

PREDICTION OF SEASON AVERAGE POL % CANE FOR A MILL

By R. G. HOEKSTRA and S. M. BAKER

Operations Research Department, Hulett's Sugar Limited, Mount Edgecombe, Natal

Abstract

The introduction of a relative cane payment system initiated the need for a method of predicting the average season pol % cane for a mill. A system of forecasting was developed, consisting of the following 2 main steps: (a) Development of a pre-season estimate before commencement of crushing, based on the weighted sum of the double exponentially smoothed value of previous seasons' averages. (b) Development of a within-season prediction once the current season is under way, based on the weighted sum of: The pre-season estimate determined in (a), the to-date average pol % cane value and the difference between the latest month's pol % cane and the to-date average pol % cane values. The relative values of these weightings change from month to month during the course of the season. A computer was used both in the development of the method and for routine forecasting.

Introduction

With the introduction of the relative cane payment system, the price per ton of pol paid out to the individual grower is based on the estimate of the season average pol % cane for the mill. Because subsequent adjustments are required to reflect the actual mill season average pol, it became essential to develop a forecasting system which would, as far as possible, provide accurate forecasts of the season average pol % cane for each mill.

This paper describes the development of prediction formulae for the season average pol % cane for each of the 5 Hulett mills: Empangeni, Felixton, Amatikulu, Darnall and Mount Edgecombe, using historical data going back to the 1951/52 season.

Source of data for the analysis

The most easily obtainable source of data for going back over a number of years, where month by month pol % cane values were available, was the monthly mill returns, as reported to the Sugar Milling Research Institute.

The following modifications had to be made:

- (a) In years prior to the 1973/74 season, sucrose % cane figures were recorded, whereas nowadays it is pol % cane. For greater consistency, the older records had to be converted to pol % cane by multiplication by the relevant correction factor for each particular mill and season.
- (b) For the relative cane payment system, the average pol % cane figure for the quota growers of the mill is required, whereas the data from which the forecasting system was developed referred to crush mill figures, which are not necessarily the same as the quota mill figures, because of diversions which regularly occur among all but the more remote mills in the South African sugar industry. The values of the parameters in the prediction equations were determined using crush mill data because of its easier availability, but in the implementation of the forecasting system, the input data used was that of the quota mills.

Basic concept

In principle, there are 2 different situations in which an average pol % cane estimate for the season is required:

- (1) A forecast made at the start of the season, before the results for any of that season's operations are known, using only data from previous seasons. This is called the *pre-season estimate*.
- (2) Once some pol % cane figures for the present season are known, these results can be used for improving the prediction of season average pol % cane. This will be known as the *within-season estimate*.

Development of the pre-season estimate

This estimate is made at or before commencement of crushing for the current season, when no information is yet available on how the pol curve for the present season is actually behaving.

The following information for basing the estimate on was considered:

(a) Average pol % cane figures for previous seasons:

Various ways of using this historical information were available namely:

(i) Fitting of a line to past data by linear regression analysis:

The drawback of this method is that the position of the fitted line is influenced just as much by values going far back as by more recent values, i.e. not sufficient weight is given to more recent data. Also, at the end of each additional season, the line must be re-fitted. These objections, although not insurmountable, complicate the matter.

(ii) Using a moving average of say the last 4 or 5 years:

This gives weight to more recent data, but has the drawback that any data further back is abruptly cut off and completely ignored.

(iii) Exponential smoothing¹:

This technique is somewhat analogous to using a moving average, but with the difference that more recent observations carry a higher weight, and the weight of older observations tapers off the further back they go, but without an abrupt cut-off as with a moving average.

Simple exponential smoothing has the drawback that, when the observations show a trend with time, either upwards or downwards, the smoothed value will lag behind the true trend. With double exponential smoothing, the slope of the trend line is also updated in an exponentially smoothed manner, thus reducing the lag.

The concepts of simple and double exponential smoothing are explained in Appendix 1.

