs, even a long run of MCMC output may not be a representative sample from the posterior distribution. This may especially occur when the posterior distribution is multimodel, as is the case in mixture models, and special techniques such as simulated tempering may be required (Celeux, Hurn and Robert, 2000). Since the starting point of the paper by Domijan, Jorgensen and Reid (2006) is the difficulty of numerically maximising the likelihood function, no guarantee can be made that the numerical aspects of the problem could be improved by a Bayesian approach. However, numerical problems may well be improved by the use of informative priors, and successful application of MCMC methods would allow model criticism without requiring further methodological innovations.

References

BRESLOW, N.E. (2003) Are statistical contributions to medecine undervalued? *Biometrics* **59**, 1–8.

CELEUX, G., HURN, M. & ROBERT, C.P. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistal Association* **95** 957–970.

DOMIJAN, K., JORGENSEN, M. & REID, J. (2006) Semi-mechanistic modelling in non-

linear regression: A case study. *Australian and New Zealand Journal of Statistics* (in press).

LUNN, D.J., BEST, N.G., THOMAS, A., WAKEFIELD, J. & SPIEGELHALTER, D.J. (2002) Bayesian analysis of population PK/PD models: general concepts and software. *J Pharmacokinetics Pharmocodynamics* **29**, 271–307.

PEARL, J. (2000) Causality: models, reasoning and inference. Cambridge University Press.

ROSS, R. (1916) An application of the theory of probabilities to the study of *a priori* pathometry, part 1. *Proceedings of the Royal Society, Series A* **92**, 204–230.

SPIEGELHALTER, D., THOMAS, A., BEST, N. & LUNN, D. (2004) WinBUGS User Manual Version 2.0.

Discussion of

"Semi-mechanistic modelling in nonlinear regression: A case study"

The fitting of nonlinear models, especially models that are derived from some simplified description of reality, is something that is dear to my statistical heart. It has certainly been interesting to see the approach described by the work under discussion (hereafter DJR), and the simulations demonstrated will undoubtedly provide some general intuition about the high-dimensional shape of the likelihood. However, I am concerned that this approach has some undesirable features that will limit its practicability and require it be be used with extreme caution lest it should mislead. I am also perplexed by DJR's lack of use of established statistical protocols.

It was unclear to me why the fundamental paradigms of statistical practice (e.g., falsification, parsimony) were not employed by DJR. An example of model development via falsification is given by a couple of PhD students (based at the Leigh Marine Laboratory, University of Auckland) who recently obtained strong empirical evidence that the existing (semi-mechanistic) model of fertilization in marine invertebrates was inadequate. This model (Vogel *et al.*, 1982) predicted that fertilization rates would increase towards 100% as the concentration of egg and sperm became sufficiently high. However, the PhD students had data that showed strong evidence of a decline in fertilization rate at extreme gamete concentrations (Franke, Babcock & Styan, 2002). This prompted a more complex model to be developed (Millar and Anderson, 2003), again on semi-mechanistic grounds. The students are currently comparing the simpler model to the more complex model, via confrontation with data, to see if the simpler model can be falsified as a general model of fertilization in marine invertebrates.

I never got the sense that the PARJIB model, or any subset of it, was truly confronted by the data. DJR's examination of correlation structure and likelihood surfaces is an examination of the likelihood in the neighbourhood of its optimum, but this is only a small part of the art of model fitting. I felt that the exposition fell short of presenting clear guidance, due to straying from the usual statistical path, and it was not clear to me that DJR achieved the fit of a parsimonious mechanistic model to these data. It is our statistical responsibility to inform the zealous client that best doesn't mean biggest.

The modelling objectives of DJR were unclear to me. In the Introduction, in the context of finding least-squares solutions to nonlinear models, DJR say *"Nevertheless, the quality of such solutions need to be evaluated by calculation of an information matrix and exploration of the likelihood surface in the neighbourhood of the solution."* This suggested to me that, rather than seeking a parsimonious model

(by mimimizing AIC, say) the objective was to fit as complex a model as possible, up to some arbitrary (unspecified) point of being over-parameterized. Indeed, in Section 5.1, DJR write *"Removing from the model one of the parameters, N_{min}, in a highly correlated pair greatly reduced the estimated variance of the other parameter, γ_N.*

A number of other issues arise from DJR not using a statistical framework for model fitting.

