USE OF MULTILAYER GROUP METHOD OF DATA HANDLING FOR PREDICTING TEA CROP PRODUCTION

A. MUSTAFI[1] AND A.S. CHAUDHURI[2]

(Received : July, 1979)

SUMMARY

The paper develops a monthly tea crop production stochastic process as functions of stochastic variables like past values of monthly tea crop production and also of both past and current values of meteorological parameters (viz. rainfall and penman's evaporation records). This involves generation of regression polynomials of optimal complexity through the use of a heuristic method due to Ivakhnenko 4 referred to as "multilayer group method of data handling". The method provides a prediction of tea crop production a month ahead of the crop's picking, In addition it helps to determine the optimum level of precipitation needed for a possible desired level of tea crop production. The method has been found to be useful and worth adoption by other scientists engaged in crop forecasting studies.

INTRODUCTION

The problem of predicting the tea crop production in advance of its picking is important from the point of view of producers, consumers, and the government alike. In this paper, an attempt has been made to predict tea crop production in Danguajhar Tea Estate. Jalpaiguri, West Bengal on the basis of data on monthly green tea leaves production, monthly rainfall and Penman's evaporation records. This involves the development of tea crop production process as functions of the past values of tea crop production and also of both past and current values of the meteorological parameters like rainfall and Penman's evaporation records. The process is rather complex and it is difficult to formulate any differential or difference equation with deductive logic. In view of this difficulty, the method applied here for this purpose is that referred to Ivakhnenko's [4|

[1] Bengal Engineering College, Howrah, West Bengal.

[2] Jalpaiguri Govt. Engineering College, Jalpaiguri, West Bengal,

Multilayer Group Method of Data Handling (GMDH) which is a heuristic method. This method involves the generation of different regression polynomials by using all possible combinations of input variables and selection therefrom of the best possible ones according to the criterion of minimum integral square error defined in section 3.5.

Brief Description of GMDH

The multilayer group method of data handling involves the use of regression polynomials as the basic means of investigation of complex dynamical systems. The relevant polynomial for prediction purposes is a regression equation which connects a value of an output variable with past or current values of all output and input variables. The regression analysis in this case helps in evaluating tye co-efficients of the polynomial by using the criterion of minimum mean square error. The polynomials are then treated in the same manner as that used for selection of seeds in agriculture as per a unique mathematical concept propagated and established by A.G. Ivakhnenko [5,6],

The salient features of GMDH as applicable in the case of multilayer selection process used in the present work are now briefly described here :

The process can by described by

$$\theta = f(x_1, x_2, ..., xn) \quad ...(1)$$

and it involves the construction of several layers of partial descriptions using two input variables at a time, *e.g.*, the first layer can be represented as :

$$y_i = f(x_j, x_k) \quad ...(2)$$

for $j=1, 2, ..., n$;

with $k=1, 2, ..., n \ (j \neq k)$

and $i=1, 2, ..., m$ where $m = \binom{n}{2}$

Likewise the second layer can best represented as

$$Zi' = g\ (y_i', Y_k') \quad ...(3)$$

for $j'=1, 2, ..., m = \binom{n}{2}$

with $k'=1, 2. ..., m \ (j' \neq k')$

and $i'=1, 2, ..., p$ where $p = \binom{m}{2}$

and so on, it being noted that m and p are the numbers of pairwise combinations of first and second layers respectively. The first step concerns the selection of input variables on the basis of strong correlation [defined in (12) and (13)]. All experimental data are divided into two parts referred to as 'training' and 'checking' sets in the ratio of approximately 2 : 1 respectively. Data with higher values of variation (defined in (16)) are kept in 'training set' and those with lower values in the 'checking set'. The use of an independent 'checking set' helps filter out unwanted noise, if any, in the input data.