(b) Rainfall up to the start of the season:

Although it was found that the pol % cane recorded for May showed some correlation with the weighted individual monthly rainfalls from October of the previous year up to the April preceding that May, there was poor correlation between the average pol for the whole of the new season and these rainfalls.

In view of its small contribution to the accuracy of the pre-season estimate, it was considered preferable rather to keep the pre-season prediction simple by leaving rainfall out altogether.

(c) Length of season:

Although length of season does have an influence on average pol % cane, at the start of the season it is usually not known when milling will stop, as this date depends mainly on unforeseeable events which occur during the season, such as climatic conditions and milling problems. Incorporation of this variable into the prediction equation would have served no useful purpose.

It was therefore decided to let the pre-season estimate be given by the latest double exponentially smoothed value of average pol % cane.

The equation had the form:

$$\hat{e}(n+1) = p_{\alpha}^{(2)}(n)$$

where $\hat{e}(n+1)$ = Pre-season estimate for season (n+1)

$p_{\alpha}^{(2)}(n)$ = Double exponentially smoothed value of average pol % cane using data up to season n inclusive, for a smoothing constant α .

The method used for determining the best value of α is described in Appendix 2. A smoothing value of $\alpha = 0,1$ was used for all mills.

Figure 1 shows the pre-season estimated values of average pol % cane for each season for Amatikulu, together with the actual values subsequently recorded.

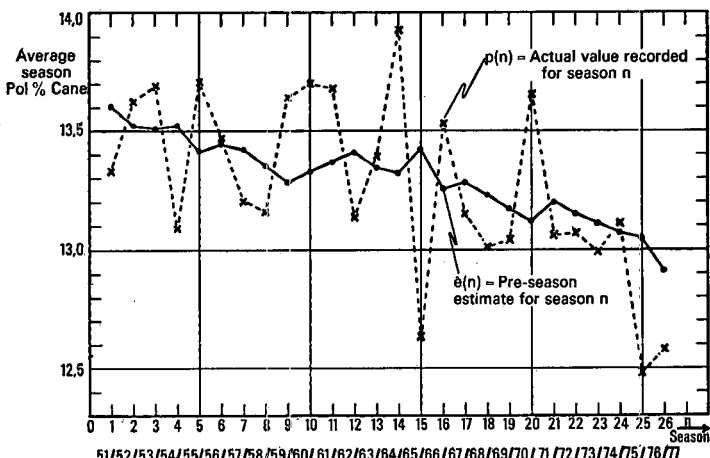


FIGURE 1 Operation of Pre-season Average Pol % Cane Prediction method for Amatikulu Mill.

Development of the within-season estimate

Once the mill is crushing and information becomes available on the pol content of the cane being crushed, it is of course possible to improve on the original pre-season estimate. The further the season progresses, the greater is the proportion of the pol milled which is no longer changeable by whatever might happen later in the season, and the more accurate the forecast becomes.

It was decided to let the within-season average pol prediction equation be a weighted sum of the pre-season estimate, the to-date average pol % cane value and the difference between the latest month's pol % cane and the to-date average pol % cane values, i.e. in the following form:

$$\hat{p} = E.\hat{e} + D.d + M.(m - d) \tag{1}$$

where \hat{p} = within-season estimate, which is the value to be predicted,

\hat{e} = previously calculated pre-season estimate and E is its weight,

d = to-date average pol % cane value and D is its weight, and

m = latest month's pol % cane value and M the weight of (m - d).

The reasons for this choice of variables are:

(a) Pre-season estimate:

At the start of the season, the values obtained from crushing are still too uncertain to attach too much significance to them as predictors of the eventual average season pol % cane, and a fair weight still has to be given to the pre-season estimate \hat{e} . The value of \hat{e} remains fixed throughout the season.

(b) To-date average pol % cane:

This value d becomes more and more important as the season progresses, because whatever pol has been produced to-date has become an unalterable fact.

