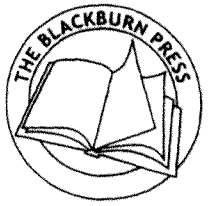
Plant and Crop Modelling

A Mathematical Approach to Plant and Crop Physiology

John H. M. Thornley
Agricultural and Food Research Council
Institute for Grassland and Animal Production
Hurley, Maidenhead, Berkshire

and

Ian R. Johnson
Department of Agronomy and Soil Science
University of New England
Armidale, New South Wales



The Blackburn Press

1
Dynamic modelling

1.1 Introduction

Science is concerned with prediction. We can predict only by virtue of having models or conceptual schemes of the world. The models we use today are those conjectures that have best survived the unremitting criticism and scepticism that are an integral part of the scientific process. There are (it is assumed) no new processes under the sun, but our models of these processes have allowed mankind to transform its corner of the universe. Science is commonsense; it is also an unpredictable, fascinating, and thoroughly human activity. These are simple truths, self-evident to the practising scientist, but they need frequent repetition, especially at a time when the scientific community is increasingly assailed by politicians or other careerists who, while enjoying the fruits of science, would deny its methods (Thornley and Doyle 1984).

Plant and crop modelling has, broadly speaking, two aims. The first, and here we nail our flag clearly to the mast, is to increase knowledge in this area of science. Historically, increasing knowledge has led to unpredicted (in detail) and enormous benefits to the human race, and there is no reason why this should not continue. The second aim might be called 'applied' and 'strategic' in today's terminology, and is directed at the solution of currently perceived problems in the short to medium term. Given present agricultural and horticultural practice, one can envisage many uses for models of plant and crop growth, which could increase efficiency, improve the environment, and generally contribute positively to life. There is then, no difficulty in defending the practice of the techniques, ideas, and approaches which we are about to expound.

Scientific knowledge is not only about observational data, but also about having a theory (or hypothesis, or conceptual scheme, or model) that corresponds to the data. It is the continual interaction between hypothesis (how we think things work) and observational data (how they actually do work) that leads to progress. With the passage of time, measurements become more accurate and more extensive; similarly we are continually widening the scope of our theories and demanding more accurate predictions from them. When comparing theory with experiment, we attempt to connect the theory to nature at as many points as possible and as precisely as possible. As a branch of science progresses from the qualitative to the quantitative, one day it may be expected to reach the point where the connections between theory and experiment are most efficiently made using the language of mathematics. It is to be emphasized that the ideas and hypotheses of the theory are not contributed by mathematics. Mathematics is used as a tool or language, enabling biological scientists to express their ideas

so that quantitative prediction is possible, and these predictions are then compared with observational data.

Agricultural and horticultural practice is based partly on tradition, partly on scientific knowledge, and partly on conjecture or guesswork. By tradition we mean an inherited folk-wisdom or set of customs where things are done because it is known that they work to a certain degree, but it is not understood why, or whether better results might be obtained by doing things a little differently. The formal knowledge of science can give a rationale for decision-taking when problems fall within the scope of current knowledge. Conjecture is also needed because, from time to time, a novel situation arises, there is no guidance from current knowledge on what to do, and yet a decision has to be taken. One purpose of agricultural research is to increase the knowledge-based component of agricultural decision-taking at the expense of the other two components. Increased knowledge does not necessarily lead to higher efficiency, but it may uncover more efficient options. With present agricultural practices, and other things remaining equal, the current efficiency of production provides a baseline from which it is only possible to move forwards. Equally important is the fact that increases in scientific knowledge allow a more rational response when other things do not remain equal, when the environment in which the farmer operates, natural or man-made (if this dubious distinction is permitted), changes.

Mathematical models can contribute to both of the aims discussed in the second paragraph of this section; that is, enlarging knowledge and helping with practical applications. Not only can models encapsulate knowledge, but, suitably programmed for the increasingly ubiquitous computer, they can also make this knowledge accessible to and usable by the non-expert. While the research worker delves ever more deeply into the minutiae and mechanisms of phenomena, it seems that technical developments will continue to make this detailed knowledge ever more easily available to the non-specialist farmer and farm adviser.

As with many things in this good life, there are models and models. Mostly, this book is concerned with dynamic deterministic models: dynamic models predict how a system unfolds with the passage of time—the time course of events; deterministic models make definite predictions (e.g. on 1 July the dry matter per unit area of the wheat crop will be 1 kg m^{-2}) without any associated probability distribution. Even dynamic deterministic models come in three types, demonstrating yet again the richness of science and the diversity of approaches possible. We call these types teleonomic, empirical, and mechanistic, although some would choose a different terminology. In terms of the organizational hierarchy of levels to be discussed later, teleonomic models look (mostly) upwards to higher levels, empirical models examine a single level, and mechanistic models look downwards, considering a level in relation to lower levels. Teleonomic models are sometimes called teleological or goal seeking. Empirical models belong to the category associated with curve fitting, regression, and applying mathematical formulae directly to observational data, usually without being constrained by scientific principles or any knowledge of mechanism. Mechanistic models are

reductionist, concerned with mechanism, and integrative; they contribute understanding, and are sometimes called explanatory. In any given investigation, the objectives of the enterprise should determine what modelling approach, if any, can be used. It is therefore important to understand how these different types of model relate to each other and to the structure of the problem, and this is the main concern of the next section.

De Wit (1970) gives an excellent early account of concepts in the crop modelling area, which are also discussed by Thornley (1976, 1980).

1.2 Hierarchical systems

Biology, including plant biology, is notable for its many organizational levels. Whereas in physics and chemistry one travels more or less directly from atomic and molecular behaviour to that of liquids and solids, in biology there are several intervening organizational entities. It is the existence of the different levels of organization that gives rise to the great diversity of the biological world. For the plant sciences, a typical scheme for the hierarchy of organizational levels is as follows.

Level	Description of level
...	...
$i+1$	crop
i	plant
$i-1$	organs
...	tissues
...	cells
...	organelles
...	macromolecules
...	molecules and atoms

(1.1a)

The levels that are of principal interest to this book are labelled $i+1$, i, and $i-1$. Using this diagram, we shall pinpoint the differences in viewpoint associated with the empirical, mechanistic, and teleonomic approaches to modelling, but first we discuss the principal properties of a hierarchical system.

Hierarchical systems have several important properties.

1. Each level has its own language, which is unique to that level. For example, the terms crop yield, leaf area, or whole-plant dry mass have little meaning at the cell or organelle levels.
2. Each level is an integration of items from lower levels. The response of the system at level i can be related to the responses at lower levels. This is scientific reductionism, and leads to mechanistic models.
3. Successful operation of a given level requires lower levels to function properly, but not vice versa. For example, if a cup is smashed to small pieces, it will no longer function as a cup, although the molecular interactions are hardly altered.

4. The higher levels provide the constraints, boundary values, and driving functions, including any inputs and outputs, to the lower levels.
5. On descending to a lower level, generally both the spatial and temporal scales become smaller; this corresponds to smaller physical size and to faster processes at the lower levels.

1.2.1 *Empirical models*

Empirical models are essentially direct descriptions of observational data, but they can, none the less, be exceedingly useful. The well-known saying 'red sky at night, shepherd's delight' has helped in the planning of many harvesting activities and family picnics. The tables that describe the tides round our coastlines are constructed by totally empirical methods. In an empirical model, any mathematical relationships that are written down are usually unconstrained by physical laws such as that of energy conservation or the laws of thermodynamics, by biological information, or by any knowledge of the structure of the system. The empirical modeller attempts to describe level i behaviour (observational data) in terms of level i attributes alone, without regard to any biological theory. The approach is primarily one of examining the data, deciding on an equation or set of equations, and fitting these to the data. Essentially, an empirical model re-represents the data, perhaps more conveniently, and no new information is acquired.

In Fig. 1.1 a simple example of an empirical mathematical model is given. The observational data shown give the response of crop yield Y to the level of fertilizer N applied. Often such data can be fitted by a three-parameter rectangular hyperbola

$$Y = Y_{max}\frac{(N_s + N)}{K + (N_s + N)} \tag{1.1b}$$

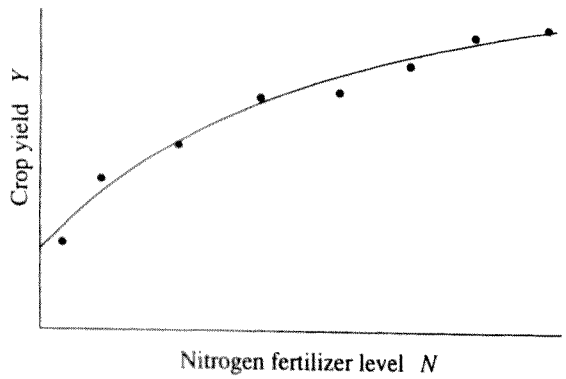


Fig. 1.1. A simple empirical model of the response of crop yield Y to nitrogen fertilizer level: •, observational data; —, eqn (1.1b).

where Y_{max} is the maximum value of yield Y at high N values, N_s is the effective residual soil nitrogen which gives a yield at $N = 0$, and K is a parameter determining the initial slope of the $Y : N$ response at $N_s + N = 0$. Also shown in Fig. 1.1 is the curve obtained by fitting eqn (1.1b) to the observational data and adjusting the parameters.

The important point to recognize is that the fitted curve (or empirical model) contains no information beyond the original data; it says nothing about the mechanisms that give rise to the response (for which a mechanistic model of the system would be needed), or why the response is as it is in terms of possible goal-oriented behaviour of the plant (we use 'why?' in the 'for what purpose?' sense); for the latter a teleonomic model would be appropriate. Empirical models can provide a powerful means of summarizing data and interpolation, and may provide a practical tool for the farmer or farm adviser. Traditionally much agricultural research has been of a descriptive nature, and a great deal of essential groundwork has been done using empirical models. However, our thesis is that the needs of the subject are changing, and it is now timely to seek mechanistic explanations and an understanding of these responses by means of models that integrate the underlying mechanisms.

1.2.2 *Mechanistic models*

The main concern of this book is with dynamic deterministic models that are concerned with mechanism and can lead to an understanding of the ith level (1.1a) that is based on component processes at the $(i - 1)$th level, and possibly at lower levels. The mechanistic modeller attempts to construct a description of level i behaviour which has some extra content of mechanism, understanding or explanation at lower levels. Mechanistic modelling is 'hard' science, and it follows the traditional reductionist method that has been so very successful in the physical sciences, molecular biology, and biochemistry. As shown in Fig. 1.2, in contrast with the empirical modeller who proceeds directly to the whole-plant variables that are of interest and may connect these in whatever way seems best to fit the data, the mechanistic modeller goes round a relatively circuitous route: under analysis and reduction he breaks the system down into components and assigns processes and properties to these components; this introduces extra variables at the $(i - 1)$th level, and additional observational data are generally also available at the $(i - 1)$th level; finally, it is by the integration of the set of equations that define the system that the responses at the whole-plant level are synthesized.

A mechanistic model of responses at a certain hierarchical level is always far more complex than an empirical model; it will generally fit the data at the ith level less well because it has many constraints built into its structure by means of the assumptions of the model. However, its content is far richer in that it applies to a greater range of phenomena, relating them to each other. Because

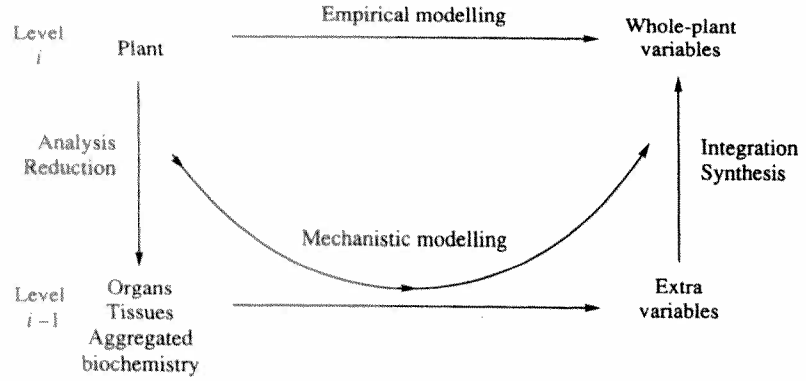


Fig. 1.2. Mechanistic and empirical modelling.

of this, a mechanistic model always offers more possibilities for manipulating and improving the system.

Some have questioned whether integration of knowledge at the $(i-1)$th level can indeed give rise to new knowledge at the ith level. Is the whole more than the sum of its parts? Are there 'emergent' properties? There is, in our view, no doubt about the answer to this question. Shakespeare was more creative than the proverbial monkey at a typewriter because of the way in which he assembled familiar words and phrases into new patterns. The same is true in music. In the sciences, the juxtaposition of well-known components to give rise to new results (or 'emergent' properties) has occurred again and again. The mathematician Turing showed that diffusion combined with chemical reaction could generate patterns—a result that has been most valuable in studies of morphogenesis (Chapter 19). The kinetic theory of gases with all its ramifications is a consequence of a few very simple assumptions at the molecular level. The theories of the atmosphere are becoming steadily more successful and yet, at base level, nothing has changed. The special theory of relativity is a consequence of attributing a constant velocity of light to all vacuum reference frames—this seems innocuous enough, and yet the results of the ensuing analysis are stupendous. The whole is more than the sum of the parts, and yet it is explainable in terms of the parts and how they interact. The post-war progress in molecular biology and biochemistry indicates that a reductionist programme is applicable to all areas of biology, even including value systems and religion (Monod 1972). An explanation of the responses, behaviour, and mechanisms of an organism does not mean that the particular combination of mechanisms that gives rise to the observed responses and behaviour could have been predicted. Even with deterministic equations, the prediction of a detailed time course may be totally unreliable, owing for instance to sensitivity to initial conditions, as with chaotic systems. The range of future possibilities may be so great that the particular time

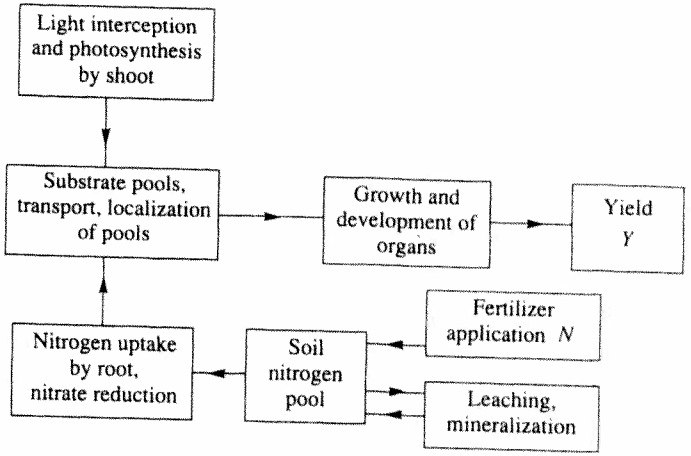


Fig. 1.3. Elements of a possible mechanistic model of response of crop yield Y to fertilizer level N.

course followed is unpredictable, and yet it is explainable in terms of established mechanisms.

In Fig. 1.1 and eqn (1.1b), an elementary empirical model of crop yield response to fertilizer is given. In Fig. 1.3, a possible mechanistic approach to this problem is shown. There are usually several ways in which a system might be analysed into components. There are no hard and fast rules. One attempts to incorporate the 'important' components in the model, to make 'reasonable' assumptions, and to strike the right balance between complexity, tractability, and realism commensurate with the overall objectives. If possible, elegance is also part of the structure. Modelling is partly an art form, although hopefully the end result is of scientific value. Metaphysics has been the precursor to much good science. Conjecture (or guesswork) is needed because there are many areas of ignorance. Individual scientists will always assign different weights to the various ingredients in the mix; hence the variety of crop models that exists. In biology, most models are simplifications and are therefore wrong at some level of detail (this is less true of models in physics and chemistry). Therefore critics can always legitimize their reasons for rejecting a model. A model should be criticized mostly in relation to the stated objectives. The strengths of the model, which should enable the modelling objectives to be met, ought to be emphasized more than the weaknesses, which should not detract appreciably from the model's performance in relation to the stated modelling objectives. However, a different investigator with different objectives might find those weaknesses unacceptable.

Frequently, people will talk about the 'complexity' of a model, and yet there is no consensus about what is meant by this term. Possible measures of complexity might be the number of state variables (p. 15), the number of parameters, the topology of the system diagram (as in Fig. 1.3) with perhaps the number of

closed loops or cycles, the level of mathematics required (partial differential equations, perturbation theory or topology may be regarded as 'complex' or 'difficult'), or even the computing power needed to generate solutions. Generally, in a 'good' model, only significant parameters are retained—those that have an appreciable effect on the solutions or the scope of the model. Thus we choose the number of parameters as the best measure of model complexity.

Sometimes, it is said that a situation is 'too complex' to model, or that not enough is known about the system to build a model. Yet often, we model because it is the only way of grappling with complexity, or we model in order to define what we do not know about the problem. Complex situations with uncertain outcomes, which are common in agriculture and biology, are ripe for modelling.

Models can also be used to calculate physiologically meaningful parameters which are not accessible to direct observation. For example, plant and root respiration parameters can be obtained by analysis (Chapter 11, Sections 11.3 and 11.5). In physics and chemistry, there is a well-established tradition of finding quantities such as the mass of the electron, or the activation energies of reactions, by indirect methods using a model.

Let us assume we have a model (empirical or mechanistic) with n parameters. The model makes predictions for the m values of some observable quantity Y, Y_i, $i = 1, 2, \ldots, m$. These predictions can be compared with the corresponding observational data $y_i, i = 1, 2, \ldots, m$. Regarding the n parameters as all adjustable and as having been adjusted to give the best fit to the observational data, the estimated mean square prediction error (MSPE) can be written as

$$\text{MSPE} = \sum_{i=1}^{m} \frac{(y_i - Y_i)^2}{m - n} \tag{1.2a}$$

Figure 1.4 shows schematically how we might expect the MSPE for empirical and mechanistic models to vary with the complexity parameter n. The empirical model (EM) is applied to a given set of observational data at the organizational level i (1.1a); as the complexity n of the model (as measured by the number of parameters) is increased, the observational data base m remains the same. Eventually the data are being over-fitted, and the MSPE, after initially decreasing, increases again. The situation with the mechanistic model (MM) is rather different. To begin with, for low numbers of parameters and with roughly the same data base, MM will always give a worse fit because it is constrained by the structure and assumptions of the model, whereas EM is 'free'. As model complexity increases (with increasing n), the observational data base increases; there are data at the level of the assumptions of the model, and the model can now be tested at the level of the assumptions $(i - 1)$ as well as at the level of its predictions (i). Thus the MSPE for MM can decrease with increasing complexity, but may now conceivably approach an asymptote. It may be argued that in comparing empirical with mechanistic models in this way, like is not being compared with like, and this is correct since the scope of the mechanistic model is allowed to

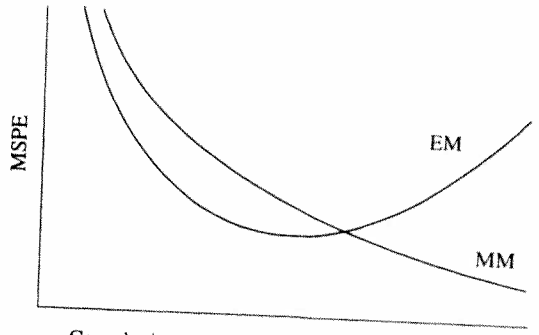


Fig. 1.4. Relative predictive ability of empirical (EM) and mechanistic (MM) models. The hypothetical dependence of the mean square prediction error (MSPE) on model complexity, represented by the number of parameters n, is shown.

increase with increasing complexity, whereas this is not the case for the empirical model. However, this is generally the manner in which empirical and mechanistic models are developed and applied.

1.2.3 *Teleonomic models*

From time to time in this book, a third class of model is used or referred to—teleonomic models. Such models may be relatively unfamiliar to the plant scientist, and although they only play a minor role currently in plant and crop modelling, this could expand, and it is important that the teleonomic viewpoint be appreciated and understood (Monod 1972) including the relationship between teleonomic models, empirical models, and mechanistic models.

