Alternative Methods of Adjusting for Heteroscedasticity in Wheat Growth Data

Economics,
Statistics, and
Cooperatives Service

U. S. Department
of Agriculture

Washington, D. C.
20250

ALTERNATIVE METHODS OF ADJUSTING FOR HETEROSCEDASTICITY IN WHEAT GROWTH DATA

By

Greg A. Larsen

Mathematical Statistician
Research and Development Branch
Statistical Research Division
Economics, Statistics and Cooperatives Service
United States Department of Agriculture

Washington, D. C. February, 1978

Table of Contents

Index of Diagrams, Tables and Figures

Acknowledgments

I would like to thank the staff of the Kansas State Statistical Office for supervising the data collection phase of the 1977 Within-Year Wheat Growth Study. A group of approximately 30 enumerators were employed and the office staff did a good job in coordinating their efforts. The key-punching, also done in the state office, was nearly perfect.

I would also like to recognize the fine job that Jack Nealon did in setting up the sampling plan and preparing all the manuals, forms and supplies necessary for the data collection.

I am especially appreciative to Mary Jiggetts for her patience and care in typing a difficult manuscript.

Introduction

The purpose of this report is to suggest some techniques which are useful in adjusting for heteroscedastic disturbances common to growth data. While all the results pertain specifically to dried wheat head weights, it is hoped that the concepts involved are general enough to be of use with other types of growth data or in non-growth situations.

Two basic approaches will be discussed; log transformations and weighted least squares. Several different log transformations will be developed most of which are of the form $\ln (Ay + B)$ where y is the dependent variable and A and B are derived constants based on various criteria. The last transformation discussed has four constants to be specified and the log is taken twice. The second approach is to weight the diagonal elements of the variance-covariance matrix of the dependent variable by some appropriate function which creates a homoscedastic disturbance term. A linear function is used which relates the expected value of the dependent variable to the population standard deviation.

The data used in this report was collected during April to July 1977 in 68 winter wheat fields distributed throughout Kansas. These fields were a subset of the wheat objective yield sample in which fields were selected with probability proportional to acreage. This was done so that field level yield observations could be weighted together equally to form a state average. Within each field there were two randomly and independently located plots. Each plot consisted of one row and was approximately five feet in length. In the case of braodcast wheat, the plots were six inches wide. Stalk counts were made in the five foot plot. Every tenth plant was tagged until a total of 30 was reached. The 30 tagged plants could lie within the five foot section or go beyond it depending on the plant density. Stalk counts averaged less than 300 plants and generally the sample extended past the five foot area.

Weekly visits were made by trained enumerators to observe when the tagged plants had heads fully emerged and when flowering occurred. As a general rule, once 80% of the tagged plants in a particular field had flowered, clipping began. A random sample of four heads per plot per weekly visit was clipped, each head was placed in an air tight plastic tube and mailed to the state lab. In the lab, wet and dry weights were determined for each head. The heads were dried for 46 hours at 150 degrees Fahrenheit. Head emergence and flowering continued to be observed until two weeks after clipping began or until all tagged plants had flowered. Heads continued to be clipped on a weekly basis until harvest.

Time since full head emergence and time since flowering were calculated for each clipped head. Since the random sample of heads to be clipped each week was pre-determined, some heads were clipped before head emergence and flowering was observed. It was determined in previous work * that time since flowering is the preferred variable. This has been verified and the time variable used in this report is time since flowering. Heads for which flowering was not observed were excluded from the data set.

The sample design that has been outlined is a two level nested design. Plots are nested within fields and heads are nested within plots. Consequently, individual head weights are not strictly independent and aggregation should take place until independent points are obtained. In the past *, this has meant averaging head weights for all plants in a field on a particular weekly visit. Since the success of the growth model depends upon the relationship between time since flowering and head weight, there is concern that this method of aggregation might adversely affect the time-growth relationship. The reason for this is that even though a set of heads might have the same clipping date, they do not necessarily have the same number of days since flowering. Aggregation averages over both time and weight. If time since flowering is fairly consistent for a particular visit, there is no problem. However, in the data to be used in this report there was as much as a two week difference in time values for a particular visit. Therefore a method of aggregating observations with similar time values was sought. Since weekly visits were made, the flowering date is actually an average between the date of the visit when flowering was first observed and the date of the previous visit. This puts a measurement error of approximately 3.5 days on either side of the flowering date. (Visits have been made every two or three days in earlier research but analysis indicated that weekly visits would be sufficiently accurate. *) With this in mind, the data was divided into time intervals in each of which the time values were assumed to be essentially the same. Head weights in each time interval were averaged together within a field without regard to plots. So long as the two plots within a field have the same number of observations, each plot will receive equal weight as the sampling plan intended. However, plots with very few observations would tend to receive less weight. Averaging within time intervals should more fully preserve the time-growth relationship than would aggregation by visit.

*Nealon, Jack 1976 The Development of Within-Year Forecasting Models for Winter Wheat. Research and Development Branch, Statistical Research Division, ESCS, USDA.

Since the sampling plan was not conceived with time interval aggregation in mind, one consequence is an increased variability in the number of observations per mean. Means are comprised of from one to thirteen observations. To reduce the effect of means with few observations, each mean was expanded by the number of observations going into it. In other words, a mean comprised of one observation would be included once while a mean comprised of eight observations would be included eight times. This gives the allusion of more data points than really exist causing a reduction in the parameter standard error estimates. The reduction is relative, however, and does not affect comparisons as long as the expanded data set is used consistently. Another consequence of expanding the data set is that fields are weighted unequally because of varying numbers of observations in each field. This violates the intention of the sampling plan but the purpose of this report is solely to investigate methods of adjusting for heteroscedasticity and not to estimate a state yield. It is suggested that the expanded data set is suitable for this purpose.

It should be pointed out that even when we do not expand the aggregated means, there is still a problem with the way in which the fields were sampled. This is because the number of visits to clip heads varied from three to seven depending upon when harvest occurred. Therefore, if no expansion is done, fields that had more visits would be more heavily favored. A possible solution to this would be to weight each aggregated mean by the inverse of the number of time intervals in the corresponding field. In this way, fields with three visits would receive the same weight as fields with seven. Ideally, we'd like to develop a sampling plan that would clip heads a certain number of days after flowering rather than clipping a random sample comprised of various stages of development on a fixed date. To control the number of heads being clipped in such a plan would not be easy and this problem remains for future study.

The basic growth model which will be referred to throughout this report is as follows:*

$$(1) \qquad y_i = \frac{\alpha}{1 + \beta\rho^{t_i}} + \varepsilon_i \text{ where } i = 1,2, \ldots,n$$

Least squares theory will be used to estimate parameters α, β and ρ. This requires the following assumptions about the nature of the model.

(a) $E(\varepsilon_i) = 0$ for all i

(b) $\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ for all i

(c) $\text{Cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$

Stated another way, assumptions (b) and (c) imply that the variance-covariance matrix of the dependent variable y is

$$(2) \qquad I_n \sigma^2 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \sigma^2 \quad nxn$$

Heteroscedasticity is defined to occur when the diagonal elements of the V-C matrix are no longer equal. In matrix notation, the V-C matrix can be expressed as

$$(3) \qquad V_n \sigma^2 = \begin{bmatrix} K_1 & & & \\ & K_2 & & \\ & & \ddots & \\ & & & K_n \end{bmatrix} \sigma^2$$

*A good discussion on this model appears in an earlier report. House, Carol C. 1977 A Within-Year Growth Model Approach to Forecasting Corn Yields. Research and Development Branch, Statistical Research Division, ESCS, USDA. pp 2-3.

In the original notation, assumption (b) above would be replaced by

(4) $$E(\varepsilon_i^2) = K_i\sigma^2 \text{ for all } i$$

In words, heteroscedasticity means that the variance of the dependent variable does not remain constant over the range of the independent variable. In the case of the growth model, the variance of the y's increases with time.*

The effect that heteroscedasticity has when standard least squares is applied is often difficult to assess. In generalized linear regression, parameter estimates tend to become unreliable depending upon the degree of the heteroscedasticity. While they remain unbiased, their standard errors are underestimated. Also, the estimate of σ^2 will no longer be unbiased.** Several different methods of adjusting for heteroscedasticity will now be discussed.

The Log Transformation

The basic growth model was fit to the data using the NLIN procedure in SAS.*** A plot of the estimated values of y overlayed on the data is shown in Figure 1. A plot of the residuals vs. time is shown in Figure 2. The presence of heteroscedasticity can be seen in the latter figure. The cone-shaped plot shows that the deviation of the data from the fitted equation increases with time. A residual plot showing homoscedasticity (and no other model deficiencies) would appear to show a random pattern. The correlation between the absolute value of the residuals and time is .28 using Pearson's R (denoted by R_p) and .31 using Spearman's Rho (a comparable nonparametric test denoted by R_s). Both are significantly different from zero. See Table 1 on page 9.

To adjust for the non-homogeneity of variance, we would like to decrease the variance as t becomes large so that the absolute value of the residuals will become uncorrelated with time. It is possible to do this with a log transformation. The natural logarithm will be used in this discussion but the logarithm to the base 10 could also be used. To visualize how this works refer to Diagram 1 on the next page. Taking the ln (y) has the effect

*If data is collected well after maturity is reached, the variance will become stable for large values of time. While this did not appear to happen with the wheat data used in this report, it was noted in some earlier corn research (House).

**Goldberger, Arthur S. 1964 Econometric Theory. John Wiley & Sons, Inc. New York, London, Sidney. pp 238-241.

***Barr, Anthony J., Goodnight, James H., et. al. A User's Guide To SAS 76. SAS Institute Inc. Raleigh, N. C. pp 193-199.

