Parameter prediction for cash flow forecasting models

Martin Skitmore

Parameter prediction for cash flow forecasting models

MARTIN SKITMORE

Department of Surveying, University of Salford, Salford M5 4WT, UK

The paper describes the application of the DHSS formula to 27 completed construction projects comprising four types – steel-framed low rise buildings, new build housing developments, housing refurbishment projects, and multi-house 'pre-paint' maintenance contracts. Application of the formula to individual projects indicates that the 'best' parameter values offer a ten-fold improvement over the published values based on project size. Similar results occur when using the best parameter values for other two parameter models.

Various approaches are considered in attempting to predict the best parameter values of the models based on known characteristics of the project. A multiple linear regression with project value, duration, and type independent variables is shown not to produce any significant improvement on standard DHSS formula predictions. However, a reduction in the number of independent variables by cross validation produces an approximately 25% improvement on standard DHSS formula forecasts outside the data base. Examination of the models derived from this analysis indicate the type of project to be of major importance.

Keywords: Cash flow, forecasts, project type, regression, cross validation.

Introduction

The model most commonly used for expenditure forecasting in the UK is the two parameter model derived by the Hudson (1978), commonly known as the DHSS formula. The values of the parameters are provided over a range of project sizes as a result of Hudson's analysis of a large number of hospital projects. Different parameter values should, according to Hudson, be used for different types of construction rather than the general practice of applying his parameter values to all types of work.

Since Hudson's work, the efficacy of the DHSS formula seems to have received little attention. Peer's (1982) analysis of a small sample of data indicated that projects within the same building type group had similar expenditure patterns even when they were of different size and had a different rate of progress. What is not clear, however, is the extent of the differences between building type groups, especially in comparison with the DHSS formula and Hudson's parameter values. Nor is it clear what kinds of building work constitute a group in terms of expenditure profiles generally.

In this paper the expenditure flows for a set of 27 completed projects is examined to assess the effect of different building type groups in relation to the DHSS and, for ease of comparison, other similar two parameter models. Firstly, the best parameter values for each

model for each project is found by a method of successive approximation. Secondly, a standard linear multiple regression is used to analyse and predict model parameters via the value, duration and project type group. Finally, a non linear technique is used to find the best set of predictors for each model for projects outside the data base. Analysis of the resulting models indicates that: 1. the inclusion of the project type grouping significantly improves the predictive power of the models, 2. all have similar predictive power, and 3. the predictive power is approximately 25% better than the DHSS formula with Hudson's parameter values.

Analysis

The data

The data consisted of details of 27 completed construction projects in the United Kingdom, together with their associated estimated monthly values. The data were obtained from one single private practice quantity surveying firm and two local authority surveying departments. All contract values were rebased to 1974 prices by means of the RICS Building Cost Information Service Tender Price Index (Appendix A).

Four project type groups were involved – steel-framed low rise buildings, new build housing developments, housing refurbishment projects, and multi-house 'pre-paint' maintenance contracts. The cumulative gross value of work executed, before adjustment for retention or fluctuations, was used since these adjustments vary between projects and may distort any model trying to reflect a general trend. In practice, of course, such adjustments can easily be applied for a specific project as an adjunct to the results provided by an expenditure model. Clearly, some knowledge of the likelihood of cost and time overruns is needed for a comprehensive treatment of the subject and it is intended to address this issue as an extension of the work described.

No adjustments were made for any inter-project variations such as winter working, industry holidays or delivery to site of steel or mechanical plant, since these adjustments were considered to be relatively small and have little effect on the results. An extension of the work to cover these aspects is also under consideration.

The models

Four alternative models were considered: 1. the DHSS formula (Hudson, 1978); 2. the Kenley–Wilson formula (Kenley and Wilson, 1989); 3. the Miskawi formula (Miskawi, 1989), and 4. the Berny–Howes formula (Berny and Howes, 1982). For comparative purposes, the following notation is adopted: V = total value of the contract (£); D = total duration of contract (days); v = percentage value complete; d = percentage of time complete; a and b are constants.

The DHSS formula

$$v = 100[x + ax^2 - ax - (6x^3 - 9x^2 + 3x)/b] \tag{1}$$

where

$$x = d/100$$

The Kenley–Wilson formula is

$$v = 100F/(1 + F) \tag{2}$$

where

$$F = e^a [d/(100 - d)]^b$$

The Miskawi formula is

$$v = (3^d/2)\sin[(\pi/200)(100 - d)]\sin(\pi d/100)\log[d + 50)/(a + d)] - 2d^3/10\,000 + 3d^2/100 \tag{3}$$

The Berny–Howes formula is

$$v = 100x\{1 + a(1 - x)(x - b)\} \tag{4}$$

   Two problems are immediately apparent with these formulae. Firstly, the Kenley–Wilson formula contains the divisor $100 - d$ which implies that $100 > d$. It is important therefore to either omit values where $d = 100$ or raise the value of 100 slightly to say 100.1. The latter method has been adopted in this study. Secondly, the Miskawi formula contains a power variable 3^d which is clearly inappropriate for d values approaching 100. This was overcome by substituting a new constant b for the d power in the formula.

Parameter estimation

Estimates of the DHSS formula constants, a and b, are provided by Hudson (1978) based on the contract value of the project. This method is termed here the standard DHSS formula. An attempt was made to estimate the values of the constants in the four formulae by method of least squares, i.e. to minimize for each project

$$\text{SSQ}(a, b) = \sum_{i=1}^{n} (v_i - u_i)^2 \tag{5}$$

where

   u_1 = actual percentage value completed;

   v_i = forecast percentage value completed; i.e.

   $v_i = 100[x_i + ax_i^2 - ax_i - (6x_i^3 - 9x_i^2 + 3x_i)/b]$ (from Equation 1) $\tag{6}$

$$v_i = 100F_i/(1 + F_i) \text{ (from Equation 2)} \tag{7}$$

$$v_i = (3^b/2)\sin[(\pi/200)(100 - d_i)]\sin(\pi d_i/100)\log[(d_i + 50)/(a + d_i)]$$
$$- 2d_i^3/10\,000 + 3d_i^2/100 \text{ (from Equation 3)} \tag{8}$$

$$v_i = 100x_i\{1 + a(1 - x_i)(x_i - b)\} \text{ (from Equation 4)} \tag{9}$$

$i = 1, 2, 3, \ldots, n$ are the valuation points recorded.

