Ann Zootech (1996) 45, Suppl, 143-152

Use of mathematics and statistics in nutrition modelling

G Ciuperca [1], R Tomassone [1,2], JP Flandrois [2,3]

[1] *Institut National Agronomique, Département Mathématique et Informatique, 16 rue Claude Bernard, 75231 Paris cedex 05;* [2] *CNRS URA 2055, Université Claude Bernard, 43 Bd du 11 novembre 1918, 69622 Villeurbanne cedex* [3] *Laboratoire de Bactériologie, Faculté de Médecine Lyon-Sud, BP 2, 69921 Oullins, France*

Summary - For over twenty years dynamic modelling has been used in the biological sciences. Broadly speaking, biologists have reproduced what has long been used in the physical sciences by researh workers and engineers. Essentially, they use dynamic system modelling and control theory. Nevertheless a random component is seldom introduced, although it plays a central role in biological systems. In this paper we illustrate the major aspects of dynamic system processing, we also try to show how developments in statistics have introduced new ideas which may be of some use in this field. The metabolism of glucose in goats is used as an illustration of numerical analysis and statistics symbiosis.

Introduction

Mathematical biology is a fast growing, even not clearly defined, subject application of mathematics. As biology becomes more quantitative, the increasing use of mathematics is inevitable. But research and applications in this field, to be useful and interesting, must be relevant from a biological standpoint. We are going to present the most important ideas, to our mind, and apply them to animal nutrition. Of course, these ideas are developed in some important text-books, recently published (Murray, 1989; Brown and Rothery, 1993; Pavé, 1994): the interested reader will increase his knowledge by consulting them.

As it is now well established «*the efficiency with which absorbed nutrients are used for fat synthesis is an important factor in determining the feed requirements of animal*» (Gill et al, 1984). The analysis of fat synthesis is the analysis of a very complex «system», surely far away of a simple mathematical modelling. A mathematical modelling is always a simplification of a real world; but as a matter of fact, this simplification imposes to an experimenter to isolate the most important effects able to describe the entire system, even if he has to know that a model is always a virtual description of a specific reality.

He may use a simplified model as an instrument to describe a complex situation; but his ambition may be more important and he may use the model to analyze situations he is not able to experiment. As an example, he may compute parameters he is unable to measure directly, such as transfer coefficient between two organs. He may also use the model to analyze new situations and try to control the dynamic of the system under consideration.

A specific example: metabolism of glucose

When studying the metabolism of glucose of a goat an experimenter may have in mind the simplified model described in figure 1. This model is a compartmental one where some compartments are hypothetical, such as the

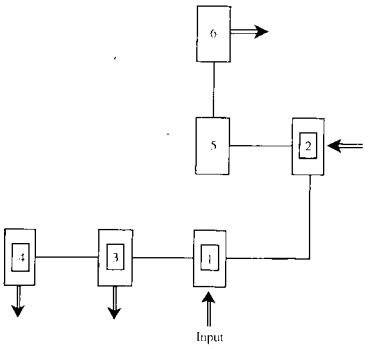

Figure 1. Metabolism of glucose in goat. →: represents an outpout from the compartment. Only the compartment number in bold bordered character may give rise to measurements.

existence of two insulin compartments. One of his aim may be to verify that they really exist with experimental data.

The aim of such a model is to analyze the dynamic of this metabolism, when some injection of glucose are made at the beginning of the experiment; the same model, with minor modification, is also used for insulin injection. The model has six compartments as shown in figure 1, their definition is the following:

1: glucose in blood
2: insulin in pancreas
3: non esterified fatty acid
4: β hydroxybutyrate
5: glycogen in a first compartment only for transfer
6: glycogen in a second compartment for stocking

To make measurements, the experimenter has only access to compartments 1, 2, 3 and 4. For more details see Sauvant and Grizard (1992), Ciuperca (1996).

Classical methodological approach

The methodological approach for such models is based on dynamic modelling which appears well suited to their processing. This sort of modelling is classical in physical sciences; the biologists have made a transposition of ideas and tools already well known and used with success by engineers. As we shall see later this analogical approach has certain limits.