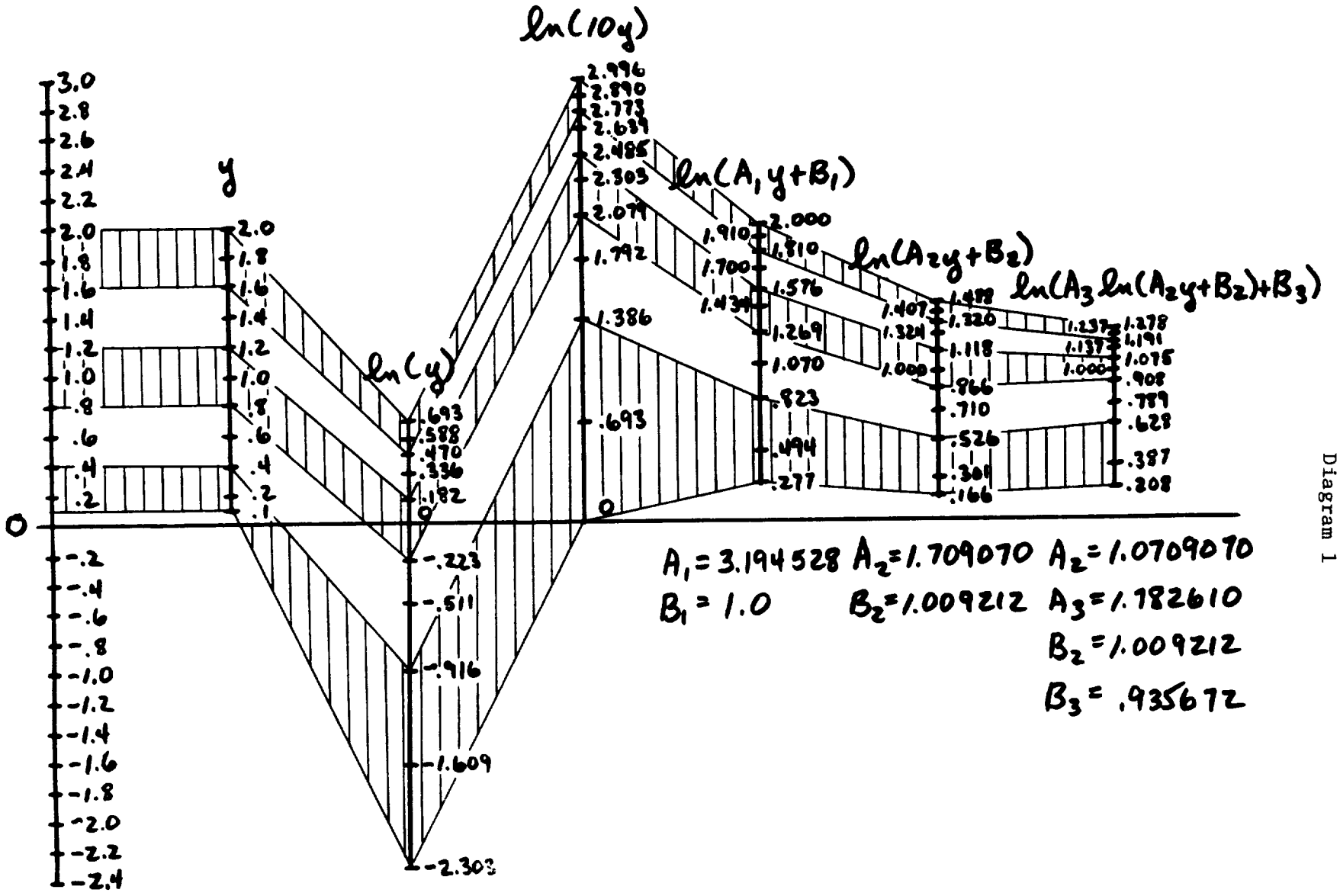

ln(10y)
y
ln(y)
ln(A1y+B1)
ln(A2y+B2)
ln(A3 ln(A2y+B2)+B3)
A1 = 3.194528
A2 = 1.709070
A2 = 1.0709070
B1 = 1.0
B2 = 1.009212
A3 = 1.782610
B2 = 1.009212
B3 = .935672

Diagram 1

of spreading values of y less than 1 farther apart and collapsing values of y greater than 1. In growth data, since smaller values of y are less variable than larger values, the ln (y) will have more consistent variability over the range of t.

Since y represents weight, it would be good to keep the transformed variable positive also. Values of y are known to be in the range .1 to 2.0 so simply using ln (10y) would produce strictly positive values. Notice that ln (10y) = ln (y) + ln (10) = ln (y) + 2.303 so that multiplying y by a constant doesn't change the spreading and collapsing effect but only moves the transformed variable up or down the scale. Refer to Diagram 1.

The discussion so far suggests that we multiply both sides of (1) by 10 and then take the log of both sides. This gives the following model:

$$(5) \qquad \ln (10y) = \ln \left[\frac{10\alpha}{1+\beta\rho^{t}} + 10\varepsilon \right]$$

To get least squares parameter estimates, the model must be expressed with an additive error term. We therefore redefine the model with additive error ε' which is different from the ε in (5).

$$(6) \qquad \ln (10y) = \ln \left[\frac{10\alpha}{1+\beta\rho^{t}} \right] + \varepsilon'$$

If we are willing to make this change in models, there will be no violation of assumptions in applying least squares theory to estimate the parameters in (6). We are primarily interested in forecasting the average value of y at the end of the growing season so α (the asymptote) is the parameter which we are most interested in estimating. Notice that the least squares estimate of α (and, for that matter β and ρ) obtained from (6) will be in terms of the untransformed dependent variable and therefore no antilog will be required. This would not be the case if, for example, we had only taken the log of the left-hand-side of (1). A plot of the estimated values of the transformed dependent variable overlayed on the transformed data is shown in Figure 3. A plot of the residuals versus time is shown in Figure 4. The residual plot shows that the heteroscedasticity has reversed itself, that is, we have over adjusted. Table 1 shows the correlation coefficients which support this conclusion. R_p is significantly different from zero at - .0645 while R_s is not significantly different from zero but is also in the negative direction. The mean square errors (MSE's) of

the unadjusted and log models are not directly comparable because the data sets are different. This is particularly true with the log transformation presently being considered because, as Diagram 1 shows, the transformed dependent variable increases for all but the very smallest values of y. A uniform increase in y (such as adding a constant) would not change the MSE but here the expansion of y values less than 1 is more than the contraction of y values greater than 1 causing the log model to have a larger MSE.

Table 1 also shows the parameter estimates with the corresponding relative standard error in terms of a percent. Although the current log model over adjusted, there was a lessening of the affect of non-homogeneous variance as evidenced by the decrease in the significance of the correlations. It can be observed that the relative standard error of $\hat{\alpha}$ increased. This seems to be consistent with the generalized linear regression situation stated earlier in which the standard error of a parameter estimate tends to be understated when influenced by heteroscedasticity. Obviously, linear theory may or may not hold in a non-linear setting so we only seek to point out the comparison. The errors for $\hat{\beta}$ and $\hat{\rho}$ decrease apparently bringing the consistency to an end. These same directional changes in the relative standard errors occur in all the adjusted models. The fact that the transformation ln (10y) over adjusted leads us to want to find a transformation which does not cause the small values of y to become more variable than the large values. As noted earlier, multiplying y by a constant prior to taking the log does not change the relative relationship between the points. Therefore, we now want to consider transformations of the form ln (Ay+B) where A and B are constants to be specified. This will cause changes in the relative relationship between the points when $B \neq 0$.

Several approaches can be used to adjust for heteroscedasticity. Small values of y could be spread out until they have variability comparable to large values of y. Or, conversely, the variability of the small values could be held relatively stable and the large values collapsed until a comparable variance was attained. Thirdly, as in the case of (6), there could be both an expanding and contracting of the y values. Obviously, it would not be desirable to increase the total variation so, a transformation maintaining the same approximate range as the untransformed data would help to keep the total variation from increasing. Therefore, A and B will be determined with this in mind.

The y values are known to lie within the interval of 0 to 2 grams. The following two constraints will keep the transformation of 0 equal to 0 and the transformation of 2 equal to 2.

$$0 = \ln (0A_1 + B_1)$$

$$2 = \ln (2A_1 + B_1)$$

TABLE 1

MODELS

	Unadj	ln (10y)	ln (A_1y+B_1)	ln (A_2y+B_2)	ln (A_3ln (A_2y+B_2) + B_3)
Regression					
MSE	.0337	.0726	.0309	.0188	.0114
R^2	.942	.980	.978	.971	.984
Param Est's					
$\hat{\alpha}$	1.0360	1.0307	1.0292	1.0293	1.0280
$\hat{\sigma}_{\hat{\alpha}}/\hat{\alpha}$	1.5	2.0	1.8	1.7	1.9
$\hat{\beta}$	4.2556	4.2159	4.1674	4.1714	4.1501
$\hat{\sigma}_{\hat{\beta}}/\hat{\beta}$	5.3	2.8	3.5	3.8	3.1
$\hat{\rho}$	.8886	.8947	.8929	.8919	.8946
$\hat{\sigma}_{\hat{\rho}}/\hat{\rho}$	.5	.4	.4	.5	.4
Correlation					
R_p	.2780	-.0645	.0770	.1297	-.0162
Prob $>\|R_p\|$	.0001	.0041	.0006	.0001	.4712
R_s	.3093	-.0172	.1119	.1626	.0240
Prob $>\|R_s\|$	.0001	.4452	.0001	.0001	.2850

Exponentiating both sides of each equation gives:

$$e^0 = (0A_1 + B_1)$$
$$e^2 = (2A_1 + B_1)$$

Solving the first equation for B_1 gives $B_1=1$.
Substituting into the second equation and solving for A_1 gives:

$$e^2 = (2A_1 + 1)$$
$$A_1 = \frac{e^2-1}{2}$$

$$A_1 = 3.194528$$

The transformation is then ln (3.194528 y + 1). The model now becomes

$$(7) \qquad \ln (A_1 y + B_1) = \ln \left[\frac{A_1 \alpha}{1+\beta\rho^t} + B_1 \right] + \varepsilon_i$$

where $A_1 = 3.194528$

$B_1 = 1$

The plot of the estimated values of the transformed dependent variable overlayed on the transformed data is shown in Figure 5. The corresponding residual plot is shown in Figure 6. The residuals show a slight positive correlation with time and Table 1 indicates that the correlation is significantly different from zero and positive. Both R_p and R_s are, however, substantially smaller than in the untransformed data.

The MSE in (7) is in the same neighborhood as the MSE in the unadjusted model which says that the total variation has not been increased. It does not necessarily indicate that the model is superior since the two models aren't directly comparable. The R^2 value from (7) indicates that more of the variability in the transformed data is being accounted for than in the untransformed data. All the R^2 values indicate that good relationships exist, however.

While (6) over adjusts, (7) does not adjust enough to cause the variances of y to become constant over the range of time. This suggests that another transformation be developed. The primary function of the growth model is to forecast what the value of α will be at the end of the season based on data collected early in the growing season. With this application in mind, it would be best not to increase the variance for small values of y which, in the early part of the growing season, are the main input into the model. Also, since α is the primary parameter to be estimated, a transformation which causes less distortion around the true asymptote would be preferable.