Co-efficients of the first layer of partial description are calculated by solving a system of normal Gaussian equations. The left hand side of the equations are set equal to the values of output at every point. After finding the values of the co-efficients, the values of the intermediate variables are obtained. Then using the data of the 'checking set' the integral square error is determined for each of the variables. Only those variables which give low error are selected for subsequent use. These variables are retained in the 'training' and 'checking' sets and the other variables are discarded. In the second layer of selection, the co-efficients of the partial descriptions of the layer are calculated and the accuracy is checked again to select the accurate intermediate variables of the layer ; $Z_1, Z_2, \ldots, Zp$. The process of selection continues so long as the integral squar error on the 'checking set' comes to a minimum and then starts increasing in the next layer.

Every intermediate variable is examined for its effect on prediction accuracy. The 'training set' is used for finding the co-efficients of the partial descriptions, whereas the 'checking set' is used to evaluate the quality of the partial descriptions. This is the basic method used for decision regularisation.

Polynomial description of the process is obtained in the form of partial description of the intermediate variables of different layers. Eliminating the intermediate variables, the complete polynomial description of the process is obtained in the form of Gabor—Kolmogorov type of polynomial as

$$= a_o + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j x_i x_j + \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} a_i a_j \bar{a}_k x_i x_j x_k + \quad \ldots(4)$$

Use of GMDH in formulating tea crop production process :

Input data :

The input data consisted of the monthly green tea leaves production at Danguajhar Tea Estate, Jalpaiguri. West Bengal, monthly rainfall and Penman's evaporation at Nagrakata Observation Centre of Tea Research Association of India. Nagrakata which is near to Danguajhar. for the period January,. 1970 to December, 1975 (Appendix—I).

Normalised power density spectra :

Firstly. all the relevant data X_k for the kth month were transformed as $x_k = \dfrac{X_k - X_{min}}{X_{max} - X_{min}}$ ...(5)

Following [2, 3, 7] the transformed data were then used to prepare the normalised power density spectra graphs of monthly tea crop production, rainfall and evaporation versus cycle per month which are shown in (Fig. 1). The procedure for this briefiy involved the following :

Observed monthly data are $x(i)$, $i=1, 2, ..., N$ Monthly mean value is

$$\bar{x}(N) = \frac{1}{N} \sum_{i=1}^{N} x(i) \qquad ...(6)$$

Auto covariance of monthly data at lag month k is

$$GA(k) = \frac{1}{N-k} \sum_{i=1}^{N-k} [x(i) - \bar{x}(N) \cdot (x(i+k) - \bar{x}(N)] \qquad ...(7)$$

where $k=o, 1. 2, ..., M ; M < \frac{1}{4} N$

The normalised covariance co-efficient is

$$RA(k) = \frac{GA(k)}{GA(0)} \qquad ...(8)$$

The estimates of the normalised power density spectra for the monthly data are given by

$$PS(w_h) = \frac{2}{\pi} \sum_{k=0}^{M} E_k . RA(k). \cos w_h k \qquad ...(9)$$

$$\text{where } w_h = 2\,\pi\, f_h\,;\, f_h = \frac{h}{2M}\,;\, 0 \leqslant f_h \leqslant \frac{1}{2}$$

and $h=0,\ 1,\ 2,\ldots,\ M.$

E_k is defined as weight for window correction

$$E_k = \begin{cases} 1\,;\ 0<k<M \\ \dfrac{1}{2},\ k=0,\ M \end{cases}$$

These raw estimates of power spectral density are smoothed by using Hamming Window to obtain the final estimates of the power spectra. The smoothed estimators of the ordinates of the power spectra are given by,

$$h=0\,;\ S(w_o)=0.54\ PS(w_o)+0.46\ PS\ (w_1)\ 0<h<M\,;$$
$$S\ (w_h)=0.23\ PS\ (w_{h-1})+0.54\ PS\ (w_h)+0.23\ PS\ (w_{h+1}) \quad \ldots(10)$$
$$h=M\,;\ S\ (w_M)=0.54\ PS\ (w_M)+0.46\ PS\ (w_{M-1})$$

From fig. (1), it is seen that the relavent spectra follow the same pattern on the basis of which it is inferred that the two meteorological parameters (*viz.* rainfall and evaporation) have great influence on monthly tea crop production.

