*Int. J. Engng Ed.* Vol. 14, No. 5, pp. 375–380, 1998
Printed in Great Britain.

0949-149X/91 $3.00+0.00

# Simplified Curve Fitting using Spreadsheet Add-ins*

E. G. JOHN
*Systems Division, School of Engineering, Cardiff University, Cardiff CF2 3TE, Wales UK.*
*E-mail: john@cardiff.ac.uk*

*This paper describes how the Solver function contained within the Microsoft Excel® spreadsheet package, can be readily employed to create a simple functional solution to the problem of generalised curve fitting. The method is illustrated with two worked examples. The approach suggested is simple to understand and apply, and only requires that any proposed model can be described using the standard mathematical and statistical functions contained within Excel. The method is thus capable of fitting a wide range of different models and, because of the simplicity of the solution approach, can be readily applied to evaluate the 'comparative goodness of fit' of several different models.*

## INTRODUCTION

FITTING an appropriate curve or model to a series of data points and then extrapolating the model is a fundamental requirement in many disciplines. Yet apart from the simplest of linear models ($y = mx + c$), curve fitting often proves to be beyond the scope of many involved in engineering, business and science. In mitigation, it is also true that little or no concerted effort, within a typical educational programme, is afforded to curve fitting in its wider application sense. Certainly there is effort directed at simple fitting using standard formulae, specially created curve fitting packages or hardwired functions on a calculator. However a broader understanding of how to fit a particular model and why that model should be selected are frequently overlooked. This paper suggests how the available standard curve fitting functions, contained in Microsoft's Excel [1] spreadsheet package, can be enhanced by employing the Solver [2] add-in to create a simple functional solution to the former of these problems. The paper also briefly considers how the latter problem can be addressed.

## OVERVIEW

Frequently in engineering, science and business, data is collected and plotted as a graphical representation of the variables involved. The next reaction is to create an association between the variables by connecting the 'points' with a line. Once drawn, the line is examined and a 'model', which 'best fits' the data points, assumed. This is then 'fitted' and used to replace the existing set of data points as 'the appropriate model'. Having thus 'modelled' the data, this model is then used to predict future values of one variable, given changes in the other (extrapolation).

The foregoing statement covertly mentions several features of the problems which often inhibit achievement and limit confidence in any subsequent extrapolation from the model. These are:

- The validity of the assumption, that the data points are, or can be, connected.
- That the model selected, is the most appropriate model for the collected data.
- That a curve-fitting method is available, usable and understandable.
- That the model parameters can be obtained.
- That present facts are reflective of future behaviour.

This paper will assume that first, the data points are connectable, and secondly, that subsequent extrapolation of the fitted model is a perfectly valid action, and will address the three other features viz:

- model selection;
- curve-fitting package availability;
- curve-fitting *per se*.

## DIFFICULTIES WITH MODEL SELECTION

A brief examination of most standard texts on data and its analysis, will show that commonly only a limited number of standard models are normally described. Those frequently mentioned are the simple linear model ($y = mx + c$) or those that can easily be transformed into a simple linear form ($y = ae^x$). Invariably the text will offer a discourse on the theory of curve fitting, followed by a derivation of the normal equations and conclude with advice on how simple curves can be either mathematically transformed or linearised

* Accepted 15 October 1998.

with the aid of specialist graph paper. The usual advice given will be to prepare a simple plot of the data, either on linear or logarithmic graph paper, confirm that the data exhibits a linear tendency, undertake the necessary transformation and then apply the least squares criteria to obtain the 'line of best fit'. Typically little or no advice will be offered on the selection, fitting or choice of model for a given data set, nor how more unusual curves can be fitted.

As an example of a more unusual model, consider the curve fitting problems associated with causality [which expresses the relationships between cause and effect in a system]. Data collected from a causality situation is frequently modelled as a hyperbolic tangent (tanh) function. Although best fits for such models can be obtained, using mathematical transformations, or linearised using the less common graph papers, very few analysts will be capable of recognising these models and even fewer will have the expertise to undertake the necessary curve fitting.

The problem of selecting the most appropriate model has been, and continues to be, one of the most difficult aspects of data modelling. Experience in data modelling is vital if good data fits are to be obtained. Experience comes from practice, coupled with a good understanding of function plots and the parameters that affect them. The best advice that can be offered to any would-be model builder, is to employ a spreadsheet system to create a whole range of simple function plots, experiment with changing the parameters and examine the resulting plot forms. By this means it is possible to experiment with a wide range of models in a short time period. The problem of selecting the most appropriate model can be readily resolved by employing the method detailed in this paper on a number of models and comparing the sum of the squared errors values for each model.

