

Mathematical Model Building with an Application to Determine the Distribution of Dursban® Insecticide added to a Simulated Ecosystem

G. E. BLAU

*Computation Research, The Dow Chemical Company,
Midland, Michigan 48640, U.S.A.*

and

W. BROCK NEELY

*Ag-Organics Product Department,
The Dow Chemical Company, Midland, Michigan 48640, U.S.A.*

I. Introduction	133
II. Model Building Techniques	134
A. Types of Mathematical Models	134
B. Model Building Procedure	135
C. The Design Problem and the Analysis Problem	138
D. The Likelihood Approach to Model Discrimination	139
E. Example of Model Discrimination by Likelihoods	143
F. Parameter Estimation Procedures	144
G. Tests of Model Adequacy	145
1. Goodness of Fit	145
2. Residual Analysis	147
H. Conclusion	148
III. The Environmental Fate and Distribution of DURSBAN® Added to an Ecosystem	149
A. Introduction	149
B. Description of the Ecosystem	149
C. Building the Model	151
D. Discussion of Results	160
E. Conclusion	162
References	162

I. INTRODUCTION

Mathematical models of different types and different levels of sophistication have been widely used in the chemical industry. These models have ranged from large plant models, used to determine the optimum operating conditions which maximize or minimize some economic criterion, to process models which predict the steady state operation of processes or the dynamic response of the process to

disturbances. A plant model usually consists of several process models. In order to obtain these process models it is necessary to have some understanding of the underlying physical and chemical mechanisms involved. This gives rise to mechanistic or phenomenological modelling. These mechanistic models should be capable of describing the basic physical and chemical steps so that a process may be both designed and operated properly.

Apart from these conventional engineering applications, mechanistic models provide valuable insight into the behaviour of any system in which chemical reactions are taking place. For example, it is vitally important to know the environmental impact that a chemical may have when added to an ecosystem. Many of these systems are very complex, so that selection of a suitable model is by no means transparent before or after data have been collected. In fact, one can frequently postulate several models which, superficially at least, represent experimental kinetic data. The problem then is to determine the constants and if possible choose between these candidate models.

This report presents a procedure for building a mechanistic model which represents an experimental reaction system. Starting with one or more plausible models, the principle of maximum likelihood is applied to the data collected in order to estimate the constants in the model and choose the best model among those originally postulated. Then, conventional statistical techniques are used to determine the suitability of this "best" model. If the model is inadequate, a technique is presented for identifying the specific limitations. Then the model builder must postulate additional physical meaningful models to accommodate this limitation and the procedure is repeated.

There are two parts to this report. In the first part, the model building procedure is developed from elementary statistical principles. The important concept of maximum likelihood is introduced and illustrated with an example. The need for proper experimental design and an iterative experimentation-analysis program is presented with examples. The second part of the paper illustrates the model building procedure by finding a model which describes the fate and distribution of DURSABAN® insecticide added to a laboratory system which simulates a pond of water.

II. MODEL BUILDING TECHNIQUES

A. TYPES OF MATHEMATICAL MODELS

In theory, it is possible to represent all the phenomena occurring in any physical system by a precise mathematical model. To do this

requires a complete description of the true scientific mechanism of each phenomenon. In practice, however, a complete description of this mechanism is not available, so approximations must be made. The extent of these approximations classifies the mathematical model representation as mechanistic, empirical or regression. For example, if one is concerned with the basic steps that take place when a chemical is introduced into an ecosystem for different conditions of the system, e.g. amount of chemical added, temperature etc., a *phenomenological* or *mechanistic* model must be used. Here each term or group of terms represents some specific phenomenon such as the formation of a metabolite or the transfer of a chemical from one compartment to another. Obviously, development of this type of model requires an extensive and carefully designed testing program. Suppose, however, that considerable data has been gathered either in the laboratory or in the field on, say, the decomposition of an insecticide with time. Then, a *regression model* may be used to condense or organize this data so that it is readily accessible. No attempt is made to add any physical significance to the individual terms of the regression models, which are simply multivariable polynomials of different degrees. A compromise between the mechanistic model and regression model is the *empirical* or, as it is sometimes called, the *quasi-mechanistic* model. In such a model, some physical meaning is attributed to the potential selection of terms for the model, although no attempt is made to identify the basic steps in the process being modelled. These models are widely used where the biological variation is high and/or the testing is minimal. A typical example is an attempt to characterize the biodegradation of chemicals by their 1/2 life. Here, the physical principle assumed is that a chemical biodegrades exponentially with time. However, the steps involved in this degradation process are left unspecified.

This paper presents statistical methods for developing mechanistic mathematical models. Many of the methods employed, however, are valid for the other model types and in most cases were originally derived from techniques for regression models. Tests will be given to help guide the model-building practitioner in deciding whether his data is of sufficient quality to justify using these more meaningful, albeit more mathematically complex, mechanistic models.

B. MODEL BUILDING PROCEDURE

It is frequently possible to postulate several physically meaningful mathematical models describing the particular system being studied. Usually these models are based on theoretical principles or intuitive insights from observations taken on analogous systems. In general, the

degree of sophistication of these models will range from complex multiparametered models to simple one-parameter models. *Model discrimination* is the statistical procedure which chooses or distinguishes among the various postulated models to find the model or models which best describe the system studied. Note that this discrimination only takes place among the set of postulated models. That is, the model selected by model discrimination may be the best of the originally postulated models but totally inadequate in describing the actual physical system. Using statistical residual analysis on the data collected, it is frequently possible to identify specific inadequacies in this "best" model. The model builder should then be able to suggest other plausible models which include one or more additional terms to accommodate the inadequacies in the original model system. Then discrimination is carried out on the new models and the process is repeated.

It is apparent from the foregoing discussion that model building is, in general, an iterative procedure. The steps may be summarized as follows:

1. Postulate one or more models to describe the physical system studied.
2. Use model discrimination techniques to identify the best model among those postulated in step 1 from experimental data collected on the system.
3. Determine whether the model identified in step 2 adequately describes the experimental data generated. If it does the procedure is terminated.
4. Use residual analysis to identify the specific inadequacies of the model selected in step 2 and suggest a new model or models to accommodate these inadequacies. Return to step 2.

This model building procedure is continued until a suitable model is found and the procedure is terminated at step 3.

As an example, consider the problem of building a model to describe the appearance and disappearance of a chemical B with time where B is formed from A . Suppose concentration-time data is available for component B only. In the absence of any prior knowledge of the chemistry of the process, the simplest model to postulate corresponds to an irreversible reaction

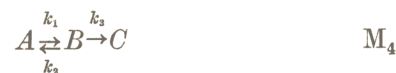


where k_1 is a reaction rate constant. By adding an additional parameter, k_2 , describing the reversible reaction between A and B , one obtains



Choosing between these two models is equivalent to determining whether or not the reverse reaction rate constant k_2 is greater than zero, i.e. $k_2 > 0$. M_2 is said to be *more complex* than M_1 since it has an additional parameter. The effect on the model discrimination method of adding k_2 to M_1 to form M_2 is analogous to the physical chemistry phenomenon of changing the degrees of freedom in a system. That is, there is twice as much flexibility in making M_2 explain the data as M_1 . This increased flexibility is reflected in the statistical criterion used to discriminate the models. For example, if M_1 and M_2 "explain the data to the same extent", the additional parameter k_2 is indeterminate and M_1 is said to adequately represent the data.

Suppose the concentration-time data for this example exhibited a maximum. Then both M_1 and M_2 would be inadequate. It would be necessary to postulate different models to explain the data and recycle through the model building procedure. Some typical models which could account for such a maximum are



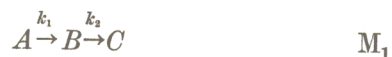
Generally, some of these models can immediately be eliminated by physicochemical reasoning. The most suitable of those remaining can be identified by the discrimination methods discussed below. If the model selected is still inadequate, additional ones can be postulated and the procedure continued until an adequate model or models is found.