This was done numerically by a method of successive approximation for each of the 27 project data sets. The results are shown in Table 1 together with the mean square error

$$MSQ_{(a,b)} = SSQ_{(a,b)}/(n - 2) \tag{10}$$

As the Table shows, no results were obtained for 13 of the Miskawi formula applications, due to the lack of a minimum least squares result within the bounds allowed ($0 \le a \le 100$). The Berney-Howes b values for projects 5, 6, 10, 24 and 26 may be regarded as spurious as, when $a = 0$, $v = x$, a straight line.

The substantial, around ten-fold, improvement brought about by using 'best' estimates of the formulae constants is certainly encouraging. It should be noted however that the method of least squares is limited to a simple quantification of the aggregated differences between the model's predictions and the observed values and therefore provides no specific information concerning counter-intuitive predictions, such as negative values at the beginning and end of the curve, that sometimes occur with such one-element polynomial models. More sophisticated multi-element models exist to overcome these problems by providing independent variables potentially capable of modelling greater extremes and variability in expenditure (e.g. Tucker, 1986; Tucker, 1988; Khosrowshahi, 1991). Bearing in mind these limitations, a further analysis was conducted aimed at investigating the relationship between the model parameters and characteristics of the projects involved.

Analysis of formulae constants

Discussion. The first approach to this was to examine the relationship between the a and b values and the $MSQ_{(a,b)}$ for each project. The $MSQ_{(a,b)}$ was calculated over a range of a and b values and plotted in the form of a contour map. Figure 1 gives the results for project 3 analysed by the Kenley–Wilson formula. Although not contributing very much to the overall analysis, the contour maps were found to be a useful aid in understanding the nature of the lack of fit of the formulae between projects, particularly the sensitivity of the a and b parameter values on model performance.

These contour maps nicely illustrate the task in hand, which is to find a meaningful grouping of similar maps. This raises two issues. Firstly, what constitutes a 'meaningful' grouping, and secondly, what measure of similarity should be used?

The answer to the first of these questions is relatively straightforward and depends on the purpose of the analysis. In our case, the object was to find some suitable predictors that would be readily available in the project preconstruction phase. The standard DHSS formula, for instance, uses the contract value as a predictor. Other obvious predictors include the project type and duration.

Table 1. Results of fitting the four formulae

Proj Nr	Standard DHSS			DHSS			Kenley–Wilson			Miskawi			Berny–Howes		
	a	b	MSQ	a	b	MSQ	a	b	MSQ	a	b	MSQ	a	b	MSQ
1	−0.07	3.20	63.4	0.04	1754.5	6.8	−0.04	1.00	6.8	no result			−0.08	−0.03	6.7
2	−0.07	3.20	10.8	−0.21	3.1	3.3	0.25	1.56	2.9	49.8	8.8	5.8	1.94	0.39	3.3
3	−0.20	4.06	10.1	−0.09	4.3	6.0	0.09	1.34	6.9	no result			1.41	0.43	6.0
4	−0.07	3.20	16.9	−0.05	6.3	5.1	0.05	1.22	5.1	no result			0.95	0.44	5.1
5	−0.24	4.93	138.6	−0.76	1773.6	7.3	0.82	1.12	5.2	47.5	7.9	9.4	0.00	−41.4	7.3
6	−0.24	4.93	135.6	−0.65	1853.2	25.1	0.72	1.12	21.9	49.5	9.1	32.2	0.00	−423.8	25.1
7	−0.24	4.93	39.0	−0.33	104.1	15.4	0.34	1.05	15.2	0.8	4.1	28.0	0.06	−5.1	15.4
8	−0.24	4.93	18.6	−0.22	2.8	4.9	0.27	1.64	3.9	49.9	9.0	9.9	2.13	0.40	4.9
9	−0.20	4.06	25.8	−0.04	8.0	9.4	0.04	1.17	9.7	no result			0.75	0.45	9.4
10	−0.07	3.20	180.7	−0.57	1807.8	10.1	0.58	0.97	5.8	6.2	4.7	17.1	0.00	−467.0	10.0
11	−0.20	4.06	26.0	−0.44	5.6	1.7	0.48	1.30	2.3	49.5	8.7	1.9	1.07	0.09	1.7
12	−0.36	5.00	20.0	−0.43	8.9	13.5	0.47	1.23	11.4	49.4	8.6	15.7	0.68	−0.13	13.5
13	−0.20	4.06	236.2	−0.89	12.5	2.9	1.03	1.33	2.8	49.7	9.6	8.6	0.48	−1.36	2.9
14	−0.20	4.06	10.8	−0.26	4.3	9.3	0.28	1.34	12.3	6.3	3.8	8.4	1.40	0.31	9.3
15	−0.07	3.20	264.7	−0.86	12.0	2.6	0.96	1.28	5.8	49.5	9.3	6.7	0.50	−1.23	2.6
16	−0.20	4.06	25.4	−0.20	21.3	1.7	0.21	1.08	1.5	no result			0.28	−0.22	1.7
17	−0.07	3.20	11.0	−0.01	5.6	0.8	0.01	1.26	1.0	no result			1.07	0.50	0.8
18	−0.07	3.20	98.5	0.01	19 107.0	8.1	−0.01	0.87	1.8	no result			−0.67	0.49	1.6
19	−0.07	3.20	11.5	−0.04	3.3	10.8	0.05	1.50	10.4	no result			1.82	0.48	10.8
20	−0.24	4.93	252.4	0.26	1921.1	28.4	−0.21	0.81	2.8	no result			−0.94	0.25	8.7
21	−0.20	4.06	8.6	−0.15	6.9	2.1	0.16	1.21	2.2	no result			0.87	0.33	2.1
22	−0.20	4.06	269.1	0.43	1917.8	24.6	−0.39	0.89	16.1	no result			−0.81	0.01	14.8
23	−0.20	4.06	181.6	0.30	1750.2	19.5	−0.30	1.02	19.5	no result			−0.08	−3.46	19.4
24	−0.36	5.00	464.4	−1.18	1848.5	57.5	1.57	1.47	32.9	49.7	10.0	50.5	0.00	−500.9	57.5
25	−0.07	3.20	153.5	−0.71	3.8	4.4	0.86	1.57	2.5	49.8	9.8	22.3	1.56	0.05	4.4
26	−0.07	3.20	27.7	0.18	3.4	4.6	−0.23	1.52	2.5	no result			0.00	−352.5	48.3
27	−0.20	4.06	17.8	−0.28	5.3	13.9	0.29	1.27	16.1	2.2	3.8	12.7	1.14	0.26	13.9
Average			81.79			8.54			7.17			—			12.16

```
 _____
|                                                                        |
| 66665555555544444444444443333333333333333344444444444445555555566666667777788889999AA |
| 66665555555544444444444443333333333333333344444444444445555555566666667777788889999AA |
| 66655555554444444444433333333333333333333334444444444455555556666667777788889999AA |
2.0| 66655555544444444433333333333333333333333333334444444455555556666667777788888999AA |
| 66555555444444444333333333333333333333333333334444444455555556666667777788889999AA |
| 65555555444444333333333333333222222222333333333333334444444455555566666777788888999AA |
| 65555554444443333333333333222222222222222223333333333333344444455555566666777788888999AA |
| 55555444444333333333222222222222222222222222233333333334444444555556666677778889999AA |
1.8| 55555444443333333222222222222222222222222222222233333333334444444555556666677778889999AA |
| 55555444433333332222222222222222222222222222222223333333334444455556666777788889999AA |
| 55554444433333332222222222222211112222222222222223333333334444445555666777788889999AA |
| 55555444433333322222222222211111111111111122222222222333333334444455556667777788999AAA |
| 55555444433333322222222221111111111111111111122222222233333334444455556666777888999AAB |
1.6| 55555444433333322222222211111111111111111111111122222222233333334444555556666777888999AAB |
| 55555444433333322222221111111111111111111111111112222222233333334444555566677788899AAAB |
| 55555444433333222222221111111111111111111111111112222222233333334444555566677788999AABB |
| 55555444433332222222211111111111111111111111111112222222233333444455556667778899AABBB |
| 65555444433332222222211111111111111111111111111112222222233333444455556667778899AAABBC |
1.4| 66555544443333322222221111111111111111111111111112222222233333444455566677788999AABBCC |
| 66555544443333322222221111111111111121111111111111222222333334444555566677788899AABBCCD |
| 66655554444333322222221111111111111111111111111112222222333334444555566667778899AAABBCCD |
| 66655554444333322222221111111111111111111111111111122222233333444455556667778899AABBCCDE |
| 76665554444333322222221111111111111111111111111111122222223333344455556667788999AABBCCDEE |
1.2| 77666655554444333322222221111111111111111111111111122222223333344455556667788999AABBCCDEEF |
| 77766655554444333322222221111111111111111111111111112222222333334444555666777888999AABBCCDEEFF |
| 88776665555444433333222222221111111111111111222222233333444555566677788899AABBCCDDEEFGG |
| 88877666555544443333322222222221111111112222222222333334444555667778899AABBCCDDEEFFGGH |
| 9988776665555444433333322222222222222222222222222233333344455556667778899AABBCCDEEFFGHHI |
1.0| 999887776655554444333332222222222222222222222233333334445556667778899AABBCCDDEEFFGGHIIJ |
| AA99888776665555544443333333333333333333333344455566677888899AABBCCDDEEFFGGHHIJKK |
| BBAA99887776665555444443333333333333333344444455556667778899AABBCCDDEFFGGHHIJJKLLM |
| CCBBAA99887776665555554444444444444444444555566666777889AABBCCDDEEFGGGHIIJKLLMN |
| DDCBBAA99988777766665555555444444455555555556666677788899AABBCCDDEEFGGHIIJKLLMNOO |
0.8| EEDCCBBAAA998887777666666555555555556666667777888999AABBCCDDEEFGGHIIJKLLMNOOPQ |
| FFEEDDCCBBAA99988888777776666666666666777778888999AABBBCCDDEFFGGHIIJKLLMNOPPQRS |
| GGFFEEDDCCBBBAAA99988888877777777788888889999AAABBBCCCDDEEFFGHHIJJKLMMNOPQQRSTU |
| IIHGGPFEEDDCCBBBAAAA999999998888889999999AAABBBCCCDDEEFFGGHIIJKKLMNNOPQRRSTUVW |
| KJJIHHGGFFEEDDDCCCBBBBAAAAAAAAAAAAAAAAAABBBBCCCDDEEEFFGGHHIJJKLMMNOPPQRSTUVVWXY |
0.6| KKJJIIHGGFFFEEDDDCCCCBBBBBBBBBBBBBBBBBBBCCCDDDEEEFFGGHHIJJKKLMNNOPQQRSTUVWXYZZ |
|                                                                        |
 _____
    -0.5       -0.3       -0.1       0.1       0.3       0.5       0.7       0.9
       -0.4       -0.2       0.0       0.2       0.4       0.6       0.8       1.0
                                        a
```