- The simple model assumed that nutrients P, K and Mg, were all at optimal supply. Is this model reasonable?

- There is no explicit notion of competing models or nested models. In consequence, I felt that advise such as *"We suggest fixing one variable from each of the most highly correlated pairs..."* would be rather nebulous in practice. In the PARJIB example, N_{min} was set to zero, but what does one do in general? For example, when fitting size-at-age curves to fish, the growth-rate parameter and maximum-size parameter typically have very high correlation, but it doesn't make sense to set either to zero, or to any particular fixed value.

- I feel that whether or not to set N_{min} to zero should be decided on the basis of model fit and parsimony.

- Looking at the bottom-right plot in Figure 5, it appears that a reduced model which set η_1 equal to η_2 would incur a small reduction in log-likelihood. It looks like this reduction is less than unity, in which case the reduced model would have a lower AIC than the model presented.

I also had concerns about the simulation approach proposed by DJR. As noted by DJR, the simulated covariates did not include correlation structure between the covariates. This seems to me a fatal flaw - it would give entirely misleading conclusions in the case of multiple linear regression with multi-collinear explanatory variables.

Eye-balling correlation matrices may not be enough to (in general) detect complex multiple correlations. Moreover, correlation matrices are parameterization dependent. In the size-at-age example I mentioned above, the high correlation between the growth-rate and maximum-size parameters can be avoided by using an appropriate re-parameterization (Schnute, 1981). So, one parameterization would suggest that a parameter needs to be fixed (somehow), but the other parameterization would not. In contrast, use of model selection via AIC will not depend on the parameterization.

References

Franke, E. S., Babcock, R. C. & Styan, C. A. (2002). Sexual conflict and polyspermy under sperm limited conditions: in situ evidence from field simulations with the free spawning marine echinoid *Evechinus chloroticus*. *Am. Nat.* **160,** 485–496.

Millar, R. B., and M. J. Anderson. (2003). The kinetics of monospermic and polyspermic fertilization in free-spawning marine invertebrates. *J. Theor. Biol.* **224,** 79–85.

Schnute, J. (1981). A versatile growth model with statistically stable parameters. *Can. J. Fish. Aquat. Sci.* **38,** 1128–1140.

Vogel, H., Czihak, G., Chang, P. & Wolf, W. (1982). Fertilization kinetics of sea urchin eggs. *Math. Biosci.* **58,** 189–216.

R. B. Millar
Dept of Statistics
University of Auckland
Private Bag 92019
Auckland
New Zealand

Invited Comment on ANZJS A031010:

K. Domijan, M. Jorgensen, and J. Reid -

“Semi-mechanistic modelling in non-linear regression: a case study”

The paper raises an underlying but central issue concerning the place of statistics in applied research, as well as detailing more technical, statistical matters. The underlying question needs proper debate because with increasing sophistication of statistical models and technique, applied research for teams of researchers from different disciplines increasingly involves formal and strongly delineated responsibilities. The core philosophical question is what role statistics has to play in such situations beyond solving purely technical problems in fitting pre-determined and pre-specified models.

Some of the core philosophical and operational issues can be summarised as follows:

1. If highly parameterised models are not intended to be substantially modified by statistical analysis, because their structure is decided by subject matter considerations, what is the role of Statistics?
2. Given this strong constraint on model amendment via statistical methods, what is the logical basis for refutation of any such model in a Popperian sense (Popper, 1992)? Is this still a statistical matter? If so, how? And if not, on what basis are scientists in the substantive discipline to make decisions to change models, or choose between them, seemingly without the benefit of data analysis and statistical tests to trim and supplement models?
3. Highly parameterised models containing submodels used to define variables to be included in the final model, raise a further level of complexity, at least statistically. How are submodels to be tested (and hence modified) in this context? How are we to establish a sound foundation for inference?
4. The statistical technicalities of fitting over-parameterised models are important, but so is the proper application of Ockham’s razor to remove redundancy. Multi-collinearity is a particular concern, especially where as here the data set is small, or limited in range, or structured so there is no variation at all for at least some variables used in the model. How are these technicalities to be dealt with in a way that does not reduce the fitting and interpretation of such models to an arcane art? Once the new methodology becomes more complex, how can (statistical and non-statistical) scientists check mechanistic and semi-mechanistic models via the modelling process, particularly where data sets are small and/or clustered (even if it proves possible that this refutation and 'amendment' process can be part of the methodology in principle)?