Usually little difficulty is experienced in generating a variety of models of varying degrees of sophistication. A good rule to follow in choosing models is to keep them as simple as possible (i.e. minimal number of parameters and degrees of freedom). In fact, the best approach is to progress from the simplest model to progressively more complex models until no further increase in complexity is warranted by experimental uncertainties in this data. This principle of going from the simple to the complex is called Ockham's razor (Solberg, 1972) or the principle of parsimony (Kittrell, 1970). A good example of this principle is the stepwise add procedure of multilinear regression analysis (Draper

and Smith, 1966). Blau *et al.* (1970, 1972a, b) have demonstrated the utility of this technique in a wide variety of model building applications.

C. THE DESIGN PROBLEM AND THE ANALYSIS PROBLEM

In the procedure described above, it is assumed that the available data collected on the system is sufficient to choose between different models. In many cases this is not true. Consider the problem of choosing between the following two chemical reaction models



where A , B and C represent three chemical species and k_1 , k_2 and k_3 represent reaction rate constants. Concentration-time data is available for component B as shown in Fig. 1. Chemically speaking, to choose

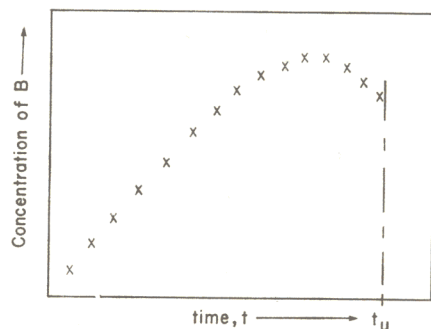


FIG. 1. Inadequate concentration time data.

between these models it is necessary to decide whether the disappearance of B occurs irreversibly, $B \rightarrow C$ (Model 1), or whether B is in equilibrium with C , $B \xrightleftharpoons[k_3]{k_2} C$ (Model 2). The key to distinguishing the models is the rate constant k_3 . That is, Model 1 is best if k_3 is zero while Model 2 is best if k_3 is nonzero. This can be expressed in statistical terms by saying, discrimination between the models is equivalent to testing the *null hypothesis* $k_3 = 0$. The data of Fig. 1 does not allow us to make this distinction between models. Even doubling the number of points between $t = 0$ and $t = t_u$ would shed no new light on the value of k_3 . What is needed, of course, is some data at times greater than t_u . Figure 2 shows two situations which might arise. If Model 1 is correct the concentration of B would drop off to zero for $t \gg t_u$. Conversely, an

equilibrium concentration greater than zero, $B_e > 0$, would be observed if Model 2 were correct. Note that only one or two additional data points may be necessary to distinguish these models provided they are located or "designed" properly, i.e. $t \gg t_u$.

The foregoing illustrates that there are two aspects to the application of the model discrimination phase (step 2 of the model building procedure). The first is the *design problem*, i.e. choosing the experimental conditions in such a way that discrimination is possible. The second is the *analysis problem*, i.e. analysing the data to assess how much discrimination has been achieved. The design problem is the more

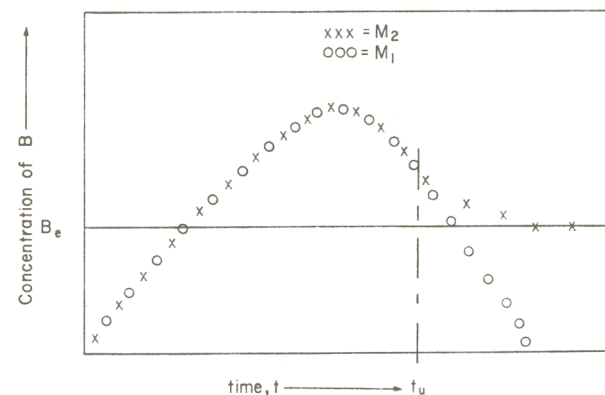


FIG. 2. Adequate concentration time data.

fundamental one. If for some reason the analysis of the data is faulty it may be repeated. However, the damage of poor design is irreparable and invalidates subsequent data analysis regardless of its level of sophistication. Considerable research effort by the scientific community has recently been expended on this design problem (Box and Hill, 1967; Reilly, 1970; Hsiang and Reilly, 1971). The methods developed rely heavily upon efficient optimization algorithms implemented on high-speed computers. It is beyond the scope of this paper to discuss these methods in detail, and the interested reader is referred to the literature. In the remainder of this paper it will be assumed that adequate designs have been employed so that the problem in choosing among models is only one of analysis.

D. THE LIKELIHOOD APPROACH TO MODEL DISCRIMINATION

Suppose that a set of models has been postulated and experimental data has been collected. In this section the statistical methodology for

using the experimental data to discriminate among the models will be presented. The methods to be discussed here are intended to be applied to models which are nonlinear in the parameters, and are not recommended for models linear in the parameters. The two most commonly used approaches are the likelihood approach and the Bayesian approach. The latter is based upon a subjective interpretation of probability (Bayes, 1763), a measure of the degree of belief that an event will happen rather than the objective interpretation in which the probability of an event is a long-term relative frequency. The Bayesian approach is readily embraced by scientists and engineers who advocate using knowledge other than that contained in the data. On the other hand, likelihood methods are claimed to have an advantage in objectivity in that they "let the data speak for themselves". Since the purpose of this paper is not to compare discrimination methods, the simpler likelihood method will be discussed. This is not an indictment against the Bayesian approach. The interested reader may wish to compare the two methods in the excellent paper by Reilly (1970).

Suppose $p(x, \theta)$ is a probability function which when given values of one or more parameters θ , allows the probability of any outcome to be calculated. For example, the binomial probability function

$$p(x, \theta) = \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} \quad (1)$$

becomes the following function of x alone

$$p(x, 1/2) = \frac{5!}{x!(5-x)!} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{5-x}$$

which is the probability of obtaining x heads in five tosses of a true coin. In this case n is the number of tosses or trials, x the number of heads, and θ the probability of a head in one toss.

Suppose now that the coin is being tested for trueness and therefore θ is unknown. If it is tossed five times and a head turns up once, this available information may be substituted into the probability function to obtain

$$L(\theta) = \frac{5!}{1!4!} \theta (1-\theta)^4 \quad (2)$$

The terminology $L(\theta)$ is used to emphasize that this is a function of θ only and is called the *likelihood function*. If a value of θ , say θ_1 , is substituted into $L(\theta)$, it gives the probability that the event which actually happened (one head in five tosses) would have if the value of θ were θ_1 . Comparing the values of the likelihood function for two different

values for θ gives the relative plausibilities of those two values in the light of the data. The comparison is carried out by examining their ratio, $L(\theta_1)/L(\theta_2)$, sometimes called the odds ratio. In the above example, $L(1/5)/L(1/2) = 2.51$. This indicates that the data obtained are 2.51 times as probable if $\theta = 1/5$ as they are if $\theta = 1/2$, and the value 2.51 can be taken as the weight of the evidence against the coin being true. ($1/5$ is the value of θ which maximizes $L(\theta)$, while $\theta = 1/2$ if the coin is true.) This would not ordinarily be taken as strong evidence that the coin is not true. A likelihood ratio of 10 is ordinarily taken as showing a real difference in plausibility, while 100 denotes strong preferences for the value of one parameter over the other (Reilly, 1970; Barnard *et al.*, 1962).

This concept of a likelihood ratio for measuring the plausibilities of different parameter values can be extended to measuring the plausibilities of different mathematical models. First, consider the problem of discriminating two Models M_1 and M_2 where M_1 , a function of two parameters, is denoted $f_1(\theta_1, \theta_2, x)$, and M_2 , a function of three parameters, is denoted $f_2(\theta_1, \theta_2, \theta_3, x)$, where x is a single independent or controlled variable. Suppose some dependent variable y is determined for n different experiments corresponding to n values of the independent variable generating the data set $\{(y_i, x_i) \mid i = 1, \dots, n\}$. If the Models M_1 and M_2 are to be used to predict the observed values of y , then

$$\begin{aligned} y_i &= f_1(\theta_1, \theta_2, x_i) + \varepsilon_i & M_1 \\ y_i &= f_2(\theta_1, \theta_2, \theta_3, x_i) + \varepsilon_i & M_2 \end{aligned} \quad (3)$$

where ε_i is the experimental error corresponding to the i th observation. For any set of parameter values, a set of differences between observed and calculated values is determined. These differences, called *residuals*, are given by

$$e_i(\theta_j) = y_i - f_j(\theta_j, x_i) \quad j = 1, 2 \quad (4)$$

where $\theta_1 \equiv (\theta_1, \theta_2)$ for Model M_1 and $\theta_2 \equiv (\theta_1, \theta_2, \theta_3)$ for Model M_2 .