## COMPUTER-ASSISTED CURVE FITTING

Since the advent of popular computing in the early 1970's, there have been a number of major developments in the availability of applications packages to undertake curve fitting. In the formative years it was common for machine manufacturers, specialist computing groups or programming language providers to provide the necessary software. These were invariably subroutines which could be called from within a programme. They were not user-friendly and often machine or software specific.

In time, as personal computing developed and the software platforms and operating environments stabilised, portable third-party software packages flourished. SPSS [3], MATHCAD [4] and MAPLE [5] are three such packages. However, they were, and still are, specialist application packages for those who have the time and knowledge to use them. However, once the more universal computing era began (integrated word-processor, spreadsheet and database applications package), the need for simple curve-fitting models and standard statistical data analysis tools, became evident. Consequently spreadsheet providers have since incorporated both powerful data analysis tools and curve-fitting add-ins into such packages. In the context of the curve fitting add-ins these were inevitably tailored to the limited demands of the typical user. Consequently those users who wished to consider and fit more complex models have had to undertake the task themselves or resort to specialist applications packages, such as those previously mentioned. Regrettably such packages are often mounted outside a normal work environment and can be problematic in effecting the import/export of data between applications.

### *Standard function Excel solutions*

Experienced users of Excel [1] will be familiar with the standard curve fitting functions and how they are evoked. Typically the user will enter the data into a range of cells, create a chart, point with the mouse to the data plotted on the chart to highlight the data series and then click on the toolbar to activate the Trendline function via the Insert menu command. The available models are:

- Linear ($y = mx + c$)
- Logarithmic ($y = a \ln x + b$)
- Polynomial ($y = b + a_1x + a_2x^2 + a_3x^3 + a_4x^4 \ldots$)
- Power ($y = ax^b$)
- Exponential ($y = ae^{bx}$)

The Trendline function permits the user to obtain a least squares fit to a given data series using any of the above models. Once undertaken, the fitted equation can be displayed on the chart.

### *The Solver add-in*

Excel Versions 5 and above, have an add-in facility included in the Tools menu, called Solver [2]. Solver is an optimisation procedure which can be used to generate solutions to a wide spectrum of linear, non-linear and integer problems. Solver finds the optimum value for a given 'Target cell' by changing the values of other cells which have been designated as 'Change cells' in the specification of the problem. The Target and (or) Change cells can also be defined within specified constraints. Thus to solve a problem Solver needs knowledge of:

- The Target cell (or objective function).
- The Change cells (or decision variables).
- The Constraints placed on either (or both) the Target and Change cells.

Once these have been defined, and the search conditions and solution parameters specified, a solution to a particular problem can then be obtained. The search conditions, resident under the Options menu of Solver, and solution parameters are designed to exert control over the time, precision and structure of the solution. Once a solution has been generated, Solver permits the user to keep, or reject, the solution found. Additionally, Solver can create reports which summarise the Sensitivity, Answer and Limits of the solution found.

*Curve fitting using Solver*

To fit a curve to a data series using the Solver add-in is simplicity itself. The only difficulty is that associated with all curve fitting, i.e., which model should be chosen. However for the purposes of this section, it is assumed that a data series containing the $x$ and $y$ values is available and that an appropriate model has been selected. Fitting the chosen model is then as follows:

1. Enter the known $x$ and $y$ values as a data series onto the spreadsheet.
2. Add a further column containing the 'assumed model'. (The model will be expressed as a formula based upon the $x$ values and copied down for each $y$ value). The parameters of the chosen model are estimated and located in any free cell(s). These are the Change cells.
3. If a visual representation is required, the Chart Wizard should be evoked and a chart drawn showing both the assumed model and the known $y$ values.
4. Add a further column which expresses the squared error between the known $y$ values and the assumed model values.
5. Sum the squared errors column in an appropriate free cell.
6. Evoke Solver by selecting the Tools menu and Solver to present the Solver dialogue box.
7. In the dialogue box enter the sum of the squared errors cell as the Target cell.
8. Set the Equal to option to Min.
9. Enter the selected Change cells to the 'By changing cells' box.
10. Include any constraints and modify the options as necessary.
11. Select the Solve button to initiate the curve fitting. The values of the assumed model parameters will then be adjusted in each of the Change cells until the Target cell value is a minimum.
12. Save the solution or reset the problem as necessary.