Figure 1 shows that the smallest group of y values (i.e. $0<t<5$) is approximately bounded by .1 and .8. So, we want a transformation that preserves this same interval thus keeping the variation relatively the same. Based on previous models, the true value of α is thought to be approximately equal to 1. This gives rise to the following two constraints.

$$\ln (.8A_2+B_2) - \ln (.1A_2+B_2) = .7$$

$$\ln (A_2+B_2) = 1$$

The first constraint can be expressed as

$$\ln \left[\frac{.8A_2+B_2}{.1A_2+B_2} \right] = .7$$

$$\frac{.8A_2+B_2}{.1A_2+B_2} = e^{.7}$$

$$.8A_2+B_2 = e^{.7}(.1A_2+B_2)$$

$$A_2(.8 - .1e^{.7}) = B_2 (e^{.7} - 1)$$

$$.598625A_2 = 1.013753B_2$$

$$A_2 = 1.693469B_2$$

The second constraint can be expressed as

$$e = A_2 + B_2$$

After substituting and solving, the two constraints give the following values for A_2 and B_2.

$$A_2 = 1.709070$$

$$B_2 = 1.009212$$

The model now becomes

$$(8) \qquad \ln (A_2 y + B_2) = \ln \left[\frac{A_2 \alpha}{1+\beta\rho^t} + B_2 \right] + \varepsilon_2$$

where $A_2 = 1.709070$

$B_2 = 1.009212$

The estimated values of the transformed dependent variable overlayed on the transformed data and the residuals are shown in Figures 7 and 8, respectively. The residuals again show a positive correlation with time. The R_p and R_s values in Table 1 are both significantly different from zero. A comparison of Figures 1 and 7 reveals that the range of the group of smallest y values (i.e. $0<t<5$) is the same and the asymptote is in the neighborhood of 1.0 for both. While the transformation did keep the variance for small values of y stable, it did not reduce the variation as y increased enough to create a homogeneous situation.

If the present criteria used to select A and B are deemed satisfactory, then any further attempt to adjust for heteroscedasticity by obtaining other A and B values would lead to a compromise in the selection criteria. There are no doubt many legitimate ways to select A and B but a feasible alternative to a compromise would be a double log transformation. This can be done by applying the same criteria to the transformed data in (8) as were applied to the original data. The two previous constraints will be the same except the transformed points will be substituted.

y	.1	.8	1.0
$\ln (A_2 y + B_2)$	.165615	.865615	1.0

The constraints now become

$$\ln (.865615A_3 + B_3) - \ln (.165615A_3 + B_3) = .7$$

$$\ln (A_3+B_3) = 1$$

Proceeding as before, the solution is

$$A_3 = 1.782610$$

$$B_3 = .935672$$

The model* now becomes

$$(9)\ \ln (A_3 \ln (A_2 y + B_2) + B_3) = \ln \left(A_3 \ln \left[\frac{A_2 \alpha}{1+\beta\rho^t} + B_2\right] + B_3\right) + \varepsilon_3$$

where $A_2 = 1.0709070$

$A_3 = 1.782610$

$B_2 = 1.009212$

$B_3 = .935672$

Figures 9 and 10 show the two plots associated with this model. The residual plot appears to indicate a very slight negative correlation. In Table 1, R_p is negative and R_s is positive but neither is significantly different from zero. The estimate of the population variance (MSE) is only one-third that of the unadjusted model. This is reasonable since by holding the variance stable for the smallest values of y, the transformation lessens the variance for all larger values of y.

*It has been pointed out that (9) has a good deal more potential than is being utilized here. We essentially applied the same pair of constraints twice to obtain values for A_2, B_2, A_3 and B_3. Conceivably, since there are four unknowns, four different constraints could be used. The problem, of course, would be to find four constraints that are solvable.

The relative standard errors in (9) increased for α and decreased for β and ρ which, as discussed earlier, is consistent in all the adjusted models. Compared to the unadjusted model, $\hat{\alpha}$ from (9) is only .8% less. While heteroscedasticity tends to cause unreliable parameter estimates, depending on the degree, the model in (9) seems to show that the unadjusted estimate of α is quite stable.

Weighted Least Squares

An alternative to the log transformation which sometimes may be effective in adjusting for heteroscedasticity is to perform a weighted least squares regression. Recall the assumption concerning the diagonal elements of the variance-covariance matrix when heteroscedasticity is present.

$$E(\varepsilon_i^2) = k_i\sigma^2 \text{ for all } i$$

If we divide (1) by $\sqrt{k_i}$, we obtain

$$(10) \quad \frac{y_i}{\sqrt{k_i}} = \frac{\alpha}{\sqrt{k_i}\,(1+\beta\rho^{t_i})} + \frac{\varepsilon_i}{\sqrt{k_i}} \quad \text{where } i = 1,2,\ldots,n$$

The disturbance term is now homoscedastic.

$$E\left(\left(\varepsilon_i/\sqrt{k_i}\right)^2\right) = E(\varepsilon_i^2)/k_i = \sigma^2$$

Equivalent sums of squares and parameter estimates are obtained if the diagonal elements of the V-C matrix of the dependent variable are multiplied by $1/k_i$. This fact provides two alternative ways of performing a weighted regression. The NLIN procedure in SAS can be applied to (10) or the weight statement can be used with the NLIN applied to (1). If the weight statement is used, the value of the weight would be $1/k_i$. Although these two alternatives give identical sums of squares and parameter estimates, there is a useful difference between the two. The first alternative provides a residual plot in terms of $y/\sqrt{k}$ while the other is in terms of the original y values. Therefore, since we need to see the residual plot and correlations to check for heteroscedasticity, the first alternative will be employed. The second alternative could be used if the variance-covariance matrix was obtained and then a test for differences among diagonal elements would test the hypothesis

of homogenious variance. At present, the option of outputting the V-C matrix in NLIN is not available.

In theory, a weighted regression will provide homoscedastic disturbance so the problem becomes one of finding a suitable function for k. A relationship needs to be found which describes the behavior of σ_t which is the true standard deviation of y at a point in time. Because of the way in which the data was collected, each data point has a possible error of $\pm$ 3.5 days on the value of time. Therefore, it would exceed the precision of the data to break it down into one or two day intervals. Since most observations were centered around a particular day of the week, the data in Figure 1 already appears grouped. The data readily divides into the seven time intervals listed in Table 2. It is then reasonable to estimate σ_t in each interval by calculating the standard deviation of the y values in each group. Denote this estimate by $\hat{\sigma}_t$.

TABLE 2

Time Interval (days)	$\hat{\sigma}_t$ (grams)	No. of obs.
$0<t\leq5$	.1115	210
$5<t\leq14$	.1487	389
$14<t\leq20$	.1852	416
$20<t\leq28$	.1847	422
$28<t\leq33$	.2210	324
$33<t\leq41$	.2197	180
$41<t$	.2593	42

In some applications, the disturbance variance, σ_t^2, is proportional to the square of the expected value of y or some linear function of it.* This would make $E(y)^2$ a candidate for k. Since E(y) is, of course, unknown, the $\hat{y}$ values obtained from a non-linear regression are the "best" estimates. Because of the reasons already stated, we want to apply the NLIN procedure to (10) and therefore a relationship for $\sqrt{k}$ is sought. Figure 1 shows that as $\hat{y}$ increases so does $\hat{\sigma}_t$. A linear regression of $\hat{\sigma}_t$ on $\hat{y}$ weighted by the number of observations in each value of $\hat{\sigma}_t$ was run to see if a reasonably good relationship exists. A $\hat{y}$ value was calculated for the median of each time interval using the parameter estimates from the unadjusted model. The regression was highly significant with $R^2 = .908$. Figure 11 shows the estimated values of the dependent variable and the seven data points. The estimated equation ** is

(11) $$\hat{\sigma}_t = .078111 + .145391\ \hat{y}$$

Since $E(\varepsilon^2) = \sigma_t^2$, the disturbance variance, it can be seen from (4) that $\sqrt{k} = \sigma_t/\sigma$. The MSE from the unadjusted non-linear regression although biased, is an estimate of σ^2. Therefore, dividing (11) by .183683 (see Table 1) gives a relationship for $\sqrt{k}$. ***

(12) $$\hat{\sigma}_t/\hat{\sigma} = .425251 + .791535\hat{y}$$

*Goldberger, p 245.

**It has been suggested by House, pp 12-13 that a step function be used to estimate σ_t. The steps would then be the σ_t values in Table 2. The time intervals in the wheat data are fairly wide for reasons already mentioned and there is some concern that a step function of this nature would not adequately describe the unknown continuous function of σ_t.

***Dividing by σ has no affect on the parameter or standard error estimates so the choice of $\hat{\sigma}$ only affects the MSE in the adjusted non-linear regression.

There are some obvious deficiencies in using (12) as an estimate of $\sqrt{k}$. Since $\hat{y}$ is estimated from a heteroscedastic model, while unbiased, it is influenced by the reliability of the parameter estimates. $\hat{\sigma}$ is no longer unbiased. So, we have a situation in which the more the heteroscedasticity, the less our ability to estimate σ_t/σ. This might suggest an iterative procedure where $\hat{y}$ and $\hat{\sigma}$ are revised based on the previous weighted regression.

To evaluate how effectively (12) adjusts for heteroscedasticity, NLIN was run on (10) using (12) as an estimate of $\sqrt{k}$. Figure 12 shows the estimated values of the transformed dependent variable overlayed on the transformed data. Figure 13 shows the corresponding residual plot. The residuals still indicate a positive correlation with time although not nearly as severe as in Figure 2. Table 3 gives the correlations which have been reduced but are still highly significant. It is interesting to note that the MSE is identical to that obtained in the unadjusted model to four decimal places. This result is mostly coincidence since, in theory, dividing by $\sqrt{k}$ should leave the true population variance σ^2. However, $\hat{\sigma}^2$ from the unadjusted model is biased and therefore the estimate of the population variance in the present model should be somewhat less. Apparently, the error in using (12) as an estimate of $\sqrt{k}$ offset the expected reduction in the bias of $\hat{\sigma}^2$. It can also be seen in Table 3 that the relative standard errors of the parameter estimates responded as they did in the log transformations.

At this point, the deficiencies cited earlier in using (12) as an estimate of $\sqrt{k}$ prompts us to try an iterative approach in an attempt to improve the estimates of σ and y used in the linear regression equation. Since computer costs increase quickly with successive non-linear regressions, at first only one iteration was used to assess the effect. The change in the estimate of population variance and the parameter estimates was very slight after one iteration. Thus, there was also very little difference in the $\hat{y}$ values. The stability of $\hat{\sigma}$ and $\hat{y}$ leads one to suspect the validity of the $\hat{\sigma}_t$ values.

From (12), it can be seen that $\hat{\sigma}_t/\hat{\sigma} = 1$ when $\hat{y}$ is approximately .73. This means that y values are increased when $\hat{y}<.73$ and decreased when $\hat{y}>.73$. Therefore, if (12) was the proper estimate for $\sqrt{k}$, the variance of y values to the left of $\hat{y} = .73$ would be increased while the variance to the right would be decreased so that the variance over the entire range would be constant. Since (12) didn't adjust enough, this indicates that the proper expression for $\sqrt{k}$ should have a steeper slope.

TABLE 3

MODELS

	Unadj	$y/(a_1+b_1\hat{y})$	Unadj*	$y/(a_2+b_2\hat{y})$*
Regression				
MSE	.0337	.0337	.0315	.0309
R^2	.9420	.9365	.9453	.9417
Param Est's				
$\hat{\alpha}$	1.0360	1.0479	1.0338	1.0510
$\hat{\sigma}_{\hat{\alpha}}/\hat{\alpha}$	1.5	1.7	1.5	1.7
$\hat{\beta}$	4.2556	4.1228	4.5903	4.3878
$\hat{\sigma}_{\hat{\beta}}/\hat{\beta}$	5.3	3.8	5.3	3.0
$\hat{\rho}$	.8886	.8925	.8858	.8915
$\hat{\sigma}_{\hat{\rho}}/\hat{\rho}$	.5	.4	.5	.4
Correlation				
R_p	.2780	.0840	.3349	.0482
Prob $>\|R_p\|$	.0001	.0002	.0001	.0333
R_s	.3093	.1375	.3409	.0754
Prob $>\|R_s\|$	.0001	.0001	.0001	.0009
		a_1 = .425251		a_2 = .189735
		b_1 = .791535		b_2 = 1.106668

*Note: Model run on a data set which excluded one of the original samples.