Let $p(\varepsilon, \theta_j, x; \psi)$ represent the joint probability density function of all experimental errors in the observed values where

$$\varepsilon \equiv (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n), \quad x \equiv (x_1, x_2, \dots, x_n)$$

and $\psi \equiv (\psi_1, \psi_2, \dots, \psi_m)$. Here ψ represents the parameters of the probability distribution, e.g. the mean and variance for the normal distribution. Under the hypothesis that the j th model is *true*, i.e. that there is no modelling error, the residuals are estimates of experimental error and may be substituted for the errors in the joint probability density function. This now gives a function depending only on the

parameters in the model and the form of the probability distribution which is the likelihood function $L(\theta_j, \psi) \equiv p(e(\theta_j), \psi)$.

If the experimental errors ε_i are uncorrelated from point to point, (Draper and Smith, 1966), the joint probability density function is the product of the individual probabilities $p_i(e_i(\theta_j), \psi)$.

That is

$$\begin{aligned} L_j(\theta_j, \psi) &= p_1(e_1(\theta_j), \psi) \cdot p_2(e_2(\theta_j), \psi) \dots p_n(e_n(\theta_j), \psi) \\ &= \prod_{i=1}^n p_i(e_i(\theta_j), \psi) \end{aligned} \quad (5)$$

Now, if the experimental errors are independent (Draper and Smith, 1966) and normally distributed with zero means and a known variance σ^2 , the individual probability density functions are

$$p_i(e_i(\theta_j), \psi) = p(e_i(\theta_j), \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{e_i(\theta_j)^2}{2\sigma^2}\right) \quad (6)$$

Substituting these values into Eqn (5) gives the likelihood function

$$L_j(\theta_j, \psi) = L_j(\theta_j) = \frac{1}{(\sqrt{2\pi\sigma})^n} \exp\left(-\sum_{i=1}^n \frac{e_i(\theta_j)^2}{2\sigma^2}\right) \quad (7)$$

which is valid for the j th model. For any particular parameter values, e.g. $\theta_j = \hat{\theta}_j$, this function gives the probability that the data set which actually was generated ($\{(y_i, x_i) \mid i = 1, \dots, n\}$) would have been generated by the j th model with parameters $\theta_j = \hat{\theta}_j$. Since M_1 has two parameters, while M_3 has three parameters, it is not possible to form a likelihood ratio for the same parameter values. Consequently, some way must be found to eliminate this "nuisance" dependence on the parameters. One way of eliminating parameters is by "maximizing them out". Thus, form the *likelihood ratio*

$$R_{12} = \frac{\max_{\theta_1, \theta_2} L_1(\theta_1, \theta_2)}{\max_{\theta_1, \theta_2, \theta_3} L_2(\theta_1, \theta_2, \theta_3)} \quad (8)$$

This is a comparison of the likelihoods of the two models at their individual best. It is simply a comparison of how well the two models can be made to fit the data, expressed in likelihood terms. To compare more than two models the maximum likelihood for each model is calculated and two-way comparisons made by examining the ratios. This is expressed by the relationship

$$R_{jk} = \frac{\max_{\theta_j} L_j(\theta_j)}{\max_{\theta_k} L_k(\theta_k)} \quad k = 1, \dots, m \quad k \neq j \quad (9)$$

where m is the number of different models being compared.

When the models have different numbers of parameters there are inherent difficulties with any discrimination method. Using the likelihood method described here, good discrimination requires that the likelihood ratio be much higher than usual if the favored model is the one with the larger number of parameters.

Now the likelihood functions $L_j(\theta_j)$ given by Eqn (7) are maximized by choosing θ_j values which minimize

$$S_j(\theta_j) \equiv \sum_{i=1}^n e_i(\theta_j)^2 = \sum_{i=1}^n \{y_i - f_j(\theta_j, x_i)\}^2 \quad (10)$$

This is the familiar least-squares criterion for estimating θ_j . In passing it should be noted that the justification for using the least-squares criterion to obtain parameter estimates is that it maximizes likelihood function when the error distribution is normal.

The maximum likelihood for the j th model can be written

$$L_j^* = \max_{\theta_j} L_j(\theta_j) = \exp(-\text{RSS}_j/2\sigma^2) \quad (11)$$

where

$$\text{RSS}_j \equiv \min_{\theta_j} S_j(\theta_j) \quad (12)$$

is the conventional *residual sum of squares* obtained with the optimal least-squares parameter estimates. Since only ratios are relevant between likelihoods, the constants which multiply all the likelihoods in a comparison set are irrelevant and have been dropped from Eqn (11).

The maximum likelihood approach is in principle easy to use. For each of the models postulated, determine the least-squares parameter estimates and the associated residual sum of squares. Then select the model with the smallest residual sum of squares and calculate the likelihood ratios relative to this model. Recall that a likelihood ratio of 10 is ordinarily taken as showing a real difference in plausibility while 100 denotes a strong preference for one model over the other. These numbers assume that the number of parameters in the models are the same. Therefore, it is necessary that the likelihood ratios be somewhat higher than usual if the favored model has a large number of parameters.

E. EXAMPLE OF MODEL DISCRIMINATION BY LIKELIHOODS

Consider the problem of choosing between the following three models (Reilly, 1970)

$$\begin{aligned} M_1: y_i &= \theta_{11}x_i + \varepsilon_i \\ M_2: y_i &= \theta_{21} + \theta_{22}x_i + \varepsilon_i \\ M_3: y_i &= \theta_{31} \exp(\theta_{32}x_i) + \varepsilon_i \end{aligned}$$

where x is a single dependent variable and y is the dependent variable. Data was collected at four different values of x giving the points shown below

x_i	y_i
0	-1.290
1	5.318
2	7.049
3	19.886

It is also known that the errors ε_i are normally distributed with means zero and variances $\sigma^2 = 1$. The residual sum of squares RSS_j for each of the models and the maximum likelihood values are shown below

Model (j)	RSS_j	KL_j^*	L_3^*/L_j^*
1	28.465	0.050	4000
2	22.473	1	202.2
3	11.853	202.2	1

The maximum likelihoods have been multiplied by a constant K in order to give them manageable values. Model 3 is obviously preferred to the other models. In fact, the data were generated artificially using Model 3 with $\theta_{31} = \theta_{32} = 1$ and $\sigma = 1$.

F. PARAMETER ESTIMATION PROCEDURES

An important part of the likelihood discrimination method is the determination of those parameter values which minimize the least squares criterion of Eqn (10). That is, it is necessary to have a procedure which will find those parameter values θ^* which

$$\underset{\theta}{\text{minimize}} S(\theta) = \sum_{i=1}^n e_i(\theta)^2 = \sum_{i=1}^n [y_i - f(\theta, x_i)]^2 \quad (13)$$

where $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$ represent the i th value of p independent variables. For models that are linear in the parameters, i.e. models of the form

$$f(\theta, x) = \sum_{k=1}^p \theta_k x_k \quad (14)$$

the parameters are readily estimated by linear least squares (Draper and Smith, 1966). To obtain the estimates θ^* , it is only necessary to solve a $p \times p$ system of linear equations, for which a unique solution is usually guaranteed.