Formulation of the process equation

Now the tea crop production process can be represented by

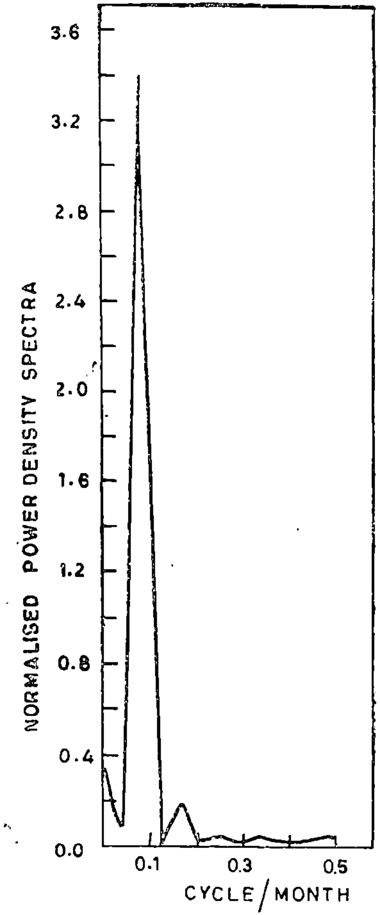
$$yk=f\,(y_{k-1},\ y_{k-2},\ldots,\ x_1,\ k,\ x_1,\ k\text{-}1,\ldots,$$
$$x_2,\ k,\ x_2,\ k\text{-}1,\ldots) \quad \ldots(11)$$

where y_k denotes the transformed tea production of current month, $x_{1,\ k}$ the transformed rainfall of the current month and $x_{2,k}$ the transformed Penman's evaporation of the current month and the subscripts $k,k-1,\ k-2,\ldots,$ refer respectively to the current month, a month preceding the current one and two months preceding the current one, and so on.

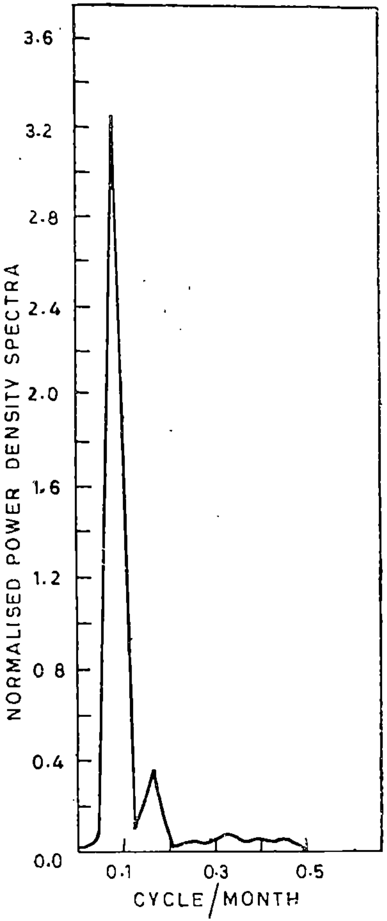
The arguments having a strong correlation with y_k are then selected for inclusion in the process equation on the basis of the correlation functions for the time shift λ defined as