(c) Pol % cane of latest month — To-date average pol % cane:

This difference (m - d) between latest and average values acts as a detector for any recent trends in the pol % cane curve.

The effect of recent rainfall was also investigated. By using the exponentially smoothed rainfall applicable to the latest month, the rainfall in earlier months could be brought into the analysis, although with a lower weighting. However, the analysis showed that, as in the case of the pre-season estimate, historical rainfall was a rather insignificant predictor.

Because the relative weightings of the various independent variables and therefore the values of their respective weights E, M and D change during the course of the season, it was necessary to let each of these weights be a function of time t, where t is chosen on some numerical time scale.

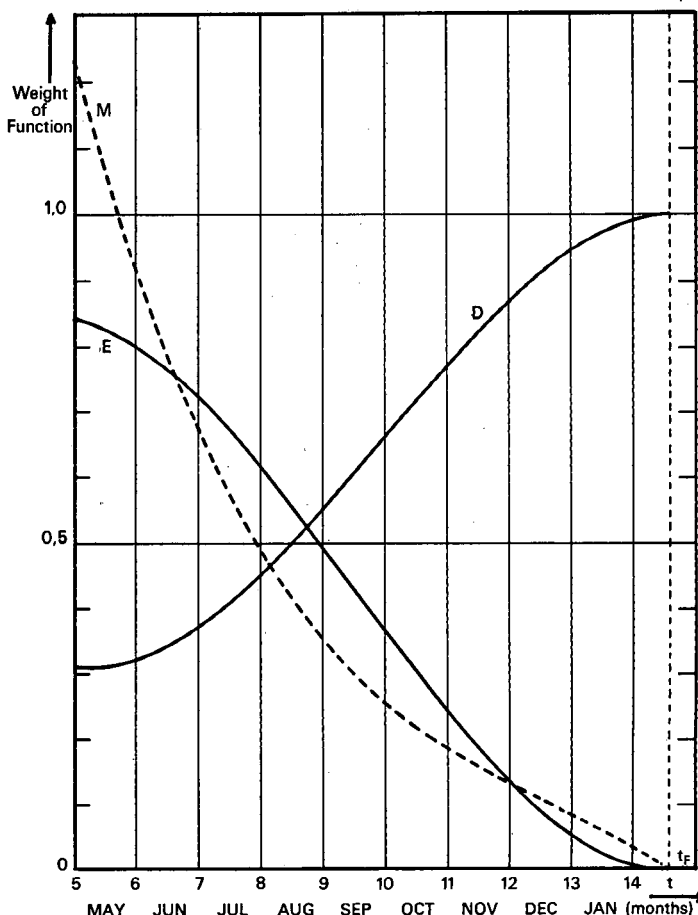


FIGURE 2 Weights of pre-season estimates (E), to-date average pol % cane (D) and (latest month's pol % cane — to-date average pol % cane) (M) as functions of time t, for Amatikulu Mill and an ending date of 17th February.

At the end or anticipated end date of the season, when $t = t_F$ say, we have the boundary condition that:

To-date average pol % cane = predicted season average pol % cane, which means $d = \hat{p}$.

Using equation (1):

$$d = E.\hat{e} + D.d + M. (m - d) \quad (2)$$

In Appendix 3 it is deduced from equation (2) that, at the end of the season,

$$E = M = 0$$

and $D = 1$.

The weight functions E and M should therefore tend towards 0 and D towards 1 as the season approaches its end.

How the structure of these functions was developed and the values of the parameters determined by multiple linear regression analysis is described in Appendix 3.

For Amatikulu mill, the within-season prediction equation was

$$\hat{p} = (0.008509 + 0.02461t - 0.001744t^2) (t_F - t) \hat{e} + (0.29051 - 0.04121t + 0.001778t^2) (t_F - t) (m - d) + [(0.01561 - 0.028035t + 0.0018418t^2 - 0.0001373t (t_F - t)) (t_F - t) + 1] d$$

Figure 2 illustrates how the values of the weights E, M and D change with time of season for the case of Amatikulu mill with an end-of-season date being 17th February, i.e. $t_F = 14.6$ on the chosen time scale.