Teleonomic models are applicable to goal-directed behaviour, and are formulated explicitly in terms of goals. These models may refer to a single level, say the ith level; they may provide a teleonomic interpretation of an otherwise empirical model; they may also refer responses at the ith level to the constraints provided by the $(i + 1)$th level (p. 5). The goals of level i can be viewed as the requirements imposed by level $i + 1$. It is the higher-level constraints which, via evolutionary pressures, can select out combinations of the lower-level mechanisms of biochemistry, which may lead to apparently goal-directed behaviour at level i. We give two examples of possible teleonomy in plant growth and development (see also Chapter 7, Section 7.6, p. 193).

Dry-matter partitioning is an aspect of most plant and crop growth models, and it can be approached by the modeller in various ways. For root : shoot partitioning all three kinds of model exist: empirical, teleonomic and mechanistic. Chapter 13 considers the topic in detail, and here we give a brief qualitative discussion to illustrate general modelling issues. The empirical approach is to

measure (say) root : shoot partitioning for a variety of situations of both environment and plant. These observational data provide the basis for an empirical model of root : shoot partitioning. However, there have been various attempts to interpret these measured responses in terms of goal-directed behaviour of the plant. An early suggestion was that the root : shoot ratio adjusts so as to maintain a constant carbon : nitrogen ratio in the plant. Another suggestion is that the root : shoot ratio and the carbon : nitrogen ratio both adjust so that they are proportional to each other. The observational data are not of sufficient quality to discriminate unambiguously between these alternative teleonomic interpretations, but nevertheless such goal-directed models have been incorporated, with a degree of success, into some crop-growth models. However, we can go beyond a direct teleonomic interpretation of observational data, and attempt to construct a teleonomic model of root : shoot partitioning by hypothesis. For example, it can be assumed that plants have evolved controls so that, when the plant is faced with a given environment, new material is allocated to the shoot and to the root in such a way that the plant attains an optimal specific growth rate in that given environment. This then leads to a teleonomic partitioning model. A goal-oriented model at the $(i-1)$th level (the organ level) has been constructed by assuming that, at the ith level (the plant level), the plant has certain 'requirements' such as optimizing the plant specific growth rate; it is of course assumed that the diversity of possible mechanisms at the lower levels is able to satisfy these requirements. The teleonomic model can be written directly in terms of the lower-level mechanisms, giving a mechanistic model which is more complicated than the teleonomic representation but is 'objective' in its operation (as are the laws of physics and chemistry).

Phyllotaxis is the second example we consider (Chapter 19). It is especially interesting because it appears to be a case where mechanistic models of the problem and teleonomic models deduced from the 'requirements' of the plant both lead to Fibonacci arrangements for leaves around a stem or for seeds in a flower-head. It is an observational fact that leaves in plants and seeds in flower-heads are often positioned so that they are separated from adjacent members by the Fibonacci angle. The teleonomic view of this is as follows. It is assumed that the plant 'requires' optimum interception of light from the azimuth by its leaves; this assumption leads to the Fibonacci angle. Also, if it is assumed that the plant 'requires' optimum packing of seeds in flower-heads (thus maximizing the number of seeds), an analysis again leads to the Fibonacci angle. Thus, in this case, the Fibonacci model is derivable from an assumption about the goals of the plant. There are two mechanistic approaches to modelling phyllotaxis—giving rise to models that in both cases generate successive primordia (leaves or seeds). One is a 'contact' theory, and works with the 'next available space'; the second is a diffusion–reaction model involving a postulated morphogen. Both are able to generate Fibonacci patterns of primordia. Thus, in this case, mechanistic considerations lead to patterns that also satisfy teleonomic requirements.

1.3 Research models and applications models

The question that is addressed in this section is this: are there important differences between models constructed for an applications purpose and models used in research? The term 'application' means here application in agricultural production. Models of plant and crop growth are, potentially at least, capable of being used by farmers, farm advisers, horticulturalists, and even enthusiastic amateurs. A good crop model that encompasses response to environment and also controllable inputs such as fertilizer levels and irrigation could provide a valuable tool for crop management.

The answer to the question posed in the first sentence above is an emphatic 'yes'. Although in time research models may evolve into applications models, when they achieve the latter state, they may cease to be of value in research. With a research model, aimed at improving understanding, much progress can be made with a model that fails to describe correctly what is happening. Part of the value (and the enjoyment) of modelling is that it is possible to speculate inexpensively and try out various ideas. Sometimes, after a perhaps long and arduous progression round and round the cycle of conjecture, prediction, and observation, one arrives at a set of hypotheses that are compatible with current observational data.

However, if an applications model is to be successful (which means being used), it must give predictions and guidance that are demonstrably better, in some way, than existing practice. All models, whether oriented towards research or applications, are based on a mix of observational data, currently accepted knowledge, and conjecture. While it may be desirable for a research model to have a high proportion of conjecture, an applications model must clearly be as firmly founded as possible on data and knowledge that are relatively secure. If an applications model were to fail, this could put into jeopardy the livelihood of many people. The failure of a research model has no such consequences.

Table 1.1 summarizes the principal differences that usually exist between research models and applications models. Since empirical models are related directly to observational data, and are often mathematically and computationally simple, applications models for farm management purposes frequently have a high content of empiricism. As increasing user sophistication demands a wider range of applicability and scope of these models, it may well be that more mechanistic models prove to be the best way forward, although such models must always be carefully evaluated under the conditions where they might be applied. Sheer empiricism can require almost infinite experimental resources in complex input-output situations. Also, empiricism on the ith level alone may be very inefficient in ignoring the ever-widening base of physiological and biochemical knowledge at the lower levels. The possible benefits of application to the problems of agriculture and the environment provide a good reason for continuing with basic and strategic research. While crop management may continue to use refined empirical models for some years, we believe that con-

Table 1.1. Research models and applications models: a comparison of the principal differences

	Research	Applications
Hypotheses	Speculative	Well-accepted
Connections to observational data	Tenuous (often)	Good
Accuracy of predictions	Variable	Good
Scope	Wide	Restricted
Complexity	Complex	Simple
Type	Mechanistic	Empirical

tinued progress in the area depends on pursuing a long-term strategy of developing more mechanistic models, although such models may often be simplified for practical application. A mechanistic model must always uncover more options than an empirical model. Until such models have been successfully constructed and evaluated, there will remain areas of ignorance that neither science nor the practitioners of agriculture can afford to leave undisturbed.

1.4 Mathematical models: objectives and contributions

There are many roles for modelling work in biology and in the plant and crop sciences. When undertaking a modelling project, the most important need is for a clearly defined and realistic set of objectives. It would be wasteful to construct a complex mechanistic model where an empirical approach would better meet the requirements. Since objectives are linked to potential contributions, we list some of the possible objectives and potential benefits associated with mathematical modelling. The abbreviations EM for empirical model and MM for mechanistic model are used.

1. Hypotheses expressed in mathematical terms can provide a quantitative description and mechanistic understanding of a biological system (MM).
2. A model requires a completely defined conceptual framework, and this may pinpoint areas where knowledge is lacking, and perhaps stimulate new ideas and experimental approaches (MM).
3. A mathematical model, especially if implemented in an easy-to-use computer program, may provide an excellent recipe by which recent research knowledge is made available to the farm manager or adviser (EM mostly, some MM).
4. Agro-economic models may highlight the benefits of new crop management techniques suggested by recent research, thereby stimulating the adoption of more efficient production methods (MM, EM).
5. Modelling may lead to less *ad hoc* experimentation, as models may make it possible to design experiments to answer particular questions, or to discriminate between alternative mechanisms (MM, EM).

6. In a system with several components, a model provides a means of bringing together knowledge about the parts, giving an integrated view of whole-system behaviour (MM).
7. Modelling can provide strategic and tactical support to a research programme, motivating scientists and encouraging collaboration (MM).
8. A model may provide a powerful means of summarizing data, and also a method for interpolation and cautious extrapolation (EM mostly, MM).
9. Observational data are becoming more precise, but also more expensive to obtain; a mathematical model may be able to make more complete use of such data (EM, MM).
10. The predictive power of a successful model can be used in many ways. For instance a model can be used to answer 'what if...?' questions. What are the consequences on crop production of halving the maintenance requirements of plant tissue? What are the effects on crop yields of changing within-plant transport resistances? However, it should be remembered that the answers given by a model are, in a sense, built into it by hypothesis. Thus a model can be used to stimulate thought, but it may be dangerous to use a model to manage a research and development programme (MM).

Any given model is only likely to contribute under two or three of these ten points. However, this list indicates the many possible reasons for undertaking modelling work.

1.5 Deterministic dynamic differential equation models

It is assumed that the state of the system under investigation at time t is defined by q variables $X_1, X_2, \ldots, X_q$; these q variables are called state variables. The q state variables are independent; that is, it is not possible to derive one of the state variables, X_1 say, from a knowledge of the values of the other state variables. The state variables represent properties or attributes of the system being considered (such as dry matter, number of cells, leaf area, starch content, etc.). The choice of state variables is the first and most important assumption that the modeller makes. The scope of the model is defined by its state variables.

The next step is to construct the q first-order ordinary differential equations that describe how the q state variables change with time t. These can be written formally as

$$\begin{aligned} \frac{dX_1}{dt} &= f_1(X_1, X_2, \ldots, X_q; P; E) \\ \frac{dX_2}{dt} &= f_2(X_1, X_2, \ldots, X_q; P; E) \\ &\vdots \\ \frac{dX_q}{dt} &= f_q(X_1, X_2, \ldots, X_q; P; E). \end{aligned} \tag{1.3a}$$

The $f_1, f_2, \ldots, f_q$ denote functions of the state variables, of a number of parame-

ters which are indicated by P, and of environmental quantities denoted by E. Equations (1.3a) are called 'rate-state' equations, and they show how the rates of change of the system state variables depend explicitly on the current values of the state variables. In plant and crop models it is helpful to indicate explicitly the presence of parameters and environmental quantities in eqns (1.3a) by P and E.

Writing $f_1(X_1, X_2, \ldots, X_q; P; E)$ does not mean that the function f_1 must contain all the state variables, parameters, and environmental quantities; the rates of change of most state variables will not depend directly upon the environment, and it would be usual for the rate of change of a given state variable to depend only upon two or three other state variables. Denoting a single state variable by X, we can write its rate-state equation as

$$\frac{dX}{dt} = \text{inputs} - \text{outputs}; \tag{1.3b}$$

the inputs are the terms that contribute positively to the rate of change of X, and the outputs contribute negatively. Each term on the right-hand side of eqn (1.3b) gives the rate of a process. In the plant sciences, all processes are either transport or chemical conversion. For example, eqn (1.3b) could take the form

$$\frac{dX}{dt} = c - \frac{kX}{K + X} + g(a - X); \tag{1.3c}$$

c stands for a constant production of X, the second term on the right for a Michaelis–Menten-like utilization of X (Chapter 2, p. 51) where k and K are parameters, and the last term stands for a diffusive-like transport of X (Chapter 4, p. 95) where g and a are parameters.

1.5.1 *Explicit time dependence*

Suppose that one of the rate-state equations (1.3a) takes the form

$$\frac{dX}{dt} = aX(b - t) \tag{1.4a}$$

where a and b are constants. Here, the time variable t appears explicitly on the right-hand side of the equation. In eqns (1.3a) as written the time variable t does not appear explicitly on the right-hand side of the equation (excluding its possible appearance in the environmental specifications in E). Generally speaking, it is not sound scientific methodology to include t directly in the rate-state equations. The system is completely specified by the set of state variables and does not 'know' what the time is, although one or more state variables may effectively be keeping track of time. However, the elimination of one or more state variables may lead to a reduced set of rate-state equations with explicit time dependence; thus explicit time dependence can be viewed as representing hidden state variables. Consequently, although the use of explicit time dependence in the rate-state equations can be viewed as a device that is sometimes convenient and can usefully simplify the model, it may impose an undesirable external constraint on the dynamics of the system.

From eqn (1.4a), we could expand the set of state variables by defining a second state variable Y by

$$Y = b - t, \tag{1.4b}$$

from which t is easily eliminated by a single differentiation, giving

$$\frac{dY}{dt} = -1 \qquad \text{with } Y = b \text{ at } t = 0. \tag{1.4c}$$

The time-independent rate-state equations are now

$$\frac{dX}{dt} = aXY \qquad \text{and} \qquad \frac{dY}{dt} = -1. \tag{1.4d}$$

Equation (1.4a) can be integrated quite easily, since the variables are separable, to give

$$X = X_0 \exp\left[\frac{at(2b - t)}{2}\right] \tag{1.4e}$$

where $X = X_0$ at $t = 0$. This function is known as the exponential quadratic (Chapter 2, eqn (2.15d), Fig. 2.12a, pp. 69, 70; Chapter 3, eqn (3.7d), p. 87), and is sometimes employed in plant-growth analysis. As used in plant-growth analysis, eqns (1.4a) and (1.4e) are rightly regarded as 'empirical' or 'curve-fitting' approaches. However, eqns (1.4d) may permit a more mechanistic biological interpretation (namely, that there is some component of the growth machinery that is steadily being depleted (see Chapter 2, Fig. 2.10, p. 66)).

The Gompertz growth equation in differential form is

$$\frac{dW}{dt} = \mu_0 W e^{-Dt} \qquad \text{with } W = W_0 \text{ at } t = 0, \tag{1.4f}$$

where W is the state variable (denoting dry matter) with initial value W_0, and μ_0 and D are constants (Chapter 3, eqn (3.5d), p. 80, and Fig. 3.4). The problem of putting eqn (1.4f) into a two-state-variable formulation without explicit time dependence is posed in Exercise 1.1.

1.5.2 *Memory functions and delays*

Consider a two-state-variable problem with state variables X and Y, as in eqns (1.4d). The differential equation for X can be written

$$\frac{dX}{dt} = f_1(X, Y), \tag{1.5a}$$

where f_1 denotes some function. Sometimes we wish to make the current value of dX/dt dependent upon the value of a variable at some time in the past, say a time period τ ago. This remembered variable could itself be X affecting dX/dt, so that eqn (1.5a) becomes (we use f_i to denote a function)

$$\frac{dX}{dt}(t) = f_2[X(t), Y(t), X(t-\tau)]. \tag{1.5b}$$

Alternatively, the remembered variable affecting dX/dt might be Y, giving

$$\frac{dX}{dt}(t) = f_3[X(t), Y(t), Y(t-\tau)]. \tag{1.5c}$$

For either of the above equations, we can denote the remembered variable as Z, which is computed from past values of one or more of the state variables of the system, and eqns (1.5b) and (1.5c) can be written simply as (omitting Y)

$$\frac{dX}{dt} = f(X, Z). \tag{1.5d}$$

The remembered variable Z is in general obtained by an integral computed over the entire past history of the system up to the present:

$$Z = \int_0^t g(X)\,dt', \tag{1.5e}$$

where g denotes the memory function and t' is a dummy time variable.

The Dirac delta function $\delta(x)$, where x is a variable, can be thought of as a very sharp spike of unit area at the origin $x = 0$, which is zero everywhere else. It is defined by the equation (x' is a dummy variable)

$$f(x) = \int f(x')\delta(x - x')\,dx'. \tag{1.5f}$$

The delta function $\delta(x)$ just selects out the value of f where the spike occurs, i.e. at $x' = x$. When the Dirac delta function is used, the discrete memory function which remembers the value of state variable X at time period τ ago is given by

$$Z = X(t-\tau) = \int_0^t \delta(t - \tau - t')X(t')\,dt'. \tag{1.5g}$$

Instead of (and perhaps more realistic than) remembering a value of X at some instant in the past, one might remember some weighted average of past values. Equation (1.5g) is replaced by

$$Z(t) = \int_0^t w(t - t')X(t')\,dt', \tag{1.5h}$$

where w is a weighting function. (See Exercise 1.2.)

In the last section, in the discussion of explicit time dependence, it was said that the system cannot 'know' what the time is, and a representation with more state variables can always be found which does not have an explicit time dependence. It is similar here for 'remembered' variables. The only way in which a system remembers a variable is through other variables, and so, again, a representation can always be constructed which does not have remembered variables but which may be considerably more complicated. For example, eqn (1.5g) can be represented by a large number of intermediate variables I_i, as in the scheme

$$X \to I_1 \to I_2 \to \cdots \to Z. \tag{1.5i}$$

This compartmental scheme, which can be regarded as a delay line or a pipe with plug flow, can be equivalent to a discrete delay (see Chapter 7, Section 7.3.4, p. 177, and eqn (7.7d) *et seq.*).

1.6 Numerical integration

For all but the very simplest models, eqns (1.3a) can only be integrated numerically. This requires specification of the initial values X_i at time $t = 0$, the parameters P and the environment E; the results of the numerical integration are the values of the state variables X_i at any subsequent time t. Computing technology has transformed numerical methods and now provides the facility to solve complex problems. There are many excellent textbooks on the subject, and here we introduce a few of the simpler methods and point out some of the pitfalls.

We begin by simplifying eqns (1.3a) and considering the single equation

$$\frac{dx}{dt} = f(x, t) \tag{1.6a}$$

with a single state variable x; f denotes a function of x and time t, and the parameters P and environment E are not shown explicitly. The process of numerical integration can be shown as

$$x(t) \to x(t + \Delta t). \tag{1.6b}$$

Given the value of x at time t, and given eqn (1.6a) so that the rate of change x can be calculated, how do we calculate the value of x at time $t + \Delta t$, where Δt denotes an increment in the time variable? A prescription of the type of (1.6b) can be applied iteratively, starting from time $t = 0$ and proceeding to any time t.

1.6.1 *Euler's method*

This is the simplest method of all. It is called a first-order method because it takes account of terms of order Δt, and the error is of order $(\Delta t)^2$. The rate of change of the state variable x at time t is by definition given by

$$\frac{dx}{dt} = \lim_{\Delta t \to 0}\left[\frac{x(t + \Delta t) - x(t)}{\Delta t}\right]. \tag{1.7a}$$

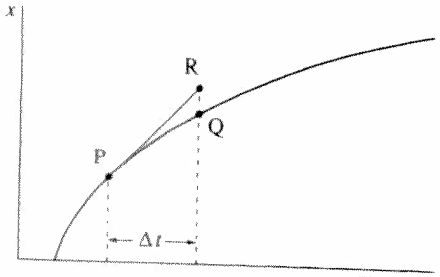


Fig. 1.5. Numerical integration by Euler's method. The curve shows the true solution to eqn (1.6a) passing through P and Q. PR is the tangent to the true solution at point P. R represents the solution predicted by Euler's method for time $t + \Delta t$, which is in error by RQ.

Substituting eqn (1.6a) into (1.7a) gives

$$f[x(t), t] = \lim_{\Delta t \to 0} \left[\frac{x(t + \Delta t) - x(t)}{\Delta t} \right]. \tag{1.7b}$$

For small values of Δt, it is approximately true that

$$f[x(t), t] = \frac{x(t + \Delta t) - x(t)}{\Delta t}. \tag{1.7c}$$

Rearranging this equation gives Euler's formula

$$x(t + \Delta t) = x(t) + \Delta t\, f[x(t), t]. \tag{1.7d}$$

An illustration of the operation of Euler's method over a single interval Δt is given in Fig. 1.5. Given the value of x at time $t = 0$, this equation can be applied iteratively to give values for x at times $t = \Delta t, 2\Delta t, 3\Delta t, \ldots$. (See Exercise 1.3.)

A first-order method such as Euler's method gives a second-order error of order $(\Delta t)^2$. This can be demonstrated by considering the Taylor series expansion of $x(t + \Delta t)$. This is

$$x(t + \Delta t) = x(t) + \Delta t \frac{dx}{dt} + \frac{1}{2!}(\Delta t)^2 \frac{d^2x}{dt^2} + \frac{1}{3!}(\Delta t)^3 \frac{d^3x}{dt^3} + \cdots. \tag{1.7e}$$

On comparing eqn (1.7e) with eqn (1.7d) and noting that $f[x(t), t] = dx/dt$, the error ε in Euler's formula (eqn (1.7d)) is given by

$$\varepsilon = \frac{1}{2}(\Delta t)^2 \frac{d^2x}{dt^2} + \text{higher-order terms}. \tag{1.7f}$$

This error is known as the truncation error. For a straight line, the second and higher derivatives are zero and Euler's formula gives an exact result. Halving the integration interval Δt reduces the error in eqn (1.7f) by a factor of 4. However, we now need to take twice as many steps to cover the same period of time, so that the resultant error per unit of time is reduced by 2.