What is clear is that although these preliminary issues, like most philosophical ones, will not be settled quickly, their answers are critical to evaluating the usefulness of this case study. Given the underlying philosophical issues remain to be settled, caution is required. Publication of this case study on semi-mechanistic models should be seen as the opening of a debate and certainly should not be seen by the wider scientific community as an endorsement (by the statistical community) of the explicit agricultural model used in the paper.

Given these preliminaries, what follows is necessarily more a set of observations and questions than answers.

The authors' contention that linear models are often descriptive, while non-linear models are usually not (because they are more likely to arise from subject matter considerations), seems to put to one side the important role of linear models in much of experimental design. In any event, in the usual nomenclature the linear and nonlinear part is the dependent variable, and non-linear models are often fitted by a sequence of linear model approximations (eg generalized least squares). From this perspective at least, the prediction structure in both types of models is identical.

The case study data used consist of 84 data points from at 12 sites, so that for some variables such as weather, cultivar, soil properties (eg N. P, K, Mg, pH) there would seem to be only 12 different values. The model fitted contains 26 parameters (see Table 2). To make fitting possible "assumed or empirically estimated functional forms" are used. If assumed, how can this be justified statistically, and if fitted empirically how many degrees of freedom are available? When a model is fitted to values that vary substantially between but not within sites (of which 12 are mentioned over the 1996-1999 period), there is usually much greater variation between sites than within sites. This has substantial effects on standard errors (via example effective sample sizes due to correlation of observations within sites). Is this problem is inherent in semi-mechanistic models in general, and how can it be incorporated into model fitting?

Sections 2.2 outlines the PARJIB model, but in a way that is completely uncritical from a statistical point of view. Use of genetic algorithms, gradient methods, profile likelihoods, linear approximations, bootstrap and setting of collinear parameters to arbitrary values are all technical matters likely to be beyond the province of the scientists who develop semi-mechanistic models. How are these technical issues to be approached in a way that does not overshadow the substantive agricultural question which is the basis of the model, and how is feedback best provided from statistician to scientist in such circumstances on what is the real biological essence?

The simulated datasets used to calculate standard errors do not fully specify the problem (eg using marginal distributions and profile likelihoods). How can the authors be certain that their simulation, since only partially specified in a statistical sense, reflects the actual (multivariate) structure in the actual data? To what extent are the profile methods useful, given they are essentially one dimensional in the context of a complicated non-linear model? And what allowance has been made in the standard errors for the range of possible (simulation) models, given the estimated standard errors are conditional on the model being correct?

In what sense are the model modifications to ensure parameters fall in plausible ranges anything more than ad hoc adjustments to the underlying model? Is there any explicit methodology available by which semi-mechanistic models can or should be formally modified, and does this meet acceptable statistical criteria?

The idea of general applicability of mechanistic and semi-mechanistic models is intuitively appealing, given the same model can be applied to different datasets. In

practice however there are complications, eg How are parameters to be estimated in models which contain interaction effects, one component of which is fixed at a constant in each particular dataset? Are parameter estimates to be 'borrowed' from other datasets in such circumstances, for example, (and on what formal statistical basis) or are effects to be set to arbitrary values (eg zero) to avoid this problem. Given such complications, how can a model be said to be a 'general' semi-mechanistic model in practice?

It is clear from the graphs and the variation in the data that the structure of the model has a considerable influence on the fitted curves. The considerable variation about the fitted model must surely remain a concern given the model has an essentially non-statistical basis, and seems somewhat immune to statistical criticism. Given that semi-mechanistic models are usually over-parameterised, what information is 'left over' from the fit to apply formal statistical tests?