The plot of the data in Figure 1 shows several relatively high observations in the first three groups which don't seem to correspond to the curve being fitted. A check was made on all the observations which appear to be considerably disjoint from the main body of the data. The observations denoted by "F" in the first group, "G" in group two and "G" in the third group all turned out to be from the same sample. The rest of the observations which were checked all came from different samples. A close examination of the sample in question revealed no discernible errors in recording or keypunching. However, the field observations relating to the time variable were highly questionable. Since this sample exerts considerable influence on $\hat{\sigma}_t$, particularly in the first two groups of data, it was deleted.

A non-linear regression using the unadjusted model was run on the resultant data set. Table 3 shows that the MSE and estimate of α decreased somewhat which was to be expected. The correlation coefficients show a slight increase in the positive direction.

In order to run a weighted regression on the altered data set, estimates for σ_t must be recalculated. Table 4 shows the standard deviation of the y values in each time interval. A linear regression

TABLE 4

Time Interval (days)	$\hat{\sigma}_t$ (grams)	No. of obs.
$0<t\leq5$	.0702	204
$5<t\leq14$	.1296	382
$14<t\leq20$	.1751	409
$20<t\leq28$	.1864	414
$28<t\leq33$	.2226	319
$33<t\leq41$	.2197	180
$41<t$	.2593	42

of $\hat{\sigma}_t$ on $\hat{y}$ weighted by the number of observations associated with each value of $\hat{\sigma}_t$ was run. The regression was again highly significant with an improved R^2 value of .948. Figure 14 shows the data along with the estimated regression line. The estimated equation is now

(13) $$\hat{\sigma}_t = .033664 + .196353\hat{y}$$

Looking back at (11), it can be seen that the slope of (13) has increased. Taken at face value, (13) should adjust for more of the heteroscedasticity. However, the change in data sets also increased the heteroscedasticity as was evidenced by the correlation coefficients in Table 3. To get an estimate for $\sqrt{k}$, (13) needs to be divided by the square root of the population variance. The estimate of this is $\sqrt{MSE}$ producing the following relationship.

(14) $$\hat{\sigma}_t/\hat{\sigma} = .189735 + 1.106668\hat{y}$$

As in (12), $\hat{\sigma}_t/\hat{\sigma} = 1$ when $\hat{y}$ is approximately .73. So, an increase in slope has occurred while the "fulcrum" remained the same. At this point, NLIN was run on (10) using (14) as an estimate of $\sqrt{k}$. Figure 15 shows the estimated non-linear equation overlayed on the transformed data. Figure 16 shows the corresponding residual plot which, visually, does not seem to depict any correlation with time. However, the correlation coefficients in Table 3 indicate a significant positive relationship. While not entirely successful it can be seen that the weighted regression using $\hat{y}$ is adjusting for nearly all the heteroscedasticity.

Table 3 also shows that the MSE of .0309 is slightly less than the corresponding unadjusted MSE of .0315. While this probably evidences a reduction in the bias of $\hat{\sigma}^2$, there is most likely an offsetting affect, as discussed earlier, because of the error in estimating the components used in obtaining (14). The estimate of α is approximately 1.7% higher than in the comparable unadjusted model.

So far, the weighted regression has utilized a relationship between the variance and the estimated values of y obtained from an unweighted regression. There also exists a strong relationship between the variance and the time variable. A linear regression of $\hat{\sigma}_t$ on t was performed and a weighted regression was run using the resultant linear equation as an estimate of $\sqrt{k}$. Results were similar to those obtained by using (12) in that a large share of the heteroscedasticity was adjusted for but correlations remained significantly different from zero. An additional problem which occurred when weighting by time was that the residuals retained a very strong correlation with $\hat{y}$. In the unadjusted model, heteroscedasticity is evidenced not only by the relationship between residuals and time but also by the correlation between residuals and the estimated y values. The residual plots using time were very similar to those using $\hat{y}$ in all the log models and the weighted regression with $\hat{y}$ and, for that reason, the residuals versus $\hat{y}$ were not shown in the Figures. Since the weighted regression using the time variable retained a strong correlation between the residuals and $\hat{y}$, it will not be presented in this report.

An Application To Forecasting

The main purpose of the growth model is to provide a forecast of dry head weight at maturity. The earlier a reliable forecast can be made, the better. Some of the previously discussed methods of adjusting for heteroscedasticity are more suited to a forecasting mode than others. Among the log transformations, (7) is the preferred model to use. It's only requirement is knowledge of the range of the data. An interval of 0 to 2 grams was used to find values for A_1 and B_1 but a simple computer program can be written to solve for A_1 and B_1 using the minimum and maximum dry head weight so that it isn't necessary to examine the data prior to using the log adjustment. If (7) doesn't adjust for all the heteroscedasticity, a double log transformation can be applied using the same logic that was employed to obtain (9). (8) would not be particularly useful in a forecasting mode since it requires some idea of the value of α. The weighted least squares regression approach is fully applicable to forecasting and we will refer to this as the YHAT adjustment.

The data set used in this example is the one used earlier in which a questionable sample was deleted. This data set and one containing only the first four weeks of data will be used to demonstrate how the dry head weight is forecasted. The fourth week of data corresponds to the first week of June, 1977. Harvest was not complete until early July and actual harvest data was available somewhat later so, with four weeks of data, we are forecasting at least one month ahead. The unadjusted model and the YHAT, log and double log adjustments were run with four weeks and all the data. The results are summarized in Table 5.

The $\hat{\alpha}$'s are consistently higher (about 9%) in all cases with four weeks of data. While this could result from a change in growing conditions during the latter part of the season, most likely the higher $\hat{\alpha}$'s are due to a shortage of data points where the growth curve begins to go asymptotic. This is the penalty of estimating α with most of the data concentrated on the lower half of the curve. The relative standard errors do indicate that the parameter estimates are fairly stable. Remember that an expanded data set was used throughout this report which inflates the degrees of freedom and reduces the magnitude of the relative standard error. Therefore, the relative standard errors in Table 5 imply more stability than really exists. When we are mainly interested in forecasting, the data set would not be expanded.

As was pointed out earlier, adjustment for heteroscedasticity increases the relative standard error of α. This makes the relative standard error a questionable criteria for selection of the preferred model in a particular situation. The relative standard error can be used as an indication of stability as the growing season progresses and more data becomes available but, beyond that, its importance should not be overemphasized.

The correlation coefficients then become the primary criteria to use in establishing a model preference. With four weeks of data, the double log adjustment is the only one that causes a nonsignificant correlation between the absolute value of the residuals and time. When all the data is included, the double log overadjusts and the YHAT adjustment becomes the preferred model. The YHAT adjustment has a significant correlation at the 5% level but not at the 1% level.

TABLE 5

	Unadj		YHAT		Log		Double Log	
	4 weeks	All	4 weeks	All	4 weeks	All	4 weeks	All
Regression								
MSE	.0212	.0315	.0215	.0309	.0194	.0301	.0160	.0213
R^2	.9411	.9453	.9374	.9417	.9625	.9714	.9772	.9868
Param Est's								
$\hat{\alpha}$	1.1339	1.0338	1.1413	1.0510	1.1220	1.0271	1.1191	1.0287
$\hat{\sigma}_{\hat{\alpha}}/\hat{\alpha}$	4.3	1.5	5.2	1.7	5.0	1.7	5.9	2.1
$\hat{\beta}$	4.5117	4.5903	4.5141	4.3878	4.4937	4.4008	4.5384	4.3736
$\hat{\sigma}_{\hat{\beta}}/\hat{\beta}$	4.6	5.3	5.0	3.0	4.8	3.3	5.9	2.6
$\hat{\rho}$	.9049	.8858	.9058	.8915	.9063	.8911	.9075	.8951
$\hat{\sigma}_{\hat{\rho}}/\hat{\rho}$	.6	.5	.6	.4	.6	.4	.6	.4
Correlation								
R_p	.4170	.3349	.1250	.0428	.2372	.1239	.0198	-.1195
Prob $>\|R_p\|$	.0001	.0001	.0001	.0333	.0001	.0001	.5238	.0001
R_s	.4307	.3409	.1745	.0754	.2578	.1337	.0623	-.0806
Prob $>\|R_s\|$	.0001	.0001	.0001	.0009	.0001	.0001	.0448	.0004

The R^2 values are high in all cases and the MSE's aren't directly comparable for reasons previously mentioned. On the basis of the correlation significances, the double log model is preferred when fitting four weeks of data while the YHAT adjustment provides the best results with the complete data set.

Conclusions

In the application of the two methods to the growth data used in this report, one comparison needs to be made. The estimate of α from (9) was 1.0280 while the weighted regression on the comparable data set (i.e. (12)) provided $\hat{\alpha} = 1.0479$, a difference of nearly 2%. The log estimate was less than the unadjusted $\hat{\alpha}$ and the weighted estimate was greater. However, it should also be recognized that the $\hat{\alpha}$ from (9) is well within a 95% confidence interval on the $\hat{\alpha}$ from (12) and vice versa. This makes it safe to say that the two adjusted $\hat{\alpha}$'s are not significantly different at the 5% level. The 2% difference in $\hat{\alpha}$'s with the complete data set is similar when only four weeks of data is used.

The log transformation and weighted regression have demonstrated the ability to adjust for at least the largest share of the heteroscedasticity. It should be stressed that the success of these two methods depends to a large extent on the particular set of data being analyzed. However, the general methods employed should be of help in analyzing other data sets.