Variables

In such models variables are quantities that change in time and, as presented by France and Thornley (1984), they can be considered under four categories:

- *state variables*: if there are m states variables, we shall note them $X_i(t)$ (or simply X_i knowing that this variable is time-dependent), i= 1,...,m; the total set will be noted as a vector in bold character:

$$\mathbf{X} = [X_1, X_2,...,X_m]^T \quad (1)$$

where the symbol $[...]^T$ represents a transposed vector. Sometimes we only know values of a subset of n state variables (n<m) at given times $t_0, t_1,...,t_N$; for presentation, there is no loss of generality to consider them as the first n **X** components. For glucose metabolism introduced previously m = 6, and n = 4. For these variables at time t_j, $X_i(t_j)$, we shall have measurements u_{ij} (the measurements for time t_0 correspond to initial conditions of the system):

$$u_{1j}, u_{2j},...,u_{nj}, (j = 1,...N) \quad (2)$$

- *rate variables*: a rate variable is a quantity per unit of time; generally, it cannot be measured instantaneously. A set of rate variables define the process at a given time; if X is the weight of an organism, a simple model may be:

$$X = X_0 + bt \quad (3)$$

the growth rate is obtained by differentiating both sides of equations (3) with respect of time t, therefore:

$$dX/dt = b \quad (4)$$

For the whole set of **X** variables, we note:

$$d\mathbf{X}/dt = [dX_1/dt, dX_2/dt,...,dX_m/dt]^T \quad (5)$$

- *auxiliary variables*: sometimes one wishes to obtain certain extra variables which are more useful for the experimenter's convenience. They may be sum of several state variables or ratio of two. The relative growth rate (1/X)dX/dt which involves a state variable (X) and its rate (dX/dt) is one of the much used auxiliary variable. As the state variables alone define the system completely, they are only additional.

- *driving variables*: driving variables are data inputs to a model; they may vary autonomously with time, if q such variables exist, we shall note them:

$$\mathbf{I} = [I_1, I_2,...,I_q]^T \quad (6)$$

They may be used to analyze the effects of a specific drug, in order to control the dynamic of the entire system. In metabolism of glucose, we have two exclusive such variables (q = 2) either input of glucose or input of insulin, for a short time beginning at t_0.

Parameters and constants

We also need the introduction of parameters and constants which will appear in the equations of models that do not vary with time. The density of water or the number of minutes in a day are typical constants; if constants status is unambiguous, it is not the same for

parameter status. The value of a Michaelis-Menten constant in enzyme kinetics may be well established, in this case it may be considered as a constant; but as the case may be it has to be determined. If there are p such parameters we shall note:

$$\mathbf{a} = [a_1, a_2,...,a_p]^T \quad (6)$$

For glucose metabolism we shall introduce ten parameters (p = 10).

Differential equations

The m state variables define the system at time t. A deterministic model consists of m first-order differential equations which describes their change in time:

$$dX_i/dt = f_i\{X_1,X_2,...,X_m; \mathbf{a},\mathbf{I},t\}, (i = 1,...,m) \quad (7)$$

where the f_i denote some generally empirical functions of the state variables **X**, of the parameters **a**, of the driving variables **I**, and perhaps of time; these functions have no to contain all the variables (state, parameters, driving). For the complete set of state variables we may write:

$$d\mathbf{X}/dt = \mathbf{F}\{\mathbf{X}; \mathbf{a},\mathbf{I},t\} \quad (8)$$

Sometimes, it is also interesting in the modelling process to make the current value dX/dt depend upon the value of X not only of time t, i.e. through X(t), but also of time τ ago, $X(t-\tau)$. An equation as dX/dt = f{X} is replaced by:

$$dX/dt = f\{X(t),X(t-\tau)\} \quad (9)$$

This is called a discrete lag, and perhaps τ has to be considered as a new parameter and could be estimated. More realistic, but more complicated, is use of a distributed lag.