b

The bounds of MSQ are defined as follows:

1: 6.900- 25.553	2: 25.553- 44.206	3: 44.206- 62.859
4: 62.859- 81.513	5: 81.513- 100.166	6: 100.166- 118.819
7: 118.819- 137.472	8: 137.472- 156.125	9: 156.125- 174.778
A: 174.778- 193.431	B: 193.431- 212.085	C: 212.085- 230.738
D: 230.738- 249.391	E: 249.391- 268.044	F: 268.044- 286.697
G: 286.697- 305.350	H: 305.350- 324.003	I: 324.003- 342.657
J: 342.657- 361.310	K: 361.310- 379.963	L: 379.963- 398.616
M: 398.616- 417.269	N: 417.269- 435.922	O: 435.922- 454.575
P: 454.575- 473.229	Q: 473.229- 491.882	R: 491.882- 510.535
S: 510.535- 529.188	T: 529.188- 547.841	U: 547.841- 566.494
V: 566.494- 585.147	W: 585.147- 603.801	X: 603.801- 622.454
Y: 622.454- 641.107	Z: 641.107- 659.760	

Fig. 1. Contour map of MSQ for Project 3.

In addressing the second question, several approaches are available. One is to conduct a cluster analysis on the $MSQ_{(a,b)}$ grid. This would group similar project maps together from which we may be able to detect some pattern relating to say project type or size. A similar alternative is a factor analysis on the maps together with the type, value and duration

characteristics of interest. This would provide factor loadings that we may be able to interpret in terms of desired groupings. Another possibility is to revert back to the best estimates of *a* and *b* for each project and use multiple regression analysis on each with the project characteristics as independent variables. Major problems exist with each of these alternatives, the cluster and factor analysis only providing a limited connection between the $MSQ_{(a,b)}$ values and the project characteristics of interest, and the regression analysis only utilizing a small part of the potential data present in the maps. Perhaps the biggest problem, however, is that all three approaches, being essentially linear models, rely on euclidian measures of the error term. In our case the error term is clearly non linear, being defined by Equations 1 to 4 above. As a result of these deliberations it was decided to utilize the method of successive approximation in the analysis of project characteristics.