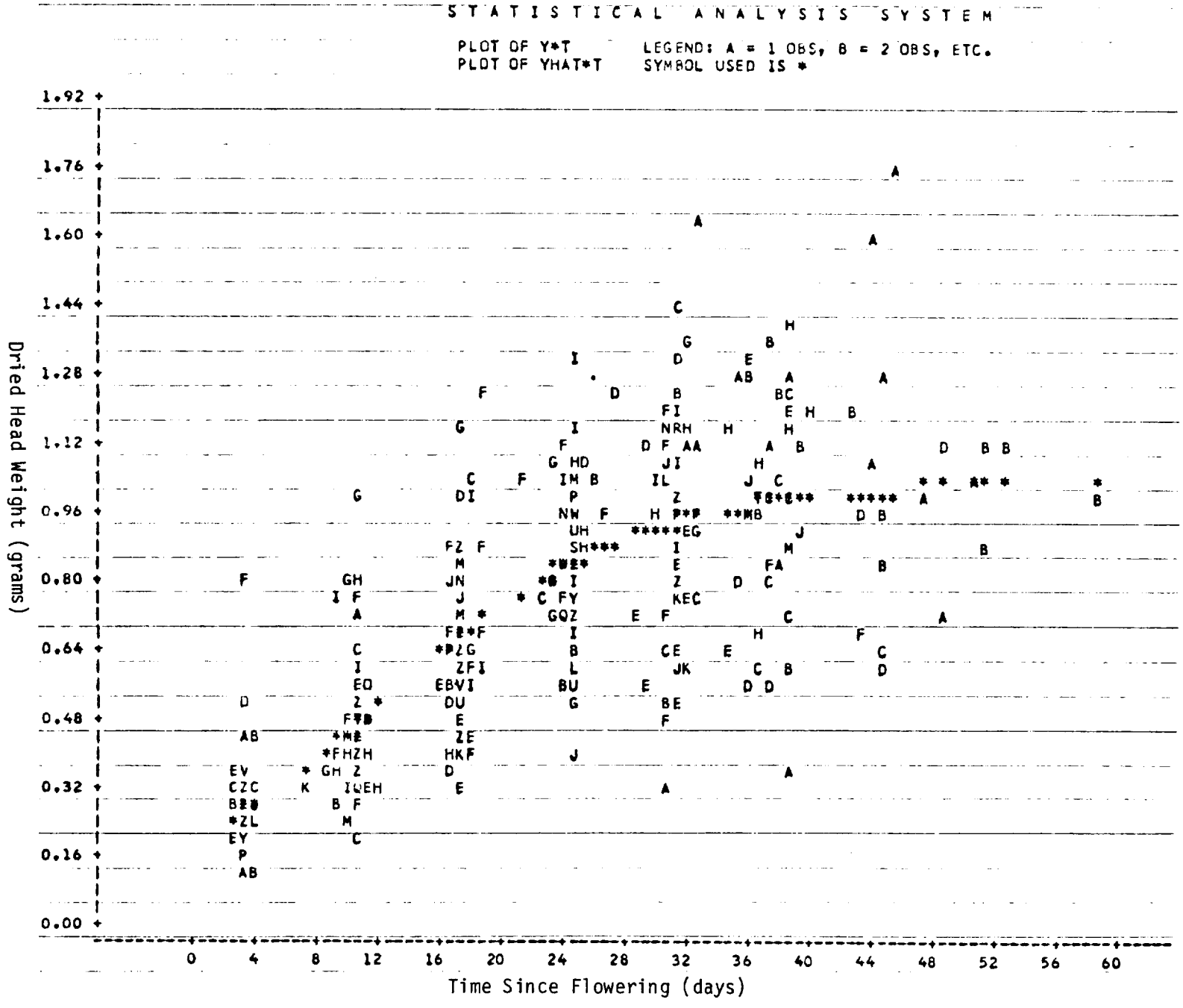

STATISTICAL ANALYSIS SYSTEM
PLOT OF Y*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF YHAT*T SYMBOL USED IS *
Dried Head Weight (grams)
1.92
1.76
1.60
1.44
1.28
1.12
0.96
0.80
0.64
0.48
0.32
0.16
0.00
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 1

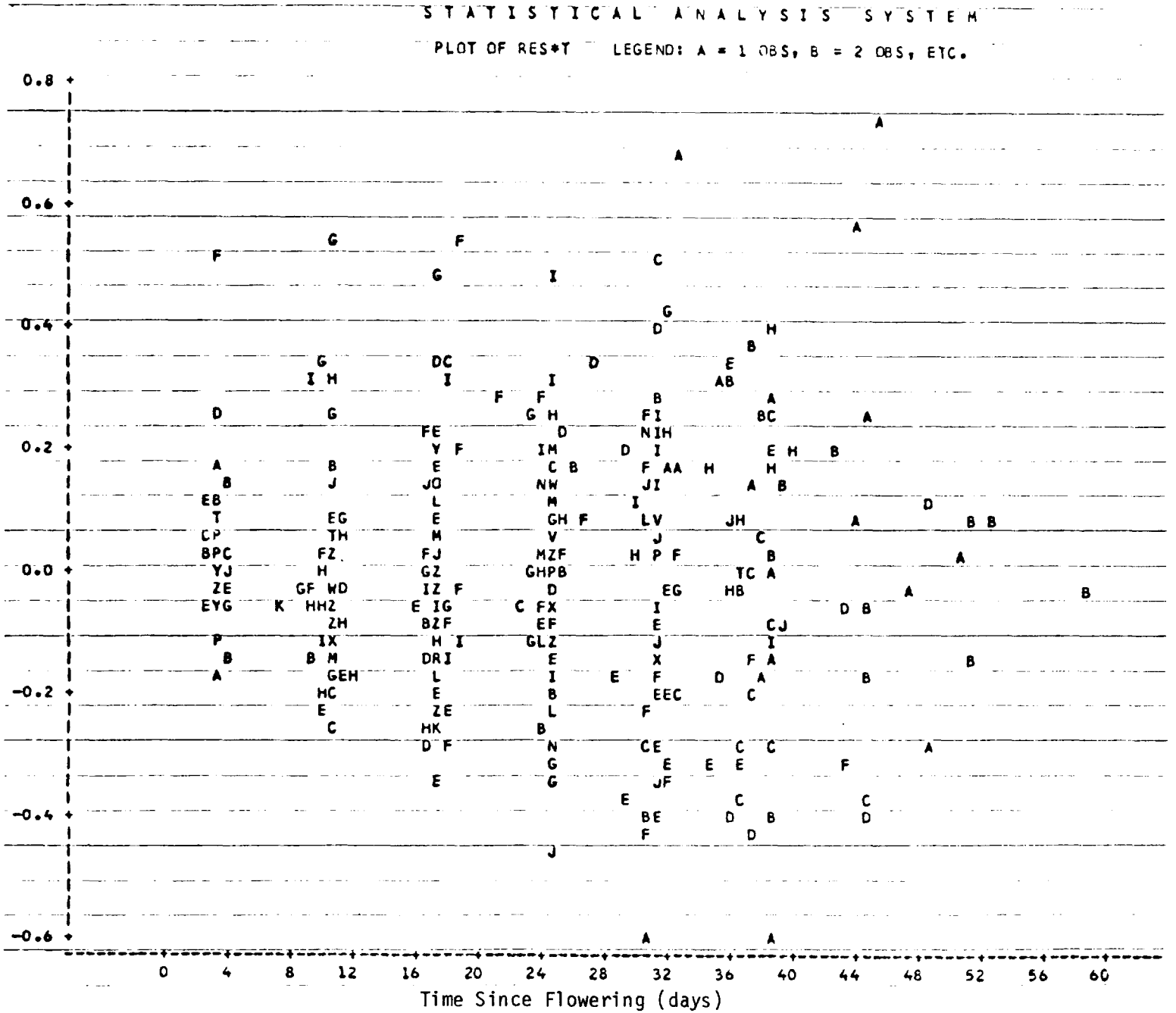

STATISTICAL ANALYSIS SYSTEM
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
Residuals (grams)
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 2

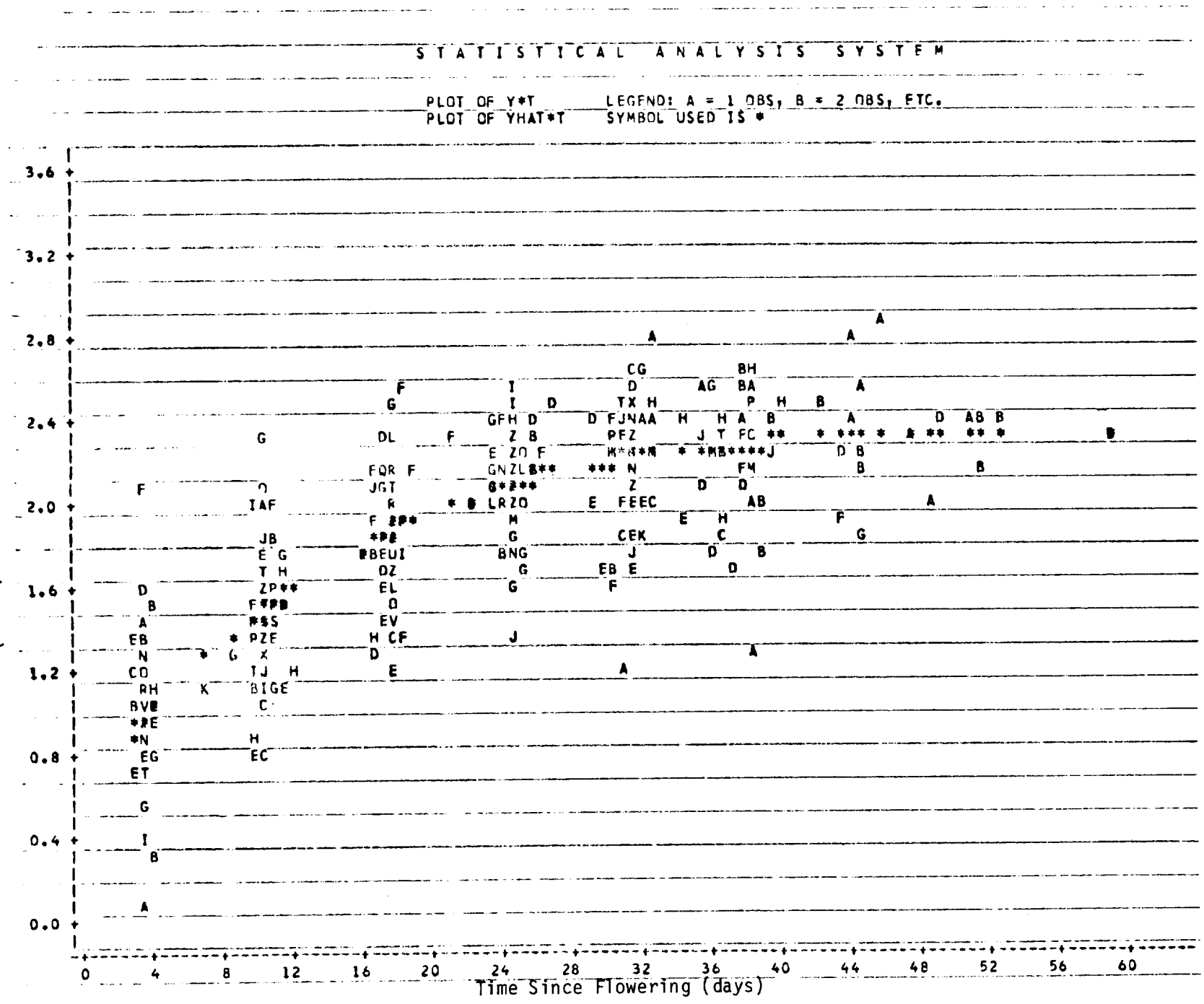

STATISTICAL ANALYSIS SYSTEM
PLOT OF Y*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF YHAT*T SYMBOL USED IS *
Dried Head Weight (grams)
3.6
3.2
2.8
2.4
2.0
1.6
1.2
0.8
0.4
0.0
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 3