Implementation

Having determined the weight functions for each of the mills, the month-by-month updating of the prediction is done by computer. Figure 3 shows the print-out where the prediction equation had been applied to the past 1976/77 season. The results for the 1976/77 season were not included in the data used for developing the prediction equations, so that the predictions shown were in no way "influenced by their own actual values".

It can be observed how the predicted value moves closer to the final value for the season and the confidence limits of the prediction become narrower as the season progresses.

The actual final values for the 1976/77 season were:

Mill	Finishing date	Average Season Pol % Cane
EM	13/2/77	12,48
FX	21/2/77	12,27
AK	17/2/77	12,58
DL	13/2/77	12,71
ME	30/1/77	12,20

HULETT'S SUGAR LTD. - PREDICTION OF SEASONAL POL % CANE. 03/05/77

FORECAST AT END OF MONTH	PRE-SEASON ESTIMATE	APRIL	MAY	JUNE	WITHIN-SEASON ESTIMATE		OCT	NOV	DEC	JAN	
					JULY	AUG					
EMPANGENI											
POL % CANE		0,00	10,71	12,17	12,78	13,04	13,74	13,44	12,92	12,44	11,70
POL % CANE TO-DATE		0,00	10,71	11,47	11,96	12,20	12,55	12,67	12,70	12,67	12,57
EXPECTED FINISHING DATE		16/01/77	23/01/77	27/01/77	29/01/77	02/02/77	09/02/77	09/02/77	14/02/77	14/02/77	14/02/77
FORECAST	12,85	0,00	12,31	12,60	12,62	12,52	12,66	12,56	12,50	12,50	12,49
90% CONF. LIMITS - UPPER	13,68	0,00	12,90	13,11	13,04	12,91	12,99	12,83	12,70	12,63	12,55
LOWER	12,03	0,00	11,72	12,09	12,20	12,13	12,33	12,30	12,29	12,37	12,43
FELIKTON											
POL % CANE		0,00	10,60	11,69	12,52	12,87	13,25	13,33	12,81	12,30	11,64
POL % CANE TO-DATE		0,00	10,60	11,11	11,68	11,97	12,27	12,42	12,47	12,45	12,37
EXPECTED FINISHING DATE		23/01/77	30/01/77	03/02/77	04/02/77	09/02/77	09/02/77	09/02/77	14/02/77	14/02/77	14/02/77
FORECAST	12,61	0,00	12,16	12,26	12,39	12,41	12,43	12,40	12,30	12,27	12,27
90% CONF. LIMITS - UPPER	13,22	0,00	12,51	12,60	12,70	12,68	12,68	12,61	12,48	12,41	12,36
LOWER	12,00	0,00	11,81	11,91	12,09	12,13	12,19	12,18	12,12	12,13	12,19
AMATIKULU											
POL % CANE		0,00	11,16	12,50	13,22	13,18	13,47	13,37	12,86	12,47	12,07
POL % CANE TO-DATE		0,00	11,16	11,68	12,22	12,42	12,65	12,75	12,76	12,72	12,66
EXPECTED FINISHING DATE		16/01/77	23/01/77	27/01/77	29/01/77	02/02/77	09/02/77	09/02/77	14/02/77	14/02/77	16/02/77
FORECAST	13,01	0,00	12,74	13,10	13,00	12,71	12,69	12,65	12,59	12,58	12,58
90% CONF. LIMITS - UPPER	13,60	0,00	13,28	13,59	13,41	13,09	13,01	12,91	12,78	12,71	12,64
LOWER	12,42	0,00	12,19	12,61	12,59	12,33	12,37	12,39	12,39	12,45	12,52
DARNALL											
POL % CANE		0,00	11,41	12,48	13,07	13,32	13,26	13,45	13,09	12,71	12,44
POL % CANE TO-DATE		0,00	11,41	11,79	12,26	12,50	12,66	12,78	12,82	12,81	12,77
EXPECTED FINISHING DATE		16/01/77	23/01/77	27/01/77	29/01/77	02/02/77	26/01/77	26/01/77	11/02/77	11/02/77	12/02/77
FORECAST	13,12	0,00	12,83	12,90	12,92	12,83	12,77	12,79	12,67	12,66	12,69
90% CONF. LIMITS - UPPER	13,84	0,00	13,46	13,46	13,38	13,27	13,08	13,02	12,88	12,80	12,75
LOWER	12,41	0,00	12,20	12,34	12,45	12,40	12,46	12,56	12,45	12,53	12,64
MT.EDGEcombe											
POL % CANE		0,00	11,00	11,80	12,18	12,78	12,81	12,67	12,34	12,48	11,83
POL % CANE TO-DATE		0,00	11,00	11,36	11,66	11,92	12,10	12,18	12,20	12,23	12,20
EXPECTED FINISHING DATE		31/12/77	15/01/77	15/01/77	15/01/77	15/01/77	15/01/77	26/01/77	26/01/77	26/01/77	26/01/77
FORECAST	12,76	0,00	12,46	12,69	12,36	12,35	12,17	12,08	12,06	12,17	0,00
90% CONF. LIMITS - UPPER	13,57	0,00	13,33	13,43	12,95	12,82	12,52	12,34	12,22	12,24	0,00
LOWER	11,95	0,00	11,58	11,94	11,77	11,89	11,82	11,82	11,90	12,11	0,00