Despite the availability of more sophisticated methods, Euler's method is of great practical value. Because of its simplicity, it is easily checked out in complete detail, and one can feel sure of just what is going on in a calculation.

Oscillations as an artefact of integration Using Euler's method, we next show how instability and oscillations can arise. Assume that eqn (1.6a) takes the form

$$\frac{dx}{dt} = -kx \qquad \text{with } x = 1 \text{ at } t = 0. \tag{1.7g}$$

k is a constant. Integration gives the exact analytical solution

$$x = e^{-kt}. \tag{1.7h}$$

Next we generate an approximate numerical solution to eqn (1.7g) using Euler's formula of eqn (1.7d) with $k = \frac{1}{2}$ and $\Delta t = 1$, to give

t	0	1	2	3	4	5	...
x	1	1/2	1/4	1/8	1/16	1/32	...

(1.7i)

However, now assume that $k = 3$. Applying eqn (1.7d) gives

t	0	1	2	3	4	5	...
x	1	−2	4	−8	16	−32	...

(1.7j)

For eqn (1.7g) it can be shown that

$k\Delta t < 1$ gives asymptotic stability

$1 < k\Delta t < 2$ gives oscillations of decreasing amplitude (1.7k)

$2 < k\Delta t$ gives oscillations of increasing amplitude.

With smaller values of Δt, the stability of the numerical solution is increased; the truncation error of eqn (1.7f) is also reduced. However, computing time increases because more iterations are required and there is an increased possibility of rounding errors which result from the fact that the computer can only store values to a certain number of decimal digits. The Δt used is a compromise between these conflicting criteria (Section 1.6.6, p. 30). Usually, to reduce the truncation error, it is more efficient to use a higher-order method than to reduce Δt. However, some problems are not easily solved with higher-order methods, as these may require the calculation of the derivatives within the time interval Δt.

1.6.2 *Trapezoidal method*

Euler's formula of eqn (1.7d) can be regarded as the first two terms of the Taylor series of eqn (1.7e). Occasionally, it is possible to calculate higher derivatives

analytically and use these directly in the Taylor series to obtain a better approximation. Thus, given

$$\frac{dx}{dx} = f(x,t), \tag{1.8a}$$

by analytic differentiation the second derivative is

$$\frac{d^2x}{dt^2} = g(x,t), \tag{1.8b}$$

and then the first three terms of the Taylor series of eqn (1.7e) are calculable, giving

$$x(t+\Delta t) = x(t) + \Delta t\, f(x,t) + \tfrac{1}{2}(\Delta t)^2 g(x,t). \tag{1.8c}$$

This second-order equation has a third-order error which is the fourth term on the right-hand side of eqn (1.7e).

More usually, it is not practical to calculate analytically using algebra the higher derivatives of the Taylor series expansion of eqn (1.7e). Essentially, the higher-order numerical integration methods operate by evaluating the first derivatives using eqn (1.8a) at different points within the integration interval, and the higher derivatives are obtained from the way in which the first derivatives change. We consider here the simplest of the second-order methods, which is known as the trapezoidal method.

Consider the second derivatives as defined by differentiating eqn (1.7a) to give (with $\dot{x} \equiv dx/dt$)

$$\frac{d^2x}{dt^2} = \lim_{\Delta t \to 0}\left[\frac{\dot{x}(t+\Delta t) - \dot{x}(t)}{\Delta t}\right]. \tag{1.8d}$$

For small Δt, therefore,

$$\frac{d^2x}{dt^2} = \frac{\dot{x}(t+\Delta t) - \dot{x}(t)}{\Delta t}. \tag{1.8e}$$

As $f(x,t) = dx/dt$, eqn (1.8e) can be written

$$\frac{d^2x}{dt^2} = \frac{f[x(t+\Delta t), t+\Delta t] - f(x,t)}{\Delta t}. \tag{1.8f}$$

Euler's method of eqn (1.7d) provides a first-order estimate x_1 of $x(t+\Delta t)$, giving

$$x_1 = x(t+\Delta t) = x(t) + \Delta t\, f(x,t). \tag{1.8g}$$

Next, substitute x_1 for $x(t+\Delta t)$ in eqn (1.8f) to give

$$\frac{d^2x}{dt^2} = \frac{f(x_1, t+\Delta t) - f(x,t)}{\Delta t}. \tag{1.8h}$$

Now take the first three terms of the Taylor series of eqn (1.7e) and use eqns (1.8a) and (1.8h) to substitute for the first and second derivatives, giving

$$x(t+\Delta t) = x(t) + \Delta t\, f(x,t) + \frac{1}{2}(\Delta t)^2\left[\frac{f(x_1, t+\Delta t) - f(x,t)}{\Delta t}\right], \tag{1.8i}$$

where the higher-order terms are omitted. On simplifying this equation, the trapezoidal method is obtained:

$$x(t+\Delta t) = x(t) + \tfrac{1}{2}\Delta t[f(x_1, t+\Delta t) + f(x,t)]. \tag{1.8j}$$

This method is a simple example of a 'predictor–corrector' method: Euler's method is used to predict $x_1 = x(t+\Delta t)$ (eqn (1.8g)), and then eqn (1.8j) is used to improve or correct this first estimate of $x(t+\Delta t)$. The trapezoidal method is second order, so that the error ε is third order. It can be shown that

$$\varepsilon = -\frac{1}{12}(\Delta t)^3\frac{d^3x}{dt^3} + \text{higher-order terms}. \tag{1.8k}$$

This truncation error is smaller than that given by Euler's method (eqn (1.7f)). (See Exercise 1.4.)

1.6.3 *Higher-order methods*

Several higher-order methods are available, such as Simpson's rule, Runge–Kutta methods, and Milne's method. These involve calculating the first derivative within or across the interval. Just as in the trapezoidal method, where in eqn (1.8d) two values of the first derivative allow the second derivative to be estimated, three values of the first derivative allow the third derivative to be estimated, and four values of the first derivative allow the fourth derivative to be estimated. With the latter, the error (see eqn (1.7e)) is then fifth order.

The fourth-order Runge–Kutta method is outlined as an example of a higher-order method. Given the function

$$\frac{dx}{dt} = f(x,t)$$

and given x at time t, at time $t+\Delta t$ this algorithm takes the form

$$x + \Delta x = x + \tfrac{1}{6}(\Delta x_1 + 2\Delta x_2 + 2\Delta x_3 + \Delta x_4), \tag{1.8l}$$

where

$$\Delta x_1 = \Delta t\, f(x,t)$$

$$\Delta x_2 = \Delta t\, f(x + \tfrac{1}{2}\Delta x_1, t + \tfrac{1}{2}\Delta t)$$

$$\Delta x_3 = \Delta t\, f(x + \tfrac{1}{2}\Delta x_2, t + \tfrac{1}{2}\Delta t)$$

$$\Delta x_4 = \Delta t\, f(x + \Delta x_3, t + \Delta t).$$

The truncation error ε is of order $(\Delta t)^5$.

1.6.4 *Variable-step methods*

Variable-step methods are popular because they enable a numerical integration to be performed efficiently in terms of the length of the time step Δt and computer time. The idea is that the time step is maximized subject to the truncation error being less than a certain value. Calculation of the error requires an extra function evaluation. The technique is outlined with respect to Euler's formula.

Euler's method gives (eqn (1.7d))

$$\Delta x = \Delta t\, f(x, t). \tag{1.8m}$$

Therefore, from eqn (1.7f) for the truncation error ε and eqn (1.8f) for the second derivative,

$$\varepsilon = \tfrac{1}{2}(\Delta t)[f(x + \Delta x, t + \Delta t) - f(x, t)]. \tag{1.8n}$$

The relative error is thus given by

$$\frac{\varepsilon}{\Delta x} = \frac{1}{2}\left[\frac{f(x + \Delta x, t + \Delta t) - f(x, t)}{f(x, t)}\right]. \tag{1.8o}$$

The integration time step Δt is then increased until the relative error is just less than a chosen value, which may depend on the word length of the computer being used.

1.6.5 *Stiff equations*

The constant k in eqn (1.7g) is often described as a rate constant. The higher the value of k, the more rapid is the rate of change of x. k has dimensions of $(\text{time})^{-1}$; thus $1/k$ has dimensions of time, and is sometimes referred to as the relaxation time of the equation. In eqn (1.7g), x falls to 1/e of its original value in a time interval of $1/k$.

Stiff equations occur when the rate constants in a system of differential equations differ markedly. The equations then have solutions whose relaxation times are very different. In many areas of biology, including the plant sciences, there may be fast biochemical processes which take milliseconds or less for their completion, at the macromolecular and cellular levels processes may take seconds or even minutes, and at the organ level (for instance, root : shoot ratios) a response may take several days. An integration interval of Δt that suits the slow processes in a model may cause unstable oscillations of increasing amplitude with the fast processes, as in (1.7j). A short time interval that gives stable integration with the fast processes may be very expensive to run over the total time period of interest. It may also be difficult to restrict the range of rate constants in a model without distorting or discarding important components of physiology or biochemistry.

Consider the process (or chemical reaction)

$$z \xrightarrow{k} \cdots, \tag{1.9a}$$

where z is a state variable and k is a rate constant. This leads to $dz/dt = -kz$, which, with $z = z_0$ at time $t = 0$, gives

$$z = z_0 e^{-kt}. \tag{1.9b}$$

Now replace (1.9a) by a two-state variable scheme

$$y \underset{k_2}{\overset{k_1}{\rightleftharpoons}} x \xrightarrow{k_3} \cdots, \tag{1.9c}$$

with state variables x and y, and rate constants k_1, k_2, and k_3. The differential equations of this system are

$$\begin{aligned} \frac{dx}{dt} &= -(k_2 + k_3)x + k_1 y \\ \frac{dy}{dt} &= -k_1 y + k_2 x. \end{aligned} \tag{1.9d}$$

Consider the case where $k_1 = k_2 = 10$, so that x and y are in comparatively rapid equilibrium with each other, and $k_3 = 1$, so that x is converted relatively slowly into other products. Some solutions, starting from several different initial values, are shown in Fig. 1.6. The solution of interest is usually the slowly developing solution, as determined by k_3 in Fig. 1.6, and in this solution x and y are in quasi-equilibrium with each other. The short spurs in Fig. 1.6 illustrate the rapid process that occurs when the system point (x, y) is started at values of x and y which are not in quasi-equilibrium: there is rapid movement due to k_1 and k_2 to re-establish the quasi-equilibrium.

As an aside, it is interesting to consider the consequences of reversing the direction of time in eqns (1.9d) and Fig. 1.6; that is, replace t by $-t$ in eqns (1.9d),

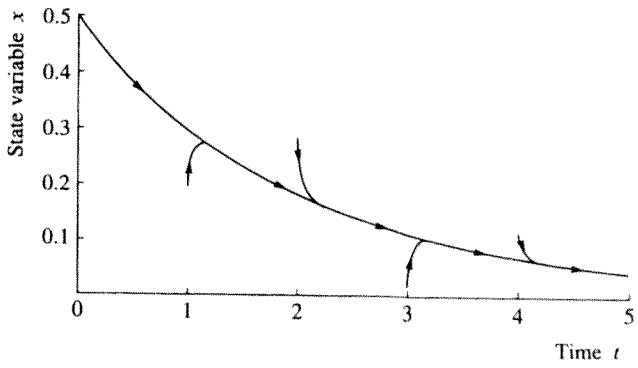


Fig. 1.6. Stiff equations. Results were obtained by integration of eqns (1.9d) with $k_1 = k_2 = 10$, $k_3 = 1$, and $\Delta t = 0.0005$, and using the fourth-order Runge–Kutta method. The 'slow' solution is calculated from initial values of $x = y = 0.5$; the 'fast' solutions shown by the short spurs were obtained by displacing the system point from the slow solution but maintaining $x + y$ at the same value.

so that the system point moves from right to left in Fig. 1.6. It will be clear that the path traced out by the system point is so sensitive to the initial values (now at $t = 5$ in Fig. 1.6) that effectively the system is indeterminate. This is reminiscent of the behaviour of the so-called chaotic systems (Cvitanovic 1984). As has been pointed out by Popper (1982) and others, even a deterministic system of equations can fail to have a determined outcome; thus nineteenth-century scientific attitudes rooted in classical (pre-quantum-mechanical) physics were in error in assuming that the universe is determined.

There are several techniques which can be applied to stiff equation problems in order to obtain solutions without using the brute-force method of a small integration interval and a lot of computing. In a given situation, one of these may well be applicable.

Combining rapidly exchanging pools Rapidly exchanging pools can sometimes be brought together. In eqns (1.9d), define z by

$$z = x + y \tag{1.9e}$$

and add the two eqns (1.9d) together to give

$$\frac{dz}{dt} = -k_3 x. \tag{1.9f}$$

If we assume that $k_1, k_2 \gg k_3$, then, since x and y are now in quasi-equilibrium under k_1 and k_2, from either of eqns (1.9d)

$$k_1 y = k_2 x. \tag{1.9g}$$

Therefore, eliminating y between eqns (1.9e) and (1.9g),

$$x = \frac{k_1}{k_1 + k_2} z. \tag{1.9h}$$

Substituting for x with eqn (1.9h) in eqn (1.9f)

$$\frac{dz}{dt} = -kz \qquad \text{where} \qquad k = \left(\frac{k_1}{k_1 + k_2}\right) k_3. \tag{1.9i}$$

This is equivalent to (1.9a), and a simplification of the problem has thus been achieved.

Reducing the fast rate constants The particular values ascribed to the fast rate constants, which cause difficulties in the integration, may have little effect on the solutions. In Fig. 1.6 the slow solutions are much the same for k_1 and k_2 values of 10, 100, or 1000. Of course varying the fast rate constants has a great effect on the rapid transients in the solutions. Some of the fluxes in the system are reduced in magnitude by reducing the values of rate constants. This may be important to the investigator. An alternative method that leaves the fluxes unaltered is described below.

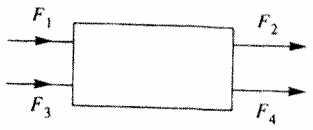


Fig. 1.7. A single compartment with state variable Q of a model is shown; the F_i are fluxes into and out of the compartment.

Increasing pool sizes Consider the 'biochemical' system shown in Fig. 1.7. Q is a state variable denoting a quantity of substance; F_1 and F_3 are fluxes of substance into the pool, and F_2 and F_4 are fluxes out of the pool. The differential equation for Q is

$$\frac{dQ}{dt} = F_1 + F_3 - F_2 - F_4. \tag{1.9j}$$

Assume that substance Q is distributed over a volume V, so that the concentration C of Q is given by

$$C = Q/V. \tag{1.9k}$$

For some applications, the steady state solutions of eqn (1.9j) may be of interest, i.e. when $dQ/dt = 0$ and $F_1 + F_3 = F_2 + F_4$. The fluxes in general depend on the concentration C, so that the quantity Q and therefore C finds a value where the fluxes are in equilibrium. Sometimes, quite small excursions from the steady state give rise to large fluxes (in relation to the pool size which is the steady state value of Q), so that dQ/dt from eqn (1.9j) is large and causes the integration to break down. This type of behaviour usually occurs with small highly labile pools, such as adenosine triphosphate (ATP) or $NADPH_2$. A way round this problem is to inflate the pool size by a multiplier m. Assume that the volume associated with substance Q is

$$V' = mV \tag{1.9l}$$

and, most important, that the initial value ascribed to the pool at time $t = 0$ is changed from $Q(t = 0)$ to

$$Q'(t = 0) = mQ(t = 0). \tag{1.9m}$$

Note that the concentration given by eqn (1.9k) is unaltered when calculated using $C = Q'/V'$. It is assumed that the reactive volumes are unchanged—these are the volumes within which the reactions take place that give rise to the fluxes F_1 etc.—so that the fluxes are unaltered, since these depend on the concentrations and the reactive volumes. Thus dQ/dt calculated from eqn (1.9j) remains the same, but now this rate of change is applied to updating Q' which is m times larger. Integration problems as in (1.7j) are less likely to arise when a rate of change of 10 is applied to a pool of 30 than when it is applied to a pool of 3, as for $m = 10$ with $Q(t = 0) = 3$.

The rate at which the steady state is approached is reduced by factor m, and so the transient behaviour is changed, but in the steady state the correct solutions are still obtained.

Elimination of labile pools In eqns (1.9e) and (1.9i) above we described how rapidly exchanging pools could be combined to give a simpler and more stable representation of the system. Another technique which could be applied to small labile pools is to assume they are vanishingly small and eliminate them from the system. Thus, referring to Fig. 1.7 and eqn (1.9j), it is assumed that

$$F_1 + F_3 = F_2 + F_4. \tag{1.9n}$$

The residual problem is that some of the fluxes may depend on the concentration C. How is this to be dealt with as Q and C are now undetermined? There are two cases to be considered.

1. All fluxes in the system are assumed independent of C: one of the fluxes in eqn (1.9n) (F_2 say) can adjust freely so that eqn (1.9n) is satisfied, and the other fluxes in the equation are determined elsewhere. This same type of balance can be obtained if it is assumed that two or more of the fluxes in eqn (1.9n) can adjust freely but in a fixed ratio: for example, $F_2 + F_4 = (1 + \lambda)F_2$, λ is assigned a value, and F_2 adjusts freely.
2. Some priority scheme is assumed. For example, it is assumed that F_4 is zero and F_2 adjusts freely up to a certain limit, after which F_2 remains constant and F_4 takes up the slack. In plant models it may (for instance) be assumed that first the apex, then the young growing leaves, and lastly the roots take up the consumption of carbohydrate, and this assumption may enable the carbohydrate pool to be eliminated from a plant model.

Special algorithms Recently a number of algorithms have been developed for the stiff differential equation problem. Essentially, they enable step lengths to be used that are considerably greater than the shortest relaxation times in the system. However, there is a computational price to pay in that a state transition matrix must be constructed and inverted. The most important point to be realized is that these algorithms are of no benefit if there are high-frequency components in the driving functions. Thus, if we drive a plant or crop model with rapidly changing light flux densities which follow the fluctuating light levels through the day, stiff algorithms do not enable us to take daily time steps and hence save computing time. It is then necessary to use the brute-force method of a sufficiently small time step and adequate computing power.

1.6.6 *Choice of integration method*

This is a matter of obtaining sufficient accuracy without using excessive computer time. Five fixed-step methods are compared in Table 1.2. The equation integrated was

Table 1.2. Comparison of integration errors (truncation + rounding) and execution times (relative units) for five different fixed-step methods for eqn (1.9o)

	Integration method									
	Euler		Adams		Trapezoidal		Simpson's		Runge–Kutta	
Δt	Time	Error (%)	Time	Error (%)	Time	Error (%)	Time	Error (%)	Time	Error (%)
0.1	1	−99	1	−32	1	−14	2	−11	2	−0.03
0.01	3	−39	3	−1	4	−0.3	5	−0.3	7	−0.2
0.001	17	−6	19	−2	29	−2	43	−2	58	−2
0.0001	160	−15	180	−15	290	−15	410	−15	570	−15

Δt is the integration interval; a word length of 32 bits was employed. The error was computed using eqn (1.9p).

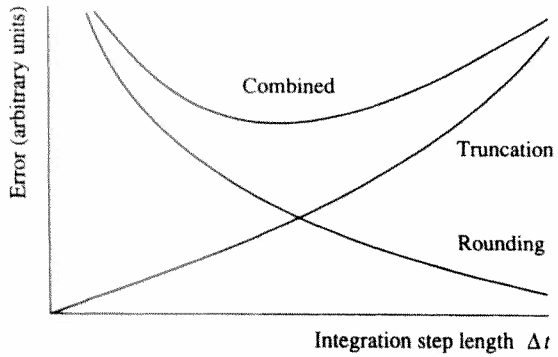


Fig. 1.8. Effect of the integration step length Δt on truncation errors and rounding errors in numerical integration by computer.

$$\frac{dx}{dt} = x \qquad \text{with } x = 1 \text{ at } t = 0. \tag{1.9o}$$

At time $t = 100$, the numerical result $x(t = 100, \text{numerical})$ is compared with the analytical result $x(t = 100, \text{analytical}) = \exp(100)$, and a percentage error is calculated as follows:

$$\text{error } (\%) = \frac{100[x(t = 100, \text{numerical}) - x(t = 100, \text{analytical})]}{x(t = 100, \text{analytical})} \tag{1.9p}$$

It can be seen from Table 1.2 that (for this particular problem and for the computer used) the best error : execution time ratio is given by the high-order Runge–Kutta method with the largest time step of $\Delta t = 0.1$. Rounding errors very quickly become important as Δt is decreased. The compromise between truncation errors and rounding errors is shown schematically in Fig. 1.8.