Model 1. Figure 2 shows the intended model conceived prior to analysis. This (T_1, T_2 and T_3) incorporates the three project characteristics of interest – project type, size (small/large), and duration (short/long) – as predictors of the *a* and *b* constants in Equations 1 to 4. For a
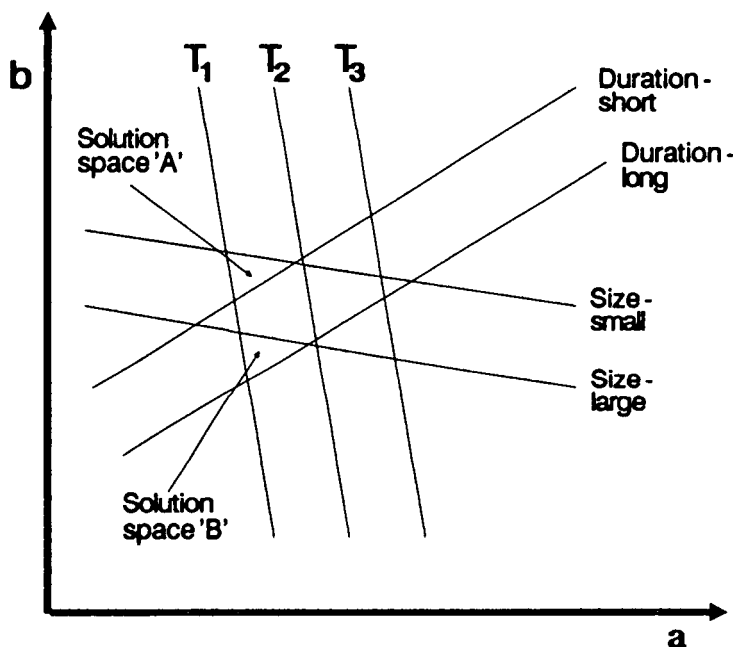

Fig. 2. Prior conceived model.

given project type, the required *a* and *b* values will lie on the line shown. Similarly for a given size and duration, the *a* and *b* values will lie on these (different) lines. Thus the correct *a* and *b* values for a project of given type, size and duration will lie somewhere in the triangle described at the intersection of the three lines (a proper statistical treatment would define the boundaries in probabilistic terms but this is beyond the capabilities of the relatively crude method (and researcher) employed here).

The implications of this in mathematical terms is that

$$a_{Tk} = \alpha_{Tk} T_k + \beta_{Tk} T_k b_{Tk} \tag{11}$$

$$a_V = \alpha_V V + \beta_V V b_V \tag{12}$$

$$a_D = \alpha_D D + \beta_D D b_D \tag{13}$$

where T_k is the type, k, of project and α and β represent the position and slope respectively of the lines.

Now by calculating the $SSQ_{(a,b)}$ over all the projects for a series of b values, the method of successive approximation can be used to estimate the values of α and β for each of the project characteristics.

The resulting α and β estimates for the Kenley–Wilson formulation are given in Table 2. Figure 3 shows the actual model for $V = 100\,000, 500\,000, 1\,000\,000, D = 100, 500, 1000$, and the project types 1 to 4. As the figure shows, the hoped for convex-like structure of Figure 2 is absent, the parallel nature of the lines indicating a general lack of orthogonality among the variables. Similar results were also obtained for the DHSS and Berny–Howes alternatives.

Table 2. Model 1 α and β estimates

Variable	α	β	SSQ_{error}
Value ($V/100\,000$)	0.137	0.206	1 278 142
Duration ($D/100$)	0.014	0.027	1 182 604
Project type 1	0.151	0.336	1 149 053
Project type 2	0.079	0.083	1 329 255
Project type 3	0.004	0.047	1 360 932
Project type 4	0.081	0.416	1 234 987

Model 2. The lack of a convenient convex-like arrangement for model 1, together with the knowledge that the solution space (if it existed) resulting from this method may not neccessarily contain optimal estimates of the required a and b constants, suggests a better model to be the more obvious regression type of approach:

$$a = \alpha_0 + \alpha_1 V + \alpha_2 D + \alpha_3 T_1 + \alpha_4 T_2 + \alpha_5 T_4 \tag{14}$$

$$b = \beta_0 + \beta_1 V + \beta_2 D + \beta_3 T_1 + \beta_4 T_2 + \beta_5 T_4 \tag{15}$$

where V is the contract value, in 100 000 units, D is the contract duration, in 100 day units, and T_1, T_2, and T_4 are dummy variables representing project types 1, 2 and 4 found in the data. It is then proposed to minimize Equation 5 as before by inserting suitable α and β values into Equations 14 and 15 simultaneously (naturally one of the project types, type 3 in this case, must be omitted from the analysis to avoid over specification).

The method of successive approximation was again employed and the necessary α and β estimates obtained. The results for the Kenley–Wilson formulation are given in Table 3 ($SSQ_{error} = 13\,981$). The presence of 12 variables for only 27 cases is not particularly satisfactory – we would normally prefer the number of cases to be at least three to four time
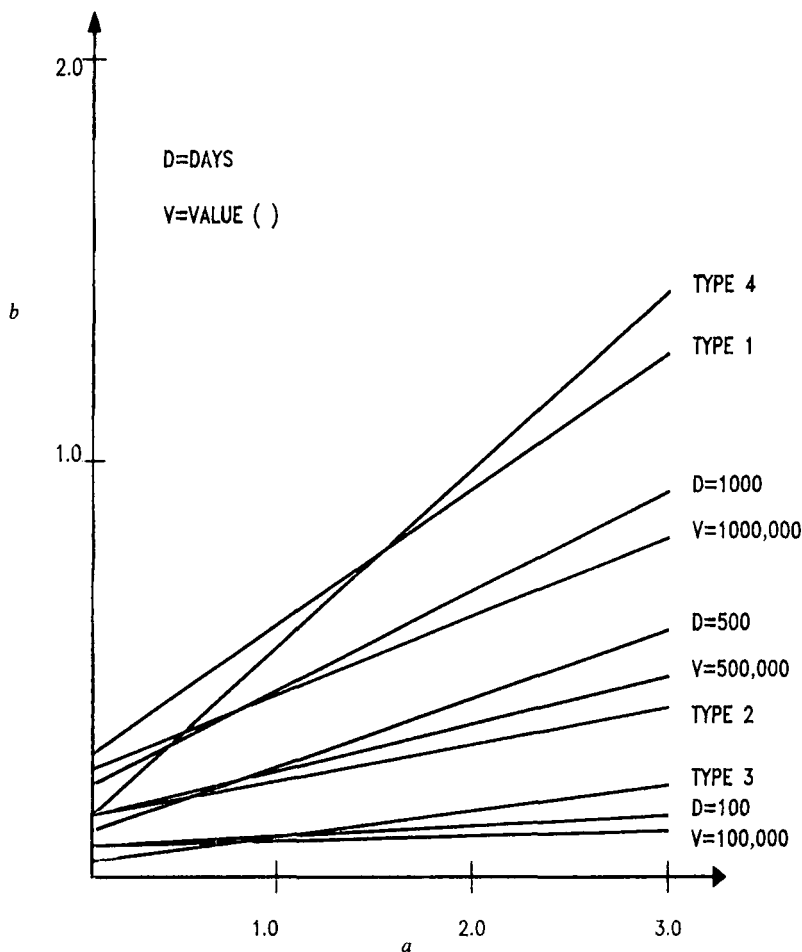
Fig. 3. Actual model.

the number of variables – but the results provide an indication of the general relative influence of the variables selected.