FIGURE 3 Example of print-out of routine prediction program, applied to 1976/77 season.

As soon as the results of the 1st month of the season are known, and a within-season estimate can be made, the confidence limits improve considerably over those of the pre-season estimate. Because the coefficient functions E, M and D in equation (2) are continuous functions of time, there is no need to wait for results up to the end of May before using the within-season estimate to obtain a more accurate basis for cane payments.

It is not desirable to up-date the season average pol % cane basis of cane payments every month, as it would cause excessive administrative work, and any cycling in the predicted value during the season would result in reducing payments one month, but increasing them again for the next. It is therefore suggested that the pol % cane value upon which the monthly payments are based should only be changed once it falls outside some chosen confidence limits of the latest estimate.

Acknowledgement

Our thanks go to Trevor Ireland for his contribution in programming the computer and in arranging the large number of runs for analysing the data.

REFERENCES

1. Johnson, L. A., and Montgomery, D. C. (1973). Operations Research in Production Planning, Scheduling and Inventory Control, John Wiley & Sons, Inc., 416-422.

Appendix 1

Exponential smoothing

Let $p(n)$ = Actual average pol % cane value for season n ,
 $n = 1, 2, \dots$
 α = Smoothing constant, where $0 < \alpha < 1$.

Define 1st order smoothed statistic $S_\alpha(n)$ for season n by the recursive relationship⁽¹⁾:

$$S_\alpha(n) = \alpha \cdot p(n) + (1 - \alpha) S_\alpha(n - 1), \quad (3)$$

$n = 1, 2, \dots$

The exponentially smoothed value $p_\alpha(n)$ for season n , using simple exponential smoothing as above, is simply

$$p_\alpha(n) = S_\alpha(n)$$

The drawback of simple exponential smoothing is that, if at any time the true trend line has a slope = b , say, the exponentially smoothed value $p_\alpha(n)$ would follow it by a lag of $(1 - \alpha) b / \alpha$ time periods (seasons). With double exponential smoothing, the slope as well as the position of the trend line are changed in accordance with the latest data value, and there is no lagging effect.