While the fourth-order Runge–Kutta method is well suited to many problems, it may sometimes give bogus results and there are some problems where it is inapplicable; these arise in crop modelling in particular. Euler's method has an advantage of being so transparent that it is easily checked out in detail. The higher-order methods can, each in its own way, exhibit unstable behaviour and generate spurious results. We advocate that the results of a numerical integration should be viewed sceptically, at least until they have been generated using a range of step lengths and if possible two or three different integration methods.

The plant and crop modeller has a special problem which can compel the use of Euler's method in preference to any higher-order methods, which all involve evaluating the first derivatives more than once per time step as in eqns (1.8j) and (1.8l). This arises because the equations of the model (eqns (1.3a)) are often driven by environmental variables denoted by E. These may be quantities such as

temperature, radiation, or rainfall, which are not available as continuous variables but only as averages over some time interval. For instance, the radiation receipt may be 10 MJ m^{-2} on day 3 of a simulation and 5 MJ m^{-2} on day 4; similarly the mean temperatures on successive days may be 22 °C and 17 °C. These discontinuities cause discontinuities in the first derivatives in eqns (1.3a) on moving from one time interval to the next. It may not be obvious how the first derivatives should be calculated within or at the two ends of a time interval. The method that many plant and crop modellers use is to measure all the environmental variables with respect to a common time interval (or refer them to a common time interval), and then to use this time interval for numerical integration using Euler's method. Many modelling languages (ACSL, CSMP; see Section 1.7.3, p. 34) allow a choice of several integration methods.

1.7 Evaluation of models

The term evaluation is used here in the wider scientific sense—as part of the general methodology of science—and, potentially at least, evaluation may be concerned with any aspect of a model. It must be stressed that the evaluation of a model is not a wholly objective process. It may be connected with properties of the model such as its utility, simplicity, elegance, economy, plausibility, goodness of fit, and appropriateness to objectives. Researchers will always attach different weights to these items, and so it is unsurprising that models in the same problem area are differently ranked by different people. Some modellers speak of model verification and validation, although this suggests a lack of understanding of the 'working hypothesis' status of a scientific theory. Popper's (1958) thesis, now widely accepted, is that theories can only be falsified, and so the term 'validation' must be presumed to mean a failed attempt at falsification. Most crop and plant models have a short life, and they are soon discarded in favour of other, usually more general, models. It is even mistaken to talk about 'validation' for an applied model with a clear practical objective, such as the efficient application of nitrogen fertilizer to grassland. All that needs to be demonstrated is that the proposed model produces better results than the current model (current practice) in a defined set of circumstances; this is sometimes referred to as the 'champion–challenger' approach. From the farmer's viewpoint, it may not matter too much that the model may be based on (what some would regard as) mistaken assumptions.

Perhaps it hardly needs to be said, but any model should be thoroughly tested for methodological correctness: the mathematical equations must correctly represent the stated biological assumptions; both assumptions and equations must be self-consistent, with the equations being dimensionally homogeneous; any algebra or analysis must be free from error; any computer code must be correct. The modelling literature is not free from such failings. Indeed, in large and complex models, it can be so difficult not to make errors that it is better to assume

that errors will be made and to adopt a self-checking method of working that traps the errors as or shortly after they are made; it may be almost impossible to find them later. In contrast with model evaluation, model testing is an objective process which has the result 'true' or 'false'.

Both evaluation and testing should be applied continuously throughout a modelling project right from the beginning. Final evaluation of a model depends upon being sure of the model's methodological correctness.

Most modelling projects can be considered in five parts.

1. The objectives of the research (p. 14): it is common for objectives to change somewhat as the project proceeds, and this is legitimate because science is intrinsically unpredictable.
2. Definition of the structure of the model: this includes the type of description (levels of aggregation), often a diagram such as Fig. 1.3, and the biological assumptions.
3. Representing 2 mathematically, and performing any subsequent analysis.
4. Solving the mathematical problem resulting from 3, usually by means of a computer program.
5. Examining and interpreting the model predictions, especially in relation to 1; attempting to simulate experimental data if suitable data are available.

Evaluation and testing are carried out at each step. The steps do overlap and interact, and it is usual to move backwards and forwards around these five items.

1.7.1 *Structure of the model*

A mathematical model represents a set of biological assumptions which are always a simplification of reality. Inevitably there are colleagues and critics who are unwilling to accept the level of simplification chosen: the model is either 'too complex' or is 'over-simplified'. Although mathematics is the servant of the sciences (for current purposes), there is little double that the assumptions of model builders are often constrained by their ability to express those assumptions mathematically. A good appreciation of the biological state of the art is essential. Much of the skill and art of modelling then lies in deciding which details can be ignored, what approximations are reasonable and appropriate, and in striking that fruitful compromise which leads to progress. Procrustean assumptions may well be needed, and colleagues who see their research area brusquely set aside may be much offended. There are no objective rules for going about this process. Conjecture is used, but it should be informed conjecture; however, it is inevitably based on the experience, skills, and judgement of the modeller.

Sometimes one sees modellers attempting biological modelling, but keeping the biology of the problem at arm's length. It rarely works: the collaborating biologist may fail to grasp what is being attempted in the mathematical representation, and the modeller may be unable to assess and rank the biological possibilities. The result is likely to be work of little value, and frustration for the concerned parties.

1.7.2 *Mathematical equations*

The accurate translation of biological ideas into mathematics requires mathematical fluency, a wide familiarity with mathematical possibilities, and a sound understanding of the biological ideas being translated. Clearly this requires either highly numerate biologists or close collaboration between biologists and people with the requisite mathematical skills. Some simple guidelines, which reduce the possibility of error and help in detecting errors, can be followed to accomplish this.

The first step is to define the symbols. It is worth giving this careful thought, since equations are much easier to read, understand, and check if similar symbols are used for similar quantities, and if the similar symbols have the same units. For example, rate constants with dimensions time^{-1} may be denoted by k_1, k_2, ..., . In a plant model the components of dry mass may be shown by W_l, W_s, and W_r for the leaf, stem, and root. Where there is a consensus in the literature about the use of certain symbols for certain quantities, the traditional symbols should be used unless there are good reasons for doing otherwise. The use of computer language notation such as FORTRAN in mathematical analysis (or in scientific papers) is, in our view, mistaken. It is less efficient, less readable, and less easily checked than the more conventional mathematical notation that has evolved over many centuries using the Latin and Greek alphabets, with upper or lower case, and subscripts and superscripts as needed. Computer notation, while still quite primitive, is rapidly evolving. Work presented in such notation may quickly become inaccessible. Indeed, many journals do not allow computer notation to be used within the body of a scientific communication.

The second step is to check dimensions. In an equation, each term must have the same dimensions as all the others. For this purpose, a symbol table should be constructed which has a verbal definition of each symbol used and also its dimensions. Sometimes it is also helpful to work out the dimensions of groups of symbols that often occur together. A single system of units (preferably SI) should be used throughout the model, even when these are not the customary units. This avoids troublesome conversion factors for quantities like grams to kilograms, or cubic metres to litres, which can very easily give rise to errors.

A third step is to check for mathematical consistency and completeness. There must be enough equations to define the problem, but the problem must not be over-defined. For instance, for a dynamic model with three state variables, three difference equations or differential equations are required. For a simple static problem with five variables, five equations are needed.

A fourth useful check is for biological consistency and completeness at the whole-system level. In many plant models (e.g. Chapter 16, Section 16.5, p. 464), carbon (C) and nitrogen (N) accounting can be carried out. We can write

$$\frac{d}{dt}(\text{total C or N in system}) = \text{system inputs} - \text{system outputs}. \quad (1.10a)$$

Internal transfers T_{ij}, say from pool i to pool j, occur twice in the mathematical equations of the model: positively in the differential equation for pool j, and negatively in that for pool i. Summing the equations should give cancellation of all internal transfers. For instance, however complicated a plant model is, some of its equations should sum to

$$\text{gross photosynthetic rate} = \text{growth rate} + \text{respiration rate} + \text{senescence rate}, \tag{1.10b}$$

where these quantities are expressed in the same units.

1.7.3 *Solving the equations of the model*

This process is usually carried out by computer. If possible, one should choose to use a modern portable language that lends itself to well-structured self-documenting programs. Variable names should be chosen with care, and should correspond to the mathematical/biological variables. The general principle to be followed is that mistakes will be made, so that programs are written so that the mistakes are easily located and corrected.

There are now some good non-procedural modelling languages available, such as Continuous Systems Modelling Program (CSMP) (Speckhart and Green 1976) and Advanced Computer Simulation Language (ACSL) (Mitchell and Gauthier Associates 1987); ACSL has been used to solve most of the dynamic models described in this book. In a non-procedural program, the program statements can be written in any order; during the compilation process, the statements are put in an executable order. To program in a non-procedural language is a liberating experience that must be experienced to be appreciated. Using such languages is easy and quick (compared with using FORTRAN), and the program can be structured according to the biology of the problem, which increases program readability enormously. While these languages are not suitable for all problems, they are ideal for dynamic deterministic models of the type of eqns (1.3a).

Where possible, self-consistency checks of the type in eqns (1.10a) and (1.10b) should be written into the program; these may pin-point programming errors or mathematical errors in model formulation. In the early runs of a program, it is often worthwhile to print out every left-hand-side quantity in the program, and sometimes errors can be located by performing a detailed check on a hand calculator working direct from the mathematical equations (not from the programmed version of these equations).

Checks should also be applied against the possibility of integration errors due to an inappropriate integration method or to too large an integration interval. The results of running the program should be reasonably stable against variation in integration method and interval. Some machines have rather a short word-length, and rounding errors can be soon encountered if very short integration intervals are used (p. 30).

1.7.4 *Comparison of model with experiment: model fitting*

After the model has been carefully tested, and is free from mathematical, computational, and numerical errors, its predictions then truly reflect the assumptions on which it is based. It is essential to evaluate a model first by examining its qualitative behaviour. If this is satisfactory, then one can proceed to a direct comparison of the model's predictions with observational data, if suitable data are available, which is often not the case. When comparing a model's predictions with observational data, a measure of goodness of fit is required, and frequently some of the parameters of the model can be estimated by optimizing the goodness of fit. This last process is called fitting, or sometimes tuning or calibrating, a model.

Fitting a model generally means adjusting some of the parameter values (shown by P in eqns (1.3a)) and perhaps some of the initial values also ($X_i(t = 0)$, $i = 1, 2, \ldots, q$) so that the predictions of the model more closely resemble the observational data; this adjustment process does not alter the structure or basic equations of the model. To some modellers this approach is unacceptable—they feel that parameters should be available from independent investigations, possibly at the level of the assumptions. This view may neglect the practical objectives and empirical content which are associated with many models. Fitting can be an important part of model evaluation, and the simple procedure outlined below has been found to be of practical value.

Consider the case where a single attribute y of the system (dry mass say) is measured at m time points (t is the time variable), to give a set of m number pairs:

$$(y_1, t_1), (y_2, t_2), \ldots, (y_m, t_m). \tag{1.11a}$$

We further assume, for simplicity, that each y_i is a mean over any replicates that are taken at the ith time point.

The state variables of the model are X_k, $k = 1, 2, \ldots, q$ (see eqns (1.3a)), and running the model on the computer predicts a value for X_k at any time point t_i. We write this as

$$X_k(t_i; P; E) \tag{1.11b}$$

to emphasize the fact that the predicted values depend on the values assigned to the parameters P. In (1.11b) the parameters P include the initial values $X_k(t = 0)$, and from now on when we refer to parameters, we include initial values in with the parameters. One of the state variables of (1.11b) may correspond directly to the experimentally measured quantity y, or it may be necessary to derive an auxiliary variable that does correspond to y. (State variables may be root, stem and leaf dry masses; from these the total dry matter is calculated (this is an auxiliary variable), and it is the predicted value of total dry matter that can be compared with experiment.) Thus the predicted variable that corresponds to observation y_i is denoted by Y_i, and the model gives a set of predicted values, one at each of the m time points t_i:

$$(Y_1, t_1), (Y_2, t_2), \ldots, (Y_m, t_m). \tag{1.11c}$$

The sets (1.11a) and (1.11c) can now be compared; a perfect fit would give $y_i = Y_i$ at every time point. Since the model is deterministic, the Y_i are obtained without any probability distribution. The environment in which the observational data were taken is the same as the environment E used within the model; if two environmental treatments were used, then there would be two sets of observational data and two sets of predicted data.

Calculation of a residual A residual r_i can be calculated according to

$$r_i = y_i - Y_i \quad \text{or} \quad r_i = \ln\left(\frac{y_i}{Y_i}\right), \tag{1.11d}$$

or by using some other measure. In the plant sciences the second of eqns (11.1d) is usually appropriate, since this weights data points with the same proportional error equally. Thus a 10 per cent error in a predicted dry mass of 0.001 kg is of the same importance as a 10 per cent error in a predicted dry mass of 1 kg. Equations (1.11d) can be summed to give a residual sum of squares

$$R = \sum_{i=1}^{m} g_i r_i^2, \tag{1.11e}$$

using a weighting factor g_i if required. The residual sum of squares R is a measure of the lack of fit of the model and it depends on the parameter values, so that we can write

$$R \equiv R(P). \tag{1.11f}$$

Consider the simplified case where the model has just a single parameter P. Best fit is obtained by adjusting P so that R is a minimum, giving

$$\frac{dR}{dP} = 0 \quad \text{and} \quad \frac{d^2R}{dP^2} > 0. \tag{1.11g}$$

Examples of a sensitive parameter (curve *A*) and an insensitive parameter (curve *B*) are illustrated in Fig. 1.9. In curve A the value of d^2R/dP^2 is large, and this defines the value of P for best fit more narrowly. It is desirable that the residual R should be reasonably sensitive to all the parameters of the model, and if there are some parameters to which R is completely insensitive, this can indicate areas in which the model might be simplified. It must be remembered that such a result depends totally on the nature of the observational data against which the model is measured.

If truncation and rounding errors (which are always present but are usually insignificant) are ignored, the predicted values $Y_i = 1, 2, \ldots, m$ are computed without error; however, the experimental data y_i, $i = 1, 2, \ldots, m$ are subject to error, and this puts a lower limit on the value of R that can be achieved by parameter adjustment. The residual sum of squares R can be divided into two

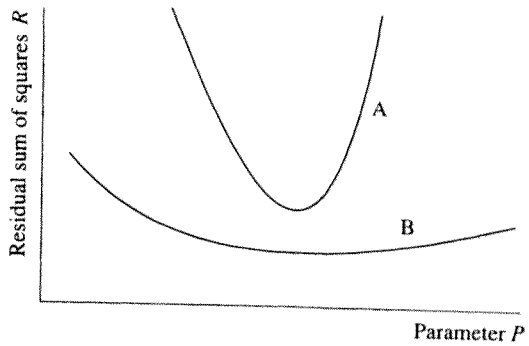


Fig. 1.9. Sensitive and insensitive model parameters: curve A, the residual sum of squares R varies rapidly as the parameter P changes, and the value of P is therefore critical to the fit obtained with the model; curve B, the value of P has little effect on R, so that fitting the model is not able to define P closely and the value assigned to P is not critical to the model.

components:

$$R = R_l + R_e \tag{1.11h}$$

where R_l is the part of the residual due to the lack of fit of the model and R_e is due to error in the experimental data. R_e has an expected value of

$$R_e = (m - n)\sigma^2, \tag{1.11i}$$

where m is the number of data points, n is the number of parameters in P which are adjusted, and σ^2 is the error variance. The parameters P only affect R_l, but if R_e is very large (adding a large constant term to the curves in Fig. 1.9) R may be rendered much less sensitive to the P values. If the experiment resulting in the data $y_1, y_2, \ldots, y_m$ can be replicated, then the error term R_e, may be known. Otherwise, an upper limit an R_e and also on σ^2 is given by the minimum value of R obtained by adjusting the parameters P.

Suppose that instead of the single set of data as in (1.11a), we have two sets, perhaps plant dry mass and leaf area, giving

$$W_i, i = 1, \ldots, m_W,$$

and

$$A_i, i = 1, \ldots, m_A. \tag{1.11j}$$

Using eqns (1.11d) and (1.11e), we can calculate two residual sums of squares, that with respect to dry mass R_W and that with respect to leaf area R_A. It is necessary to combine these in order to carry out parameter adjustment; this can be achieved using

$$R = \frac{R_W}{\sigma_W^2} + \frac{R_A}{\sigma_A^2} \tag{1.11k}$$

where the σ_W^2 and the σ_A^2 are the respective error variances. The parameters can now be adjusted with respect to the combined residual R. If the error variances are not known, then clearly eqn (1.11k) cannot be constructed, although for fitting it is still necessary to combine the residuals in some way. If the second of eqns (1.11d) is used to calculate the residuals r_i, then R_W and R_A are independent of the dimensions of W or A and of any scaling factors. If it is further assumed that dry mass W and area A have the same coefficient of variation, which is often a reasonable assumption, then the combined residual can be written

$$R = R_W + R_A, \tag{1.11l}$$

and the parameters can be adjusted so that R in eqn (1.11l) is minimized.

The goodness of fit can be estimated by comparing the residual sum of squares due to lack of fit of the model (R_1 in eqn (1.11h)) with the error residual sum of squares R_e using the F test (see Exercise 1.5). An estimate of the error term is only available if a replicated experiment has been performed; if this is not the case, only a qualitative and subjective assessment of the fit of the model to the data may be possible.

Confidence intervals for fitted parameters The dependence of the residual sum of squares R on the n parameters P_j, $j = 1, 2, \ldots, n$ can be expressed by

$$R = R(P_1, P_2, \ldots, P_n). \tag{1.11m}$$

Computer methods are generally used to search for the minimum of R with respect to the P_j using an optimization procedure. The function R, whose minimum one is trying to find, is known as the objective function. The gradient vector is the set of first partial derivatives of R with respect to the parameters P_j, $i = 1, 2, \ldots n$, namely $\partial R/\partial P_1$, $\partial R/\partial P_2, \ldots, \partial R/\partial P_n$. At a minimum,

$$\frac{\partial R}{\partial P_1} = \frac{\partial R}{\partial P_2} = \cdots = \frac{\partial R}{\partial P_n} = 0. \tag{1.11n}$$

The matrix of second partial derivatives of the objective function is called the Hessian matrix, and its elements take the form

$$H_{jj'} = \frac{\partial^2 R}{\partial P_j \partial P_{j'}}. \tag{1.11o}$$

At the minimum all the eigenvalues of the Hessian matrix must be positive. Essentially this means that the second derivatives with respect to the parameters are positive, and R increases in whatever direction one moves from the point defined by $\partial R/\partial P_j = 0$, $j = 1, 2, \ldots, n$. The matrix in (1.11o) is obtained by numerical differentiation, for which computer procedures are available.

Let the matrix G with elements $G_{jj'}$ be the inverse of the Hessian matrix H with elements $H_{jj'}$. This can be represented symbolically by

$$G = H^{-1}, \tag{1.11p}$$

although it is usual to use a computer procedure to obtain the inverse matrix.

The number of degrees of freedom is denoted by ν. If there are m data points and n adjustable parameters, ν is given by

$$\nu = m - n, \tag{1.11q}$$

although if there are two data sets as in (1.11j), ν is given by

$$\nu = m_W + m_A - n.$$

An unbiased estimate of the variance V of parameter P_j is

$$V(P_j) = \frac{R}{\nu} G_{jj}, \tag{1.11r}$$

and an unbiased estimate of the covariance C of parameters P_j and $P_{j'}$ is

$$C(P_j, P_{j'}) = \frac{R}{\nu} G_{jj'}. \tag{1.11s}$$

If P_j^* is the true value of the parameter P_j, then the $100(1 - \beta)$ per cent confidence interval of P_j^* is

$$P_j \pm [V(P_j)]^{1/2} t_{\beta,\nu} \tag{1.11t}$$

where $t_{\beta,\nu}$ is the 100β percentage point of the t distribution with ν degrees of freedom (France and Thornley 1984, pp. 282–283). (See Exercise 1.5.)