Auto-correlation function,

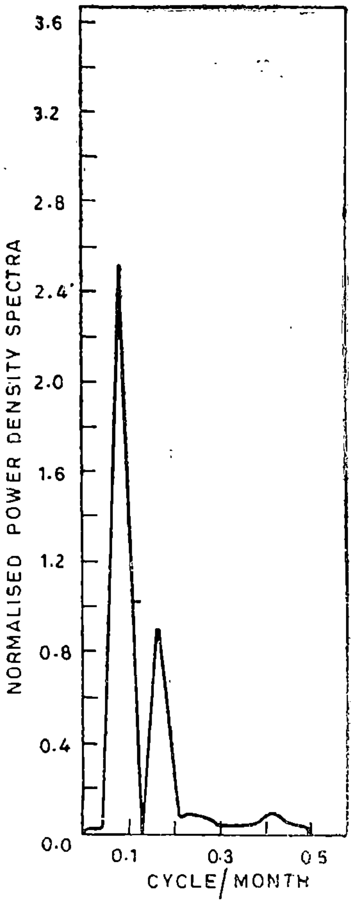
$$Kyy\ (\lambda) = \frac{\sum\limits_{i=1}^{N-\lambda} (y\ (i)-\bar{y})\ (y\ (i+\lambda)-\bar{y})}{\left[\sum\limits_{i=1}^{N-\lambda} (y\ (i)-\bar{y})^2 \sum\limits_{j=1+\lambda}^{N} (y\ (j)-\bar{y})^2\right]^{\frac{1}{2}}} \quad \ldots(12)$$



(a) TEA CROP



(b) RAINFALL



(c) EVAPORATION

and
Cross-correlation function,

$$K_{yx}(\lambda)=\frac{\sum\limits_{i=1}^{N-\lambda}(y(i)-\bar{y})(x(i+\lambda)-\bar{x})}{\left[\sum\limits_{i=1}^{N-\lambda}(y(i)-\bar{y})^2\sum\limits_{j=1+\lambda}^{N}(x(j)-\bar{x})^2\right]^{\frac{1}{2}}} \qquad ...(13)$$

where (bar) indicates the mean value and N is the number of data points.

After such selection of arguments as have strong correlation with montly tea crop production, the process equation then becomes:

$$Y_k=f(y_{k-1}, y_{k-6}, y_{k-12}, x_{1,k-5}, x_{1,k-11}, x_{2,k}, x_{2,k-4}, x_{2,k-10}) \qquad ...(14)$$

or denoting these respective arguments as:

$$y_{k-1}=x'_1,\ y_{k-6}=x'_2,\ y_{k-12}=x'_3,\ x_{1,k}=x'_4,\ x_{1,k-5}=x'_5,$$

$$x_{1,k-11}=x'_6,\ x_{2,k}=x'_7,\ x_{2,k-4}=x'_8,\ x_{2,k-10}=x'_9$$

and the output as $y_k=y$, the process equation becomes,

$$y=f(x'_1, x''_2\ x'_3, x'_4, x'_5, x'_6, x'_7, x'_8, x'_9) \qquad ...(15)$$

Construction of the 'training' and 'checking' sets:

Baased on values of D^2 referred to as a measure of variation between the points of interpolation and calculated from the expression

$$D^2=\frac{1}{n}({x'_1}^2+{x'_2}^2+...+{x'_n}^2) \qquad ...(16)$$

where in the present case $n=9$, the data with higher values of D^2 were kept in the 'training set' and those with lower values of D^2 in the 'checking set'. Actually 20 out of 60 data points with lower values of D^2 were kept in the 'checking set'.

First layer selection :

There are now $\binom{9}{2}=36$ possible combinations of selecting 2 arguments at a time out of 9. For every such combination, the partial regression equation is written in the form.

$$y_a=\alpha_{0a}+\alpha_{1a}x'_b+\alpha_{2a}x'_c+\alpha_{3a}x'_b\ x'_c+\alpha_{4a}x'^2_b+\alpha_{5a}x'^2_c \qquad ...(17)$$

where $a=1,2,\ldots,36$ while b and c are indices for all 36 combinations. And these, therefore lead to 36 systems of normal Gaussian equations with matrices of order 6×6. The co-efficients α's are then estimated by solving the normal equation systems constructed from the 'training set data. For estimating the co-efficients it is assumed that the equation error is small being distributed with zero mean, constant variance and also uncorrelated with inputs. The second assumption is that the inputs and outputs are known exactly without any measurement error (1).