Numerical problem

Numerical integration

Given the parameters values **a**, the driving variables **I** and the initial conditions:

$$X_i(t = 0), (i = 1,...,m) \quad (10)$$

the differential system can be solved by integrating equations (7) or equivalently (8); the solution is a set of predicted values at different time values. Seldom, the system can be solved analytically; the only classical case arises when parameters are linearly introduced in the model, as it occurs in classical linear compartmental analysis. In this case (8) may be written:

$$d\mathbf{X}/dt = \mathbf{A}\,\mathbf{X} + \mathbf{k}(t) \quad (11)$$

where **A** is a m*m matrix of coefficients (parameters and constants), where state variables do not appear. In this case the analytical solution is (Tomassone et al, 1993):

$$X_i = b_{0i} + b_{1i}exp(l_1t) + b_{2i}exp(l_2t) + ... + b_{mi}exp(l_mt) \quad (12)$$

In (12) l_1, l_2,...,l_m are the eigen values of **A**. Except for this case, it is necessary to use numerical integration techniques to obtain solutions:

$$X_i(t) = X_i(\mathbf{a},\mathbf{I},t) \quad (13)$$

Now good computer softwares are doing this quite easily, but a lot of difficulties still exist and the experimenter cannot ignore them. One of the most important is linked with the modelling process itself. A typical animal model may represent conversions between molecules (they occur in few milliseconds), constructions of new membranes (hours or days) and productions of new organs (weeks). A numerical integration must take account of this: we have, or perhaps the computer software has, to choose an integration interval. An interval that suits the slow process will give rise to instable and increasing oscillations with the fast one; conversely, if it suits for the fast one, it may be too long in time for the slow one. This is known as stiff equations.

The fitting to data

As noted previously, parameters have seldom perfectly known values, and the integration results may give some important distortion when compared to real data. In this case the model has only a qualitative interest; sometimes it may be sufficient.

But if we have measurements u_{ij} (j=1,...,N) at time values t_j, it may seem interesting to compare them to numerical values $\hat{u}_{ij}$ obtained by integration namely $X_i(\mathbf{a},\mathbf{I},t_j) = \hat{u}_{ij}$, or more concisely $X_i(t_j)$. The way to obtain the best predicted values for state variables **X** is therefore to consider **a** parameters as unknown and to try to obtain the best

estimation $\hat{\mathbf{a}}$ by minimizing an objective criterion like:

$$S(\mathbf{a}) = \Sigma_{ij} w_{ij} [X_i(t_j) - u_{ij}]^2 \quad (14)$$

where $X_i(t_j)$ are values obtained in (12) when $t = t_j$, and w_{ij} are weight depending of observations. In physical modelling this is known as «calibrating» or «tuning», in statistics as «estimation». This procedure may be interpreted as a statistical modelling process, where the values obtained by numerical integration are considered as random variables U_{ij} obtained by the following model:

$$U_{ij} = X_i(\mathbf{a}, \mathbf{I}, t_j) + \varepsilon_{ij} \quad (15)$$

where u_{ij} is a realization of this variable and ε_{ij} are random variables such that $E\{\varepsilon_{ij}\} = 0$ and $var\{\varepsilon_{ij}\} = \sigma^2$. The difficult question concerns their independence which is seldom admissible.

We must note here that another approach is to introduce directly in (8) random components, this give rise to stochastic differential equations. Theoretically, this approach is surely more interesting and full of promise, but we don't know practical applications (Gard, 1987).

As noted by France and Thornley (1984) an epistemological problem concerns the simultaneous processes of dynamic modelling and of statistical estimation. Some scientists think that the parameters should be known from independent investigations. To our mind, it would be a pity not to use all the information contained in an experiment. On the contrary, it is surely important to take account of it to have more insight in the modelling process itself, to validate it, and for sure to improve it.

The role of statistic ideas

We are going to use statistical analysis to delineate some useful ideas to apply in dynamic modelling.