1.7.5 *Sensitivity analysis*

Sensitivity with respect to observational data Consider a model with a single adjustable parameter P which has been fitted to a data set by minimizing a residual sum of squares R with ν degrees of freedom (eqns (1.11d), (1.11e), and (1.11q)). The variance $V(P)$ of P is given by (eqns (1.11o), (1.11p), and (1.11r))

$$V(P) = \frac{R}{\nu} \frac{1}{\partial^2 R/\partial P^2}. \tag{1.12a}$$

To compare the effects of different parameters on model performance, a dimensionless quantity is required that is independent of the absolute value of a parameter; the variance $V(P)$ of P has the same units as P^2 and is therefore not acceptable. The coefficient of variation $\mathrm{CV}(P)$ of V is such a quantity, and is defined by

$$\mathrm{CV}(P) = \frac{[V(P)]^{1/2}}{P}. \tag{1.12b}$$

In Fig. 1.9, curve A has a high value of curvature $\partial^2 R/\partial P^2$, giving reduced values of $V(P)$ and $\mathrm{CV}(P)$; a low value of $\mathrm{CV}(P)$ indicates that the model is sensitive to the parameter P. Similarly, curve B in Fig. 1.9, which denotes an insensitive parameter, has a low value of curvature and a higher value of $\mathrm{CV}(P)$.

Now consider a model with several parameters P_j, all of which affect the performance of the model. Analogous to eqn (1.12b), the coefficient of variation

of P_j can be calculated by

$$\mathrm{CV}(P_j) = \frac{[V(P_j)]^{1/2}}{P_j}, \tag{1.12c}$$

where the variance $V(P_j)$ of P_j is obtained from eqn (1.11r). The coefficients of variation can be used to rank the parameters—a low value of $\mathrm{CV}(P_j)$ denotes that parameter P_j has a considerable effect on the fit of the model to that particular data set, and vice versa. It must be emphasized that the results of a sensitivity analysis depend entirely on the data set being used, and different data sets can give very different results. There is, therefore, nothing 'objective' or absolute about the method, which may say more about the data set than about the model. Nevertheless, the technique can be useful, especially when preparing a model for applications purposes which may require defining the precision required of the parameters of the model and specifying its range of validity (by this we mean a certain specified level of predictive accuracy). Experience with plant and crop models suggests that values of $\mathrm{CV}(P_j)$ in the range 0.05–0.2 are reasonable.

Ranking the parameters by means of a sensitivity analysis may indicate ways in which a model might be simplified. A high value of $\mathrm{CV}(P_j)$ means that the parameters P_j has little effect on the predictions of the model. It may be possible to remove such a parameter from the model. However, there may be good reasons, biological or physiological, for retaining the parameter, even though it has little effect on the model's predictive performance.

Sensitivity with respect to model predictions So far we have considered sensitivity analysis in terms of the effects of a parameter P_j on the goodness of fit as measured by a residual. It may be more pertinent to evaluate the sensitivity of some quantity Q to the parameter P_j; we denote this by $S(Q, P_j)$. For example, we may be concerned with the yield of a crop at maturity or at a particular time during the growing season. Suppose that a model has been evaluated and is judged satisfactory. Some of the parameters of the model will be physiological/genetic and some will be environmental. Some of the environmental parameters, such as the timing and quantity of fertilizer applications or irrigation, may be within the control of management. Objectives for programmes of plant breeding, genetic manipulation, or management priorities can be formulated more effectively if it is possible to rank the parameters P_j according to their effects on (say) yield Y. A dimensionless measure of the sensitivity of yield Y to a parameter P_j is

$$S(Y, P_j) = \frac{\partial Y}{\partial P_j}\frac{P_j}{Y} \approx \frac{\delta Y}{Y}\frac{P_j}{\delta P_j}. \tag{1.12d}$$

δP_j denotes a small finite change in the parameter P_j, and δY is the change that this produces in Y. For computing the sensitivity $S(Y, P_j)$ of Y and P_j, a 5 per cent parameter increment is usually sufficient, so that $\delta P_j/P_j = 0.05$. If a 5 per cent change in P_j gives a 5 per cent change in Y, then $S(Y, P_j) = 1$. Parameters with $S(Y, P_j) > 1$ have larger effects on yield, and vice versa.

An example of the use of a model to rank parameters is given by Thornley, Hurd, and Pooley (1981, Table 1); they rank the parameters of a leaf-growth model in terms of their marginal contribution to the carbon budget of the plant.

Finally, it should be noted that the sensitivity of a model, as defined by eqn (1.12d), is very similar to the economist's concepts of demand elasticities and cost elasticities. If the price of some good is p and x is the number of goods sold per unit time, then a demand elasticity E can be defined by

$$E = \frac{\delta x/x}{\delta p/p} = \frac{\delta(\ln x)}{\delta(\ln p)}. \tag{1.12e}$$

The prefix δ denotes a small increment in the variable. Comparison of eqns (1.12d) and (1.12e) shows that they are identical in structure.

1.8 The presentation of models

The continued health and progress of science depends upon the existence of open channels of communication. Publication is the means by which scientists put their efforts before the scientific community, which is then free to ignore, refute, modify, or applaud the contribution. However, even the brief history of science reveals many examples of attempts by the 'scientific establishment' to suppress innovative contributions; examples of such censorship include theories of continental drift and the origin of the solar system. Fortunately the pluralistic and fragmented nature of science has thwarted such approaches, although the inefficiencies caused by 'establishment' attitudes should not be underestimated (Lock 1986).

The modeller of plant and crop growth, and indeed of wider agricultural problems, will doubtless have run up against some of the problems that arise in attempting to publish modelling papers. Our reason for writing this section is to help modellers fight the battles with which they may be faced. Just as the weak administrator likes to construct and rely upon an extensive set of rules, rather than have some guiding principles and then treat each problem on its merits, so some journals like to decide a quite detailed 'policy', which can then be used as a blunt instrument to decide what is and what is not acceptable to the journal. Of course, it will always be difficult for any journal to strike the right balance between material that is not worth publishing and encouraging the growing points of the subject, which will often seem heretical or of little value to many and will inevitably be difficult to assess (Lock 1986).

Some biological journals will not accept papers that are 'purely theoretical' and some will not accept papers that are 'purely observational', even though these papers may be highly relevant to the field of scholarship of the journal. However, to exclude purely theoretical papers (papers without direct reference to observational data) is really as unwise as to do the converse, i.e. to exclude

purely observational papers which have no interpretative content, mathematical or otherwise. Not only does this indicate a misunderstanding of the very nature of science and its components of observation, speculation (or hypothesis), and deduction (perhaps by means of a mathematical model), but in other areas of science one can readily see that all these different types of science, including the extremes, have made profound contributions. Examples of purely theoretical work, sometimes untested and untestable for many years, include Einstein's theory of general relativity, Maxwell's electromagnetic theory, and the prediction of the neutrino; all these were enormously stimulating to the subject. Examples of pure observation are early work in atomic and molecular spectroscopy, superconductivity, and antibiotic activity. It would have been a great loss if these contributions had not won publication, and one could speculate endlessly on what might be missing from current science as a result of editorial policy. Of course many papers lie at neither extreme, and they contain a mix of observational data and interpretation, sometimes by means of a model.

The referees and editors of some biological journals sometimes demand of modelling papers the satisfaction of criteria that are seldom applied to other papers, and they may also require the meeting of objectives of their own choosing (such as an extensive 'validation'; see p. 31). It should be remembered that science progresses mostly by quite modest steps, and this is true of theoretical work as well as experiment. Small but useful contributions are often noticed far away, where further work may be stimulated. Just as experiments may lead to further experiments, so may models lead to other models, experiments may stimulate models, and models may give rise to directed experimentation. Further, one may construct different models of the same system for different purposes, and confronting a model with experimental data may or may not be needed to meet the modelling objectives. Although the desirability of comparing models with experimental data is self-evident, anyone who has attempted to do this will be aware of (a) the difficulties of finding suitable data, (b) the operational problems of comparing complex mechanistic models covering two or more hierarchical levels with data, and (c) the inconclusiveness of such procedures even when carried out. Indeed, in some areas, such as modelling the geochemical carbon cycle or the 'nuclear winter' scenarios, comparison of theory with experiment may be impossible, yet there is little doubt about the value of these models. It seems to be of more scientific value to examine qualitatively the trends and patterns predicted by the model; the topology of the responses is of greater significance than precise numerical values, at least in the current state of the art. Perhaps the most important thing is that the objectives stated by the author(s) should be given close scrutiny, and one should assess to what extent these objectives are legitimate and are subsequently met.

1.8.1 *Requirements for publication*

It seems that editorial policy and authors should be concerned with five principal items:

(i) clarity;
(ii) economy;
(iii) methodological correctness;
(iv) not a trivial repetition of already published work;
(v) accuracy.

Referees and editors who pass judgements that go beyond these five points are assuming an authority that belongs properly to the scientific community as a whole. Work that satisfies the above criteria should be published so that evaluation by the scientific community can proceed unhindered.

Clarity is essential if the work is to be understood, and if it cannot be understood, at least by a few who can communicate it further, it will not have any value. Readability and clarity can usually be assessed by workers in the same general area, and a specialist in the particular topic is not necessarily required; indeed, specialists may be so concerned with what is being said, that they find it difficult to evaluate how it is being said.

Economy is needed simply because journal space is expensive. Clarity and economy may sometimes be in conflict, but more often they go together.

Methodological correctness is usually easy to assess. In many areas of both mathematics and experimental technique there is a wide acceptance of a basic methodology, and in these areas the work should be free from error. For the mathematical modeller, this might include algebra, calculus, and numerical analysis; for an experimentalist this could include the measurement of temperature and dry mass, and many chemical techniques.

The fourth item—that the work should not be a trivial repetition of work that has already been published—is self-explanatory. All work stands on what has gone before, and to give continuity, comprehensibility, and context, there will always be some repetition. However, there must be some non-trivial aspect of the newly reported work which is different from previous work. This may be interpreted as a requirement for 'originality', although what is original may be a subject for much debate; for instance, a synthesis of existing concepts may be unoriginal at one level, but may lead to novel insights. Modelling is often about the integration of ideas—the whole is more than the sum of the parts, although it is explainable in terms of those parts and how they fit together (Section 1.2.2, p. 8).

With regard to the fifth and last item, accuracy, authors should realize that if they allow minor errors of referencing, style, tabulation, equation layout, and typing to appear too frequently in the final manuscript, this must cast substantial doubts on the thoroughness and correctness of the work, and greatly reduces its overall credibility.

There is no wholely objective way of scoring a piece of work with respect to these attributes or of combining the attributes into a single score. One manuscript may be rather obscure but highly original, whereas another might present little new at the ideas level but may be a lucid and accessible exposition of a difficult topic; the originality of the latter lies in its lucidity. Many different types of contribution play a valuable role in furthering science. The scientific community

has its own way of sorting out the wheat from the chaff. It is possible that a wider adoption of 'open refereeing' would be beneficial. This would encourage referees to be more objective in their criticisms. It would not be possible to express opinions or prejudices under the shelter of anonymity; in science (and indeed elsewhere) there should be no place for secret judgements. For most of us, it is a continual struggle to regard our views and attitudes as 'working hypotheses', but open refereeing could help greatly in this respect.

Exercises

1.1. The Gompertz growth equation in differential form is

$$\frac{dW}{dt} = \mu_0 W e^{-Dt} \qquad \text{with } W = W_0 \text{ at time } t = 0.$$

W is a state variable denoting dry matter with initial value W_0; μ_0 and D are constants. Write this single differential equation with explicit time dependence as two differential equations for two state variables. Suggest a biological interpretation.

1.2. The compartmental scheme $X \rightarrow Z \rightarrow \cdots$ where the rate constants into and out of compartment Z are both k has the differential equation $dZ/dt = k(X - Z)$. Show that the integral equation

$$Z(t) = k \int_{-\infty}^{t} X(t') \exp[-k(t - t')]\, dt'$$

is equivalent to the differential equation above. This is equivalent to a remembered variable as in eqn (1.5h), with the weighting function equal to $k \exp[-k(t - t')]$.

1.3. Use Euler's method of eqn (1.7d) to integrate numerically the differential equation $dx/dt = -x$ with $x = 1$ at time $t = 0$. Use a time interval $\Delta t = 0.1$, and perform the calculation over three time steps up to $t = 0.3$. Also check your results against the analytical solution.

1.4. Use the second-order trapezoidal method of eqn (1.8j) to rework the integration in Exercise 1.3. Note the greatly increased accuracy.

1.5. A model has been fitted to 84 data points by adjusting four parameters to minimize the log residual sum of squares R (eqns (1.11d) and (1.11e)), obtaining R(minimum) = 0.8. Calculate the mean residual sum of squares and estimate the average relative error (or lack of fit) between prediction and observation. Assume that the error residual (total) is found to be 0.4 with 50 degrees of freedom. Is the model giving an acceptable fit to the data at the 10 per cent probability level?

1.6. Suppose that the fractional carbon content f_C of plant tissue is defined by the equation $W_C = f_C W$, where W_C (kilograms of carbon) and W (kilograms of total dry matter) denote the masses of carbon and total plant dry matter respectively. Derive the units of f_C. Can these units be simplified? What are the units of leaf area index (LAI) and can these be simplified?

2
Some subjects of general importance

2.1 Introduction

In this chapter we consider some topics which are of general importance in the plant sciences, and are useful for much of the subsequent material of the book. We begin by considering the basic system of units and conversion factors. This is an area where confusion can often arise, but can easily be avoided. We then consider some of the principles of enzyme kinetics, since it is possible to derive from enzyme-kinetic considerations several functions which are of considerable utility to the modeller, as they are mathematically well behaved and have biologically interpretable parameters. An added advantage of understanding the basic concepts of enzyme kinetics is that it gives some insight into the underlying biochemical processes involved in plant and crop modelling. The final topic covered is cell division and organ growth, which is relevant to the difficult area of differentiation and development, and again provides useful background knowledge for modelling plant and crop processes.

2.2 Units and conversion factors

Confusion that arises from the choice of units can cause unnecessary problems in plant and crop modelling, and in science in general. For example, single-leaf photosynthesis may have units mg CO_2 (m^2 leaf)$^{-1}$ s^{-1}, and crop yield may have units kg (dry matter) ha^{-1}. The mixture milligrams and kilograms for mass, and square metres and hectares for area within the model invites error and confusion. It is important to adhere to a consistent set of units throughout any model. Any conversions to what may seem more appropriate units should be made at the end of, or preferably outside, the model. By doing so, the model is independent of conversion factors, and the modeller is in no doubt as to the units of quantities within the model. Another important advantage of this approach is that it facilitates the essential check for dimensional consistency of equations which should always be done.

In this book, we use the now almost universally accepted International System of Units (SI) (Royal Society 1975), with one exception (discussed below) in regard to the mole as the amount of substance. The basic units for mass, length, and time, are kilogram (kg), metre (m), and second (s). Derived units for quantities such as energy (joule (J)), pressure (pascal (Pa)), and force (newton (N)) are all defined in terms of these base units. For dynamic models of crop growth over the growing cycle of the crop the day may be the natural description of time and

in such cases will be used, where

$$1 \text{ day} = 86\,400 \text{ s}. \tag{2.1a}$$

2.2.1 *Relative molecular mass and kilogram mole*

We shall first define these units as they will be used in this book, and then relate them to the SI definitions. In so doing we shall highlight the weaknesses of that system.

A kilogram mole (kg mol) is defined as that amount of a substance which contains N_A units, or entities of the substance, where N_A is Avogadro's number. Choose a reference substance where the mass of one entity is m_r kg. By definition

$$N_A m_r = r \text{ kg (kg mol)}^{-1}, \tag{2.2a}$$

where r is the mass of 1 kg mol of the reference substance. In (2.2a), either N_A or r can be assigned an arbitrary numerical value. We select a reference substance with known m_r, define r, and then derive the corresponding N_A. Following convention, the reference substance is taken to be carbon 12, and r is defined as 12 kg (kg mol)$^{-1}$, so that

$$N_A m_{^{12}C} = 12 \text{ kg (kg mol)}^{-1}, \tag{2.2b}$$

and this defines N_A as

$$N_A = 6.022 \times 10^{26} \text{ (kg mol)}^{-1}. \tag{2.2c}$$

It follows from these equations that the kilogram mole is defined as the amount of substance of a system which contains as many elementary entities as there are in 12 kg of carbon 12; that is, the mass of 1 kg mol of carbon 12 is 12 kg. Now consider a substance with n entities, each of mass m kg. The total mass M of the substance is

$$M = nm \text{ kg}. \tag{2.2d}$$

The quantity Q of the substance, measured in kilogram moles, is

$$Q = \frac{n}{N_A} \text{ kg mol}. \tag{2.2e}$$

We can now define the *molar mass* or *relative molecular mass* μ as

$$\mu = \frac{M}{Q} \text{ kg (kg mol)}^{-1}, \tag{2.2f}$$

which, using eqns (2.2d) and (2.2e), is equivalent to

$$\mu = N_A m \text{ kg (kg mol)}^{-1}. \tag{2.2g}$$

Eliminating N_A, using eqn (2.2a), we can write this as

$$\mu = \frac{m}{m_r/r} \text{ kg (kg mol)}^{-1}. \tag{2.2h}$$

Now, if carbon 12 is used as the reference, then this states that the relative molecular mass is the ratio of the average mass per molecule of the natural isotopic composition of the elements to 1/12 of the mass of an atom of carbon 12. It is important to note here the factor 1/12 actually has units kg mol kg^{-1} (eqns (2.2a) and (2.2b)) and the relative molecular mass has dimensions kg (kg mol)$^{-1}$. The relative molecular mass of any substance can now be calculated. For example, $\mu_{CO_2} = 44.0098$, which means that a kilogram mole of CO_2 is 44.0098 kg of CO_2 (the value is not exactly 44 since there are other isotopes apart from carbon 12 and oxygen 16).

In these definitions, carbon 12 has been used as the reference. However, the choice of carbon 12 is arbitrary, and oxygen 16, hydrogen 1, or any other elemental isotope could be used. This follows from Avogadro's law that a unit volume at standard temperature and pressure contains the same number of molecules, regardless of what those molecules are. For example, if oxygen 16 were used then, since

$$\frac{m_{^{16}O}}{m_{^{12}C}} = \frac{16}{12}, \tag{2.2i}$$

eqn (2.2b) would become

$$N_A m_{^{16}O} = 16 \text{ kg (kg mol)}^{-1}, \tag{2.2j}$$

and the values for the relative molecular mass for any substances would be unchanged.

It is not always appreciated that relative molecular mass is not dimensionless, but has dimensions kg (kg mol)$^{-1}$, although this is apparent from the above equations. The impression scientists generally have is that a mole is the molecular weight in grams; converted to the present units, a kilogram mole is the relative molecular mass in kilograms. This statement is a special case of eqn (2.2f) which can be rewritten

$$M = \mu Q, \tag{2.2k}$$

so that if $Q = 1$ kg mol, then M is *numerically* equal to μ.

Now consider the SI definition of the mole and the reasons why we feel it to be inadequate. The SI definition of the mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kg of carbon 12 (Royal Society 1975, p. 22). The SI definition of relative molecular mass is the ratio of the average mass per atom (molecule) of the natural isotopic composition of an element (the elements) to 1/12 of the mass of an atom of the nuclide ^{12}C (Royal Society 1975, p. 15). In this definition, it follows that the factor 1/12 has units mol g^{-1} which is equivalent to kg mol kg^{-1}. There is an obvious inconsistency here, in that the gram appears, and this is due to the SI definition of the mole. It follows that the relative molecular mass of carbon 12 is 0.012 kg (mol)$^{-1}$, which is clearly unsatisfactory. Another example is, say, oxygen, which according to this definition has a relative molecular mass of 0.032 kg (mol)$^{-1}$ (or, more accurately, 0.0319988 kg (mol)$^{-1}$), although most scientists would give the

value as 32. If the kilogram mole is used, as defined above, then the units are entirely consistent; for example, the relative molecular mass of oxygen is 31.9988 kg (kg mol)$^{-1}$. A further point to note is that, according to our set of definitions, Avogadro's number N_A is a factor of 10^3 greater than as defined by the SI system (Royal Society 1975, p. 44).