In conclusion, there are a range of questions this paper leaves and must leave unanswered, but the authors are to be thanked for raising and answering some of them. Despite uncertainty both in the statistical underpinnings of semi-mechanistic models and the expected role of any statistician in teams of researchers looking at such complex models remaining open questions, the authors have done statisticians a service in bring these issues more explicitly to our attention.

References

Popper, K. R. (1992). *Conjectures and Refutations: The Growth of Scientific Knowledge*, 5th Edition, Routledge and Keegan Paul, London.

Stephen Haslett
Professor of Statistics
Institute of Information Sciences and Technology
Massey University
PO Box 11222
Palmerston North
New Zealand

16 October 2005

Replies to comments on "Semi-mechanistic modelling in nonlinear regression: a case study"

We are grateful for the well-considered and thought-provoking contributions of the reviewers. Rather than reply to each single point made by each reviewer, we have chosen to group together their comments under generic headings and discuss those.

1 Purpose of the case study

The case study has prompted thought on a wide variety of issues that confront statisticians and biological scientists. We could not hope to address all of those, and the first parsimonious principle we invoked in the paper was to define its purpose.

It is worth repeating that purpose here. The paper sets out "*...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. We discuss 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.*"

In places, Millar in particular has assumed that our purpose was to develop and defend a semi-mechanistic model. That is not correct.

2 Collaboration between mathematical modellers and statisticians

In this and similar cases, the main question that causes a mathematical modeller to seek help from a statistician boils down to "how reliable is this model?" As Plummer and Welsh comment, at the level of overall model performance this question can be answered by examining how well the model's yield simulations and predictions match the experimental values. Reid et al. (2002) had already done that for the initial model fit (by genetic algorithm). Naturally, when sufficient data are available, the fitting of a model provides an opportunity for examining some of the functional forms used in the model. In this case study, the initial question asked of the statistician was to assess how reliable are the individual parameter values from that earlier process of fitting by genetic algorithm (or later alternatives to it). It is when undertaking parameter estimation at this level that we must confront the "*core philosophical and operation issues*" referred to by Haslett and others.

In building and testing empirical models, we suggest the statistician's role is paramount. In developing and applying mechanistic and semi-mechanistic models, we submit that the statistician's role should be more interactive and complementary with that of the scientist.

This case study covers a part *but not all* of the scientific process that seeks to develop and apply mechanistic and semi-mechanistic models in the biological sciences. The paper concentrates mostly on the role of the statistician in the fitting and testing part of the process – that is a critical role, but not necessarily the most important throughout the process as a whole.

The process is advanced considerably by the nature of the hypothesis testing that goes into the statistician's analysis of model performance. That testing is as much part of the scientific process as gathering the experimental data itself. The process of scientific judgement that sets up the semi-mechanistic models in the first place could often do better if statisticians were involved very early on, but there are often valid and important scientific and behavioural reasons why that might not occur. We argue that the statistician's input at any stage should be positively sought by the scientists. The scientists must be prepared for their pride to be dented and work schedules affected as their models are overhauled and perhaps rejected as inconsistent with the observations or unjustifiably complex. In turn, the statisticians must be prepared to accept that even if a much simpler model gives a statistically more acceptable fit to the available data there may be good reasons for persisting with a model structure that fits accepted theory of how a system works. Examples of the latter include:

1. More or better quality data could be expected soon, or by publishing the model the scientist hopes to solicit previously unavailable data;
2. Adoption of a simpler, more empirical model can introduce other complications – not the least being the implicit or explicit recommendation to abandon scientific knowledge that may be amply supported by other experiments and which is important elsewhere in the research project;
3. Recognition that the available data is inadequate to properly test a theoretical model, especially one based on current perceived wisdom, is an important incentive for better experimental science;
4. The attachment that scientists may have to the mechanistic elements of their models, even in the face of less than resounding statistical endorsement, centres on the philosophical principle of distinguishing description from explanation. Empirical models (based solely on statistically derived relationships) describe datasets. Mechanistic and semi-mechanistic models seek to ascribe causality to relationships observed in nature. Their developers are attracted to the promise of quantifiable causal relationships that explain observations (ascribe causality) and give confidence that the model has relevance beyond just describing the dataset used for fitting.