## EXAMPLE PROBLEMS

Consider the following causality problem taken from the electronics industry. The data represents the number and categories of fault observed in a sample of 85 rejected printed circuit boards (Table 1).

From the data collected in Table 1, the cumulative % causes (categories) and effects can be determined as shown in Table 2.

Table 1.

| Fault category | Number counted |
|---|---|
| Incomplete function (IF) | 5 |
| Missing items (MI) | 4 |
| Surface defects (SD) | 22 |
| Surface cracks (SC) | 1 |
| Overall finish (OF) | 2 |
| Misshapen set (MS) | 8 |
| Poor assembly (PA) | 43 |
| Total | 85 |

A graph of cumulative % causes against cumulative % effects (Fig. 1) can now be drawn and for the purposes of this paper, it will be assumed that the data is to be modelled using the positive half of the hyperbolic tangent (tanh) function. The hyperbolic tangent (tanh) function can be mathematically expressed as:

$$y = (1 - \mathrm{e}^{-x/h})/(1 + \mathrm{e}^{-x/h})$$

and modelled as:

$$y = \tanh(x/h)$$

where $h$ is the shaping constant of the curve. Thus the problem is to find the value of the shaping constant ($h$) which causes the assumed model to best fit the raw data.

*Setting up the causality problem in Excel*

The data from Table 2, pertaining to cum. % of cause and effect is entered into the spreadsheet, together with the assumed model and the squared error column. The squared error values are then summed into an appropriate free cell and an initial estimate of ($h$) for the assumed model tanh $(x/h)$ entered in another free cell. This is shown in Fig. 2 and for illustrative purposes a graph of the raw data and assumed model is also drawn.

Having entered the problem into the spreadsheet, the next stage is to evoke the Solver add-in and minimise the sum of the squared errors by adjusting the initial estimate of ($h$).

The Excel spreadsheet and Solver dialogue box are shown in Fig. 3. The Solver dialogue box has the target cell set to $D$10 (initial value 711.23) and this will be minimised by changing the value in cell $D$11 (initially set at 18).

When the Solve button is activated these are

Table 2.

| Ranked fault category | Number counted | Cum. % of causes ($x$) | Cum. % of effects ($y$) |
|---|---|---|---|
| Poor assembly (PA) | 43 | 14.28 | 50.58 |
| Surface defects (SD) | 22 | 28.57 | 76.46 |
| Misshapen set (MS) | 8 | 42.85 | 85.86 |
| Incomplete function (IF) | 5 | 57.14 | 91.74 |
| Missing items (MI) | 4 | 71.42 | 96.4 |
| Overall finish (OF) | 2 | 85.71 | 98.79 |
| Surface cracks (SC) | 1 | 100.00 | 100.00 |
| Total causes = 7 | Total = 85 | | |

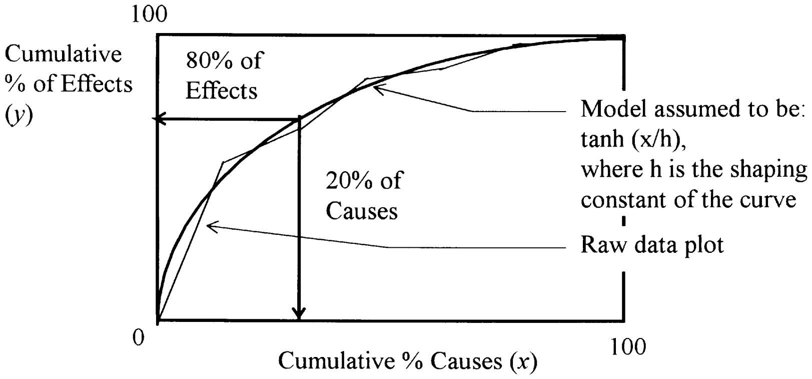
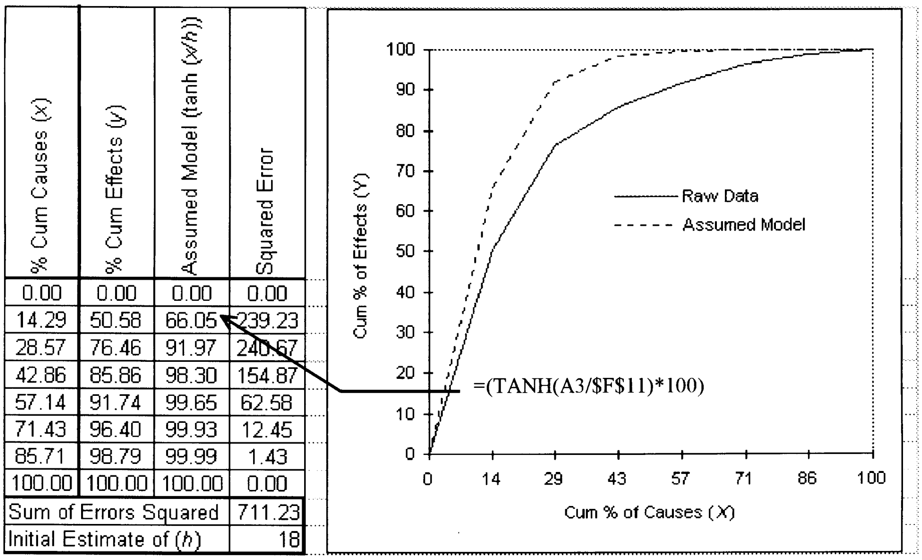