Design of experiment

In a statistical analysis, the first step is to make an optimal experiment to obtain the best estimation, generally the most precise and stable estimators. Here, this aspect concerns the good choice of times for measurements. It is well known that in a Michaelis-Menten model:

$$X = V_{max}\, t / (k_M + t) \quad (16)$$

$$\text{where } a_1 = V_{max},\ a_2 = k_M$$

if we want to obtain precise estimations for both a_1 and a_2 we must have observations for t values near from a_2 and for high t values (theoretically to infinity !). If we don't do this, the estimations will be less precise. The problem is the same for dynamic modelling, but more complicated (Vila, 1985).

A specific problem may also occur: if we have to take a blood specimen at different times, it is physiologically impossible to take it at short intervals; so we have to impose some constraints to time values as the difference between two blood-taking is greater than a specified value, say 4 minutes.

One of the major problem in non-linear situation is that designing an efficient experiment will require knowledge of parameter, but the purpose of the experiment is to generate data to yield parameter estimated ! The experiments to be considered have two fundamental stages: a static design in the initial one, followed by a fully adaptive sequential one in which the design points are chosen sequentially and using parameter estimates based on available data (Chaudhuri and Mykland, 1993).

Reformulations when collinearities are present

One of the major problem with statistical models, either linear or non-linear, is the instability of parameters. As it is known the minimization of $S(\mathbf{a})$ in (13) is done by iteration; each step consists in finding the solution of linear model (regression model). The final iteration gives a solution $\hat{\mathbf{a}}$ and an usual variance estimate:

$$var\{\hat{\mathbf{a}}\} = \sigma^2\, \Omega^{-1} \quad (17)$$

where Ω (a p*p matrix) is computed through the derivatives of S respect to each component of $\mathbf{a}$. If this matrix is ill-conditioned the estimators have large variances and are highly correlated.

The simplest idea is to obtain linear combinations of initial p parameters through the computation of principal components of Ω, and to consider only the first ones associated

with the largest eigen values. If this number is k (k<p), the p-k remaining may be computed as linear combinations of the first k. The non-linear estimation is simplified (Box et al, 1973; Simonoff and Chih-Ling Tsai, 1989). After having estimated these k parameters, it is easy to compute the p-k others with their associated covariance structure.

Resampling techniques and influential data

The influence of some observations may modify drastically the estimations. As it is always difficult to have another experiment to validate the first result, the idea is to use the jakknife technique, deleting each observation one at a time to obtain jackknifed estimations. This procedure, which is time consuming, is now quite easy with the increase speed of computers. Jackknife introduces pseudo-values which are of the primest interest to detect observations with great influence on vector parameters **a** of the model (or on functions of these parameters).

Analysis of variance strategy (ANOVA)

Sometimes we have several animals introduced in the same experiment, and one of the aims of the analysis may be to analyze the variability between animals. We may use some ANOVA-like strategy, trying to see if a common model may be applied to each animal. In this case we are in a multivariate (MANOVA) situation, and we may use pseudo-values of the parameters obtained during jackknife estimation as «pseudo-data» to compare animals (Tomassone et al, 1993).

Prediction

One of usefulness of a model is its ability to explore situations where experiments are not possible. In metabolism of glucose, we have said that only four compartments are attainable by measurements. But the two others, as elements perhaps hypothetical of the global system, may be analyzed and their dynamic may be computed; their evolution may give insight into what cannot be measured.

Of course, predictions are possible, and used with statistical assumptions, may give most probable values for a dynamic associated with confidence bands for evolution (Audrain and Tomassone, 1994).

Example

If we look at the model described in section «A specific example: metabolism of glucose», we may construct a set of six differential equations. We don't give here the details of these equations which may be found elsewhere (Ciuperca, 1996). As an example, the differential equations for fatty acid in plasma and for ß hydroxybutyrate are the following:

$$dX_3/dt = c_1[1+(10^3X_2/a_1)^{a_{10}}]^{-1} - a_2 X_3 - a_3 X_3$$
$$dX_4/dt = a_2 X_3 - a_4 X_4$$

where c_1 is a constant with known value. The entire system is highly non-linear. The links between compartments are indicated in table I, where the ten coefficients are identified.