It should be clear from our response above that we do not consider it acceptable that, by the time of the statisticians input, the model is considered subject to only minor adjustment.

3 Why was the subject model so complex?

Questions and comments to this effect were raised by all of the reviewers. This paper describes the process from the initial involvement of the statisticians after the initial model was derived from a mix of mechanistic knowledge and empirical experience based on a wide range of previously published experiments.

Haslett and Millar ask why more statistical testing was not done with a view toward simplifying the model. This is a process that the statistical analysis has indicated should be a priority in the ongoing scientific process. Even so, the following points may be noted.

1. The model is based on individual response curves for N, P, K, or Mg. There is no serious doubt amongst crop agronomists that these elements are related to crop yield. It is open to further investigation whether the forms adopted by PARJIB are the best to model these response curves.
2. It is also open to doubt whether the function $q_{nut} = \max\left(0, 1-\sqrt{\left(1-q_N\right)^2+\left(1-q_P\right)^2+\left(1-q_K\right)^2+\left(1-q_{Mg}\right)^2}\right)$ used to combine the individual nutrient curves into a multi-nutrient function is the most appropriate. There is no more suitable equation in the crop physiology literature, but this case study indicates a search for alternatives would be very helpful. Ideally this function should be replaced by a nonparametric form, perhaps constrained to be monotonic in each argument. However this would raise the effective number of parameters in the model (which is already troublesomely large for the amount of data that experimenters can hope to amass).
3. The maize data was from an aggregation of experiments, not all of which were necessarily designed to measure response to all four nutrients. It is notoriously difficult to reliably locate experiments that will show substantial responses to applications of some fertilizers (in NZ, background soil levels of nutrients such as Mg can be in or close to the flat portion of the yield vs nutrient supply curve). If stress due to a particular nutrient does not vary enough in the data, it will not be possible to obtain reliable estimates of the parameters in its response curve. Thus we can be confident that all four nutrients play a role in predicting maize yield, but not confident that we can estimate the response parameters for each nutrient. We chose to concentrate on the response to Nitrogen alone in our "simple" model. If a formal test showed that the "full" model was not significantly better than the simple we would not conclude that P, K, and Mg were unimportant for maize yield; only that variations in stresses for these nutrients were not important in the data analysed. Were formal tests felt to be of interest, notwithstanding these comments, then a score test - see, for example, Rayner (1997) - has the advantage of only requiring quantities local to the reduced maximum likelihood estimator. It does not require the successful fitting of the full model which is difficult to achieve when the number of parameters is large in relation to the data. As there are 16 models to consider corresponding to the choice of the subset of {N, P, K, Mg } used, it seems in principle that model selection, for example using AIC as suggested by Millar, may be a more

appropriate methodology than hypothesis testing. However it seems very hard to fit more than a few of these 16 models. As noted by Millar, the reduced model with $\eta_1 = \eta_2$ seems attractive with the maize data, but with other crops this has been found not to be so.

In this case study, the *a priori* reasoning that lead to the complex model has been challenged further by the statistical analysis. In the case study, some discussion centred around the correlation structure of the parameter estimates, and showed that for each nutrient, two parameters in particular are highly correlated and cannot be estimated independently with the sorts of field data that we can reasonably expect to obtain. As a result, difficulties of fitting the parameter X_{min} (where X may be N, P, K or Mg) with real world data are unlikely to go away, and model formulations without these parameters are currently being investigated. Furthermore, the difficulties experienced in fitting the terms for the "broken stick" style response of scaled yield to soil pH and water deficit have prompted moves to simplify the model using continuous, differentiable, ramp functions. These are not *ad hoc* model adjustments, but reflect considered mechanistically-based responses to statistical challenges.

We are rather concerned at Millar's implication that the objective was to fit as complex a model as possible. As our above replies make clear that is absolutely incorrect. Maybe this notion arose from a misunderstanding of the role of the "simplified model" in our case study. The use of simplified models like this is to aid the initial search for better methods for fitting and analysis. If an approach is unwieldy and works poorly with the simplified model then it certainly should take a low priority for use on the "full" model that the biometrician is presented with by the scientist.