We use the term relative molecular mass, which is that recommended by the Royal Society (1975) and replaces the traditional term *molecular weight*. Relative molecular mass is preferred as it gives quite an accurate indication of the definition of the term. The term molecular weight is inappropriate and should be discarded; a weight is a force and has dimensions of newtons (N). *Molar mass* would be acceptable, although it is better to use only one term.

2.2.2 *Concentration*

The concentration C of a substance, is given in units of kilogram moles of a substance per cubic metre (kg mol m^{-3}). This is 10^3 times the SI definition. Note that the unit of 1 kg mol m^{-3} is exactly equivalent to the older unit of 1 gram molecule per litre, which can be convenient.

From the gas laws, the concentration of any gas is

$$C = \frac{P}{RT} \tag{2.3a}$$

where P is the pressure (Pa), T is the temperature (K), and R is the gas constant (8314 J K^{-1}(kg mol)$^{-1}$). Thus at normal temperature (273.15 K) and pressure (101 325.0 Pa, equivalent to 0.76 mHg), the concentration of any gas is

$$C = 0.044\,618 \text{ kg mol m}^{-3}. \tag{2.3b}$$

At arbitrary temperature and pressure, therefore, the concentration is

$$C = \frac{273.15}{T}\frac{P}{101\,325.0}0.044\,618 \text{ kg mol m}^{-3}. \tag{2.3c}$$

We also use the term concentration in another sense in crop models to define the substrate status of the plant or crop. This is discussed below.

2.2.3 *Density*

The density ρ of a substance is the number of kilograms of the substance per cubic metre (kg m^{-3}). This is exactly the same as the SI unit. The quantities density, relative molecular mass, and concentration are related by the equation

$$\text{density} = \text{relative molecular mass} \times \text{concentration}. \tag{2.3d}$$

2.2.4 *Carbon dioxide concentrations and densities*

The concentration and density of CO_2 are obtained directly from eqns (2.3c) and (2.3d). Plant scientists generally refer to CO_2 concentrations in units of parts per

million (ppm). Parts per million, defined as volume parts per million, is related to concentration and density, as defined above, by

$$C = \frac{\text{ppm}}{10^6}\frac{273.15}{T}\frac{P}{101\,325.0}0.044\,618 \text{ kg mol CO}_2 \text{ m}^{-3} \tag{2.4a}$$

and

$$\rho = \frac{\text{ppm}}{10^6}\frac{273.15}{T}\frac{P}{101\,325.0}1.963\,6 \text{ kg CO}_2 \text{ m}^{-3}, \tag{2.4b}$$

where the relative molecular mass μ_{CO_2} of CO_2 is 44.009 8 kg (kg mol)$^{-1}$.

The term parts per million has two deficiencies. First, it is not clear whether the definition is kilograms per million kilograms or molecules per million molecules, although traditionally it is taken to be the latter (as in the above equations) and this problem can, in part, be overcome by using the term volume parts per million (vpm). The second, and more serious, problem is that photosynthesis depends upon the absolute number of CO_2 molecules per unit volume, and not just the proportion of CO_2 molecules in the air. In any model, therefore, it is more appropriate to use the definition of concentration above. There are times when parts per million may be useful as it is quite easy to visualize—for example, when talking of the general increase in atmospheric CO_2 levels during this century from around 300 to 340 ppm—but in mathematical models it should be avoided.

2.2.5 *Plant composition*

The simplest means of defining plant composition is to consider plant dry matter. Although this may be quite sufficient for many purposes, it soon becomes necessary when modelling plant processes to incorporate the metabolic function of the various components within the plant. For many purposes it is helpful to separate the plant into substrate and structure. Structure comprises the cell wall material—mainly cellulose and hemicellulose—and the 'machinery' of the plant—protein. The remainder of the plant is taken to be substrate and this includes labile compounds such as glucose and amino acids. Clearly this is a simplification. However, it is the logical step from considering plant dry matter alone and, as is apparent from many of the models considered in this book, does permit considerable progress in plant and crop modelling.

The structural dry matter will generally be denoted by W_G and substrate by W_S, although other subscripts will be introduced to define, for example, shoot and root dry matter, or carbon and nitrogen substrate. A useful technique for defining the plant status with respect to a particular substrate is to look at the substrate concentration, which differs from the definition of concentration given above, and is given by

$$W_S/W_G. \tag{2.5a}$$

The dimensions of concentration in this context are kg substrate (kg structure)$^{-1}$.

2.2.6 *Water potential*

Water potential an important variable when considering the water status of a plant or crop. Any physical system will attempt to minimize its potential energy, and this is the basis for the understanding of many physical problems. For example, if a chain is suspended from two points (not necessarily at the same height), then the profile of the chain can be derived by calculating the shape required to minimize the potential energy (the curve is the well-known catenary given by the hyperbolic cosine function). In an equivalent way, in any system which involves water, the water will flow so as to minimize its potential energy. Later in the book (Chapter 14 and 15) we look at transpiration by a crop canopy and crop water use, and we need to use the concept of the water potential, which is the energy of the water in the crop, soil, or air.

An adequate definition of water potential for most plant and crop studies is the amount of energy required to transport a unit mass of pure water from a reference state to its position in the system. Water potential is therefore measured relative to some reference height, which is usually taken to be ground level. Water potential can be considered, in simple terms, as the energy per unit mass of water, and has units of joules per kilogram ($J\ kg^{-1}$).

Other units often used are the pascal (Pa) (or more commonly the kPa) and the bar, which are pressure units. The pascal is the SI unit of pressure, and has dimensions of force per unit area ($N\ m^{-2}$) which is equivalent to $J\ m^{-3}$. Consequently, if water potential is defined in pascals, this is energy per unit volume rather than per unit mass. Since a given mass of water can occupy a different volume, depending on the temperature and pressure, it is more appropriate to use $J\ kg^{-1}$. The density of water depends on temperature and takes its maximum value of $1000\ kg\ m^{-3}$ at 4 °C. Using this value, $J\ kg^{-1}$ and the pascal are related by

$$1\ J\ kg^{-1} \equiv 10^3\ Pa = 1\ kPa. \tag{2.6a}$$

The bar is not an SI unit, but is a unit of pressure in the cgs (centimetre-gram-second) system which preceded the SI system. The bar is defined as

$$1\ bar = 10^5\ Pa, \tag{2.6b}$$

so that

$$1\ J\ kg^{-1} \equiv 0.01\ bar. \tag{2.6c}$$

There are perhaps three main reasons why the bar is still retained in some quarters, despite its not being an SI unit. The first is that 1 bar is approximately equal to standard atmospheric pressure, which is 101 325 Pa = 1.013 25 bar. The second is that plant water potentials generally lie in the range 0 to −20 bar, which is an easy range to work with. The final reason is that, in practice, water potential is often measured by measuring a pressure exerted by the water in the

system. In all branches of science, quantities are measured both directly and indirectly, and there is little justification for ascribing units on this basis.

2.3 Useful responses derived from enzyme kinetics

We now consider some of the basic ideas of enzyme kinetics both as an introduction to the subject and in order to derive several useful equations for the plant and crop modeller. The equations that are derived can be used both as direct models for the reaction types considered and also in a more qualitative sense to represent aggregated processes. For example, the biochemistry of protein synthesis from sugars and amino acids is complex but, for some modelling purposes, it may be appropriate to represent this by using a bi-substrate Michaelis–Menten equation. The reaction schemes and equations considered below cover all the equations of this type that are used in this book. For a more complete discussion of the topic, including a historical perspective of its development, the interested reader should consult Dixon and Webb (1979).

2.3.1 *Michaelis–Menten equation (rectangular hyperbola)*

The most widely applied model of enzyme kinetics is the Michaelis–Menten equation. The reaction scheme is

$$E + S \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES \overset{k_{+2}}{\rightarrow} E + P, \tag{2.7a}$$

where E, S, and P indicate the enzyme, substrate, and products of the reaction respectively; k_{+1}, k_{-1}, and k_{+2} are the rate constants for the reactions, where the plus and minus signs refer to forward and reverse reactions. ES denotes the enzyme-substrate complex. In the steady state the concentration of ES is constant, so that the rate of production of ES must equal the rate of degradation of ES:

$$k_{+1}[E][S] = (k_{-1} + k_{+2})[ES] \tag{2.7b}$$

where the square brackets denote concentrations. If E_0 is the total concentration of the enzyme present, which does not vary with time, then

$$E_0 = [E] + [ES] \tag{2.7c}$$

(that is, the total number of kilogram moles of enzyme is constant). Combining eqns (2.7b) and (2.7c) and rearranging leads to

$$[ES] = \frac{k_{+1}[S]E_0}{k_{+1}[S] + k_{-1} + k_{+2}}. \tag{2.7d}$$

This gives the concentration of substrate molecules which are combined with (adsorbed onto the surface of) the enzyme molecules in terms of the substrate concentration [S] and the total enzyme concentration E_0. A similar equation regarding the adsorption of gas molecules onto solid surfaces, which is known

as Langmuir's isotherm, was obtained many years prior to the work of Michaelis and Menten.

If the speed of the steady-state reaction is v, then

$$v = k_{+2}[ES] = \frac{k_{+1}k_{+2}[S]E_0}{k_{+1}[S] + k_{-1} + k_{+2}}, \tag{2.7e}$$

which can be written

$$v = \frac{v_m[S]}{K + [S]} \tag{2.7f}$$

where

$$v_m = k_{+2}E_0 \qquad \text{and} \qquad K = \frac{k_{-1} + k_{+2}}{k_{+1}}. \tag{2.7g}$$

v_m is the maximum speed of the reaction which occurs when all the active sites on the enzyme molecules are occupied by substrate molecules. K is known as the Michaelis–Menten constant, and is the value of the substrate concentration for half-maximal speed $v = \frac{1}{2}v_m$. Equation (2.7f) is illustrated in Fig. 2.1. If $k_{+2} \ll k_{-1}$, then $K \approx k_{-1}/k_{+1}$ and the first stage of the reaction is virtually in equilibrium. In this case eqn (2.7d) becomes

$$\frac{[ES]}{E_0} = \frac{[S]}{[S] + K}, \tag{2.7h}$$

and K can be regarded as a binding constant (or dissociation constant) in that it indicates the proportion of enzyme E that is bound to the substrate S; thus a low value of K means that the enzyme has a high affinity for the substrate, and the reaction saturates at low concentrations of substrate.

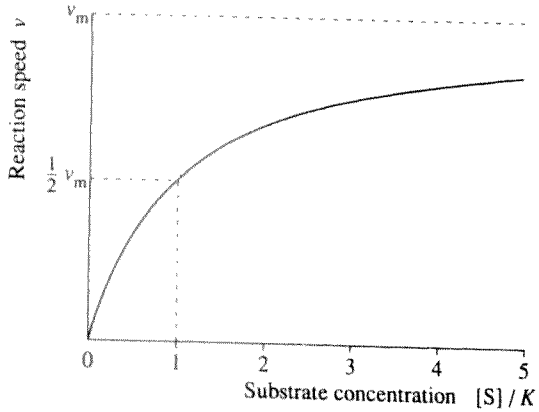


Fig. 2.1. The Michaelis–Menten equation (2.7f) for the reaction speed v as a function of substrate concentration [S]. There is an asymptote at $v = v_m$, and $v = \frac{1}{2}v_m$ when $[S] = K$.

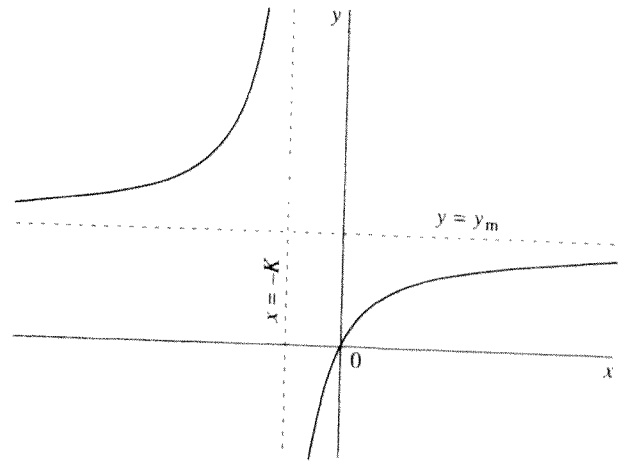


Fig. 2.2. The rectangular hyperbola (eqn (2.7i)). The asymptotes $y = y_m$ and $x = -K$ are shown.

Equation (2.7f) is known as a rectangular hyperbola. The name is apparent from the illustration in Fig. 2.2 where the equation

$$y = \frac{y_m x}{K + x} \tag{2.7i}$$

is presented. There are two branches to the curve, both approaching asymptotes defined by the lines

$$y = y_m \qquad \text{and} \qquad x = -K. \tag{2.7j}$$

These two asymptotes are at right angles to each other, and this is the basis for the name rectangular hyperbola. Equation (2.7f) is a particular case of the curve for the range $x > 0$.

An alternative formulation for the rectangular hyperbola is

$$y = \frac{\alpha x y_m}{\alpha x + y_m}, \tag{2.7k}$$

where it is readily shown that the constant α is the initial slope of the curve:

$$\frac{dy}{dx}(x = 0) = \alpha. \tag{2.7l}$$

The two forms presented for the rectangular hyperbola (eqns (2.7f) and (2.7k)) are mathematically equivalent in that K and α are related by

$$K = y_m/\alpha. \tag{2.7m}$$

The choice of which form to use will depend on the nature of the problem. For example, if it is convenient to prescribe the substrate level at which the reaction speed is half its maximum, then eqn (2.7f) is appropriate, whereas if it is more appropriate to incorporate the reaction rate at low substrate levels, then eqn (2.7k) might be preferred. Both forms are used throughout this book.

2.3.2 *Threshold response equations*

The rectangular hyperbola does not have a point of inflexion. However, many biological processes show this type of behaviour. An equation of this form can be derived by modifying the Michaelis–Menten reaction scheme (eqn (2.7a)) to

$$E + 2S \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES_2 \overset{k_{+2}}{\rightarrow} E + P. \tag{2.8a}$$

In this case molecules of substrate S can combine with the enzyme E at two sites. Using the approach of the previous section, it can be shown that the speed of this reaction is given by (Exercise 2.1)

$$v = \frac{v_m[S]^2}{K^2 + [S]^2} \tag{2.8b}$$

where v_m and K^2 are given by eqns (2.7g). This equation is analogous to eqn (2.7f) and is illustrated in Fig. 2.3: the initial slope is zero, the asymptote is $v = v_m$, $v = \frac{1}{2}v_m$ when $S = K$, and there is a point of inflexion at $S/K = 1/\sqrt{3}$ (Exercise 2.1).

Equation (2.8a) can be generalized to

$$E + nS \underset{k_{-1}}{\overset{k_{+1}}{\rightleftharpoons}} ES_n \overset{k_{+2}}{\rightarrow} E + P \tag{2.8c}$$

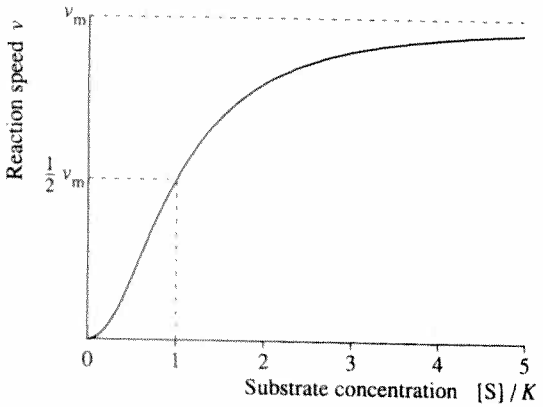


Fig. 2.3. The sigmoidal response curve (eqn (2.8b)) for the reaction speed v as a function of substrate concentration [S]. As for the Michaelis–Menten equation, there is an asymptote at $v = v_m$, and $v = \frac{1}{2}v_m$ when $[S] = K$.

where n is a positive integer. According to this scheme, there are n active sites per enzyme molecule for the substrate. Proceeding as above generates the family of curves given by (Exercise 2.2)

$$v = \frac{v_m[S]^n}{K^n + [S]^n}. \tag{2.8d}$$

Again, v_m and K^2 are given by eqns (2.7g) and $v = \frac{1}{2}v_m$ when $S = K$. The equation for v is illustrated in Fig. 2.4 for several values of n: when $n = 1$, the curve is the Michaelis–Menten rectangular hyperbola (eqn (2.7f)) as illustrated in Fig. 2.1 and there is no point of inflexion; for $n = 2$, the curve is the same as that illustrated in Fig. 2.3 and there is weak sigmoid behaviour; as n increases the sigmoid behaviour becomes more and more pronounced, until in the limit $n \to \infty$ a step function is obtained. For all values of n ($n \geqslant 1$) the asymptote is v_m. For $n \geqslant 2$ the initial slope is zero, and there is a point of inflexion at (Exercise 2.2)

$$\frac{[S]}{K} = \left(\frac{n-1}{n+1}\right)^{1/n}. \tag{2.8e}$$

Although eqn (2.8d) has been derived from the reaction of eqn (2.8c), for some applications it can be regarded as being a useful empirical equation and there is no reason why n need be integral.

Equation (2.8d) and Fig. 2.4 show a type of 'switch-on' behaviour where the sharpness of the switching characteristic depends on the value of n. A similar relation can be constructed to represent 'switch-off' behaviour, and has the form

$$v = \frac{v_m K^n}{K^n + [S]^n}. \tag{2.8f}$$

This equation is illustrated in Fig. 2.5. v_m now occurs at $[S] = 0$, and v

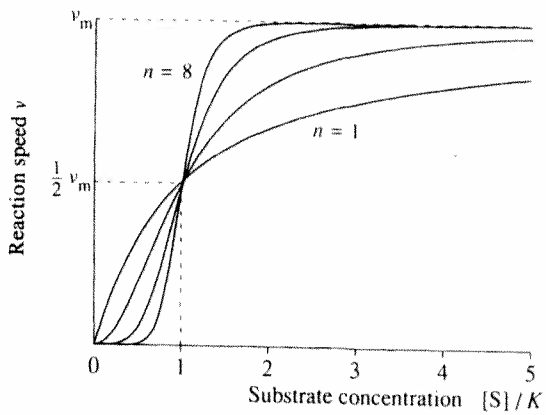


Fig. 2.4. Sigmoidal 'switch-on' response (eqn (2.8d)): $n = 1$ and $n = 8$ are indicated, and the intermediate curves are for $n = 2$ and $n = 4$. All curves have an asymptote at $v = v_m$, and $v = \frac{1}{2}v_m$ when $[S] = K$.

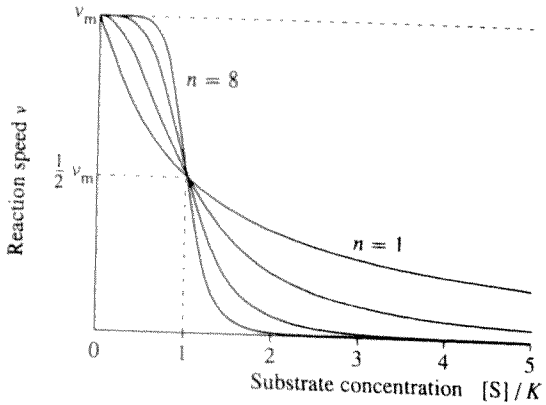


Fig. 2.5. Sigmoidal 'switch-off' response (eqn (2.8f)): $n = 1$ and $n = 8$ are indicated, and the intermediate curves are for $n = 2$ and $n = 4$. For all curves $v = v_m$ when $[S] = 0$, and $v = \frac{1}{2} v_m$ when $[S] = K$.

approaches zero as [S] increases for all values of n ($n \geqslant 1$). The point of inflexion is still given by eqn (2.8e) (Exercise 2.2).