The accuracy of every variable y_a is calculated by using the 'checking set' data only. From all the variables, we choose 9 more accurate ones which give low values as per integral square error criterion. the integral square error being defined as

$$ISE=\frac{\sum\limits_{i=1}^{N_1}(y_i\,(\text{checking})-y_i\,(\text{model}))^2}{\sum\limits_{i=1}^{N_1}(y_i\,(\text{checking}))^2} \qquad ...(18)$$

where N_1 is the number of 'checking set' data points.

Other layers selection :

Firstly, 9 intermediate variables of y-layer chosen from the first layer give 36 combinatlons of two arguments of y-layer. Again in the second layer these become,

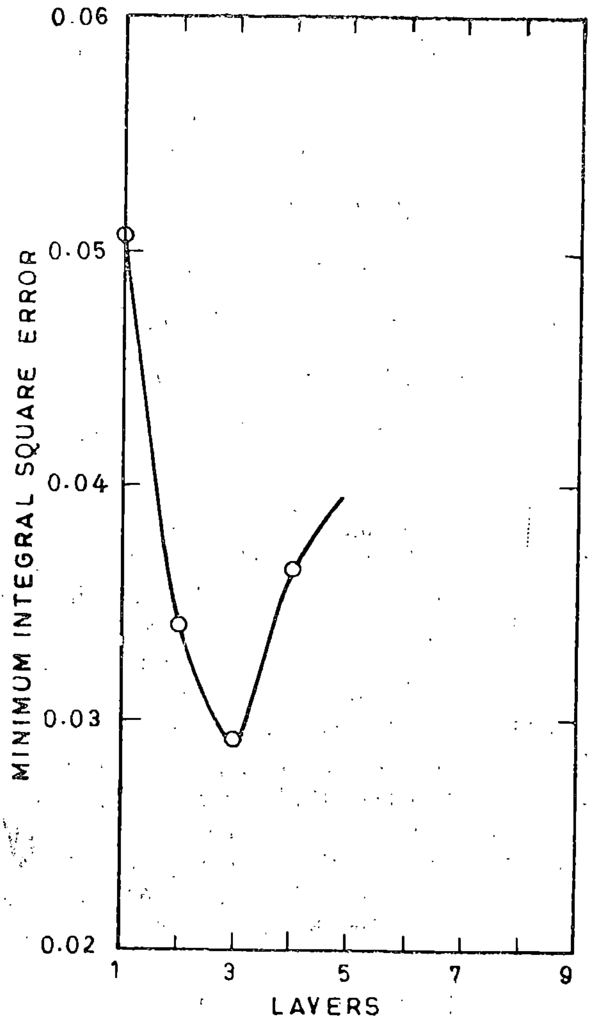
$$Z_a=\beta_{0a}+\beta_{1a}y_b+\beta_{2a}y_c+\beta_{3a}y_by_c+\beta_{4a}y^2_b+\beta_{5a}y^2_c \qquad ...(19)$$

where $a=1,2,\ldots,36$ while b and c are the indices of all 36 combinations. Calculation of the co-efficients β's and estimation of the accuracy of the variables z_a are rcpeated as in case of y_a.

The 9 z_a variable are then chosen for the next layer u_a.

$$u_a=\gamma_{0a}+\gamma_{1a}z_b+\gamma_{2a}z_c+\gamma_{3a}z_bz_c+\gamma_{4a}z^2_b+\gamma_{5a}z^2_c \qquad ...(20)$$

In this way layer after layer is tested for accuracy by using the 'checking set' data and on the basis of minimum integral square error criterion as explained earlier. For all layers, the variables on the left hand side of the equation are kept equal to the value of the output variable. The minimum error is obtained in the *u*-layer, for *u*, *ISE*=0.0293, and afterwards in the *v*-layer the error starts increasing. The changes of minimum integral square error for diffcrent layers are shown in (Fig. 2).