Table 3. Model 2 coefficient estimates (Kenley–Wilson formula)

Coefficient	Estimate	Coefficient	Estimate
α_0	0.126	β_0	0.862
α_1	−0.043	β_1	0.862
α_2	−0.023	β_2	−0.061
α_3	0.525	β_3	0.364
α_4	0.254	β_4	0.453
α_5	0.482	β_5	0.372

Reliability. As the most important application of the analysis is in forecasting values of the constants, 'error of forecast' or 'error rate' estimates are likely to be the most relevant and useful.* Two types of error rate estimates are available – parametric and non parametric. In this case a non parametric error rate estimate was considered to be most appropriate due to 1. the small sample size, as the coefficient estimates are likely to be biased, and 2. although the least squares approach used here is known to be robust for mild departures from the implicit assumptions of the approach, i.e. normally distributed errors, the parametric estimates of the error rates of the approach are not necessarily robust in themselves.

Non parametric error rate estimators, and in particular, resampling methods are known to produce improved estimators of the error rates by appropriate bias correction of the apparent (error of prediction) error (McLachlan, 1987).

Three possible resampling methods were considered; 1. cross validation, where one case is omitted in turn from the model derivation and the resulting coefficients applied to that case; 2. the jackknife method, where one case is omitted in turn from the model derivation and the resulting coefficients applied to the other cases; and 3. the bootstrap method, where the coefficients are used to generate simulated data from which a second set of coefficients are obtained. For predictive applications, the cross validation method 1. has the most intuitive appeal as, with the non time series data of this nature, each error value can be thought of as a real error that may arise in the 'real world' practice of forecasting.

For the data under study, cross validation involves the repeated estimation of regression coefficients for a series of selected case removals, in this case with one project removed for each estimation – a total of 27 repeats. Thus α and β estimates were made as described above but with project 1 removed from the data base. The a and b values were calculated from Equations 14 and 15, and the $MSQ_{(a,b)}$ error obtained for project 1. This was repeated for all the projects in turn. The results for the Kenley–Wilson formulation are shown in Table 4.

Each row in this Table give the results of applying to that case the parameter estimates obtained by analysis of the data for all the other cases. The overall average MSQ is 80.34, a slight improvement over the standard DHSS equivalent of 81.79 (Table 1).

Variable parsimony. Previous experience of cross validation suggests that a reduced number of variables may increase the predictive ability of the model (Skitmore and Patchell, 1990). Ideally we should wish to extract the subset of variables which produces the lowest overall average MSQ. Some of the standard computer based statistical analysis procedures are able to provide this for more routine problems. Unlike standard parametric procedures however, where the significance of variables entering in and exiting from the equation may be readily obtained, the use of non parametric error rate estimates together with the non linear model is a major computational task.

To find the best subset involves the consideration of all possible combination of 1 to 12 variables, i.e.

$$\sum_{r=1}^{13} 13!/[r!(13-r)!] = 8191$$

As each combination involves the computation of a Table similar to Table 4, a task currently

---

*These measures of error concern the relationship between the actual values of a case variable and the value predicted by the model where the model is derived from a data base which does not include that case.

Table 4. Results of cross validation analysis

Case	α_0	α_1	α_2	α_3	α_4	α_5	β_0	β_1	β_2	β_3	β_4	β_5	a	b	MSQ
1	0.04	−0.74	0.09	0.29	−0.04	0.32	0.88	0.71	−0.04	0.28	0.35	0.30	0.00	1.06	8.36
2	0.04	−0.73	0.09	0.29	−0.03	0.32	0.91	0.71	−0.05	0.30	0.36	0.30	0.26	1.34	9.13
3	0.06	−0.77	0.08	0.30	0.00	0.31	0.86	0.73	−0.03	0.27	0.32	0.31	0.23	1.19	17.76
4	0.05	−0.71	0.08	0.30	−0.01	0.33	0.89	0.75	−0.05	0.30	0.41	0.31	0.12	1.44	14.31
5	0.04	−0.75	0.09	0.29	−0.04	0.21	0.88	0.70	−0.04	0.28	0.36	0.30	0.42	1.14	62.15
6	0.04	−0.75	0.09	0.29	−0.04	0.25	0.88	0.71	−0.04	0.28	0.36	0.34	0.45	1.18	54.79
7	0.03	−0.82	0.10	0.27	−0.07	0.44	0.88	0.71	−0.04	0.28	0.35	0.34	0.76	1.15	77.10
8	0.05	−0.70	0.08	0.30	−0.02	0.37	0.87	0.64	−0.03	0.27	0.34	0.24	0.52	1.11	107.68
9	0.07	−0.78	0.08	0.30	0.03	0.31	0.90	0.69	−0.04	0.28	0.38	0.30	0.26	1.27	27.19
10	0.04	−0.77	0.09	0.30	−0.05	0.31	0.82	0.54	0.00	0.30	0.26	0.27	0.70	1.30	32.12
11	0.04	−0.73	0.09	0.30	−0.03	0.32	0.88	0.68	−0.03	0.26	0.35	0.30	0.49	1.19	5.44
12	0.05	−0.75	0.09	0.30	−0.03	0.32	0.87	0.71	−0.03	0.26	0.34	0.30	0.53	1.08	21.60
13	0.01	−0.83	0.11	0.21	−0.10	0.30	0.88	0.70	−0.04	0.27	0.35	0.30	0.33	1.18	156.27
14	0.03	−0.76	0.09	0.34	−0.05	0.31	0.88	0.69	−0.04	0.26	0.36	0.30	0.62	1.16	60.78
15	0.05	−0.85	0.10	0.19	−0.04	0.30	0.88	0.74	−0.04	0.28	0.35	0.31	0.39	1.23	101.65
16	0.02	−0.73	0.09	0.31	−0.02	0.34	0.86	0.70	−0.03	0.29	0.36	0.32	0.05	0.94	18.94
17	0.01	−1.31	0.13	0.27	0.01	0.27	0.88	0.58	−0.02	0.26	0.34	0.28	−0.55	1.26	96.37
18	0.05	−0.74	0.09	0.29	−0.04	0.31	0.90	0.71	−0.04	0.26	0.34	0.28	0.05	1.02	11.21
19	0.05	−0.72	0.09	0.29	−0.04	0.32	0.84	0.66	−0.03	0.32	0.39	0.34	0.07	0.96	69.89
20	0.07	−0.75	0.09	0.28	−0.04	0.30	0.90	0.68	−0.04	0.27	0.35	0.29	0.08	0.93	69.89
21	0.05	−0.69	0.08	0.37	−0.01	0.33	0.87	0.68	−0.03	0.28	0.35	0.30	0.55	1.20	53.53
22	0.12	−0.81	0.09	0.22	−0.09	0.24	0.92	0.65	−0.03	0.25	0.33	0.26	0.16	0.97	148.45
23	0.10	−0.81	0.09	0.23	−0.09	0.25	0.88	0.66	−0.03	0.27	0.35	0.29	0.17	0.93	111.89
24	−0.09	0.24	−0.03	0.69	0.34	0.69	0.90	0.81	−0.06	0.32	0.42	0.33	−0.20	0.69	932.62
25	0.10	−0.82	0.08	0.30	−0.11	0.30	0.89	0.72	−0.04	0.28	0.36	0.30	0.04	1.42	171.67
26	−0.14	−0.54	0.13	0.25	−0.09	0.37	0.84	0.63	−0.02	0.26	0.33	0.30	0.48	1.53	132.05
27	0.02	−0.72	0.09	0.29	−0.07	0.32	0.88	0.69	−0.03	0.28	0.35	0.30	0.15	1.27	23.06
Average															80.34