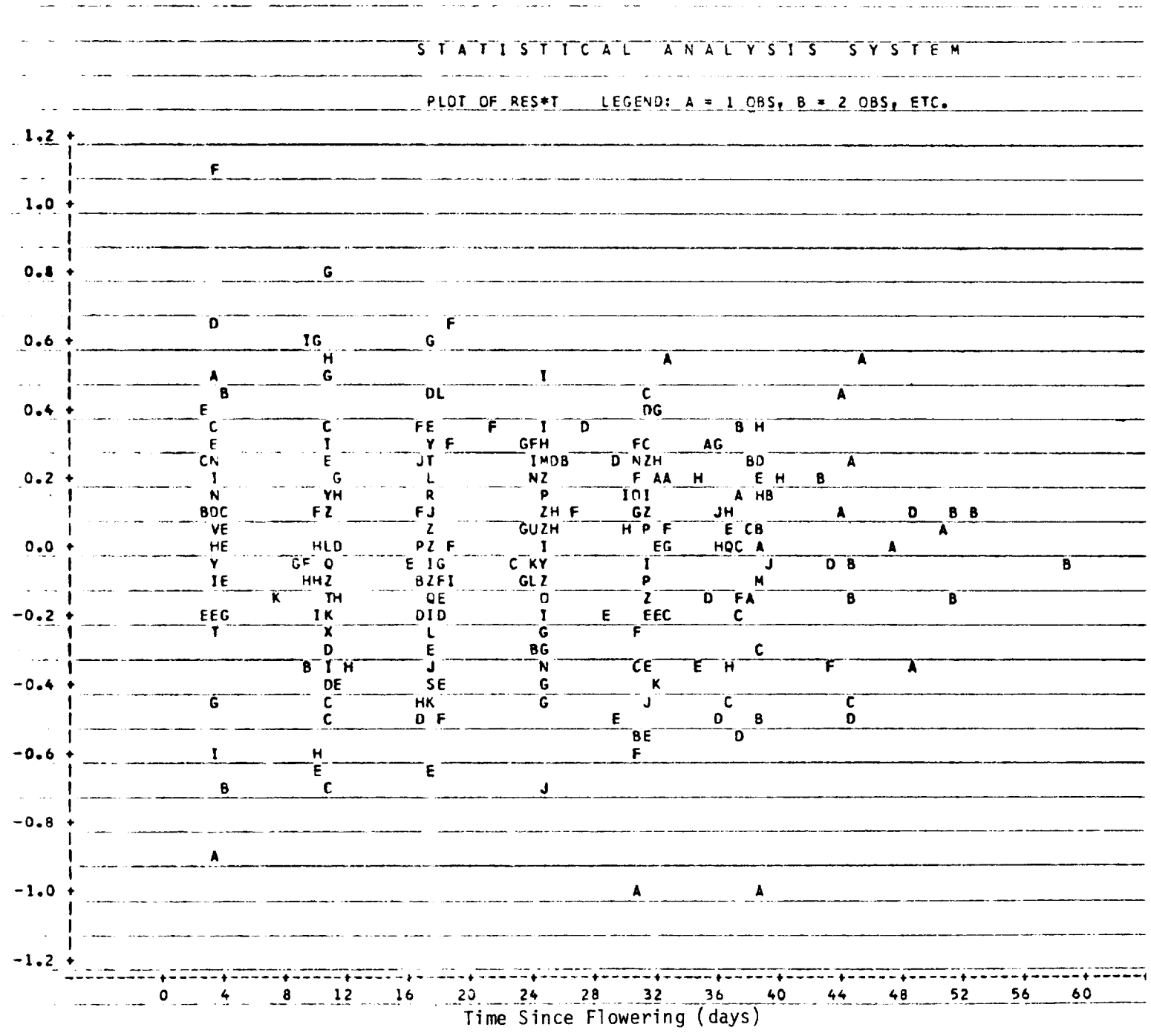

S T A T I S T I C A L A N A L Y S I S S Y S T E M
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
Residuals (grams)
1.2
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
-1.2
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
Time Since Flowering (days)

Figure 4

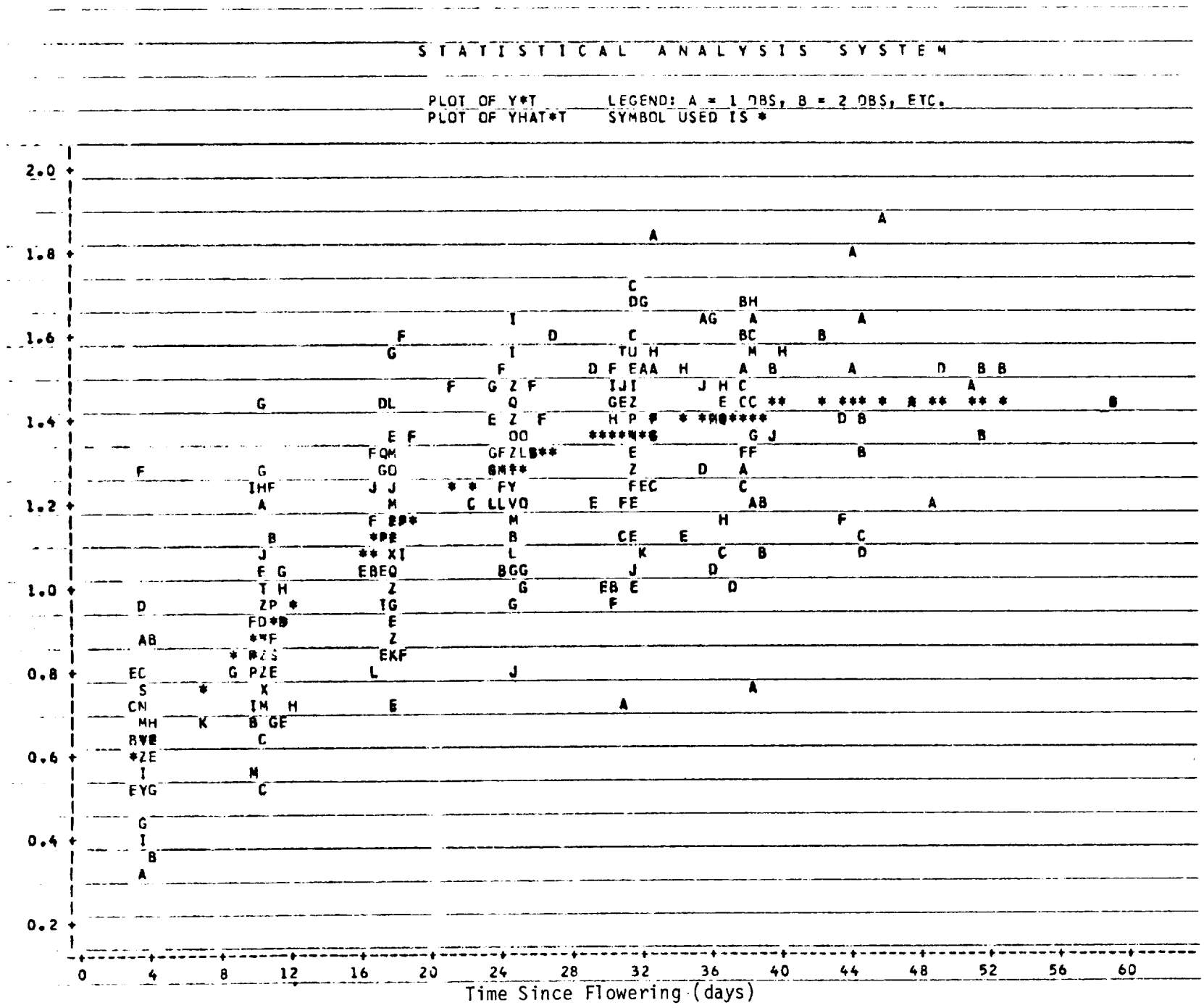

S T A T I S T I C A L A N A L Y S I S S Y S T E M
PLOT OF Y*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF YHAT*T SYMBOL USED IS *
Dried Head Weight (grams)
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
Time Since Flowering (days)

Figure 5

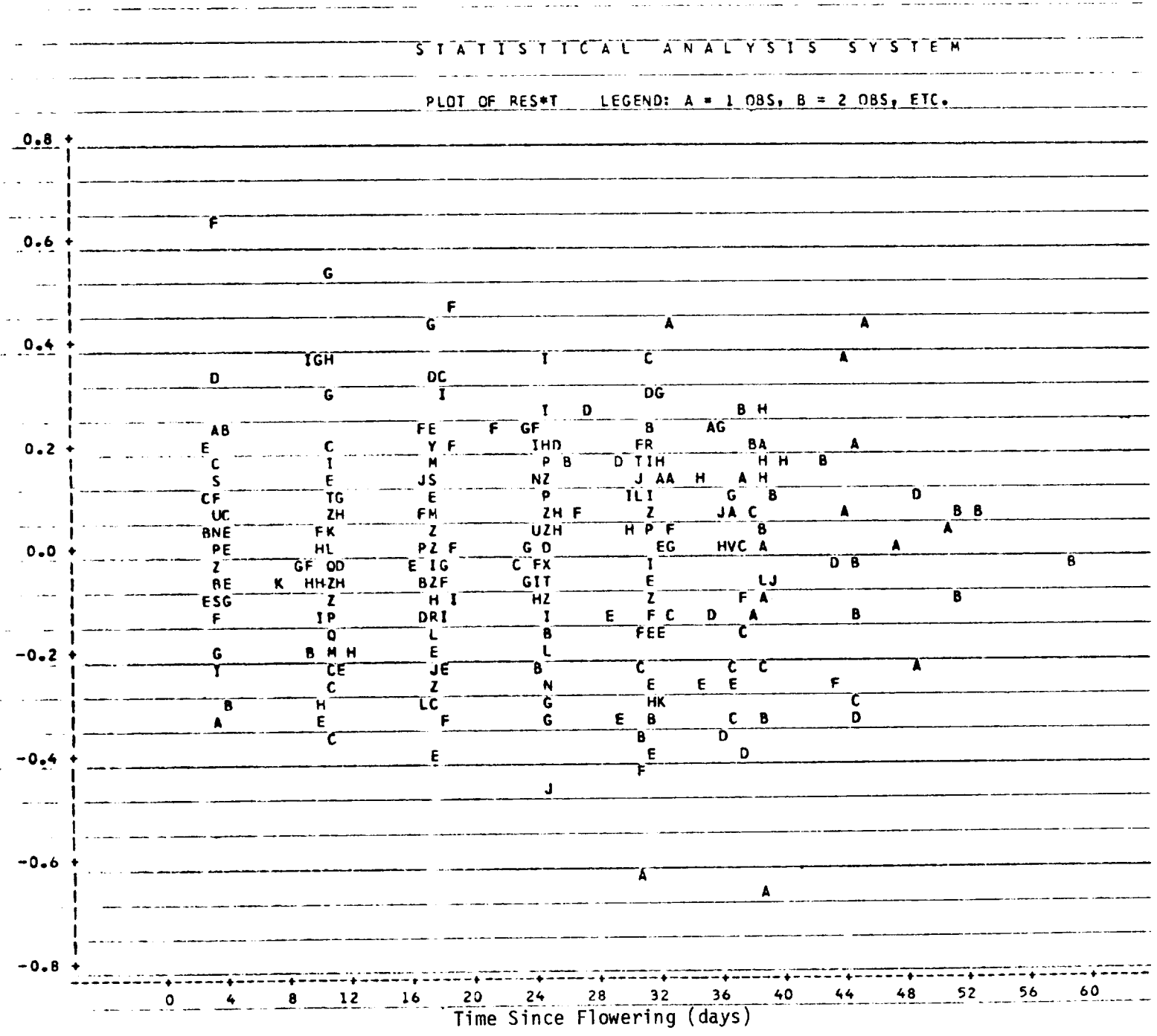

STATISTICAL ANALYSIS SYSTEM
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
Residuals (grams)
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 6