The initial data are the four different concentrations (X_1,X_2,X_4 and X_5) at ten

Table I. Links between compartments. A link is identified by a «*»; the 10 numbers show the coefficients of vector a to estimate.

From\To	1	2	3	4	5	6	Out	In
1	*	6,8			*			I
2	7,9	*	1,10				5	I
3			*	2			3	
4				*			4	
5	*	*			*	*		
6					*	*		

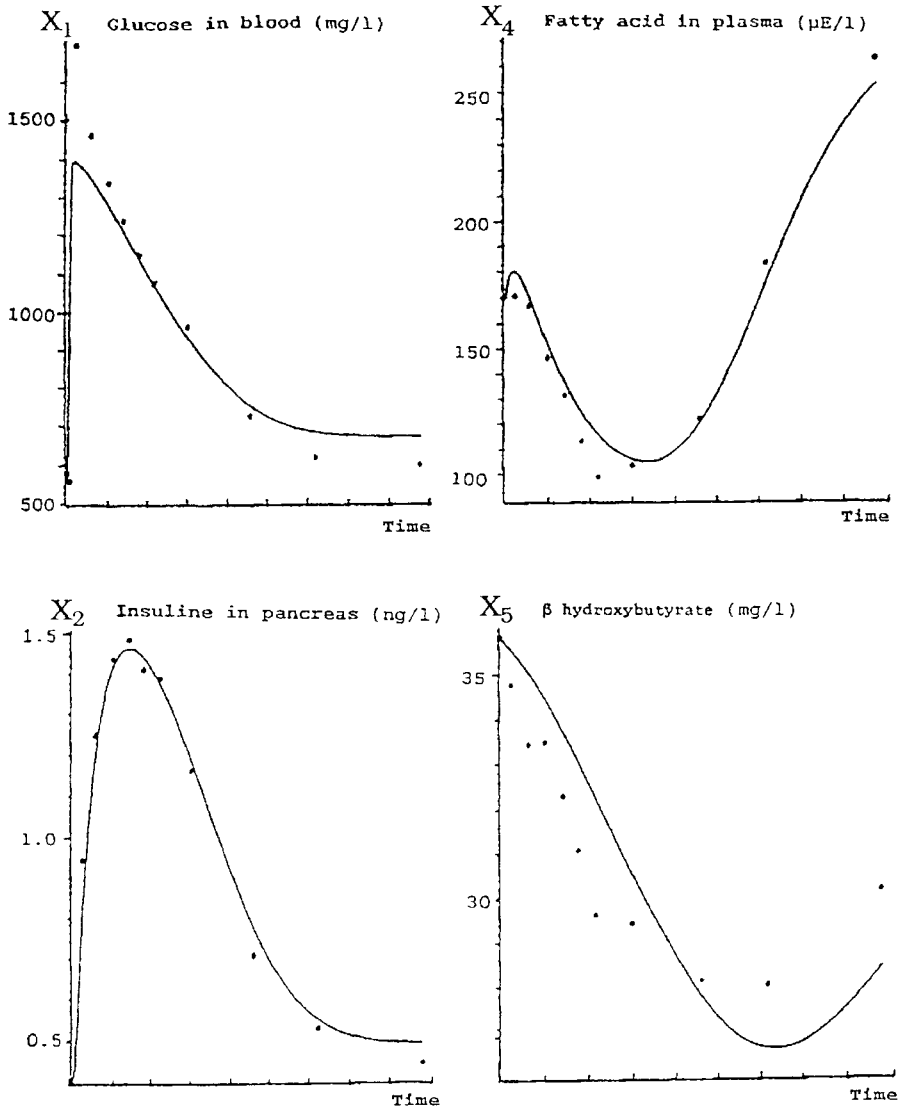

Figure 2. Metabolism of glucose: results of the integration for the four concentrations, with data points (■). The curves come from the integrated values. $X_i(t)$, for t from 0 to 90 minutes (the scale for time is 10 minutes).