Results :

The monthly tea crop production process has been identified by the polynomials as shown below :

$$y=-0.016852+1.058209\,z_{30}+0.1662367\,z_{26}$$
$$+108.5559\,z_{30}z_{26}-55\;21898\,z_{30}^2-53.53989\,z_{26}^2$$

$$z_{30}=-0.030785+0.753067\,y_9+0.211311\,y_{30}$$
$$-1.662558\,y_9y_{30}+0.837753\,y_9^2+0.917495\,y_{30}^2$$

$$z_{26}=-0.0180489+0.956433\,y_{17}-0.036179\,y_{30}$$
$$-3.132089\,y_{17}y_{30}+1.390793\,y_{17}^2+1.857717\,y_{30}^2$$

$$y_9=-0.119043+0.471011\,x'_2+1.334358\,x'_3$$
$$-0.475267\,x'_2\,x'_3-0.307475\,x'^2_2-0.228290\,x'^2_3$$

$$y_{30}=1.04006-0.961599\,x'_5-0.561185\,x'_9$$
$$-0.48578\,x'_5x'_9+1.018227\,x'^2_5-0.210997\,x'^2_9$$

$$y_{17}=0.107056+0.784304\,x'_3-0.279867\,x'_5$$
$$+0.141815\,x'_3x'_5+0.093925\,x'^2_3+0.230375\,x'^2_5 \quad \ldots(21)$$

Illustration :

(*i*) Now the transformed data for the month of September, 1971 are $x'_2=0.0306$, $x'_3=0.8097$, $x'_5=0.1379$ and $x'_9=0.2469$ (these variables are explained in section (3.3). Inserting these values in the polynomial description of monthly tea crop production (21), the transformed value of the tea crop production for the month of September, 1971 is 0.7642. When converted, the value is 562.42 tonnes. The actual production is 559.35 tonnes.

(*ii*) Let us consider another example. The transformed data for the month of September, 1974 are $x'_2=0.3426$, $x'_3=0.8033$, $x'_5=0.1107$ and $x'_9=0.1017$. The corresponding predicted value for tea crop production obtained from polynomial description (21) is 0.8610, When converted the value is 633.67 tonnes. The actual tea crop production for the month of September, 1974 is 652.08 tonnes.

It is clearly evident that a month ahead predictions of tea crop production obtained with the help of polynomial description (21) of the crop production process approach very near to the actual production figures.

CONCLUSION

The polynomial description for monthly tea crop production can be used to ascertain the optimum amount of precipitation required to adhere to a planned production target.

Among the meteorological parameters only rainfall and evaporation are taken into account to obtain the polynomical description. However, other meteorological parameters for example, sunshine hour and wind velocity may be considered. It is hoped that when all the factors are considered the model can b϶ used as a handy tool for the man in the field.

This method of identification can also be extended to the identi fication of the inter-actions of meteorological parameters on other agricultural processes mainly rice, wheat, jute, cotton, etc.

ACKNOWLEDGEMENT

Authors extend their appreciation to the referee of this paper for his valuable suggestions and comments. The partial financial support of Tea Reseach Association of India is gratefully acknowledged. The assistance and co-operation of Danguajhar Tea Estate and the Computer Centre of Bengal Engineering College is gratefully appreciated.

REFERENCES

Box G.E.P. : 'Use and abuse of regression', *Technometrics*, Vol. 8, 625-629 Nov. 1966.

Box G.E.P. and Jenkins G.M.: '*Time series analysis, forecasting and Control*', Holden-Day, 1970.

Chaudhuri A.S. : 'Prediction of monthly tea crop', published in Russian in *the Herald* of Kiev Poly Technical Institute, USSR, 1979.

Ivakhnenko A.G. : 'Heuristic self-organisation in problems of engineering cybernetics', *Automatica*, Vol. 6, 207-219, 1970.

Ivakhnenko A.G. and Kappa Yu.V. : 'Group method of data handling for the solution of various problems of cybernetics', Second International Federation of Automatic Control Symposium on Identification and Process Parameter *Estimation*, Prague, paper 2.1, 1970.