The 2nd order smoothed statistic $S_\alpha^{(2)}(n)$ for season n is recursively defined by

$$S_\alpha^{(2)}(n) = \alpha \cdot S_\alpha(n) + (1 - \alpha) S_\alpha^{(2)}(n - 1), \quad (4)$$

$n = 1, 2, \dots$

where $S_\alpha(n)$ and α have the same meanings as before.

The double exponentially smoothed value $p^{(2)}(n)$ for season n is given by

$$p_\alpha^{(2)}(n) = 2S_\alpha(n) - S_\alpha^{(2)}(n), \quad (5)$$

$n = 1, 2, \dots$

To start the above relationships off at $n = 1$, suitable starting values $S_\alpha(0)$ and $S_\alpha^{(2)}(0)$ for the right-hand sides of the equations (3) and (4) have to be determined. A straight line is fitted to all the data either by hand or by linear regression.

Letting its slope be B and its intercept at $n = 0$ be A , it can be shown that

$$\begin{aligned} S_\alpha(0) &= A - (1 - \alpha) B / \alpha \\ S_\alpha^{(2)}(0) &= A - 2(1 - \alpha) B / \alpha \end{aligned}$$

Appendix 2

Determination of the smoothing constant for the pre-season estimate

- Let $p(n)$ = Average pol % cane recorded for season n
 $p_\alpha^{(2)}(n)$ = Double exponentially smoothed value of average pol % cane for season n , for a smoothing constant α .
 $\hat{e}(n + 1)$ = Pre-season estimate for season $(n + 1)$, based on historical information up to season n inclusive.

The double exponential smoothing technique is based on the assumption that the trend line to-date is a straight line, which could but not necessarily would have a slope. Normally, an extrapolation of the latest double exponentially smoothed value into the future would be made along the slope of the fitted line, but in the case of average seasonal pol % cane there did not seem any valid reason why any declining trend which might have been observed in recent years should necessarily persist into the future.

For this reason it was decided to horizontally extrapolate the latest double exponentially smoothed value to provide the pre-season estimate for the next season, i.e.

$$\hat{e}(n + 1) = p_\alpha^{(2)}(n) \quad n = 1, 2, \dots \quad (6)$$

The task was to determine the value of smoothing constant α which would give the best (lowest) sum of squared errors between pre-season predicted value $\hat{e}(n + 1)$ and actual value $p(n + 1)$ for the next season $(n + 1)$,

i.e. the value of α which would minimize

$$\sum_n \left(p(n + 1) - p_\alpha^{(2)}(n) \right)^2$$

For a given value of α , the series of values $\{p_\alpha^{(2)}(n)\}$ was calculated for the range of n , up to the penultimate season for which the results were known. The values of $\{p(n + 1)\}$ were already known, and the sum of squared deviations was calculated. This process was repeated for various values of α and for each mill.

In calculating the series of values $\{p_\alpha^{(2)}(n)\}$, it was found necessary in the determination of $S_\alpha(0)$ and $S_\alpha^{(2)}(0)$ to slightly offset the values of A and B (refer to Appendix 1) from what they would have been for a linear regression fit to the historical data, thereby testing the ability of the chosen value of α to correct the initially incorrectly chosen trend, without tending too much to follow random deviations of the recorded $\{p(n)\}$.

Generally, the best results were obtained around $\alpha = 0.1$, which was chosen as the value of smoothing constant for all the mills.

Using that value of α , the series $\{p_\alpha^{(2)}(n)\}$ was re-calculated, this time without the aforementioned off-setting, thus providing the pre-season estimate for the next season, as per equation (6).