requiring approximately 6 minutes cpu time, a full analysis of this kind would take over 34 days of computing time!

Several parametric approximating methods are available which may be utilized to reduce the computational burden. The most well known of these are forward, backward, and step wise regression. Forward regression involves the sequential selection and retention of individual variables most contributing to the reduction of the SSQ error term. Backward regression involves the sequential deletion of individual variables least contributing to the reduction in the SSQ error term. Stepwise regression is a combination of both forward and backward regression.

The use of all these methods in parametric problems has a common difficulty. As each variable is added to the equation, the SSQ error always reduces thus making the choice of termination criteria an important issue. With non parametric approaches, such as the cross validation method used in this study, the termination criteria is simply the reduction in SSQ error, as both forward and backward methods produce reductions that seem to tend to an optimal like solution.

Two methods were used in this analysis 1. forward regression to minimal SSQ error, followed by a stepwise regression of backward to minimum, forward to minimum etc; 2. backward regression to minimal SSQ error, followed by a step wise regression of forward to minimum, backward to minimum etc.

For the DHSS formulation, the forward regression resulted in the progressive inclusion of β_0 (99.29), α_3 (75.34), α_5 (65.24), β_4 (60.20), and β_3 (59.98) whilst the backward regression eliminated only α_4 (75.00). The progressive mean square error is given in brackets. The Kenley–Wilson formulation forward regression resulted in the inclusion of β_0 (99.80), α_0 (74.26), β_1 (67.79), α_3 (64.39), α_5 (61.94) and β_4 (61.21), followed by the backward exclusion of α_0 (60.25), whilst the backward regression eliminated α_2 (70.19), α_0 (64.10), α_1 (61.64) and α_4 (59.63). The Berny–Howes formulation forward regression resulted in the inclusion of α_0 (72.21), β_3 (68.72), β_5 (66.50), β_0 (66.09) and α_4 (61.81), whilst the backward regression eliminated β_2 (81.71) and β_1 (71.07). The resulting best models are summarized in Table 5.

Several points of interest emerge from this analysis of the parsimonious models:

Table 5. Summary of regression results

Formula	Lowest MSQ	Variables in equation
DHSS	59.98	$\alpha_3\alpha_5\beta_0\beta_3\beta_4$
Kenley–Wilson	59.63	$\alpha_3\alpha_5\beta_0\beta_1\beta_3\beta_4\beta_5$
Berny–Howes	61.81	$\alpha_0\alpha_4\beta_0\beta_3\beta_5$

1. All models give similar results in terms of reliability with an MSQ improvement over the standard DHSS model (81.79 MSQ) of 26.8, 27.1 and 24.4 per cent for the DHSS, Kenley–Wilson, and Berny–Howes formulations respectively;

2. All models include all the project type measures;

3. Only the Kenley–Wilson model utilizes the contract value measure;

4. None of the models include α_2 or β_2 – the duration related measures.

Revised models. The coefficient estimates of the revised models are shown in Table 6. The standard error values of the coefficients were not available but the mean square errors associated with the models were computed for comparative purposes. The resulting parameter estimates and mean square errors for each of the projects are shown in Table 7. Comparison with Table 1 confirms the improved predictive ability of the models over the standard DHSS model. It is of interest to note that the revised DHSS model, in contrast with the standard DHSS model does not contain any 'size' parameters.

Table 6. Coefficient estimates

Coefficient	DHSS	Kenley–Wilson	Berny–Howes
α_0	—	—	0.268
α_1	—	—	—
α_2	−0.498	—	—
α_3	—	0.533	—
α_4	−0.499	—	1.188
α_5	210.18	0.539	—
β_0	210.18	0.835	0.378
β_1	—	0.455	—
β_2	−372 094	—	—
β_3	−999 790	0.221	−1.737
β_4	—	0.288	—
β_5	—	0.262	−1.750

Conclusions

The DHSS formula, since Hudson, seems to have received little testing on hospital or any other projects. This paper has described its application to some data concerning the expenditure flow on four types of projects – steel-framed low rise buildings, new build housing developments, housing refurbishment projects, and multi-house 'pre-paint' maintenance contracts. Comparison of the formula applied to individual projects using the recommended parameter values with the best parameter values, indicates that tenfold improvements may be made. Using the best parameter values for some other two parameter models produced similar results.

Various possibilities were considered in attempting to predict the best parameter values for the two parameter models based on known characteristics of the project. A multiple linear regression with project value, duration, and type independent variables failed to produce any significant improvement to the standard DHSS formula. However, a reduction in the number of independent variables by means of a cross validation approach produced an approximately 25% improvement on the standard DHSS formula by all the alternative models. Examination of the models derived from this analysis indicated the type of project to be a major contributor to this improvement with contract value being of lesser importance.