Ivakhnenko A.G. : 'Polynomial theory of complex systems', *IEEE Transc. on Systems, Man and Cybernetics*, Vol. SMC-1, No. 4, October 1971.

Kashyap R.L. and Rao A.R. : 'Analysis, construction, validation of stochastic models for monthly river flows', *Tech. Rep.* CE-HYD-73-1, December, 1973, School of Civil Engineering, Purdue University, West Lafayetite, Indana U.S.A.

Appendix-1

Monthly Green Leaves Production, Rainfall, Evaporation, January 1970 to December 1975

Serial No.	*Months*	*Tea Crop Production in Tonnes*	*Rainfall in mm*	*Evaporation in mm.*
1	*2*	*3*	*4*	*5*
1.	January '70	0.0	11.8	67.0
2.	February	2.50	26.6	90.2
3.	March	65.2	13.4	136.6
4.	April	274.69	231.6	145.1
5.	May	391.46	359.1	175.2
6.	June	504.56	1018.7	133.9
7.	July	607.44	1287.2	126.1
8.	August	579.25	775.4	151.8
9.	September	595.94	920.8	120.5
10.	October	503.21	74.5	134.8
11.	November	361.02	3.1	100.1
12.	December	59.25	0.0	75.2
13.	January '71	0.0	6.1	69.6
14.	February	0.0	2.1	93.1
15.	March	24.74	5.4	149.8
16.	April	302.86	414.1	134.9
17.	May	304.40	263.6	161.2
18.	June	510.4	840.1	125.5
19.	July	650.67	974.7	140.6
20.	August	646.72	697.1	123.2
21.	September	559.35	418.1	125.7
22.	October	735.96	472.3	119.8
23.	November	277.87	42.3	90.7
24.	December	70.30	15.2	74.4
25.	January '72	0.0	6.6	71.0
26.	February	1.60	31.0	82.9
27.	March	123.32	15.2	145.8
28.	April	291.92	199.1	161.4
29.	May	347.83	715.8	165.9
30.	June	456.61	888.8	144.9
31.	July	550.19	1563.3	124.9
32.	August	601.25	879.5	150.4
33.	September	637.87	698.9	128.9
34.	October	675.29	241.4	120.9

(table contd.)

1	2	3	4	5
35.	November	385.01	8.2	100.0
36.	December	109.02	0.0	71.0
37.	January '73	0.0	4.3	59.2
38.	Febuary	8.9	29.7	91.5
39.	March	237.28	11.9	140.4
40.	April	167.89	120.7	171.9
41.	May	386.29	517.5	165.5
42.	June	431.99	1086.6	125.6
43.	July	554.5	701.7	146.2
44.	August	625.25	751.1	150.8
45.	September	594.86	449.8	119.9
46.	October	623.78	349.5	90.3
47.	November	396.13	1.3	101.8
48.	December	125.0	3.1	71.0
49.	January '74	0.0	42.1	62.7
50.	February	8.57	0.00	93.0
51.	March	252.16	40.9	135.8
52.	April	507.27	173.0	149.8
53.	May	494.82	400.0	164.8
54.	June	541.84	917.5	138.8
55.	July	671.4	1105.2	117.6
56.	August	670.44	696.0	131.8
57.	September	652.08	554.0	122.0
58.	October	663.01	290.0	126.4
59.	November	513.43	0.0	103.9
60.	December	135.44	0:0	68,4
61,	January '75	0.0	0.0	72.9
62.	February	10.45	24.4	84.9
63.	March	158.15	8.2	150.0
64.	April	334.16	49.6	159.1
65.	May	451.55	165.9	167.7
66.	June	462.21	931.5	152.2
67.	July	663.62	1260.8	117.6
68.	August	616.08	291.2	150.7
69,	September	637.6	596.5	121.6
70.	October	689.07	204.7	130.1
71.	November	406.01	0.0	98 9
72.	December	142.99	0.0	70.8