Appendix 3

Derivation of within-season prediction equation

It was already decided that the weights E, M and D of the weight functions in equation (2) should be functions of time t , i.e.

$$E = \sigma_E(t), \quad M = \sigma_M(t), \quad D = \sigma_D(t),$$

Equation (2) thus reads:

$$\hat{p} = \varphi_E(t) \cdot \hat{e} + \varphi_D(t) \cdot d + \varphi_M(t) \cdot (m - d) \quad (7)$$

- where \hat{p} = Estimate of season average pol % cane
 \hat{e} = Pre-season estimate of pol % cane
 d = To-date season average pol % cane
 m = Pol % cane for month preceding present month
 t = Present date during season on basis:
 $t = 5.0$ for 1st May,
 $t = 12.5$ for 16th December,
 $t = 13.0$ for 1st January,
 $t = 14.25$ for 8th February, etc.
 t_F = Finishing date of season, on same time scale as t .

Each of these functions $\varphi_E(t)$, etc. must conform to the following conditions:

- (i) It must be linear in its parameters, so as to allow determination of these by multiple linear regression analysis.
- (ii) It must conform to the boundary condition that, if $t = t_F$, then $\hat{p} = d$

Equation (1) then becomes:

$$d = \varphi_E(t_F) \cdot \hat{e} + \varphi_D(t_F) \cdot d + \varphi_M(t_F) \cdot (m - d) \quad (8)$$

Relationship (8) is an identity of each of e , d and $(m - d)$, i.e. it must hold for whatever the values of these variables are at time t_F . From this it follows that:

$$\varphi_E(t_F) = \varphi_M(t_F) = 0,$$

$$\varphi_D(t_F) = 1$$

Conditions (i) and (ii) are both satisfied if:

$$\begin{aligned} \varphi_E(t) &= E_0 \cdot (t_F - t) + E_1 \cdot t \cdot (t_F - t) + E_2 \cdot t^2 \cdot (t_F - t) \\ \varphi_M(t) &= M_0 \cdot (t_F - t) + M_1 \cdot t \cdot (t_F - t) + M_2 \cdot t^2 \cdot (t_F - t) \\ \varphi_D(t) &= 1 + D_0 \cdot (t_F - t) + D_1 \cdot t \cdot (t_F - t) \\ &\quad + D_2 \cdot t^2 \cdot (t_F - t) + D_F \cdot t \cdot (t_F - t)^2 \end{aligned}$$

where $E_0, E_1, E_2, M_0, M_1, M_2, D_0, D_1, D_2$ and D_F are the parameters (constants) which have to be determined.

One could continue expanding the above functions for still higher powers of t and $(t_F - t)$, but subsequent analysis showed that only those above were generally significant.

Substituting into equation (7), taking the term d arising from $\varphi_D(t) \cdot d = d + D_0(t_F - t) \cdot d + \dots$ to the left-hand side and letting the observed value $(p - d)$ for each month and season be the dependent variable and $(t_F - t) \cdot \hat{e}$, $t \cdot (t_F - t) \cdot \hat{e}$, $t^2 \cdot (t_F - t) \cdot m$ etc. be the independent variables, the parameters E_1, E_2, M_1 , etc. could be determined by applying multiple linear regression analysis to the following equation:

$$\begin{aligned} p - d &= E_0 \cdot (t_F - t) \cdot \hat{e} + E_1 \cdot t \cdot (t_F - t) \cdot \hat{e} + E_2 \cdot t^2 \cdot (t_F - t) \cdot \hat{e} \\ &\quad + M_0 \cdot (t_F - t) \cdot (m - d) + M_1 \cdot t \cdot (t_F - t) \cdot (m - d) + \\ &\quad M_2 \cdot t^2 \cdot (t_F - t) \cdot (m - d) + D_0 \cdot (t_F - t) \cdot d + D_1 \cdot t \cdot (t_F - t) \cdot d \\ &\quad + D_2 \cdot t^2 \cdot (t_F - t) \cdot d + D_F \cdot t \cdot (t_F - t)^2 \cdot d \end{aligned} \quad (9)$$

It should be noted that the fitted equation must be constrained to pass through the origin, because if $t = t_F$, all the independent variables will be zero, and from equation (9) it follows that $(p - d) = 0$.