Unfortunately, most meaningful mechanistic models are nonlinear in

the parameters. Here it is necessary to apply iterative parameter estimation procedures. That is, a sequence of parameter estimates $\theta^1, \theta^2, \dots, \theta^s, \dots$ are generated which eventually converge to the optimum. This presents numerous complications such as initial guesses of θ^1 to institute the sequence, efficiency and effectiveness of convergence algorithms, multiple minima in the least-squares surface, and poor surface conditioning (Rosenbrock and Storey, 1966). A discussion of these topics is beyond the scope of this paper and are mentioned only to inform the reader that these problems exist. Nonlinear least-squares parameter estimation is a nontrivial task. The paper by Bard and Lapidus (1968) discussed the merits of several of the different algorithms as they relate to maximum likelihood estimation.

G. TESTS OF MODEL ADEQUACY

After likelihood discrimination has chosen the best model from the set of candidate models, it is still necessary to test the suitability of this model to describe the data. Then a method is needed to identify any specific limitations in the model so that the model builder may modify the existing model to overcome these limitations. Although several new methods exist (Blau *et al.*, 1972a, b), they do not supplant the more conventional tests of model adequacy of classical statistical theory, i.e. the goodness of fit test and tests of residuals.

1. Goodness of fit

A goodness of fit test compares the amount of variability between the differences of predicted and experimental values, i.e. the residual sum of squares, with the amount of variability in the data itself. This comparison allows the model builder to determine whether the overall model is adequate. If the model being considered is correct, the residual for the i th data point using the least-squares estimates θ^* ,

$$e_i(\theta^*) = y_i - f(\theta^*, x_i),$$

will be a measure of experimental error. A measure of the total amount of variation unaccounted for by the model is the residual sum of squares

$$RSS = \sum_{i=1}^n e_i^2(\theta^*) = \sum_{i=1}^n [y_i - f(\theta^*, x_i)]^2 \quad (15)$$

It is a direct result of the orthogonality property of linear least squares (Draper and Smith, 1966), that

$$\sum_{i=1}^n y_i^2 = \sum_{i=1}^n f^2(\theta^*, x_i) + RSS \quad (16)$$

Equation (16) states that the total amount of variability in the data

called the *crude sum of squares*, $\sum_{i=1}^n y_i^2$, is equal to the total amount of variability which can be accounted for by the model, called the *sum of squares due to regression*, $\sum_{i=1}^n f^2(\theta^*, \mathbf{x}_i)$, plus the residual sum of squares. Associated with each source of variation is a certain number of degrees of freedom, which is used to attribute more information to, say, 100 data points than to five data points. In particular, if n data points are used, the crude sum of squares possesses n degrees of freedom. The predicted values estimated by the model with p parameters have p degrees of freedom while the remaining $n-p$ degrees of freedom are possessed by the residual sum of squares.

If several data points have been taken at the same settings of the independent variables, then a measure of the inherent error in the data is given by the *pure-error sum of squares*

$$\sum_{j=1}^k \sum_{u=1}^{n_j} (y_{ju} - \bar{y}_j)^2$$

where

$y_{11}, y_{12}, \dots, y_{1n_1}$ are n_1 repeat observations of \mathbf{x}_1

$y_{21}, y_{22}, \dots, y_{2n_2}$ are n_2 repeat observations of \mathbf{x}_2

$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$

$y_{k1}, y_{k2}, \dots, y_{kn_k}$ are n_k repeat observations at \mathbf{x}_k

and $\bar{y}_j = (y_{j1} + y_{j2} + \dots + y_{jn_j})/n_j$ is the average of all the repeated or replicated points of \mathbf{x}_j . Since the residual sum of squares measures the amount of variability as seen by the model, and the pure error sum of squares is a true measure of error in the data, it follows that the inability of the model to fit the data is given by the difference of these two quantities which is appropriately called the *lack-of-fit sum of squares*

$$\sum_{i=1}^n [y_i - f(\theta, \mathbf{x}_i)]^2 - \sum_{j=1}^k \sum_{u=1}^{n_j} (y_{ju} - \bar{y}_j)^2 \quad (17)$$

For simplicity assume there are r replications at q different settings of the independent variables, then the pure error sum of squares possesses $q(r-1)$ degrees of freedom (one degree of freedom being used to estimate \bar{y}_j); while the lack-of-fit sum of squares possess $n-p-k(r-1)$ degrees of freedom, which is the difference between the degrees of freedom of the residual sum of squares and the pure-error sum of squares.

The quotients obtained when the sum of squares discussed above are

divided by their degrees of freedom are called *mean squares*. The pure error mean square

$$\text{PEMS} = \sum_{j=1}^k \sum_{u=1}^{n_j} (y_{ju} - \bar{y}_j)^2 / q(r-1)$$

is a measure of experimental error independent of the validity of the model employed. Therefore, a test of whether a model is adequate can be made by determining the ratio of the lack-of-fit mean square,

$$\text{LOFMS} = \sum_{i=1}^n [y_i - f(\theta^*, \mathbf{x}_i)]^2 - \sum_{j=1}^k \sum_{u=1}^{n_j} [y_{ju} - \bar{y}_j]^2 / [n-p-k(r-1)]$$

to the pure error mean square. If the ratio is large, it suggests the model inadequately fits the data. Using the F statistic to quantify the magnitude of this ratio, the test of inadequacy is usually written

$$\frac{\text{LOFMS}}{\text{PEMS}} > F_{\alpha}[n-p-q(r-1), q(r-1)] \quad (18)$$

where $(1-\alpha) 100$ is the confidence level in percent for rejecting the hypothesis that the model is adequate. The F statistic is tabulated in almost every statistics reference text.

If an independent estimate of pure error is available, say s^2 with u degrees of freedom, then the test for adequacy of the model simply becomes the ratio of the residual mean square to this measure of pure error. That is, the model is said to be inadequate if

$$\frac{\text{RSS}/n-p}{s^2} > F_{1-\alpha}[n-p, u] \quad (19)$$

at the $(1-\alpha) 100$ percent confidence level.

2. Residual analysis

The goodness of fit test provides information about the overall ability of the model to fit the data. It can also be used to test the importance or contribution of certain terms in the model towards providing the overall fit of the data. However, these methods do not identify the specific limitations of the model. In particular, even though the overall goodness of fit is quite acceptable, more subtle model inadequacies may exist. These inadequacies can often be detected through an analysis of the residuals of the model.

As defined by Eqn (4), a residual is the difference between the observed and predicted values of the dependent variable. If the model is correct, the residual for any point is solely attributable to experimental error. Therefore, plots of this residual versus any independent

variable should exhibit all the characteristics of this error, such as being random with zero mean. However, if the model is inadequate, the residual will not be random and possibly biased above or below zero when plotted against some independent variable. Several methods have been suggested for preparing the most revealing residual plots (Kittrell, 1970; Draper and Smith, 1966). Consider the following three typical methods:

(a). *Predicted value residual plots.* A plot of the residual $e_i(\theta^*)$ versus the predicted value $f(\theta^*, x_i)$ can indicate whether the model truly represents the data. For example, residuals that are generally negative at low predicted values and positive at high predicted values indicate a model inadequacy even though it may have passed the goodness of fit test. These plots can also provide information about the assumption of constant error variance made in the maximum likelihood approach. If the residuals continually increase or decrease in such plots, a non-constant error variance is indicated and either a weighted least squares analysis should be conducted (Kittrell, 1970) or a transformation must be found to stabilize the variance (Box and Cox, 1964).

(b). *Independent variable residual plots.* By plotting the residuals versus the independent variable values, it is possible to identify which of the variables in the model is causing the residual trends that occur in the predicted value residual plots. The nonconstant error variance described above also is exhibited in these plots and can provide useful information for developing a weighting function.

(c). *Overall residual plots.* If one plots the frequency of occurrence of the rounded values of the residual against the magnitude of the residual, it is possible to assess the normality of the error if the model is correct. Also these plots test the assumption made earlier that the mean of the error distribution is zero. Basically this plot allows an approximate check on the assumptions made in the development of the least squares analysis from the theory of maximum likelihood.

H. CONCLUSION

The preceding sections have presented a methodology for building a mathematical model of some physical system from experimental data collected on the system. Again, it is important to re-emphasize that the best approach to model building is by carrying out the experimentation and analysis programs iteratively. Nothing is more frustrating than trying to obtain information about a system after the experimentation program has been terminated and the existing data are inadequate.