Table II. Different estimations of the four parameters. Standard errors are indicated in parenthesis.

Estimation with parameter	10 parameters	4 parameters	Jackknife
a_1	0.034 (0.018)	0.037 (0.002)	0.043 (0.003)
a_2	0.0029 (0.00037)	0.0027 (0.00031)	0.0025 (0.00021)
a_5	0.301 (0.016)	0.224 (0.006)	0.207 (0.003)
a_7	0.56 (0.03)	0.60 (0.02)	0.59 (0.01)

different times for 24 goats. The values at time $t_0 = 0$ are considered as the initial conditions. The times (in minutes) were the following: 2, 6, 10, 14, 18, 22, 30, 46, 62, 88

As a first step, the means of measurements serve as data to be fitted, these data represent a sort of «mean-goat». The results of integration are illustrated on figure 2. They show a quite good fitting, even if the jump at the beginning for X_1, due to glucose injection input, was difficult to manage.

The quite large instability of estimation is corrected by the computation of eigen vectors of the estimated covariance matrix of the **a**. Four coefficients (a_1,a_2,a_5,a_7) are enough to describe correctly the system, their covariance matrix is now well conditioned.

To show the differences between different estimation procedures (and even if these numerical values have no sense for the reader in this context) the estimations of the four retained coefficients are given on Table II. It appears clearly that the reduction in the number of parameters induces a decrease of their standard error. The jackknife estimation introduces no real improvement in the estimations themselves, but indicates that measurements at some times may be crucial in the estimation: t_9 for a_1, t_{10} for a_2, t_7 for a_5 and t_{10} for a_7, as it may be noticed on figure 3.

Of course, the six deleted parameters are also computed as linear combinations of the four used in estimation stage; their standard error is also computed.

On figure 4 we may see the estimated glucose concentration for every 24 goats; the same procedure was applied and the parameters were reestimated. This permits to identify specific goats for which the different state variables were badly fitted (as goat line 2, column 3). The experimenter may come back to his data and try to find a biological explanation, and perhaps to delete this goat.

Using the obtained estimations to specify what could be an optimal design, the times given in table III were obtained with two different constraints (1 minute or 4 minutes between blood-taking).

The result is clear for the experimenter: he has to make the first six measurements at the beginning of the experiment and the last ones immediately after t = 66 (for 1 minute constraint) or t = 55 (for 4 minutes constraint). The criterion for the optimality is a D_{10}-optimal design, a design with ten observations; the criterion is better for 1 minute than for 4. This

Table III. Times obtained with two different contraints (1 minute or 4 minutes between blood taking)

time	1	2	3	4	5	6	7	8	9	10
real design	2	6	10	14	18	22	30	46	62	88
optimal design constraint 1mn	8	9	11	12	13	20	66	67	68	69
optimal design constraint 4mn	4	8	12	16	21	27	55	59	64	68