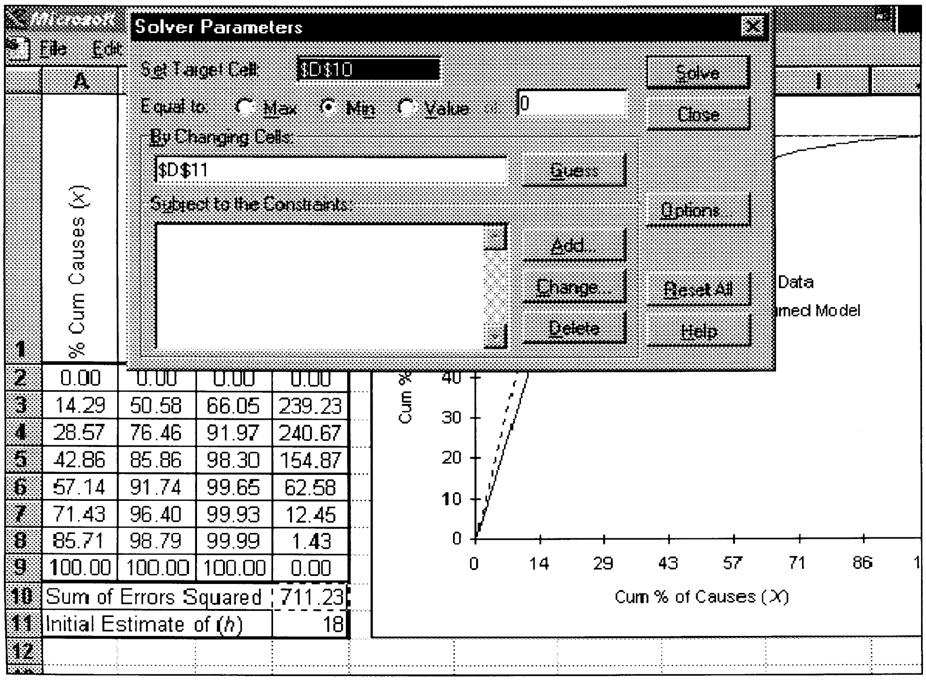
Fig. 1.



| % Cum Causes (x) | % Cum Effects (y) | Assumed Model (tanh (x/h)) | Squared Error |
|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.00 |
| 14.29 | 50.58 | 66.05 | 239.23 |
| 28.57 | 76.46 | 91.97 | 240.67 |
| 42.86 | 85.86 | 98.30 | 154.87 |
| 57.14 | 91.74 | 99.65 | 62.58 |
| 71.43 | 96.40 | 99.93 | 12.45 |
| 85.71 | 98.79 | 99.99 | 1.43 |
| 100.00 | 100.00 | 100.00 | 0.00 |
| Sum of Errors Squared | | | 711.23 |
| Initial Estimate of (h) | | | 18 |

=(TANH(A3/$F$11)*100)

Fig. 2.