Another important point is the importance of properly designed experiments. Certain statistical assumptions relative to the distribution of the experimental errors are inherent in applying the statistical techniques to analyse the data collected. Proper experimental design will provide some indication of the validity of these assumptions. If the assumptions are invalid the data can be transformed and this transformed data can be analysed. The importance of knowing this distribution of the experimental error or *error structure* cannot be overemphasized. Although it may require more experimental measurements, the probability of building a meaningless or overly sophisticated mathematical model will be minimized.

III. THE ENVIRONMENTAL FATE AND DISTRIBUTION OF DURSBAN® ADDED TO AN ECOSYSTEM

A. INTRODUCTION

An important environmental problem is the determination of the ultimate fate and distribution of a chemical introduced into an ecosystem. Numerous phenomena take place simultaneously in such a situation. Hence, a true mathematical model describing each step of the process would be extremely complex. It is important, however, to try and find a suitable model to identify the most important chemical, physical and biological phenomena taking place and to predict the long-term environmental consequences.

The example chosen for study in this paper concerns the addition of a chemical agent to a laboratory system which simulates a pond of water. Some of the phenomena that need to be included are the distribution and partitioning of the agent between the water and soil that may be present. In addition to these, consideration must also be given to the absorption, metabolism and excretion of the agent by the various aquatic species.

B. DESCRIPTION OF THE ECOSYSTEM

Smith *et al.* (1966) published some studies on the distribution and fate of a new agent for the control of insects, DURSBAN® insecticide. The active ingredient of DURSBAN®, 0,0-Diethyl 0-(3,5,6-trichloro-2-pyridyl) phosphorothioate, was labelled with radioactive carbon ^{14}C in the pyridyl ring and added at a level of 1 mg/6 gal in a 10-gallon glass jar. This aquarium contained 2 in. of soil (13.3% organic matter), plants (salvinia, anacharis, milfoil and water cucumber) and 45 goldfish.

Samples of the various components were analyzed for radioactivity at different time periods after addition of the DURSBAN®. A summary of this data taken from the paper of Smith *et al.* (1966) is presented in Table I and plotted in Fig. 3.

The experimental setup described above was disassembled before this model building program was initiated so that additional experimentation was impossible. Therefore, knowledge of the underlying error structure must be based on existing replicate analysis and subjective interpretation of the experimentalist. From independent measurements

TABLE I

Distribution of ^{14}C DURSBAN® in the ecosystem

Time after DURSBAN® addition (h)	Percent radioactivity in the three components of the ecosystem		
	Fish	Soil and plants	Water
0	0	0	100
1.5	15.2	35.2	49.7
3.0	19.0	46.0	28.3
4.0	19.3	56.0	24.5
6.0	20.7	61.0	18.3
8.0	23.0	60.5	17.0
10.0	24.2	59.3	18.2
24.0	21.2	51.5	26.5
48.0	23.0	38.3	34.5
72.0	22.7	38.3	39.5
96.0	20.5	36.3	43.0
120.0	17.3	38.3	44.5

made in the system but not reported in Table I, it may be concluded that:

1. Measurements of ^{14}C in the three components are independent of the different components.
2. Measurements of ^{14}C for any one component are independent of other measurements of that component.
3. The measurement errors are approximately the same for each component.

If one assumes that the errors are normally distributed with zero means and constant variance for each of the components, then the single response likelihood analysis of Section II E is readily extended to this multiresponse case (Kittrell, 1970). Here, for example, the residual sum

of squares defined by Eqn (15) is the sum of the residual sum of squares for the three components. This total residual sum of squares can be used to determine the lack of fit sum of squares. It will also be informative to analyze the residuals of the individual components. Such an analysis provides valuable insights into particular limitations of the model.

The data in Table I have been transformed into percentages from the crude radioactivity measurements. Although the error structure defined in the preceding paragraphs is also transformed, the variability in the

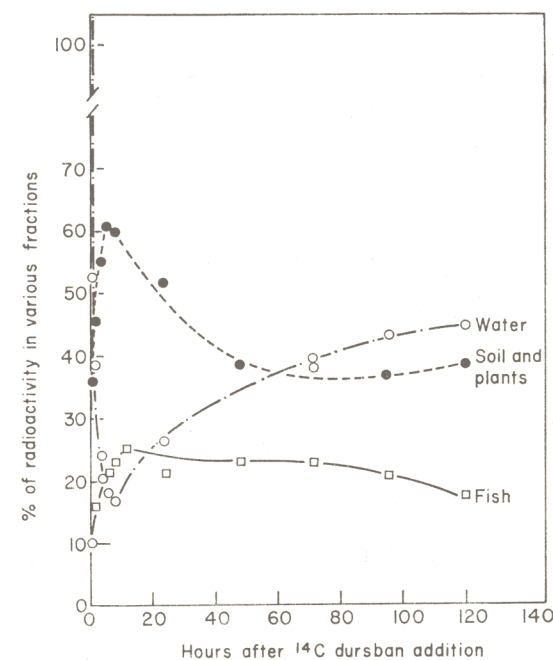


Fig. 3. Distribution of ^{14}C Dursban in the ecosystem.

original data is so small that the effects of this transformation are minimal. It will be assumed, therefore, that error properties 1-3 are still valid and that for each component, the error is normally distributed with zero means and a constant standard deviation of 1%.

C. BUILDING THE MODEL

The simplest model which can be postulated to explain the data of Table I is to assume (i) that an equilibrium exists between the chemical

in the water and the soil and plant constituents, and (ii) a direct uptake of the chemical by the fish. This model can be represented symbolically as follows:



where

$A = {}^{14}\text{C}$ in the water
 $B = {}^{14}\text{C}$ in the soil and plants
 $C = {}^{14}\text{C}$ in the fish

and k_1, k_2, k_3 are reaction rate constants in h^{-1} . It is further assumed that all the steps or reactions are first order. Mathematically, therefore, the model is represented by the following differential equation system

$$\begin{aligned} \frac{dx_A(t)}{dt} &= -k_1 x_A(t) + k_2 x_B(t) - k_3 x_C(t) \\ \frac{dx_B(t)}{dt} &= k_1 x_A(t) - k_2 x_B(t) \\ \frac{dx_C(t)}{dt} &= k_3 x_B(t) \end{aligned} \quad (21)$$

with initial conditions $x_A(0) = 100$, $x_B(0) = 0$ and $x_C(0) = 0$. Here, $x_A(t)$, $x_B(t)$ and $x_C(t)$ are the percentages at time t of A , B and C respectively with the restriction that

$$x_A(t) + x_B(t) + x_C(t) = 100 \quad (22)$$

Using a nonlinear parameter estimation program, it is possible to find the parameter values $k_1^* = 0.510$, $k_2^* = 0.800$ and $k_3^* = 0.00930$ which best describe the data of Table I. Corresponding to these parameters, the overall residual sum of squares is $RSS = 5374$. Residuals for each of the three components measured can be calculated. Since an independent estimate of error is available, i.e. $s^2 = 1$ for all three measurements, the lack of fit relation 19 can be applied directly with the numerator degrees of freedom $n - p = 36 - 3 = 33$ to give

$$\frac{RSS/(n-p)}{3 \cdot s^2} = \frac{5373/33}{3} = 54.3 > F_{0.05}(33, 20) = 1.44$$

Since this ratio is considerably greater than the tabulated F value, the model is totally inadequate. By a residual analysis it might be possible to identify the specific inadequacies in the model. Figure 4 is a plot of the residuals for each of the measured components versus the independent variable time. This residual plot reveals the following

discrepancies: (i) initially the model predicts a higher proportion of ${}^{14}\text{C}$ in the water and a lower proportion in the fish, (ii) after 72 h the model predicts the opposite of (i), and (iii) the model predicts low proportions of ${}^{14}\text{C}$ in the soil and plants throughout the experiment.