The work described in this paper provides an interesting example of the benefits of assessing the efficacy of regression models by reference to forecasts outside the data base of the analysis. In using the cross validation approach, the model is developed and tested

Table 7. Results with the revised formulae

Proj Nr	DHSS			Kenley–Wilson			Berny–Howes		
	a	*b*	MSQ	*a*	b	MSQ	*a*	*b*	MSQ
1	0.00	210.00	7.53	0.00	1.01	7.43	0.27	0.38	10.42
2	0.00	− 372 095	71.29	0.00	1.43	21.90	1.46	0.38	7.00
3	0.00	− 372 095	35.42	0.00	1.21	12.87	1.46	0.38	8.55
4	0.00	− 372 095	18.66	0.00	1.36	9.45	1.46	0.38	14.44
5	− 0.50	210.18	33.95	0.54	1.13	32.91	0.27	− 1.37	35.62
6	− 0.50	210.18	34.40	0.54	1.13	34.40	0.27	− 1.37	38.44
7	− 0.50	210.18	28.42	0.54	1.14	29.53	0.27	− 1.37	29.07
8	− 0.50	210.18	122.68	0.54	1.14	107.35	0.27	− 1.37	107.48
9	0.00	− 372 095	17.49	0.00	1.24	10.80	1.46	0.38	22.91
10	− 0.50	− 1 000 000	12.03	0.53	1.20	26.64	0.27	− 1.36	17.38
11	− 0.50	− 1 000 000	19.03	0.53	1.18	7.68	0.27	− 1.36	12.12
12	− 0.50	− 1 000 000	23.73	0.53	1.08	21.97	0.27	− 1.36	18.85
13	− 0.50	− 1 000 000	73.99	0.53	1.12	69.37	0.27	− 1.36	71.67
14	− 0.50	− 1 000 000	53.69	0.53	1.18	40.72	0.27	− 1.36	45.24
15	− 0.50	− 1 000 000	56.32	0.53	1.21	56.90	0.27	− 1.36	53.82
16	0.00	210.18	21.43	0.00	0.93	27.33	0.27	0.38	14.80
17	0:00	210.18	14.68	0.00	1.21	1.41	0.27	0.38	9.28
18	0.00	210.18	8.69	0.00	0.98	6.19	0.27	0.38	14.95
19	0.00	210.18	55.38	0.00	0.98	61.01	0.27	0.38	44.73
20	0.00	210.18	62.93	0.00	0.89	37.86	0.27	0.38	85.34
21	− 0.50	− 1 000 000	61.10	0.53	1.17	52.32	0.27	− 1.36	55.71
22	0.00	210.18	95.26	0.00	0.92	83.28	0.27	0.38	116.20
23	0.00	210.18	58.31	0.00	0.92	56.95	0.27	0.38	70.80
24	0:00	210.18	727.09	0.00	0.85	696.30	0.27	0.38	705.90
25	0.00	− 372 095	215.02	0.00	1.42	189.33	1.46	0.38	107.42
26	0.00	− 372 095	59.48	0.00	1.51	15.76	1.46	0.38	52.21
27	0.00	− 372 095	61.83	0.00	1.25	44.77	1.46	0.38	19.20
Average			67.73			53.04			49.92

simultaneously on this criteria. This clearly has considerable intuitive appeal. In addition, the fortuitous convex-like nature of the cross validation procedure when applied to these data provides an unexpected and welcome means of automatic variable parsimony. Future work will extend the method to more complex multi-variable models incorporating time-cost overruns, seasonal variations, etc.

References

Berny, J. and Howes, R. (1982) Project management control using real-time budgeting and forecasting models, *Construction Papers*, **2**, 19–40.

Hudson, K.W. (1978) DHSS expenditure forecasting method, *Chartered Surveyor: Building and Quantity Surveying Quarterly*, **5**, 42–5.

Kenley, R. and Wilson, O.D. (1989) A construction project net case flow model, *Construction Management and Economics*, **7**, 3–18.

Khosrowshahi, F. (1991) Simulation of expenditure patterns of construction projects, *Construction Management and Economics*, **9**, 113–32.

McLachlan, G.J. (1987) Error rate estimation in discriminant analysis: recent advances. In *Advances in Multivariate Statistical Analysis*, (edited by A.K. Gupta), D. Reidel Publishing Company, pp. 233–52.

Miskawi, Z. (1989) An S-curve equation for project control, *Construction Management and Economics*, **7**, 115–24.

Peer, S. (1982) Application of cost-flow forecasting models. *Proceedings of the American Society of Civil Engineers*, Journal of the Construction Division, Vol. 108, no CO2, June, pp. 266–31.

Skitmore, R.M. and Patchell, B. (1990) Developments in contract price forecasting and bidding techniques. In *Quantity Surveying Techniques: New Directions*, (Edited by P.S. Brandon), BSP Professional Books, pp. 75–120.

Tucker, S.N. (1986) Formulating construction cash flow curves using a reliability theory analogy, *Construction Management and Economics*, **4**, 179–88.

Tucker, S.N. (1988) A single alternative formula for the Department of Health and Social Security S-curves, *Construction Management and Economics*, **6**, 12–23.

Appendix

Case data (Contract values rebased to 1974 prices by means of the RICS Building Cost Information Service Tender Price Index)

Project	Value	Days	Type	v^a	d^b	v	d	v	d	v	d
1	379 996	266	3	8.68	11.28	23.64	24.44	39.77	35.71	44.18	46.24
				54.43	58.65	70.77	72.56	84.81	83.08	97.36	97.74
				100.00	100.00						
2	682 681	863	2	6.16	5.56	7.06	8.23	8.64	12.28	10.08	15.64
				12.56	19.70	15.19	22.60	18.77	25.38	23.34	30.01
				29.60	33.49	36.47	37.43	42.06	40.67	45.52	44.38
				50.75	47.28	60.05	51.33	63.92	54.92	69.81	58.52
				73.93	61.88	78.18	64.89	80.88	68.37	84.93	72.31
				88.35	75.43	91.49	79.26	94.03	83.55	94.79	86.10
				97.30	89.92	98.76	93.05	100.00	100.00		
3	1 888 123	379	2	4.64	6.07	7.68	11.87	16.60	20.32	25.42	28.50
				34.26	35.88	47.47	46.97	54.67	54.35	65.41	63.32
				81.23	71.24	89.38	78.36	98.38	87.86	99.56	96.83
				100.00	100.00						
4	526 024	552	2	4.26	7.79	8.03	11.59	13.70	18.12	21.98	23.19
				28.97	28.44	34.09	33.51	38.96	41.12	43.07	45.47
				49.55	50.72	56.83	54.89	63.43	62.50	71.71	68.66
				79.25	72.83	86.45	77.72	91.56	84.06	95.71	89.31
				97.77	94.38	100.00	100.00				
5	82 003	255	4	12.65	9.98	40.66	25.06	53.18	32.59	60.67	42.13
				69.49	51.22	77.16	57.43	86.14	70.73	88.41	79.16
				90.88	85.81	98.28	90.91	100.00	100.00		

[a] v = percentage value complete.