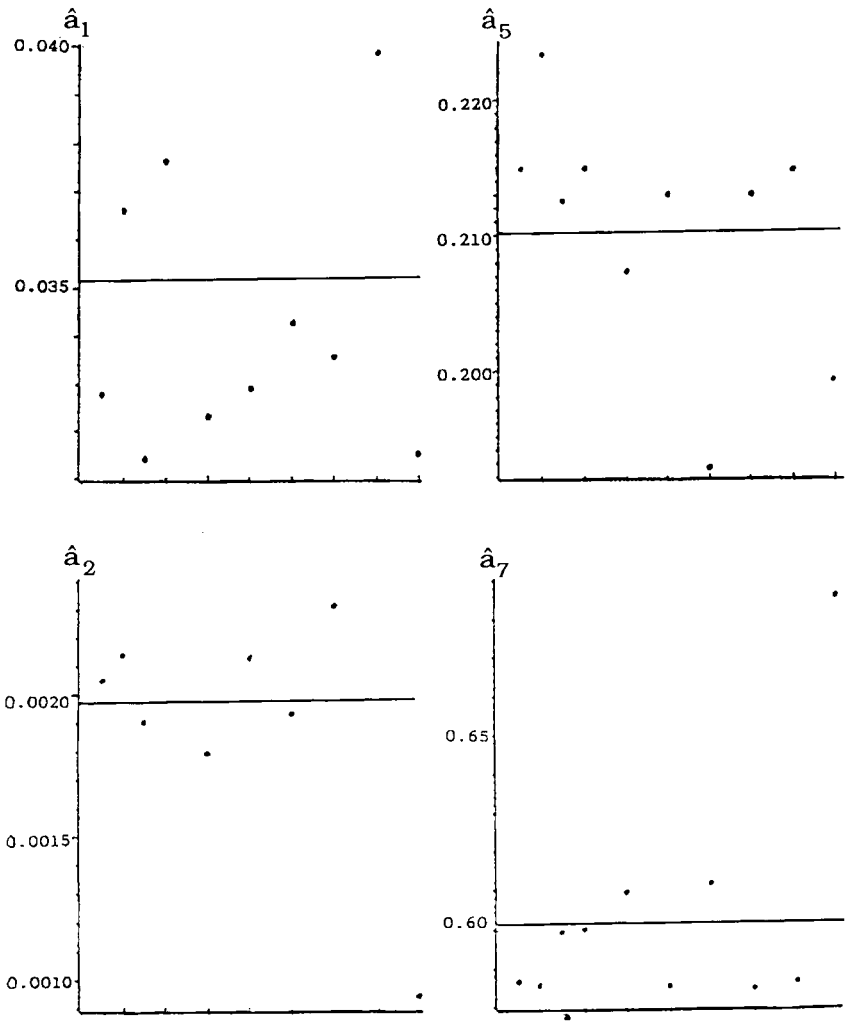

Figure 3: Metabolism of glucose: jackknifed estimations for the four parameters. Each point corresponds to a deleted time (1 to 10) with estimations indicated by «*»; the global estimation is indicated by an horizontal line.