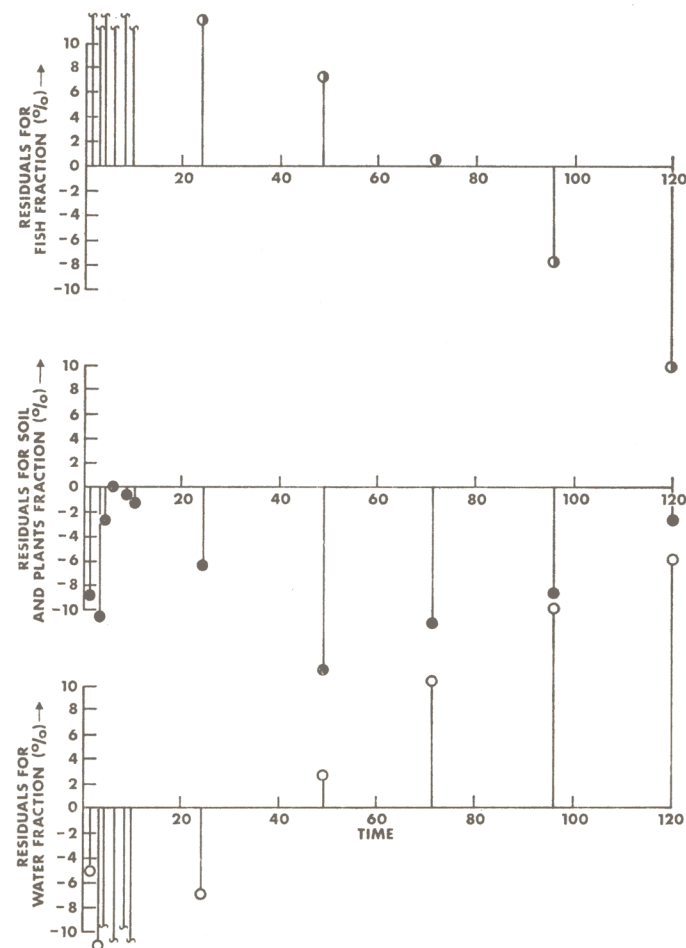
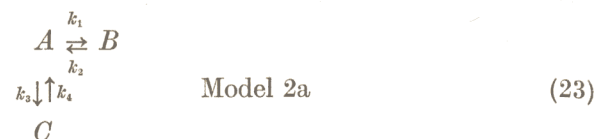


FIG. 4. Independent variable residual plot for Model 1.

The next step in the modelling process is to use this residual analysis to postulate a better model. The large negative residuals observed in predicting the ${}^{14}\text{C}$ proportion in the fish after 80 h, confirm a major limitation of Model 1. That is, Model 1 predicts an ever increasing proportion of ${}^{14}\text{C}$ in the fish. To compensate for this trend, Model 2 is

postulated where the chemical in the fish is excreted from the fish either (a) unchanged

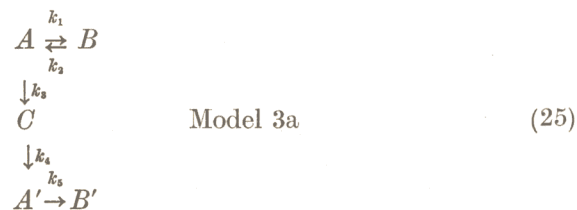


or (b) metabolized and excreted as a new entity A'



The results of fitting the differential equations corresponding to these models to the experimental data is presented in Table II. First, note that both forms of Model 2 indicate a lack of fit so that additional modifications will be necessary. Secondly, the two forms of Model 2 can be compared with themselves and Model 1 by calculating likelihood ratios. The likelihoods for the different models are shown in Table III. Obviously, Model 2b is superior to both Models 1 and 2a, indicating that the existence of the entity A' is highly probable. Residual plots for both forms of Model 2 are shown in Figs 5 and 6 respectively. Relative to the other components, the residuals for the ^{14}C in the fish are reasonable, although considerably larger than expected from experimental variations alone. However, the residuals for the best-to-date Model 2b indicate that the Model predicts higher proportions of ^{14}C in the soil and plants than indicated by the data during the first 60 h, and lower proportions during the last 40 h. Analogously, the predicted water proportions during the first 60 h are lower than indicated by the data and higher during the last 40 h.

In order to bring the proportions between soil and plants and water into agreement, the next step is to give the entity A' access to the soil and plants. Thus two forms of a new Model 3 were examined. The first is a simple uptake of A' by the soil and plants



while the second postulates an equilibrium relationship

TABLE II
Parameter estimation and lack of fit analysis

Model number (j)	Number of parameters (p)	Optimal parameter estimates (k^*)	Residual sum of squares (RSS _j)	Lack of fit mean square (LOFMS)	LOFMS Error	$F_{0.05}(n-p, 20)$
1	3	$k_1 = 0.510$ $k_2 = 0.800$	537.4	162.8	54.3	2.01
2a	4	$k_1 = 0.493$ $k_2 = 0.299$	196.4	61.4	20.5	2.02
2b	4	$k_1 = 0.286$ $k_2 = 0.0277$	84.8	26.9	8.97	2.02
3a	5	$k_1 = 0.337$ $k_2 = 0.069$ $k_3 = 0.104$	208.3	6.72	2.24	2.03
3b	6	$k_1 = 0.338$ $k_2 = 0.069$ $k_3 = 0.104$	207.9	6.93	2.31	2.04
4a	7	$k_1 = 0.338$ $k_2 = 0.0515$ $k_3 = 0.136$ $k_4 = 0.0788$	58.6	2.02	0.673	2.05
4b	6	$k_1 = 0.336$ $k_2 = 0.0572$ $k_3 = 0.124$	79.4	2.64	0.880	2.04



In these models, a new entity B' distinct from B is assumed. The results of fitting these models to the data and the calculated likelihoods are shown in Tables II and III respectively. These models exhibit a lack of fit of the data. However, the likelihood ratios L_{3a}/L_{2b} and L_{3b}/L_{2b} show a marked improvement by Model 3 over Model 2 in fitting the data. Since the likelihood ratio L_{3a}/L_{3b} is approximately unity, it is impossible to discriminate between the two forms of Model 3. In other words, the reverse reaction $B' \rightarrow A'$ does not improve the ability of the

TABLE III
Likelihood analysis

Model number (j)	Residual sum of squares (RSS _j)	Maximum likelihood for model j (L _j [*])	Likelihood ratio (L _{3a} [*] /L _j [*])
1	5374	1.44×10^{-389}	3.98×10^{385}
2a	1964	6.95×10^{-193}	8.24×10^{137}
2b	848	4.16×10^{-62}	1.37×10^{57}
3a	208.3	8.37×10^{-16}	6.84×10^{10}
3b	207.9	8.95×10^{-16}	6.40×10^{10}
4a	58.6	5.73×10^{-5}	1
4b	79.4	1.79×10^{-6}	32.0

model to explain the data so that $k_6 = 0$. The residuals for Model 3a are plotted in Fig. 7. A comparison of Figs 5 and 7 shows the striking improvement in predictability of Model 3 over Model 2. These residuals show that the consequences of bringing the proportions of chemical in the water and soil and plants into better agreement have decreased the ability to predict the proportions in the fish. Further, it appears that a low prediction of chemical in the fish is accompanied by a high prediction of chemical in the water.