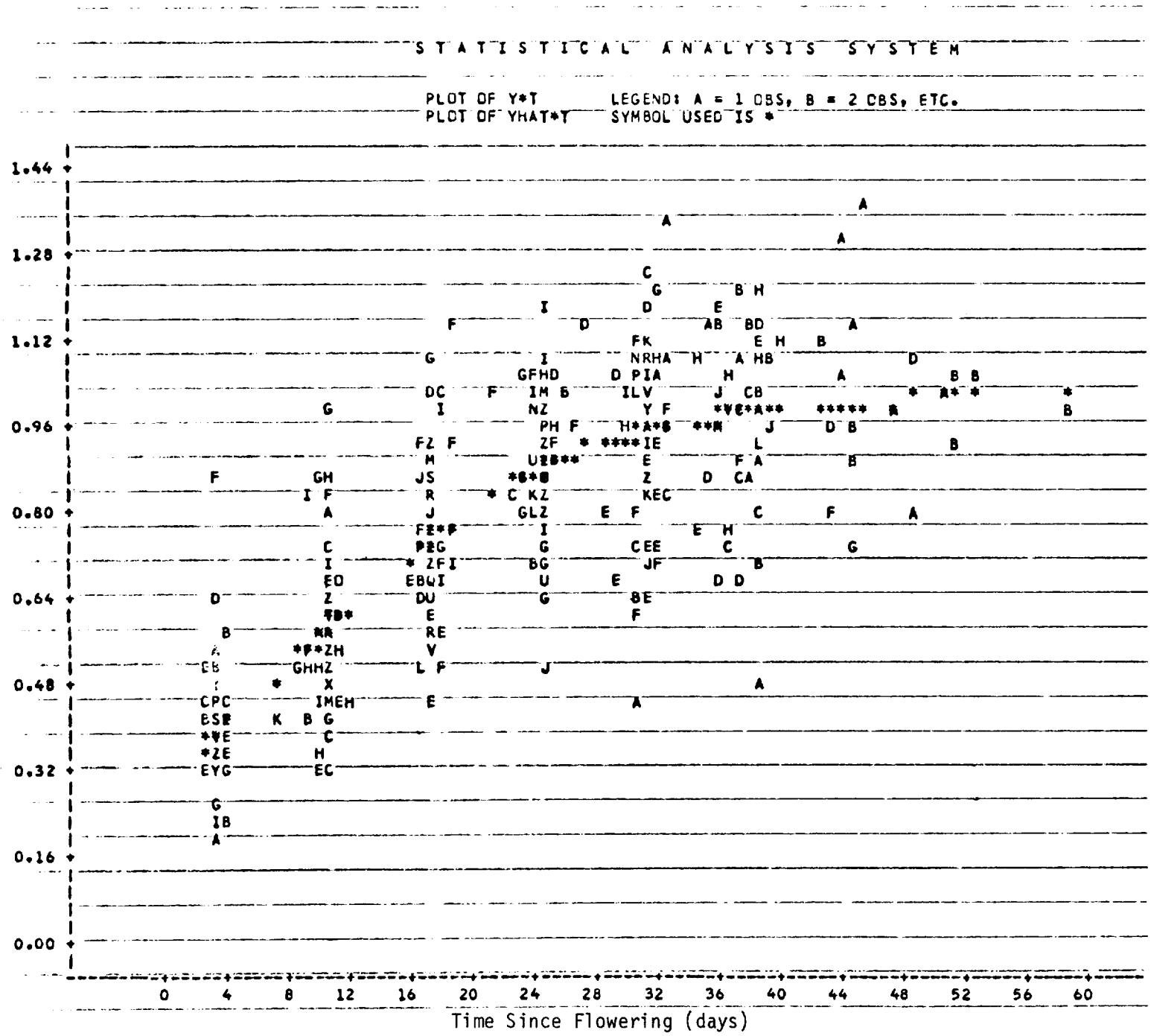

S T A T I S T I C A L A N A L Y S I S S Y S T E M
PLOT OF Y*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF YHAT*T SYMBOL USED IS *
1.44
1.28
1.12
0.96
0.80
0.64
0.48
0.32
0.16
0.00
Dried Head Weight (grams)
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 7

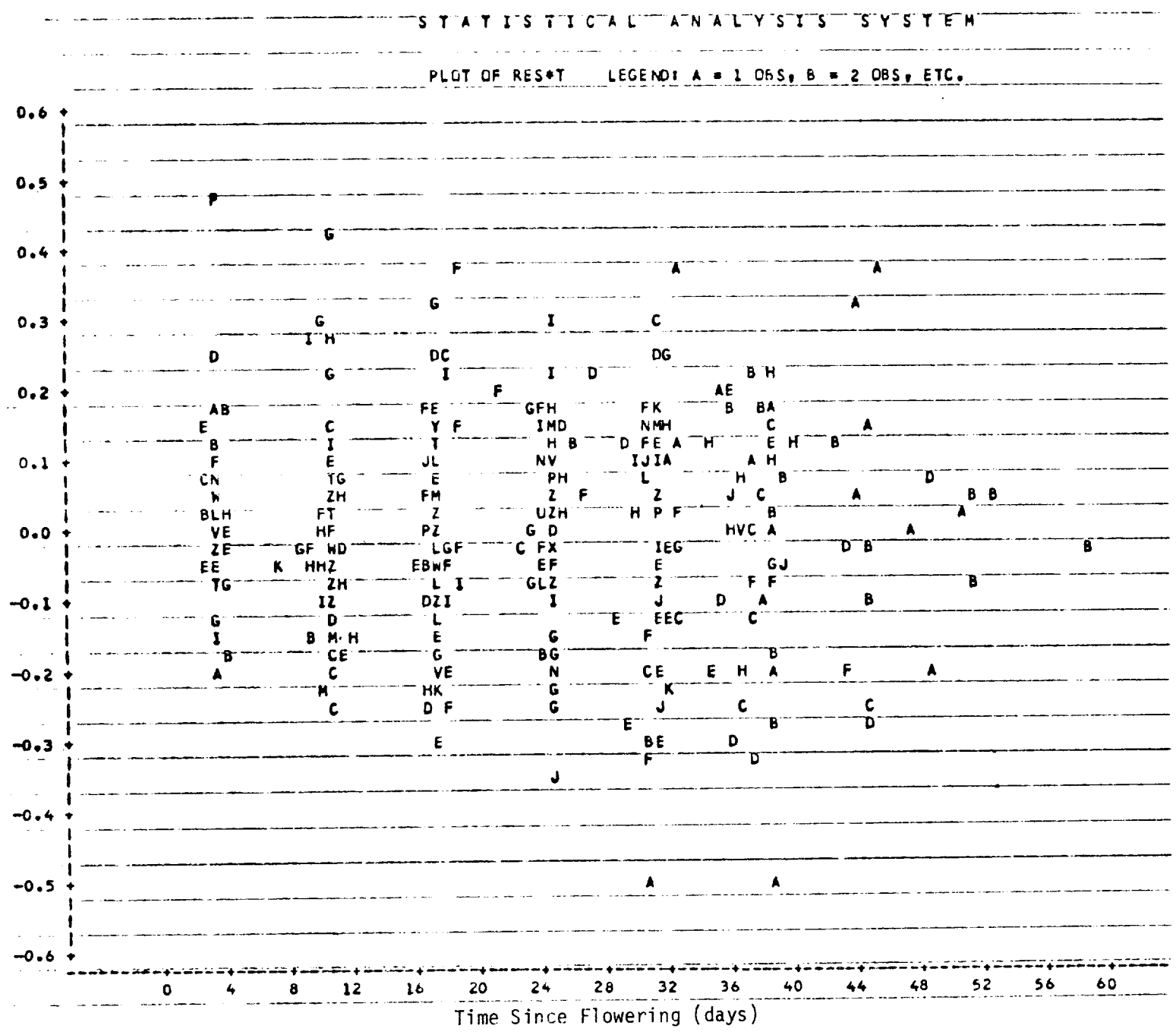

STATISTICAL ANALYSIS SYSTEM
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
Residuals (grams)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 8