4 What bases were there for refutation of the model?

Comments by Millar and Haslett on this matter link to the above discussion about the statistician's role in collaborations with mathematical modellers. Millar in particular doubts that "the PARJIB model, or any subset of it, was truly confronted by the data." A way in which such a confrontation can occur is to have a parameter estimated at a value that is inconsistent with its mechanistic meaning. A confrontation of this kind almost occurred in the paper: the parameter N_{min}, was estimated to have a negative value (-762.8265). This is inconsistent with the meaning of N_{min}, as the smallest rate of supplied Nitrogen at which positive yield is observed. The hypothesis N_{min},= 0 may be tested using either a likelihood ratio test or a score test. A likelihood ratio test gives F = 0.595 and a p-value of 0.44 when referred to the F(1,75) distribution. A score test gives a chi-squared statistic of 0.535 and a p-value of 0.46 when referred to the χ^2_1 distribution. Thus here the data does have the chance to refute the mechanistic interpretation of this parameter but fails to do so.

We may take this a little further by inverting the likelihood ratio test to construct a one-sided confidence interval for N_{min}. By this means we obtained that $N_{min} \leq 6.6$ with 98% confidence, though encountering numerical difficulties when attempting to raise this bound to obtain higher confidence. Such results do have the potential to cast doubt on mechanistic models if the bounds seem unreasonable to subject-area scientists.

Actual supplied rates of Nitrogen are well above zero, so the region in which the curve fails to model growth is not found in the data. Nevertheless the reason for this phenomenon merits further attention. This is typical of the way in which mechanistic models are challenged by the data. An interesting example is mentioned by Richards (1959) concerning Von Bertalanffy's growth model which was mechanistically derived by Von Bertalanffy leading to constraints on a "shape" parameter. Richards demonstrated that for many real animal growth datasets the constraints were violated. He recommended dropping the constraints, leading to a family of growth curves as empirical models but whose mechanistic status was now dubious.

Popper's falsifiability criterion requires that for a statement to be a scientific one it must, in principle, be capable of falsification. Statistics is often seen as the guardian of the scientific method and Haslett asks how it can uphold this role without, in this case, questioning the PARJIB model. Now it is not hard to decisively reject a statistical model with few parameters given a very large data set purportedly generated by the model. But this is much more difficult when considering a highly parameterized model with a modest amount of data.

In fact the ultimate test for complex models like PARJIB is whether they work for their intended purpose. In the case of PARJIB this will be whether the model simulates or predicts yield well (see Section 2 above) and if it is found to be useful in making recommendations on fertilizer application rates for agricultural crops. Brun et al. (2001) point out, speaking of *"large environmental simulation models which are thought to summarize current scientific knowledge in a mathematical language",* that *"In many case studies, these models are (successfully) applied without questioning the model structure."* Indeed it is often the case that no statistical advice is sought or given in either the development or application of such models. We do not see this situation as desirable but, if it is to be changed, there is a need for flexibility and understanding between statisticians and modellers. Overly rigid attitudes could well be counterproductive.

In other words we must accept that science is not necessarily built up paper by paper with each new fact confirmed against data by statistical gatekeepers. Sometimes it is necessary to work with a model or theory over time and many sets of data to see whether it stands or falls.

5 A better way to fit the model?

Plummer recommends the adoption of a Bayesian framework for investigations of this kind. Newman suggests that this might present a method of handling weak unidentifiability that would be preferable to outright fixing of parameters.

We began this work faced with a complex model whose parameters had been estimated from a modest set of data using a computer-intensive "genetic algorithm". Our aim was to attempt to fit at least a submodel to the data using classical nonlinear regression so that some inferences could be made about the parameters. It soon became clear that it was difficult to make any progress towards fitting anything without making use of background information about such matters as the relative importance of the nutrients and sensible ranges of values for the parameters of interest. It is also noteworthy that we found the Levenberg-Marquardt algorithm, a close cousin of Ridge Regression, to be a useful tool in arriving at least-squares fits of several of the models considered. As Lindley and Smith (1972, section 5.3) show for linear models, Bayesian formulations can lead to estimators of the same form as Ridge Regression estimators.