2.3.3 *Non-rectangular hyperbola*

The non-rectangular hyperbola is a useful generalization of the rectangular hyperbola as it is a more versatile curve. One way of deriving it is to assume that there is an enzyme–substrate reaction which can be described by a rectangular hyperbola, but that now the substrate has to diffuse across some boundary from an external region to the site of the reaction. In this case, the aim is to derive an expression for the speed of the reaction in terms of the external substrate concentration. As an example, this type of scheme might be regarded as representing a reaction that takes place within the root system of a plant where the substrate has to diffuse across the root cell membranes. The speed of the reaction is

$$v = \frac{v_m[S_i]}{K + [S_i]}, \tag{2.9a}$$

where the subscript i indicates the internal substrate. It is assumed that the substrate diffuses across some boundary to the site of the reaction, and that the system is in the steady state. This means that $[S_i]$ is constant so that the rate of utilization of substrate at the site of the reaction must equal the rate of diffusion across the boundary, and hence

$$v = \frac{[S] - [S_i]}{r}, \tag{2.9b}$$

where [S] is the external substrate concentration and r is a resistance. Using eqns

(2.9a) and (2.9b) to eliminate $[S_i]$ lead to the quadratic equation

$$v^2 - v\left(\frac{K + [S]}{r} + v_m\right) + \frac{v_m[S]}{r} = 0, \tag{2.9c}$$

which can be factorized to give

$$\left(v - \frac{[S] + K}{r}\right)(v - v_m) = \frac{Kv_m}{r}. \tag{2.9d}$$

This defines a pair of curves with asymptotes

$$v = \frac{[S] + K}{r} \quad \text{and} \quad v = v_m. \tag{2.9e}$$

[S] is only physiologically defined for $[S] \geqslant 0$ and v in the range $0 \leqslant v \leqslant v_m$. However, it is instructive to look at the solutions to eqn (2.9d) over the whole range of values of [S] and v, and so the curves given by eqn (2.9d) are presented in Fig. 2.6 where the full curve represents the physiologically realistic part of the solution while the broken curves denote the unrealistic solutions.

The non-rectangular hyperbola can be written in other ways. The formulation given by eqns (2.9c) and (2.9d) is useful because it has been derived from a simple model, although it may not always be the most convenient. An alternative is to write (replace [S] by x, v by y, v_m by y_m, $1/r$ by α/θ, and K by $y_m(1 - \theta)/\alpha$)

$$\theta y^2 - (\alpha x + y_m)y + \alpha x y_m = 0. \tag{2.9f}$$

For $\theta = 0$ this reduces to a simple rectangular hyperbola given by

$$y = \frac{\alpha x y_m}{\alpha x + y_m}, \tag{2.9g}$$

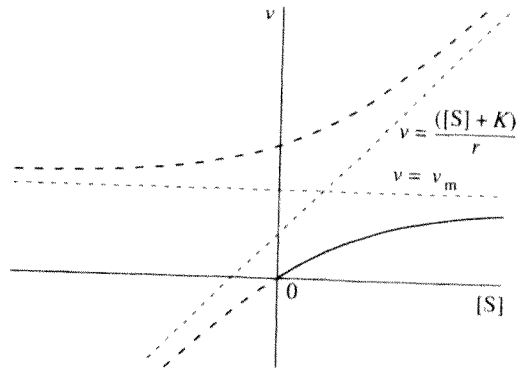


Fig. 2.6. The non-rectangular hyperbola given by the solutions to the quadratic equation (2.9c). The asymptotes $v = v_m$ and $v = ([S] + K)/r$ are indicated. The broken curves indicate the physiologically unrealistic solutions.

and for $\theta = 1$ it factorizes to

$$(y - \alpha x)(y - y_m) = 0 \tag{2.9h}$$

so that y is given by the two straight lines

$$y = \begin{cases} \alpha x & x \leqslant y_m/\alpha \\ y_m & x > y_m/\alpha. \end{cases} \tag{2.9i}$$

For θ lying in the range $0 < \theta < 1$, the physiologically realistic solution for y is the lower root of eqn (2.9f) which is

$$y = \frac{1}{2\theta}\{\alpha x + y_m - [(\alpha x + y_m)^2 - 4\theta\alpha x y_m]^{1/2}\}. \tag{2.9j}$$

It can be shown that in the limit $\theta \to 0$ eqn (2.9j) reduces to eqn (2.9g) and for $\theta = 1$ it becomes eqn (2.9i) (Exercise 2.3). For all values of θ in the physiologically sensible range $0 \leqslant \theta \leqslant 1$, the initial slope of the curve is

$$\frac{dy}{dx}(x = 0) = \alpha \tag{2.9k}$$

and the asymptote is (Exercise 2.3)

$$y(x \to \infty) = y_m. \tag{2.9l}$$

The non-rectangular hyperbola for several values of θ in the range $0 \leqslant \theta \leqslant 1$ is illustrated in Fig. 2.7.

The extra parameter θ in the non-rectangular hyperbola gives more control over the response than is the case for the rectangular hyperbola. In general, when looking at responses which increase without a point of inflexion to an asymptote, there are three basic features of the curve. The first is the initial slope, the second

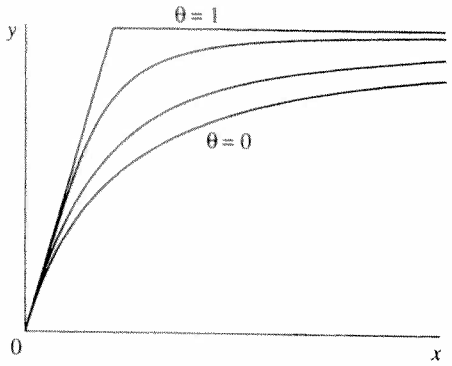


Fig. 2.7. The non-rectangular hyperbola (eqn (2.9j)): $\theta = 0$ and $\theta = 1$ are indicated and the intermediate curves are for $\theta = 0.5$ and $\theta = 0.9$. When $\theta = 0$ eqn (2.9j) reduces to the rectangular hyperbola (eqn (2.9g)) and for $\theta = 1$ it becomes two straight lines (eqn (2.9i)).

is the asymptote, and the third is the sharpness of the response i.e. how rapidly it approaches the asymptote. With the two-parameter rectangular hyperbola there is only control over two of these features. With the non-rectangular hyperbola, however, the extra parameter gives control over all three aspects of the response, which is often useful.

2.3.4 *Bi-substrate Michaelis–Menten equation*

Consider now a reaction depending on two substrates. We make the simplifying assumption that all reactions are in equilibrium, and that the order of the reactions is random. The system can be represented by the reaction scheme

$$\mathrm{E} + \mathrm{A} \rightleftharpoons \mathrm{EA} \tag{2.10a}$$

$$\mathrm{E} + \mathrm{B} \rightleftharpoons \mathrm{EB} \tag{2.10b}$$

$$\mathrm{EA} + \mathrm{B} \rightleftharpoons \mathrm{EAB} \tag{2.10c}$$

$$\mathrm{EB} + \mathrm{A} \rightleftharpoons \mathrm{EAB} \tag{2.10d}$$

where A and B are the substrates. The equilibrium equations corresponding to these reactions are

$$[\mathrm{E}][\mathrm{A}] = K_\mathrm{A}[\mathrm{EA}] \tag{2.10e}$$

$$[\mathrm{E}][\mathrm{B}] = K_\mathrm{B}[\mathrm{EB}] \tag{2.10f}$$

$$[\mathrm{EA}][\mathrm{B}] = K'_\mathrm{B}[\mathrm{EAB}] \tag{2.10g}$$

$$[\mathrm{EB}][\mathrm{A}] = K'_\mathrm{A}[\mathrm{EAB}]. \tag{2.10h}$$

K_A, K_B, K'_B, and K'_A are the binding constants for the equilibrium reactions. It is assumed that the product of the scheme is formed by the reaction

$$\mathrm{EAB} \xrightarrow{k} \mathrm{E} + \mathrm{P}, \tag{2.10i}$$

so that the speed of formation of P is

$$v = k[\mathrm{EAB}]. \tag{2.10j}$$

The aim is to derive an expression for v in terms of [A] and [B]. Equations (2.10e)–(2.10h) must therefore be used to eliminate [E], [EA], and [EB] to give [EAB] from which v is given by eqn (2.10j). However, since the total concentration of enzyme is constant

$$E_0 = [\mathrm{E}] + [\mathrm{EA}] + [\mathrm{EB}] + [\mathrm{EAB}], \tag{2.10k}$$

there are five equations for the four unknowns [E], [EA], [EB], and [EAB]. For the equations to be consistent, it can be shown that (Exercise 2.4)

$$K_\mathrm{A}K'_\mathrm{B} = K_\mathrm{B}K'_\mathrm{A} = K_\mathrm{AB}, \text{ say.} \tag{2.10l}$$

[EAB] can now be derived (Exercise 2.4) and v is given by

$$v = \frac{v_m}{1 + K'_B/[B] + K'_A/[A] + K_{AB}/[A][B]}, \tag{2.10m}$$

where

$$v_m = kE_0. \tag{2.10n}$$

Equation (2.10m) describes the speed of the reaction in terms of the substrate levels. If the concentration of either substrate is kept constant, then the response to variation in the other substrate is a rectangular hyperbola. For example, if v is plotted as a function of [A], the initial slope of the curve is

$$\frac{v_m B}{K'_A[B] + K_{AB}} \tag{2.10o}$$

and the asymptote is

$$\frac{v_m[B]}{[B] + K'_B}. \tag{2.10p}$$

Similar expressions apply when v is expressed in terms of [B].

For the special case where each substrate can only combine with the enzyme at a specific site on the enzyme and this is unaffected by the other substrate,

$$K_A = K'_A \quad \text{and} \quad K_B = K'_B, \tag{2.10q}$$

and eqn (2.10m) (using eqn (2.10l)) simplifies to

$$v = \frac{v_m}{(1 + K_A/[A])(1 + K_B/[B])}. \tag{2.10r}$$

This type of equation is very useful for representing processes that depend on the supply of two substrates, and is simpler than eqn (2.10m).

2.3.5 *Inhibitors*

The final equations to be considered are the generalization of the Michaelis–Menten theory, which led to the rectangular hyperbola eqn (2.7f), to include inhibitors. Inhibitors which manifest themselves as increasing the K parameter without altering v_m in the Michaelis–Menten equation are termed competitive; those which do not influence K but result in a decrease in v_m are known as non-competitive. There are several other forms of inhibitors which involve combinations of these effects, and these are discussed by Dixon and Webb (1979). We restrict attention to the two simplest forms known as fully competitive and fully non-competitive inhibitors. The two equations derived below (eqns (2.11h) and (2.12j)) provide useful means of representing reactions where a substrate is required but the reaction is inhibited by other components. As for the bi-substrate Michaelis–Menten equation, it is assumed that all reactions are in equilibrium.

Fully competitive inhibitor The system is represented by the reaction schemes

$$E + S \rightleftharpoons ES, \tag{2.11a}$$

$$E + I \rightleftharpoons EI, \tag{2.11b}$$

where S is the substrate and I the inhibitor. It is apparent from this reaction scheme that if an enzyme molecule combines with a molecule of the inhibitor it is completely unavailable to the substrate. The equilibrium equations for these reactions are

$$[E][S] = K_S[ES] \tag{2.11c}$$

and

$$[E][I] = K_I[EI], \tag{2.11d}$$

where K_S and K_I are the binding constants for the equilibrium reactions. As for the previous schemes, the product P of the reaction is formed by

$$ES \xrightarrow{k} E + P, \tag{2.11e}$$

and hence the speed of formation of P is

$$v = k[ES]. \tag{2.11f}$$

Again it is assumed that the total concentration of enzyme E_0, is constant, so that

$$E_0 = [E] + [ES] + [EI]. \tag{2.11g}$$

Eliminating [E] and [EI] from eqns (2.11c), (2.11d), and (2.11g) to obtain [ES] and substituting in eqn (2.11f) gives (Exercise 2.5)

$$v = \frac{v_m}{1 + (K_S/[S])(1 + [I]/K_I)}, \tag{2.11h}$$

where v_m, given by

$$v_m = kE_0, \tag{2.11i}$$

is the maximum speed of the reaction and occurs in the limit

$$\frac{K_S}{[S]}\left(1 + \frac{[I]}{K_S}\right) \to 0 \tag{2.11j}$$

which requires

$$\frac{[S]}{K_S} \gg 1 + \frac{[I]}{K_I}. \tag{2.11k}$$

This implies that the substrate is non-limiting and is in such supply that it occupies virtually all the sites on the enzyme, thus preventing any effect of the inhibitor. It is clear, therefore, that eqn (2.11h) corresponds to the Michaelis–Menten equation but with

$$K = K_S\left(1 + \frac{[I]}{K_I}\right). \tag{2.11l}$$

Fully non-competitive inhibitor For the fully competitive inhibitor, we saw that the enzyme could combine with either the substrate or the inhibitor. If it combined with the latter, then it was completely unavailable to the substrate for the formation of products. The type of inhibitor known as fully non-competitive is one where an intermediate complex [ESI] may be formed, and where the affinity of enzyme for either the substrate or the inhibitor is unaffected by the presence of the other. The reaction scheme is now

$$E + S \rightleftharpoons ES \tag{2.12a}$$

$$E + I \rightleftharpoons EI \tag{2.12b}$$

$$ES + I \rightleftharpoons ESI \tag{2.12c}$$

$$EI + S \rightleftharpoons ESI. \tag{2.12d}$$

The equilibrium equations are

$$[E][S] = K_S[ES] \tag{2.12e}$$

$$[E][I] = K_I[EI] \tag{2.12f}$$

$$[EI][S] = K_S[ESI] \tag{2.12g}$$

$$[ES][I] = K_I[ESI], \tag{2.12h}$$

where there are only two binding constants since, as stated above, the affinity of the enzyme for either S or I is independent of the presence of the other. As for the fully competitive inhibitor in the previous section, the formation of the product P of the reaction is given by eqn (2.11e), so that eqn (2.11f) defines the speed of formation of P. Equations (2.12e)–(2.12h) must now be used to eliminate [EI] and [ESI] in order to obtain [ES] and hence the reaction speed v. Although there are four equations and only three unknowns, the equations are consistent since dividing either eqn (2.12e) by eqn (2.12f) or eqn (2.12g) by eqn (2.12h) gives

$$\frac{[S]}{[I]} = \frac{K_S\,[ES]}{K_I\,[EI]}. \tag{2.12i}$$

Proceeding as above, it can be shown (Exercise 2.5) that the reaction speed v is given by

$$v = \frac{v_m}{(1 + K_S/[S])(1 + [I]/K_I)}. \tag{2.12j}$$

The effect of the inhibitor in this case is to reduce the speed of the reaction by a factor of $1 + [I]/K_I$, which is unaffected by the level of [S]. Also,

$$\text{as } [S] \to \infty,\ v \to v_m^* = \frac{v_m}{1 + [I]/K_I}, \tag{2.12k}$$

where v_m^* is the maximum speed of the reaction when the inhibitor is present. K_S is now equivalent to K in eqn (2.7f) for the Michaelis–Menten equation in that

$$\text{when } [S] = K,\ v = \tfrac{1}{2}v_m^*, \tag{2.12l}$$

so that eqn (2.12j) is equivalent to the Michaelis–Menten equation (2.7f) with v_m replaced by v_m^*.

2.4 Cell division and organ growth

We now turn our attention to the growth of an organ in terms of the cell division within that organ. Since this book is primarily concerned with processes at the plant and crop levels in the hierarchical structure discussed in Chapter 1, this topic is not of central importance, but it is useful to have some understanding of this lower level in the hierarchy. We restrict attention to the dynamics of cell division and do not consider cell growth in itself. However, it is important to remember that cell division and growth are different processes, which may occur separately or together. For further discussion of cell growth see Thornley (1981).

2.4.1 *Cell division of a purely meristematic culture*

Consider a culture comprising only meristematic (dividing) cells. It is assumed that the cells divide by binary fission, where τ_d is the time interval between divisions, so that the maximum age of any cell is τ_d, and this is taken to be constant for all cells. This assumption that τ_d is constant for all cells can be relaxed, and the theory has been presented by Powell (1956). Let the equilibrium age distribution of the cells be

$$\phi(\tau),\ \text{where } 0 \leqslant \tau \leqslant \tau_d \text{ and } \int_0^{\tau_d} \phi(\tau)\,d\tau = 1, \tag{2.13a}$$

where ϕ has units of time^{-1}. It is assumed that this is independent of time so that for any time t the proportion of cells of a given age will be constant. If M is the cell number at time t, then the number of cells with ages lying in the interval τ to $\tau + d\tau$ is

$$M(t)\phi(\tau)\,d\tau. \tag{2.13b}$$

The number of cells that divide in a time interval dt is

$$M(t)\phi(\tau_d)\,dt. \tag{2.13c}$$

With binary fission and assuming that all the progeny cells are meristematic, the increment in cell number is

$$dM = M\phi(\tau_d)\,dt. \tag{2.13d}$$

If we define the cell number growth constant ν (time^{-1}) by

$$\nu = \phi(\tau_d), \tag{2.13e}$$

then

$$\frac{dM}{dt} = \nu M \tag{2.13f}$$

and

$$M = M_0 e^{\nu t} \tag{2.13g}$$

where M_0 is the cell number at time $t = 0$. This is the usual equation for exponential growth of cell numbers. Now, since τ_d is the time between cell divisions, exactly all the cells present at time t will have doubled by time $t + \tau_d$, so that

$$M(t + \tau_d) = 2M(t) \tag{2.13h}$$

which, substituted in eqn (2.13g), leads to

$$\nu = \frac{\ln 2}{\tau_d}. \tag{2.13i}$$

Equations (2.13g) and (2.13i) define the cell population $M(t)$ in terms of the initial cell number M_0 and the time interval τ_d between cell divisions. $M(t)$ is illustrated in Fig. 2.8.

Now consider the age distribution function $\phi(\tau)$ (eqn (2.13a)). The number of cells at time t of age τ per unit age interval is

$$M(t)\phi(\tau). \tag{2.13j}$$

After an incremental increase $d\tau$ in time, the number of cells of age $\tau + d\tau$ per

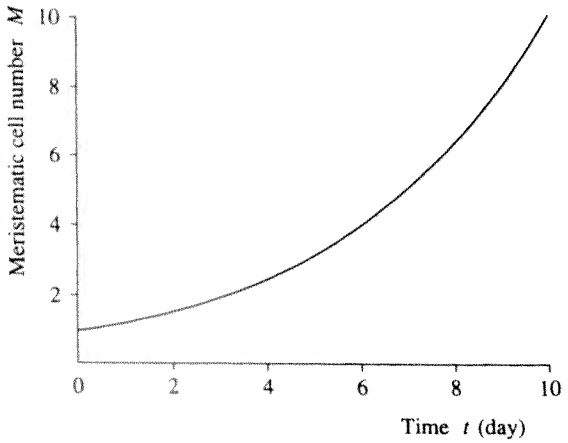


Fig. 2.8. Meristematic cell number $M(t)$ for exponential growth (eqns (2.13g) and (2.13i)). The initial value M_0 is arbitrarily taken to be unity, and $\tau_d = 3$ days (so that $\nu = 0.23$ day^{-1}).

unit age interval is

$$M(t + d\tau)\phi(\tau + d\tau). \tag{2.13k}$$

However, the cells of age τ at time t are exactly the same cells of age $\tau + d\tau$ at time $t + d\tau$, so that (2.13j) and (2.13k) are equal, i.e.

$$M(t)\phi(\tau) = M(t + d\tau)\phi(\tau + d\tau). \tag{2.13l}$$

To proceed, we use the Taylor series which is defined as follows:

$$f(x + h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) \ldots, \tag{2.13m}$$

where the primes denote derivatives with respect to x and $n! = n(n-1)(n-2)\ldots 3.2.1$, which is referred to as n factorial. Expanding the terms on the right-hand side of eqn (2.13l) using the Taylor series gives, neglecting terms of order $d\tau^2$ and higher,

$$M(t)\phi(\tau) = \left[M(t) + d\tau\frac{dM}{dt}(t)\right]\left[\phi(\tau) + d\tau\frac{d\phi}{d\tau}(\tau)\right], \tag{2.13n}$$

which reduces to (again neglecting the terms in $d\tau^2$)

$$\frac{d\phi}{d\tau} = -\phi\frac{1}{M}\frac{dM}{dt}. \tag{2.13o}$$

Substituting from eqn (2.13g) and integrating gives

$$\phi(\tau) = \phi_0 e^{-\nu\tau}, \tag{2.13p}$$

where ϕ_0 denotes $\phi(\tau = 0)$. From this equation it can be seen that young cells predominate. Indeed, setting $\tau = \tau_d$ and using eqn (2.13i), the physiologically obvious result

$$\phi(\tau_d) = \tfrac{1}{2}\phi_0 \tag{2.13q}$$

is obtained; that is, the number of cells about to divide is half the number that have just divided. ϕ_0 is derived by noting that

$$M(t) = \int_0^{\tau_d} M(t)\phi(\tau)\,d\tau \tag{2.13r}$$

and hence

$$\int_0^{\tau_d} \phi(\tau)\,d\tau = 1, \tag{2.13s}$$

from which it immediately follows that

$$\phi_0 = 2\nu. \tag{2.13t}$$

$\phi(\tau)$ is illustrated in Fig. 2.9.