It may be possible to improve the distribution of chemical between the fish and water by postulating that (a) the chemical in the fish partitions into a second compartment (e.g. the flesh), or (b) the entity A' in Model 3a is in equilibrium with the fish. Modifying Model 3a to

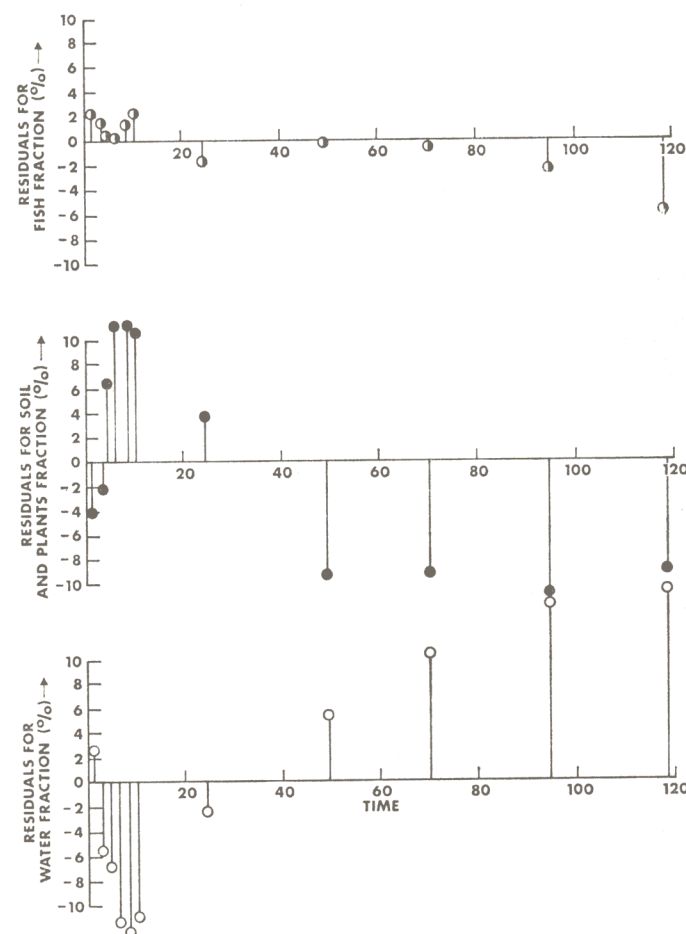
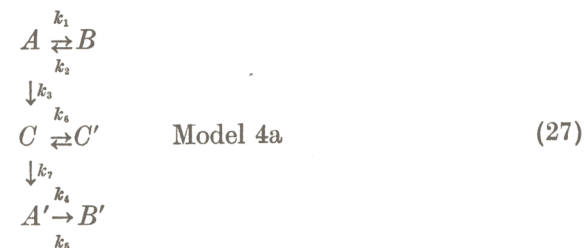
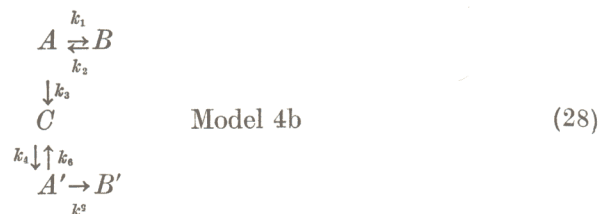


Fig. 5. Independent variable residual plot for Model 2a.

include a second compartment in the fish gives



where C' is the proportion of ^{14}C in the second compartment. Model 4b is obtained simply by making the step $C \rightarrow A'$ reversible



These models were fitted to the data and the results are presented in Tables II and III. Both of these models adequately describe the data according to the lack of fit criterion. The residual plots shown in Figs 8 and 9 do not reveal any major discrepancies in the chemical distributions among the major components, although the residuals are some-

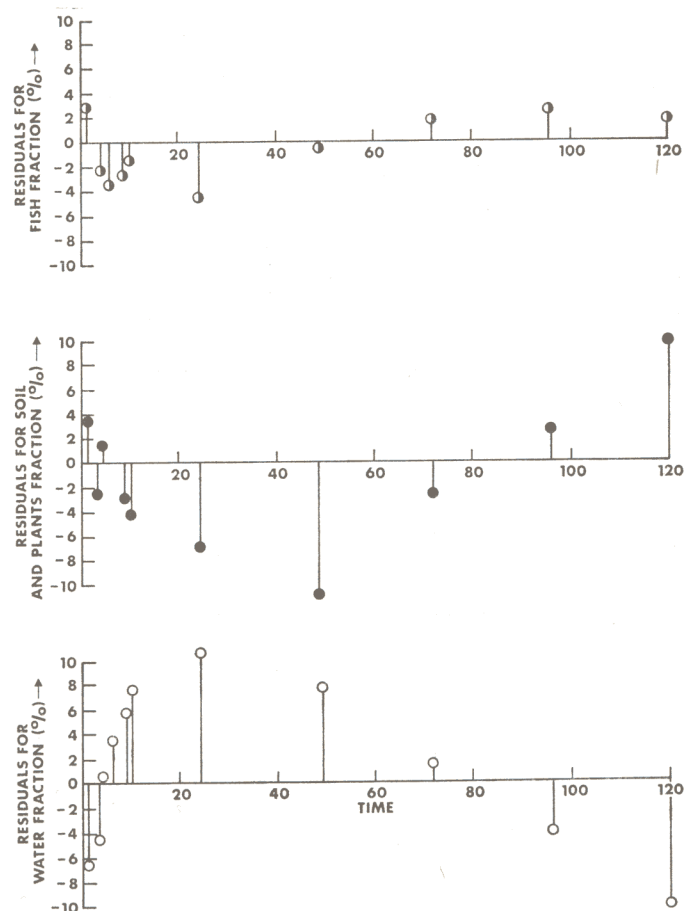


FIG. 6. Independent variable residual plot for Model 2b.

what larger for Model 4b than for Model 4a. Non-parametric statistical tests indicate that the residuals are indeed plausible estimates of normally distributed experimental error (Draper and Smith, 1966). Based on residual analysis alone, either form of Model 4 is valid and further refinement of the model to better explain the data is not

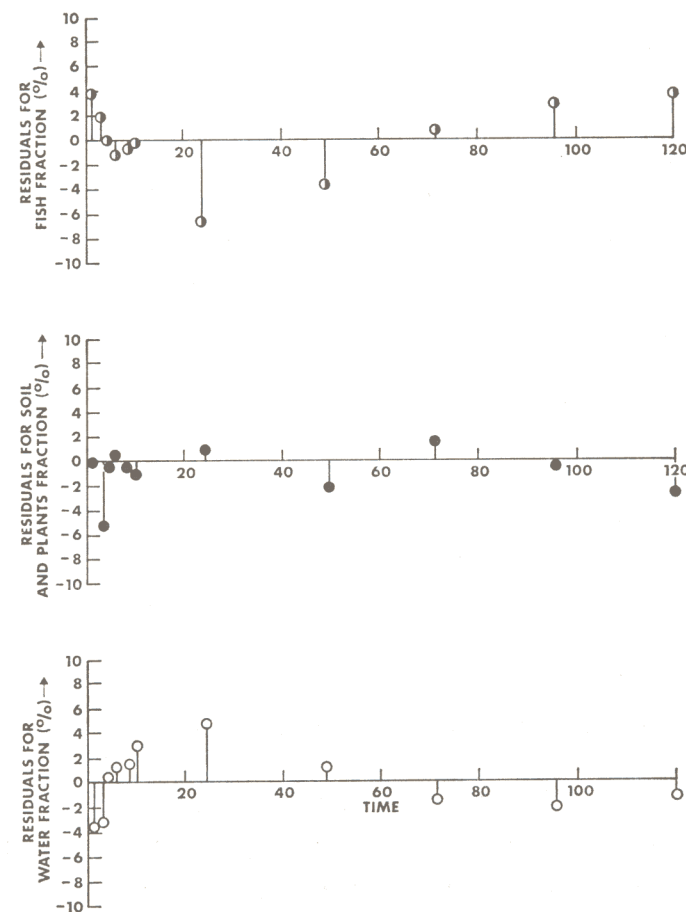


FIG. 7. Independent variable residual plot for Model 3a.

warranted. The likelihood ratio $L_{4a}/L_{4b} = 32.0$ implies a preference for Model 4a over 4b. This is equivalent to saying that there is strong evidence that a second compartment is set up in the fish. Since the value is less than 100, however, it is difficult totally to reject Model 4b without additional experimental work.