As Plummer notes, mechanistic models adapt well to Bayesian methods because there are usually good intuitions available about parameters in them which can be translated into priors. Welsh also mentions MCMC methods as an alternative to Genetic Algorithms, but Plummer points out that they by no means abolish all numerical difficulties. Prior distributions, however, will help with the parameter practical identifiability problem in a manner similar to the way in which Ridge Regression tames multicollinearity. It may also be noted that the method followed in this paper, of generating simulated data closely following the model, may be of use in tandem with MCMC model fitting as a means of exploring just how good the data need to be before MCMC output becomes trustworthy. Reversible jump MCMC (Green, 1995) would be a potential way of selecting between submodels of PARJIB with differing sets of nutrients included.

We also note that the objection of unfalsifiability brought against mechanistic models can be brought against Bayesian prior distributions. We suspect that the answer to this objection may be the same in both cases: falsifiability is there, but not necessarily with the present set of data. It may be necessary to seek further evidence.

6 Testing of sub models and use of *a priori* information

Science proceeds by building on previous results as well as by challenging them. This case study raises a few issues about the former process. At some point we are obliged to use previous models or results as *a priori* information instead of expanding the model-fitting process to include all constants or parameters involved. For example, it seems absurd to estimate π when fitting a model that involves it. However, in this case study, should we have subjected the sub-model for potential yield to the same scrutiny as the nutrient response model PARJIB itself? (We chose to trust the peer review process that led to publication of the potential yield

sub-model.) Between these extremes, mechanistic and semi-mechanistic models often use fixed parameter values taken from other research results.

Haslett notes that there are underlying philosophical and statistical issues associated with such “borrowing” of parameter values from other datasets and using them as fixed values. Newman and Welsh raise the question of the bias introduced in other parameters by fixing the values of a set of parameters. Newman points out that this can be can be investigated through simulation. Whether this is worth doing depends on whether the parameters are being estimated for their own sake or whether it is the predictions from the model that are important. If the focus is on the latter, then the effect of fixing parameters on the fit of the model could be investigated using score tests.

To some statisticians it might seem to go against the grain to give a parameter a fixed value. However statisticians commonly do equivalent things with very little angst. We may choose a logit or a probit link in a binomial generalized linear model; we may choose an additive error term to be normally distributed rather than, for example, having the shape of a Student-t with 4 degrees of freedom; we may choose a growth curve to be of Gompertz or logistic form rather than estimate the shape parameter of the Richards family of curves which contains both as members. The Box-Cox versus Bickel-Doksum controversy on the estimation of a transformation parameter makes interesting reading here (Bickel & Doksum, 1981; Box & Cox, 1982). Box and Cox defend, successfully we believe, their recommendation that inferences be made conditionally on the transformation used without attempting to account for the uncertainty in the transformation parameter.

Welsh queries the expression "to adequately test the model at the mechanistic level". What was meant was to obtain sufficiently precise estimates of the model parameters to make possible meaningful comparisons with the subject-area literature.

7 Fitting using simulated datasets

The initial motive for the use of simulated data in this paper was simply to fit at least some of the model parameters using standard nonlinear regression algorithms. The hope was that by generating a very large set of data closely following the model but otherwise similar to the maize data, we would create a situation that could be handled by the Gauss-Newton algorithm. Inspection of the parameter correlation matrix of this fit to the simulated data might then suggest a subset of the full parameter set whose correlations were not too close to ± 1 and more generally for which the corresponding submatrix was not too close to singularity. With luck the same set of parameters would be estimable by Gauss-Newton starting from an initial estimate and with remaining parameters held constant at their initial values. (In the paper, the initial parameter values were the ones obtained from a Genetic Algorithm.) Welsh gives a good summary of our approach. Newman points out that near-rank-deficiency in the parameter variance-covariance matrix need not be due to highly correlated pairs of parameter estimates. This is, of course, true, but these pairs are what one would first look for. If a fit to the real data could then be achieved by holding one member of some of these pairs constant, there would be no motivation to look for higher-order linear dependencies between the parameter estimates from the simulated data.

Matching the covariance structure of maize data is not essential for the purpose above. Generating covariates independently is likely to lead to a simulated data set that understates the parameter correlations when the model is fitted to it. But such a simulated data set will be easier to fit using Gauss-Newton, and it will have done its job if it then leads to the identification of a parameter subset that can be fitted with the maize data.