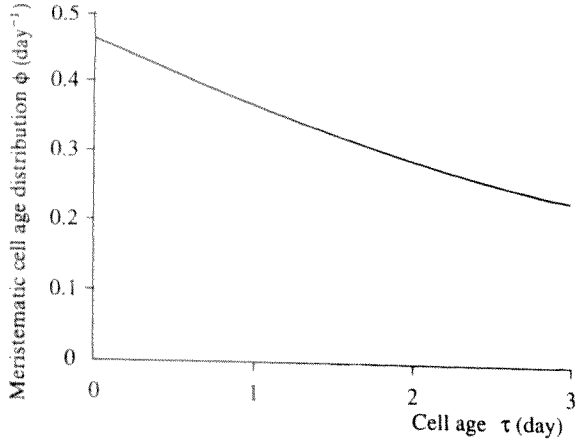


Fig. 2.9. Age distribution $\phi(\tau)$ of meristematic cells, corresponding to Fig. 2.8, and given by eqns (2.13p) and (2.13t). Note that $\phi(\tau_d) = \frac{1}{2}\phi_0$ (recall that $\tau_d = 3$ days).

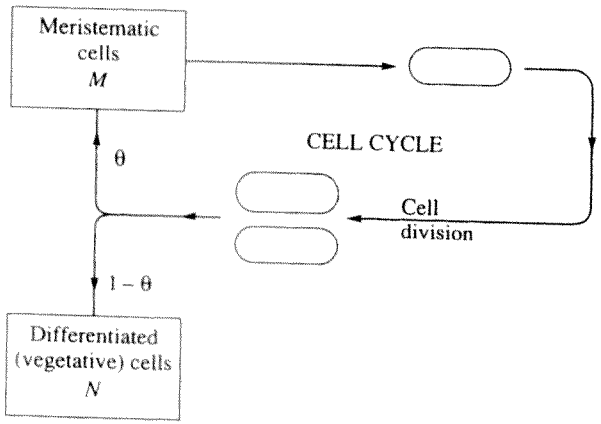


Fig. 2.10. Scheme for meristematic cell division and the production of differentiating cells.

2.4.2 *Cell division and its cessation*

Consider now the case where only a proportion θ, $0 \leqslant \theta \leqslant 1$, of newly divided cells continues to divide and the remainder are removed from the meristematic system. (This is a different θ from that of (2.9f).) This could arise for several reasons. Some of the newly divided cells may simply not be viable. Alternatively, some of these cells may be committed to a vegetative pathway of development which excludes cell division. This is likely to follow a period of purely meristematic growth. For convenience, we shall refer to non-meristematic cells as differentiating cells. This scheme, adapted from Thornley (1981), is illustrated in Fig. 2.10. The differential equation for M is now

$$\frac{dM}{dt} = \lambda v M, \tag{2.14a}$$

where the parameter λ accounts for the fact that not all cells continue to divide. λ is constrained by

$$\lambda \leqslant 1 \tag{2.14b}$$

since M cannot exceed the value for a purely meristematic culture. For constant λ, which corresponds to a constant value of θ, eqn (2.14a) integrates to

$$M(t) = M_0 e^{\lambda v t}. \tag{2.14c}$$

This equation, which is analogous to eqn (2.13g), involves the new parameter λ. To derive λ it is necessary to consider the age distribution function $\phi(\tau)$. Integrating eqn (2.13o) for $\phi(\tau)$, combined with eqn (2.14c), gives

$$\phi(\tau) = \phi_0 e^{-\lambda v \tau} \tag{2.14d}$$

for the age distribution of the cells, which is analogous to eqn (2.13p). Clearly, ϕ_0 will be affected by the proportion of cells that become non-meristematic, and so both λ and ϕ_0 must now be derived.

Applying the constraint (2.13s), ϕ_0 can be calculated as (Exercise 2.6)

$$\phi_0 = \frac{\lambda v}{1 - e^{-\lambda \ln 2}}. \tag{2.14e}$$

λ is derived as follows. Meristematic cells of age τ_d are continually dividing and producing cells of age zero. When the cells divide, a fraction θ of these newly formed cells remain meristematic, while the remainder are no longer meristematic. Thus

$$\phi_0 = 2\theta\phi(\tau_d), \tag{2.14f}$$

from which it can be shown that (Exercise 2.6)

$$\lambda = 1 + \frac{\ln \theta}{\ln 2}. \tag{2.14g}$$

The total number M of meristematic cells is therefore (eqn (2.14c))

$$M(t) = M_0 \exp\left[v\left(1 + \frac{\ln \theta}{\ln 2}\right)t\right], \tag{2.14h}$$

and the age distribution is

$$\phi(\tau) = \phi_0 \exp\left[-v\left(1 + \frac{\ln \theta}{\ln 2}\right)\tau\right]. \tag{2.14i}$$

$M(t)$ and $\phi(\tau)$ are illustrated in Fig. 2.11. For $\theta = 1$ the population is purely meristematic and the solutions for $M(t)$ and $\phi(\tau)$ are identical with those of the previous section. If $\frac{1}{2} < \theta < 1$, M is an increasing function and young cells predominate. When $\theta = \frac{1}{2}$, $\lambda = 0$ so that $M(t) = M_0$ and

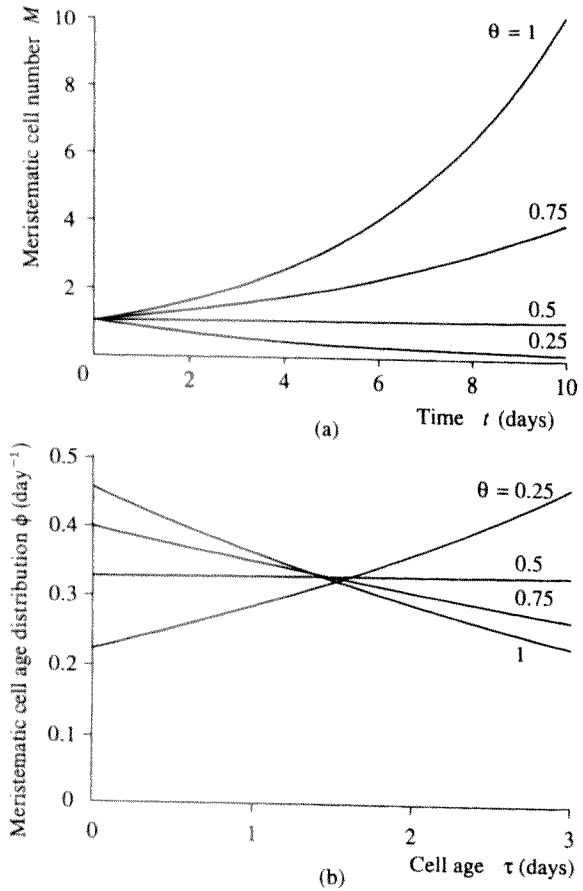


Fig. 2.11. (a) Meristematic cell number $M(t)$ as given by eqn (2.14h) for θ values of 1, 0.75, 0.5, and 0.25 as indicated ($M_0 = 1$ and $\tau_d = 3$ days, so that $\nu = 0.23$ day^{-1}); (b) corresponding age distribution function $\phi(\tau)$ (eqn (2.14i)) (note that when $\theta = 0.5$, $\lambda = 0$, and $\phi = \phi_0 = 1/\tau_d$ is constant).

means that the meristematic cell population and its age distribution are constant. For $0 < \theta < \frac{1}{2}$, $\lambda < 0$ and in this case $M(t)$ is a decreasing function with old cells predominating.

It should be noted that eqn (2.14e) for ϕ_0 is indeterminate when $\lambda = 0$. However, in this case $\phi(\tau)$ can be derived directly from the differential equation (2.13o) to be

$$\phi(\tau) = \phi_0, \tag{2.14j}$$

and from the constraint (2.13s), ϕ_0 is readily shown to be

$$\phi_0 = 1/\tau_d. \tag{2.14k}$$

It can be shown (Exercise 2.6) from eqn (2.14e) that

$$\lim_{\lambda \to 0} \phi_0 = 1/\tau_d, \tag{2.14l}$$

which is consistent with eqn (2.14k).

So far θ has been taken to be constant, although this will not be the case in a determined organ, such as a leaf, where meristematic activity eventually ceases. Following Thornley (1981) it is simply assumed that the time course of θ is given by

$$\frac{d\theta}{dt} = -D\theta, \tag{2.15a}$$

and taking D (which has dimensions of time^{-1}) to be constant, this has solution

$$\theta = e^{-Dt}. \tag{2.15b}$$

Substituting in eqn (2.14g) for λ gives

$$\lambda = 1 - \frac{Dt}{\ln 2}, \tag{2.15c}$$

and hence, integrating eqns (2.14a) and (2.13o), $M(t)$ and $\phi(t, \tau)$ become

$$M(t) = M_0 \exp\left[\nu\left(1 - \frac{Dt}{2 \ln 2}\right)t\right] \tag{2.15d}$$

and

$$\phi(t, \tau) = \phi_0 \exp\left[-\nu\left(1 - \frac{Dt}{\ln 2}\right)\tau\right]. \tag{2.15e}$$

Equation (2.15d) for $M(t)$ is generally referred to as an exponential quadratic function. It is symmetric about a maximum value given by

$$M(t_m) = M_0 \exp(\tfrac{1}{2}\nu t_m) \tag{2.15f}$$

which occurs at time

$$t_m = \frac{\ln 2}{D}. \tag{2.15g}$$

It is convenient to express $M(t)$ and $\phi(\tau)$ in terms of t_m rather than D, in which case eqns (2.15d) and (2.15e) become

$$M(t) = M_0 \exp\left[\nu\left(1 - \frac{t}{2t_m}\right)t\right] \tag{2.15h}$$

and

$$\phi(t,\tau) = \phi_0 \exp\left[-\nu\left(1-\frac{t}{t_m}\right)\tau\right]. \tag{2.15i}$$

Note that ϕ_0 is still defined by eqn (2.14e) but is now a function of time t, through the dependence of λ on t in eqn (2.15c), and is given by

$$\phi_0 = \frac{(1-t/t_m)\nu}{1-\frac{1}{2}\exp[(t/t_m)\ln 2]}. \tag{2.15j}$$

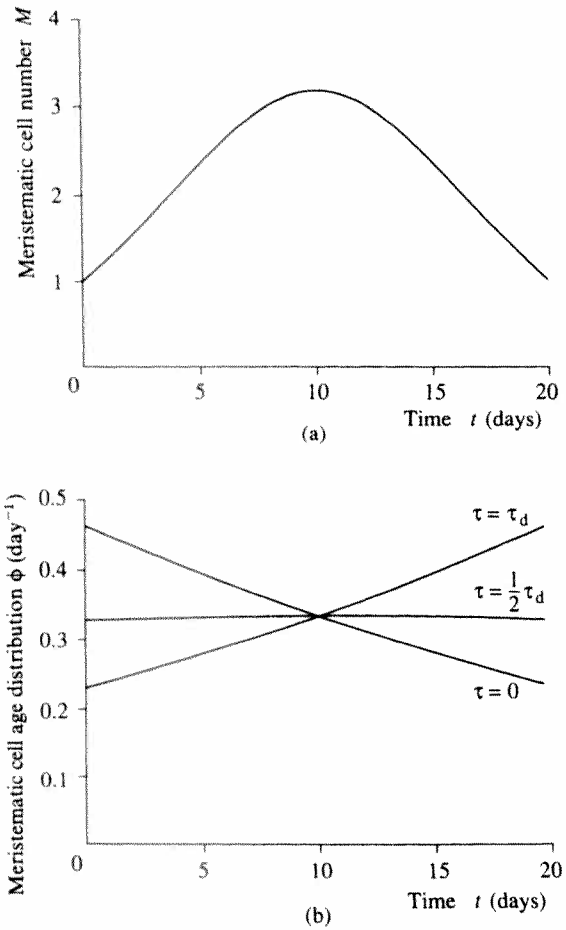


Fig. 2.12. (a) Meristematic cell number $M(t)$ as given by eqn (2.15h) with $M_0 = 1$ and $\tau_d = 3$ days, so that $\nu = 0.23\ \text{day}^{-1}$ and $t_m = 10$ days; (b) corresponding age distribution function $\phi(t,\tau)$ (eqn (2.15i)) of cells of age 0, $\frac{1}{2}\tau_d$, and τ_d in response to time t.

For $t = t_m$, $\lambda = 0$ and ϕ_0 defined by eqn (2.14e) is indeterminate as discussed above. In this case, eqn (2.14k) again applies, i.e. $\phi_0 = 1/\tau_d$. When computing ϕ care should be taken to avoid any problems of this nature. $M(t)$ and the corresponding age distribution $\phi(t,\tau)$, for $\tau = 0, \frac{1}{2}\tau_d, \tau_d$, are illustrated in Figs 2.12. It can be seen from Fig. 2.12b that initially young cells predominate whereas, as would be expected, this is reversed as time progresses.

2.4.3 *Production of differentiating cells*

The analysis of the previous section gives rise to non-meristematic or differentiating cells. If there are no differentiating cells and the population is purely meristematic, then the rate of production of these cells is given by eqn (2.13f). In the situation where differentiating cells are produced, eqn (2.14a) defines the rate of production of meristematic cells. The rate of production of differentiating cells is therefore the difference between eqns (2.13f) and (2.14a) which, denoting these cells by N and using eqn (2.14g) for λ, is given by

$$\frac{dN}{dt} = -\frac{\ln\theta}{\ln 2}\nu M. \tag{2.16a}$$

Substituting for θ, M, and D from eqns (2.15b), (2.15h), and (2.15g) gives

$$\frac{dN}{dt} = \left(\frac{\nu M_0}{t_m}\right) t \exp\left[\nu\left(1-\frac{t}{2t_m}\right)t\right]. \tag{2.16b}$$

At time $t = 0$ there are no differentiating cells, so that $N(t=0) = 0$. It is an interesting and challenging mathematical problem to show that eqn (2.16b) can be integrated to give (Exercise 2.7)

$$\frac{N}{M_0} = 1 - \exp\left[\nu\left(1-\frac{t}{2t_m}\right)t\right] + \left(\frac{\pi\nu t_m}{2}\right)^{1/2}\exp\left(\frac{1}{2}\nu t_m\right) \times\left\{\operatorname{erf}\left(\frac{1}{2}\nu t_m\right)^{1/2} + \operatorname{sgn}(t-t_m)\operatorname{erf}\left[\left(\frac{\nu}{2t_m}\right)^{1/2}|t-t_m|\right]\right\}, \tag{2.16c}$$

where the error function is defined by

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z \exp(-x^2)\,dx, \tag{2.16d}$$

the sign function by

$$\operatorname{sgn}(z) = \begin{cases}\text{sign of } z & z \neq 0,\\ 0 & z = 0,\end{cases} \tag{2.16e}$$

and

$$|z| = \text{modulus of } z. \tag{2.16f}$$

In practice, it is perhaps easier to integrate eqn (2.16b) numerically rather than use the analytical solution (2.16c). The ideal language for doing so is ACSL, which

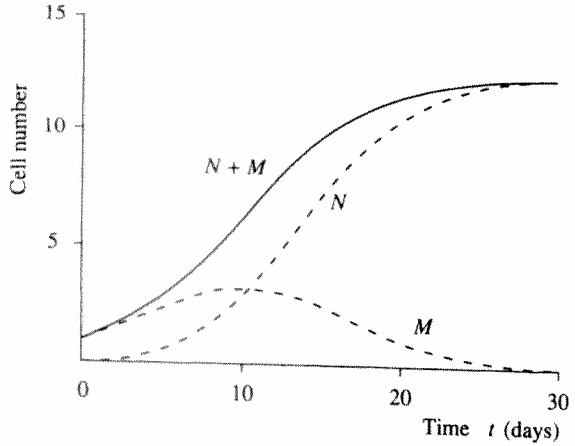


Fig. 2.13. Meristematic cell number $M(t)$, differentiating cell number $N(t)$ (broken curves as indicated), and total cell number $N + M$ (full curve) for the model of Section 2.4.3. $M(t)$ is given by eqn (2.15h) and $N(t)$ by eqn (2.16c). Parameter values are $M_0 = 1$, $\nu = 0.23\ \text{day}^{-1}$, and $t_m = 10$ day.

was discussed in Chapter 1. The solutions for M (identical with that of Fig. 2.12a), N, and the total cell number $M + N$ are illustrated in Fig. 2.13, where it can be seen that the growth of the total cell number follows the expected sigmoidal pattern.

Exercises

2.1. (a) Derive eqn (2.8b) for the reaction scheme of eqn (2.8a).
(b) Show that eqn (2.8b) has a point of inflexion at $S/K = 1/\sqrt{3}$.

2.2. (a) Derive eqn (2.8d) for the reaction scheme of eqn (2.8c).
(b) Show that eqn (2.8d) has a point of inflexion given by eqn (2.8e).
(c) Show that eqn (2.8f) also has a point of inflexion given by eqn (2.8e).

2.3. (a) Show that in the limit $\theta \to 0$ eqn (2.9j) (the non-rectangular hyperbola) reduces to eqn (2.9g).
(b) Show that eqn (2.9j) is equivalent to eqn (2.9i) when $\theta = 1$.
(c) Show that the initial slope of the non-rectangular hyperbola (eqn (2.9j)) is α (eqn (2.9k)).
(d) Show that the asymptote of the non-rectangular hyperbola (eqn (2.9j)) is y_m (eqn (2.9l)).
Hint. For part (a) use the binomial expansion in the form

$$(1 + x)^n = 1 + nx + \frac{n(n-1)}{2!}x + \frac{n(n-1)(n-2)}{3!}x^3 + \cdots \quad |x| < 1 \text{ and } n \neq -1, \quad \text{(E2.3a)}$$

where

$$m! = m(m-1)(m-2)\ldots 3.2.1 \quad \text{(E2.3b)}$$

and is termed m factorial. Note that, although not required here, when $n = -1$ the series expansion is

$$(1 + x)^{-1} = 1 - x + x^2 - x^3 + \cdots. \quad \text{(E2.3c)}$$

2.4. (a) Derive the constraint (2.10l) for the bi-substrate Michaelis–Menten reaction scheme.
(b) Derive eqn (2.10m) for the speed of the bi-substrate Michaelis–Menten reaction scheme.

2.5. (a) Derive eqn (2.11h) for the speed of the Michaelis–Menten reaction with a fully competitive inhibitor.
(b) Derive eqn (2.12j) for the speed of the Michaelis–Menten equation with a fully non-competitive inhibitor.

2.6. (a) Show that ϕ_0 is given by eqn (2.14e) for the scheme illustrated in Fig. 2.10, as discussed in Section 2.4.2.
(b) Derive eqn (2.14f) for λ for this model.
(c) Show that in the limit $\lambda \to 0$, ϕ_0 as defined by eqn (2.14f) reduces to (2.14l). To do so, use the series expansion for e^x:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots. \quad \text{(E2.6a)}$$

2.7. Derive eqn (2.16c) for the time course of the number of differentiating cells for the model presented in Section 2.4.3. This is quite a difficult problem.