[b] d = percentage of time complete.

Project	Value	Days	Type	v^a	d^b	v	d	v	d	v	d
6	80 108	255	4	7.06	9.55	24.93	16.99	33.90	23.99	47.84	30.57
				57.39	38.85	63.01	45.22	75.40	51.17	79.47	70.06
				84.23	77.92	89.79	84.93	91.10	92.99	100.00	100.00
7	95 305	375	4	8.62	7.73	17.43	14.71	27.17	21.95	36.26	28.93
				42.64	39.15	53.52	44.64	62.99	51.62	70.43	57.11
				74.28	71.82	75.59	76.06	93.25	82.79	100.00	100.00
8	91 355	204	4	6.95	10.40	18.20	27.17	38.85	38.73	54.38	47.98
				80.35	67.63	96.40	84.97	100.00	100.00		
9	247 347	442	2	4.67	5.20	10.54	11.99	19.84	17.65	25.03	26.24
				32.31	33.26	37.43	40.27	45.70	48.42	51.40	55.43
				63.10	61.76	76.95	69.91	84.84	76.02	89.32	84.16
				94.40	89.37	98.18	95.93	100.00	100.00		
10	32 3717	650	1	8.01	6.36	14.77	10.70	23.20	15.04	29.94	18.29
				41.37	24.96	46.96	29.30	51.24	33.64	54.26	38.14
				58.63	43.72	61.17	47.13	64.22	51.94	67.93	56.74
				71.88	62.33	75.85	67.13	79.07	71.47	84.58	76.59
				88.43	82.48	91.62	85.89	97.14	89.92	99.54	95.50
				100.00	100.00						
11	272 052	409	1	6.96	8.56	16.33	15.65	26.33	22.25	37.82	34.23
				47.41	39.36	56.36	44.74	67.24	55.26	74.31	61.61
				81.29	68.46	90.73	77.26	96.05	85.57	97.90	92.42
				100.00	100.00						
12	57 493	262	1	13.55	13.74	25.53	22.52	44.79	35.50	57.03	48.47
				77.26	61.45	85.11	70.99	85.95	83.97	100.00	100.00
13	143 440	214	1	22.84	14.49	41.53	27.57	65.18	42.52	81.39	57.01
				91.82	75.23	99.62	88.32	100.00	100.00		
14	280 346	496	1	1.65	3.43	7.88	9.48	13.78	16.33	25.81	25.40
				35.33	31.25	41.05	36.69	45.94	44.15	50.78	49.80
				61.32	56.85	71.65	62.90	78.74	68.35	90.67	75.20
				96.67	81.25	98.08	87.50	98.77	93.95	100.00	100.00
15	342 743	465	1	13.12	6.24	19.92	12.04	28.93	20.00	40.52	27.10
				51.17	33.12	60.12	39.35	68.50	48.17	74.06	53.55
				83.53	60.22	88.36	65.16	91.24	73.12	96.92	80.43
				97.62	87.10	99.14	96.34	100.00	100.00		
16	217 335	212	3	18.38	18.40	33.21	28.30	53.03	48.10	66.01	61.32
				81.18	75.94	94.30	91.04	100.00	100.00		
17	832 869	405	3	1.68	1.98	7.18	10.62	13.86	18.77	23.02	27.41
				28.67	34.32	39.60	41.23	48.38	48.15	58.00	56.79
				68.79	63.70	77.19	72.35	84.55	79.26	89.96	86.17
				96.35	94.81	100.00	100.00				
18	316 025	267	3	9.40	7.87	19.42	16.48	30.02	28.84	44.70	42.70
				55.23	53.56	62.20	65.54	71.65	76.03	86.80	89.51
				100.00	100.00						
19	327 272	297	3	5.65	7.74	19.54	26.60	23.20	34.34	46.39	43.77
				57.60	55.56	84.58	74.41	94.48	87.54	100.00	100.00
20	117 184	121	3	27.98	29.75	60.49	66.94	81.23	90.08	100.00	100.00
21	260 746	394	1	9.24	9.39	13.72	16.75	21.96	24.87	32.85	32.23
				40.72	39.59	52.86	47.97	60.90	55.08	67.65	63.96
				76.25	70.30	85.42	77.16	91.67	85.28	94.96	93.15
				100.00	100.00						
22	197 198	221	3	5.12	10.86	23.12	19.00	32.01	42.53	45.16	57.47
				56.87	69.68	70.85	77.83	86.40	92.76	100.00	100.00
23	195 997	242	3	2.50	9.92	21.34	17.36	30.44	38.84	40.83	52.48
				56.70	63.64	67.45	71.07	81.59	84.71	100.00	100.00
24	43 960	55	3	22.31	15.36	46.64	21.61	70.55	39.06	84.12	46.61
				87.32	65.36	89.69	81.25	100.00	100.00		

$^a v$ = percentage value complete.

$^b d$ = percentage of time complete.

Project	Value	Days	Type	v^a	d^b	v	d	v	d	v	d
25	656 715	766	2	2.37	2.74	7.93	6.79	10.20	11.88	14.29	15.54
				20.73	18.80	25.08	21.28	32.70	27.02	36.34	28.72
				42.95	34.07	52.05	38.64	57.78	42.43	64.99	47.52
				70.31	49.87	76.46	54.05	79.41	56.27	83.64	62.14
				88.41	65.80	90.93	68.93	94.81	74.93	96.04	78.20
				97.09	82.64	97.43	86.42	98.96	90.86	99.45	96.34
				100.00	100.00						
26	841 063	923	2	1.02	3.36	3.34	5.96	4.43	9.75	6.32	12.89
				8.53	16.58	11.09	19.50	12.98	23.19	14.11	27.09
				18.32	30.01	21.07	32.39	26.50	36.08	29.37	39.44
				31.79	42.25	37.56	46.15	42.58	49.30	46.66	52.98
				53.85	56.12	59.01	59.15	61.44	63.27	67.09	66.31
				73.84	69.56	79.76	72.59	82.71	75.73	87.54	78.76
				90.65	81.58	93.68	85.48	95.42	88.62	97.25	91.55
				98.99	95.23	100.00	100.00				
27	279 650	436	2	4.85	8.72	18.09	15.37	26.40	24.08	30.03	29.59
				37.51	37.16	48.29	40.37	55.45	50.69	61.46	58.49
				71.10	65.14	83.72	69.95	94.91	78.67	96.04	85.32
				97.79	93.81	100.00	100.00				

$^a v$ = percentage value complete.
$^b d$ = percentage of time complete.