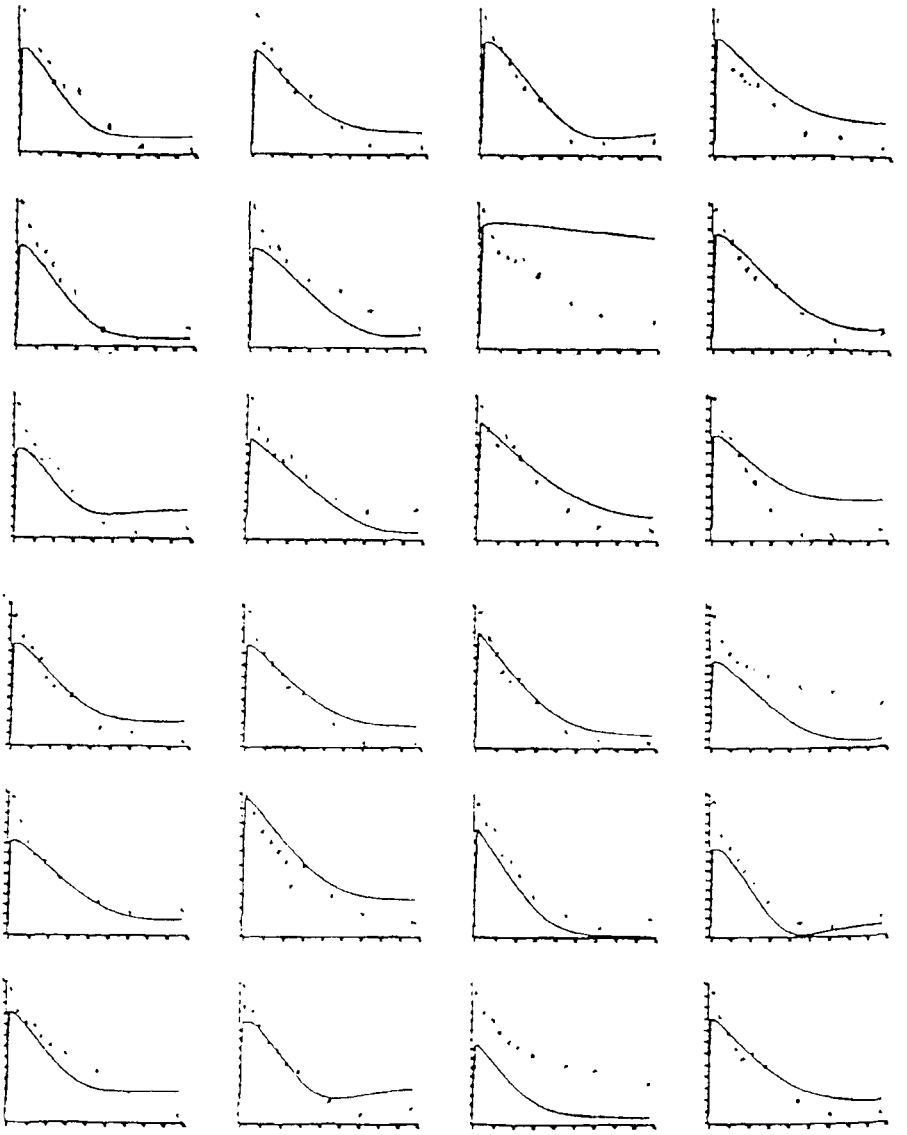

Figure 4. Metabolism of glucose: Individual curves for the 24 goats (the scale is the same as in figure 2).

information given to the experimenter indicates that he has to choose times at short intervals at two crucial periods for the dynamic analysis of metabolism of glucose.

This result is perfectly coherent with the influence analysis in which last times were more influential than the others.

Conclusions

As every work in modelling, these results must be taken with caution. They furnish to the experimenter some guidelines he has to integrate in his own experimental strategy. It is not evident that such differential models are perfectly adequate for his purpose. Nevertheless, even poor results may give some insight on what he has to do to improve his work.

Literature cited

Audrain S, Tomassone R (1994) Prediction domain in nonlinear models. *In: Proceedings of International Conference on Linear Statistical Inference LINSTAT'93* (T Calinki, R Kala, ed) Kluwer Academic Publishers, 147-58

Box GEP, Hunter WG, MacGregor JF, Erjavec J (1973) Some problems associated with the analysis of multiresponse data. *Technometrics* 15(1), 33

Brown D, Rothery P (1993) *Models in biology: mathematics, statistics and computing.* Wiley, New-York

Chaudhuri P, Mykland PA (1993) Nonlinear experiments: optimal design and inference based on likelihood. *JASA* 88, 538-46

Ciuperca G (1996) *Modélisation du métabolisme du glucose.* Thèse INAPG, in preparation.

France J, Thornley JHM (1984) *Mathematical models in Agriculture.* Butterworths, London

Gard TC (1987) *Introduction to stochastic differential Equations.* Marcel Dekker, New York

Gill EM, Thornley JHM, Black JL, Oldham JD, Beever DE (1984) Efficiency of utilization of absorbed energy in ruminants. *Brit J Nut* 623-47.

Murray JD (1989) *Mathematical biology.* Springer-Verlag, Berlin

Pavé A (1994) *Modélisation en biologie et en écologie.* Aléas, Lyon

Sauvant D, Grizard J (1992) Bases d'un modèle décrivant la régulation du métabolisme du glucose de la chèvre en lactation. *Ann Zootech* 41, 115-116

Simonoff SJ, Chih-Ling Tsai (1989) The use of guided reformulations when collinearities are present in non-linear regression. *Appl Statist* 38, 115-26

Tomassone R, Dervin C, Masson JP (1993) *Biométrie: modélisation de phénomènes biologiques.* Masson, Paris

Vila JP (1985) *Etudes et comparaisons de critères de plans d'expériences optimaux pour l'estimation des paramètres d'un modèle de régression non linéaire.* Thèse Univ Paris-Sud, Orsay