Solver Parameters

Set Target Cell: $D$10

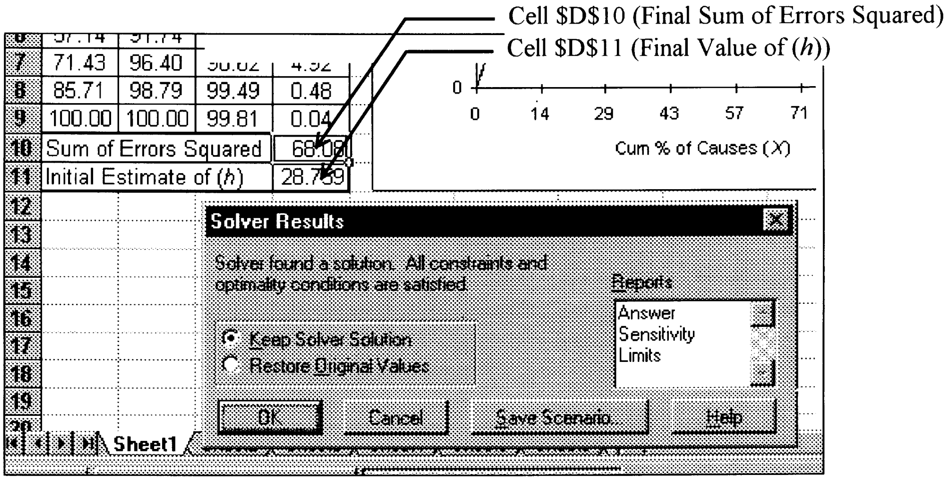
Equal to: ○ Max ● Min ○ Value of: 0

By Changing Cells: $D$11

Subject to the Constraints:

| | A | | | |
|---|---|---|---|---|
| 1 | % Cum Causes (x) | | | |
| 2 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3 | 14.29 | 50.58 | 66.05 | 239.23 |
| 4 | 28.57 | 76.46 | 91.97 | 240.67 |
| 5 | 42.86 | 85.86 | 98.30 | 154.87 |
| 6 | 57.14 | 91.74 | 99.65 | 62.58 |
| 7 | 71.43 | 96.40 | 99.93 | 12.45 |
| 8 | 85.71 | 98.79 | 99.99 | 1.43 |
| 9 | 100.00 | 100.00 | 100.00 | 0.00 |
| 10 | Sum of Errors Squared | | | 711.23 |
| 11 | Initial Estimate of (h) | | | 18 |
| 12 | | | | |

Fig. 3.



Cell $D$10 (Final Sum of Errors Squared)

Cell $D$11 (Final Value of (*h*))

| | | | | |
|---|---|---|---|---|
| 6 | 57.14 | 91.74 | | |
| 7 | 71.43 | 96.40 | 98.02 | 4.92 |
| 8 | 85.71 | 98.79 | 99.49 | 0.48 |
| 9 | 100.00 | 100.00 | 99.81 | 0.04 |
| 10 | Sum of Errors Squared | | | 68.08 |
| 11 | Initial Estimate of (*h*) | | | 28.759 |

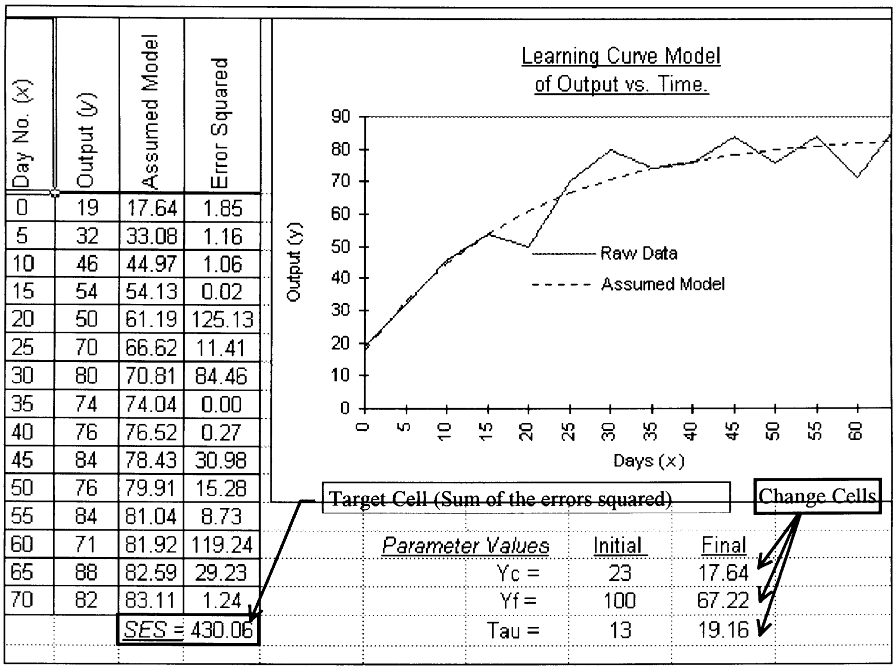
Solver Results

Solver found a solution. All constraints and optimality conditions are satisfied.