However, there is another reason why we might want to replicate the covariance structure of the covariates. If we generate a large set of simulated data that matches the covariate covariances of the maize data and closely follows the model, we might then construct a scaled sensitivity matrix, as defined by equation (6) of Brun et al. (2001), on the basis of the parameter variance-covariance matrix of a Gauss-Newton fit to the simulated data. We could then employ the methodology described in section 4 of Brun et al. (2001) for identification of influential parameters and parameter subsets with good collinearity properties. We could build up such a simulated data set by jittering the covariate vectors of the original observations, or with the help of a Cholesky factorization of the maize covariate variance-covariance matrix. We are pleased that Newman has drawn attention to the methodology of Brun et al. (2001) which we feel has great value for partially mechanistic statistical modelling.

Matching the covariance structure of the data would also be important if we were submitting simulated data to MCMC estimation as suggested in section 5 of this response.

8 Miscellaneous

We were unable to reproduce Newman's problem with the signs of some coefficients in the parameter estimate correlation matrix with the simulated data. Although not reported in this paper, Domijan (2002, Chapter 6) fitted several variants of the PARJIB model using AD Model Builder. It was necessary to modify some of the model equations to ensure that the fitted values were differentiable functions of the parameters. Compared to the Levenberg-Marquardt algorithm the results were good: both algorithms reached the same parameter estimates but AD Model Builder did so much more rapidly.

In one respect our models have not been complex enough: Welsh notes our arbitrary assumption of independent additive homoscedastic error terms, and Haslett notes that the observations at a particular site will be correlated for many reasons. We were indeed guilty of this oversimplification. Our fitted models, however, can be starting points for the fitting of more realistic models. For example we can introduce random site effects to the model. The table below compares the parameter estimates and their standard errors from the fixed effects model reported on in Table 4 with the estimates from the mixed model with random site effects. The mixed model was fitted using an EM algorithm with the site effects treated as missing data.

TABLE 6

Estimates of the eight parameters in the 'simple model' with random site effect.

Parameter	Estimate ($\hat{\theta}_j$)	Std error	Estimate ($\hat{\theta}_j$) with random site effect
N_{opt}	16.248	3.95	16.238
γ_N	0.679	0.26	0.595
ξ_{N1}	0.218	0.10	0.192
ξ_{N2}	0.255	0.10	0.543
β	0.600	0.11	0.713
D_{lim}	0.222	0.12	0.362
η_1	0.612	0.12	0.244
η_2	0.650	0.08	0.679

It is also possible to consider models in which some parameters vary randomly over sites although for some parameters this may conflict with their mechanistic understanding. When contemplating these classes of models, a Bayesian approach would seem increasingly attractive.

9 Ongoing value of the statistical insights

Finally, the reviewers' comments on the case study have stimulated us to consider further the ongoing role of the statistician in model development. As mentioned above (Section 2) the statistician's role does not end with the first fitting and appraisal of model performance. It needs to be emphasised that the statistical criticisms of model structure and fit of the available data should be used when designing future experiments to gather the extra data needed for model improvement.

10 Additional References

BICKEL, P. J. & DOKSUM, K. A. (1981). An analysis of transformations revisited. *J. Amer. Statist. Assoc.* **76**, 296-311.
BOX, G. E. P. & COX, D. R. (1982). An analysis of transformations revisited, rebutted. *J. Amer. Statist. Assoc.* **77**, 209-210.
DOMIJAN, K. (2002). Estimation of Parameters in a Large Nonlinear Crop Performance Model. M.Sc dissertation, University of Waikato.
GREEN, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika.* **82**, 711-732.
LINDLEY, D. V. & SMITH, A. F. M. (1972). Bayes Estimates for the Linear Model.
J. Roy. Statist. Soc. Ser. B. **34**, 1-41.
RAYNER, J. C. W. (1997). The asymptotically optimal tests.*The Statistician.* **46**, 337-346.
RICHARDS, F. J. (1959). A flexible growth function for empirical use. *J. Experim. Bot.*. **10**, 290-300.