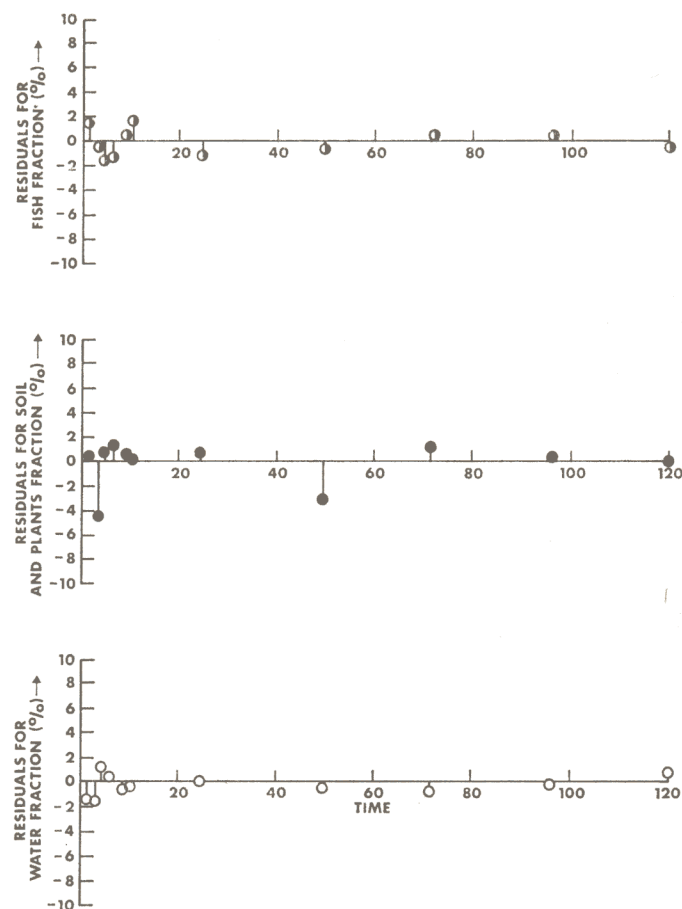


FIG. 8. Independent variable residual plot for Model 4a.

D. DISCUSSION OF RESULTS

The final model that emerges from this analysis is the following:

1. There is a rapid equilibration between the applied DURSABAN® and the soil and plant system. This step was also seen in the work reported by Smith *et al.* (1966).
2. This is followed by a slower uptake of the insecticide by the fish.
3. Once in the fish the material is metabolized and excreted. The metabolite is probably the pyridinol which was identified in the water at the termination of the 120 h exposure (Smith *et al.*, 1966).
4. The liberated pyridinol in the water is again taken up by the soil and plants.

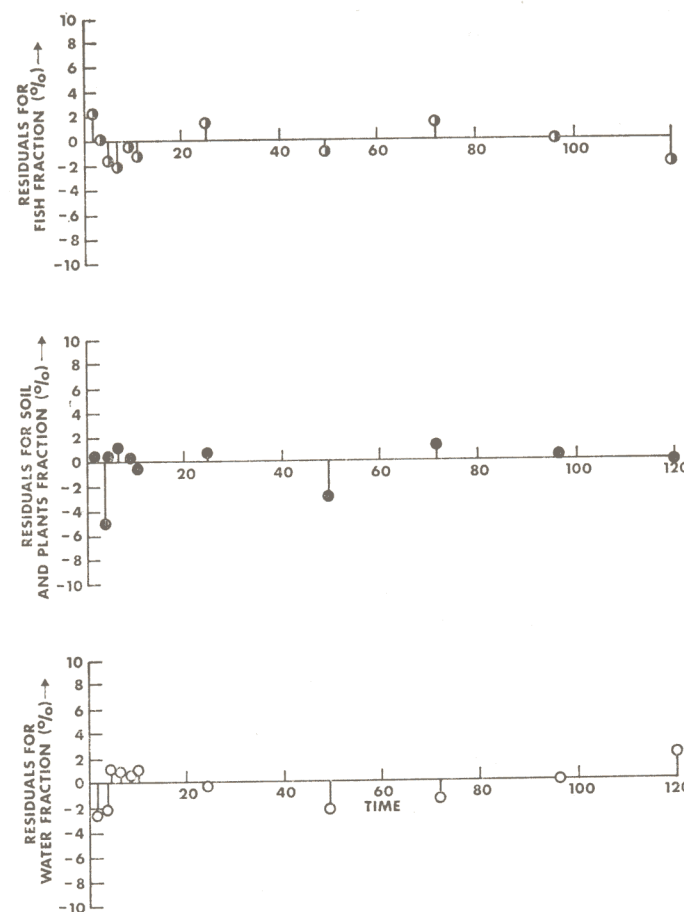


FIG. 9. Independent variable residual plot for Model 4b.

5. The best fit was obtained with Model 4a which includes a partitioning of the material between two compartments in the fish. Again the data obtained by Smith would tend to substantiate this step in that these authors demonstrated a partitioning between the viscera and the meat.

6. Finally, Models 4a, 4b and 3a all indicate that the final sink for the added Dursban is the soil and plants. This last item is very important since Smith (1966) has shown that the 3,5,6-trichloro-2-pyridinol is metabolized readily by plants and will ultimately be degraded to CO_2 , NH_3 and H_2O . Such a situation would imply that there is no persistence of Dursban in this particular ecosystem.

The fast initial absorption of the insecticide by the soil and plants has an added advantage in that this particular sink acts as a reservoir for the slow release of Dursban. This feature gives added long-term protection for the control of mosquito larvae in polluted waters. Schaeffer and Dupras (1970) demonstrated that a similar series of events occurred in a field trial.

E. CONCLUSION

The model building exercise in this paper has generated a picture of the distribution pattern of DURSBA[®] when added to a pond of water. Furthermore, the picture that emerges is compatible with what is known about the insecticide.

REFERENCES

- Bard, Y. and Lapidus, L. (1968). Kinetic analysis by digital parameter estimation. *Catalysis Reviews*, **2** (1), 67-112.
- Barnard, G. H., Jenkins, G. M. and Winsten, C. B. (1962). The likelihood inference and time series. *Jl R. statist. Soc. Ser. A*, **125**.
- Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *The Philosophical Transactions* **53**. Reprinted in *Biometrika* **45** (1958).
- Blau, G. E., Klimpel, R. R. and Steiner, E. C. (1970). Equilibrium constant estimation by nonlinear optimization. *Ind. Eng. Chem. Fundam.* **9**, 334-339.
- Blau, G. E., Klimpel, R. R. and Steiner, E. C. (1972a). Equilibrium constant estimation and model distinguishability. *Ind. Engng Chem. Fundam.* **11**, 372-3.
- Blau, G. E., Klimpel, R. R. and Steiner, E. C. (1972b). Parameter estimation and model distinguishability of physicochemical models at chemical equilibrium. *Can. J. chem. Engng* **50**, 324-332.
- Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation. *Jl R. statist. Soc. Ser. B*, **26** (2), 211-252.
- Box, G. E. P. and Hill, W. J. (1967). Discrimination among mechanistic models. *Technometrics*, **9** (1), 57-71.
- Draper, N. R. and Smith, H. (1966). "Applied regression analysis." John Wiley and Sons, New York.
- Hsiang, T. and Reilly, P. M. (1971). A practical method for discriminating among mechanistic models. *Can. J. chem. Engng* **50**, 865-871.
- Kittrell, J. R. (1970). Mathematical modelling of chemical reactors. *Adv. chem. Engng.* **8**, 97-183.
- Reilly, P. M. (1970). Statistical methods in model discrimination. *Can. J. chem. Engng.* **48**, 168-173.
- Rosenbrock, M. M. and Storey, C. (1966). "Computational Techniques for Chemical Engineers." Pergamon Press, Oxford.
- Schaeffer, C. H. and Dupras, E. F. Jr. (1970). Factors affecting the stability of Dursban in polluted waters. *J. econ. Ent.* **63**, 701-705.
- Smith, G. N. (1966). Basic studies on Dursban insecticide. *Down to Earth* **22**, 3-7.

- Smith, G. N., Watson, B. S. and Fisher, F. S. (1966). The metabolism of [¹⁴C] O, O-diethyl O- (3, 4, 6-trichloro-2-pyridyl) phosphorothioate (DURSBA[®]) in fish. *J. econ. Ent.* **59**, 1464-1475.
- Solberg, J. J. (1972). Principles of system modelling. *Proc. Int. Symp. Systems Engng and Analysis*, 67-74.