● Keep Solver Solution

○ Restore Original Values

Fig. 4.



| Day No. (x) | Output (y) | Assumed Model | Error Squared |
|---|---|---|---|
| 0 | 19 | 17.64 | 1.85 |
| 5 | 32 | 33.08 | 1.16 |
| 10 | 46 | 44.97 | 1.06 |
| 15 | 54 | 54.13 | 0.02 |
| 20 | 50 | 61.19 | 125.13 |
| 25 | 70 | 66.62 | 11.41 |
| 30 | 80 | 70.81 | 84.46 |
| 35 | 74 | 74.04 | 0.00 |
| 40 | 76 | 76.52 | 0.27 |
| 45 | 84 | 78.43 | 30.98 |
| 50 | 76 | 79.91 | 15.28 |
| 55 | 84 | 81.04 | 8.73 |
| 60 | 71 | 81.92 | 119.24 |
| 65 | 88 | 82.59 | 29.23 |
| 70 | 82 | 83.11 | 1.24 |
| | | SES = | 430.06 |

Target Cell (Sum of the errors squared)

Change Cells

| Parameter Values | Initial | Final |
|---|---|---|
| Yc = | 23 | 17.64 |
| Yf = | 100 | 67.22 |
| Tau = | 13 | 19.16 |

Fig. 5.

changed to 68.08 and 28.759 respectively (the graph is also updated automatically during the Solve phase). Figure 4 shows the Solver results box and the changed values in the two cells $D$10 and $D$11. Figure 4 also shows the options contained within the Solver results box. (The Keep or Restore solution facility together with the available Report options.)

*Curve fitting a learning curve problem*

Using the approach described for the causality problem, Fig. 5 shows data drawn from another field of activity. In this instance the model is associated with learning rates and is described as the Time Constant Model [6]. It is defined as:

$$y_t = y_c + y_f(1 - e^{-t/\tau})$$

where $y_t$ is the output at any time $t$

$y_c$ is the initial output

$y_f$ is the expected improvement

$\tau$ is the time constant.

In this scenario, unlike the causality model, there are three parameters to be estimated, $(y_c, y_f, \tau)$. However, the solution approach is exactly the same. The data series is entered into the spreadsheet, together with the assumed model. The

Table 3.

| Parameter | Initial | Final |
|---|---|---|
| $y_c$ | 23 | 17.64 |
| $y_f$ | 100 | 67.22 |
| $\tau$ | 13 | 19.16 |

individual squared errors are calculated and cells are selected into which the model parameters and sum of the squared errors are entered. Solver is evoked and the Target and Change cells identified (see Fig. 5). The Solve button is activated and the solution is generated. The initial and final parameter values are shown in Table 3.

## CONCLUSIONS

This paper has described a novel method of using the Solver function within Excel. The approach is simple to understand and apply, and is capable of curve fitting a whole range of different models. It also has the advantage that several different models, for a given data series, can be easily investigated (thus easing the model selection dilemma). This then permits the model with the minimum sum of squared errors, of those tested, to be selected as the most appropriate. Model sensitivity and error data are also generated within the Solver add-in.

## REFERENCES

1. *Microsoft Excel Users Guide*, Version 5, Microsoft Corporation, Document Number XL 57926-0694
2. The Solver Add-in, Chapter 29, *Microsoft Excel Users Guide*, Version 5, Microsoft Corporation, Document Number XL 57926-0694
3. Norman H. Nie, *SPSS: Statistical Package for the Social Sciences,* McGraw-Hill New York (1975).
4. *MATHCAD: Users Guide,* Mathsoft Inc., Cambridge Mass (1995).
5. W. C. Bauldry, *et al., Linear Algebra with MAPLE,* John Wiley (1995).
6. G. W. Hindmarch and D. R. Towill, *Theory and Application of the Time Constant Learning Curve Model,* ORSA/TIMS, Puerto Rico (October 1974).

**Elwyn John**, M.Sc., C.Eng., FIEE, after a formal education in Production Engineering and several years industrial experience, now lectures in the Manufacturing Systems Division of the School of Engineering at University of Cardiff. Whilst there he has developed a wide experience base in manufacturing and has published in several diverse fields including production scheduling, non-destructive testing, work measurement, cellular manufacturing and operations management and has current interests in quality, packaging and in engineering education. He is active in consultancy and has an interest in the professional activities of the IEE, and is a past member of its Council and of Professional Group2.