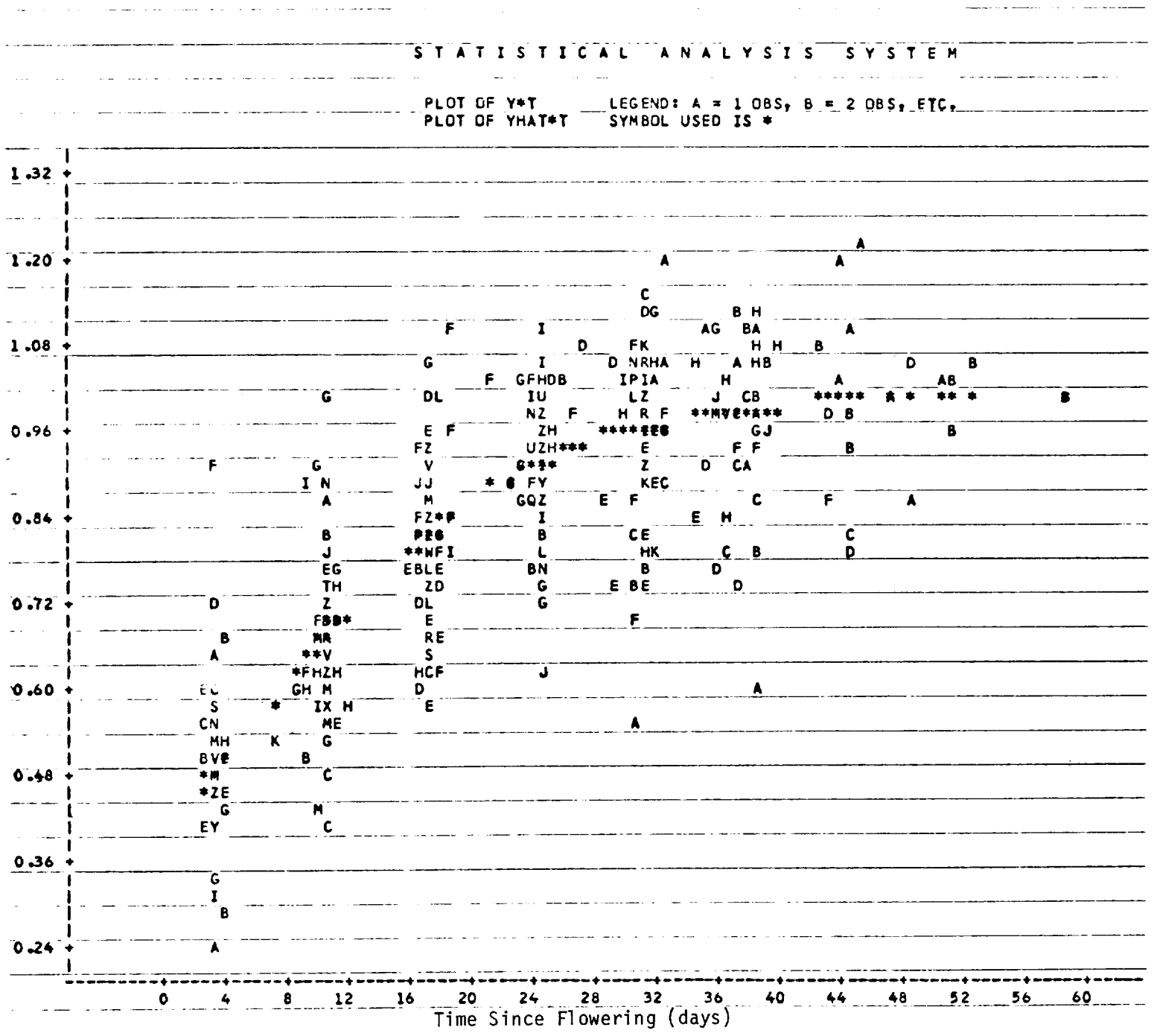

STATISTICAL ANALYSIS SYSTEM
PLOT OF Y*T
PLOT OF YHAT*T
LEGEND: A = 1 OBS, B = 2 OBS, ETC.
SYMBOL USED IS *
Dried Head Weight (grams)
1.32
1.20
1.08
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
Time Since Flowering (days)

Figure 9

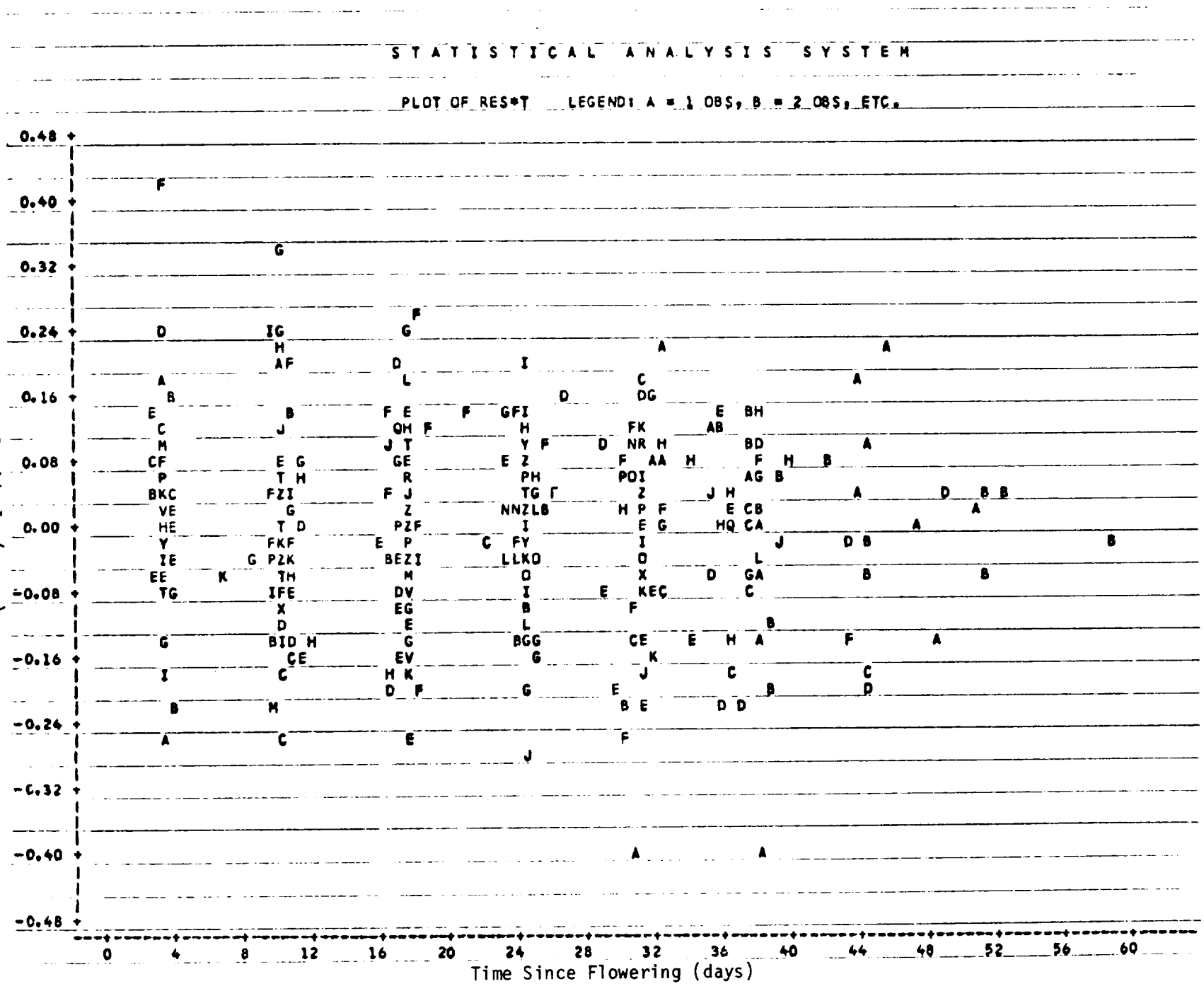
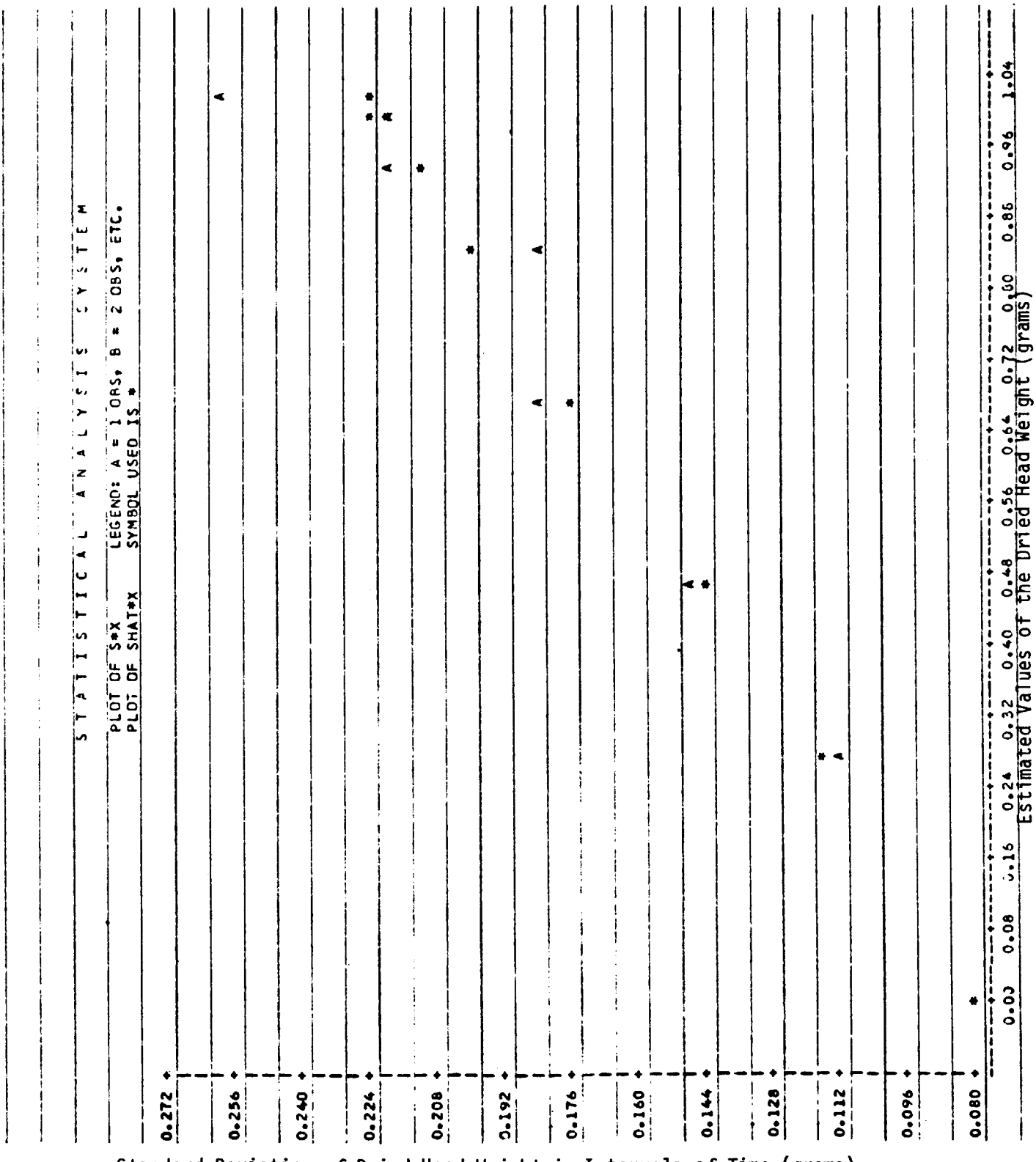

STATISTICAL ANALYSIS SYSTEM
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
0.48
0.40
0.32
0.24
0.16
0.08
0.00
-0.08
-0.16
-0.24
-0.32
-0.40
-0.48
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
Residuals (grams)
Time Since Flowering (days)

Figure 10

Figure 11


STATISTICAL ANALYSIS SYSTEM
PLOT OF S*X LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF SHAT*X SYMBOL USED IS *
0.272
0.256
0.240
0.224
0.208
0.192
0.176
0.160
0.144
0.128
0.112
0.096
0.080
0.00 0.08 0.16 0.24 0.32 0.40 0.48 0.56 0.64 0.72 0.80 0.88 0.96 1.04
Estimated Values of the Dried Head Weight (grams)
Standard Deviation of Dried Head Weight in Intervals of Time (grams)

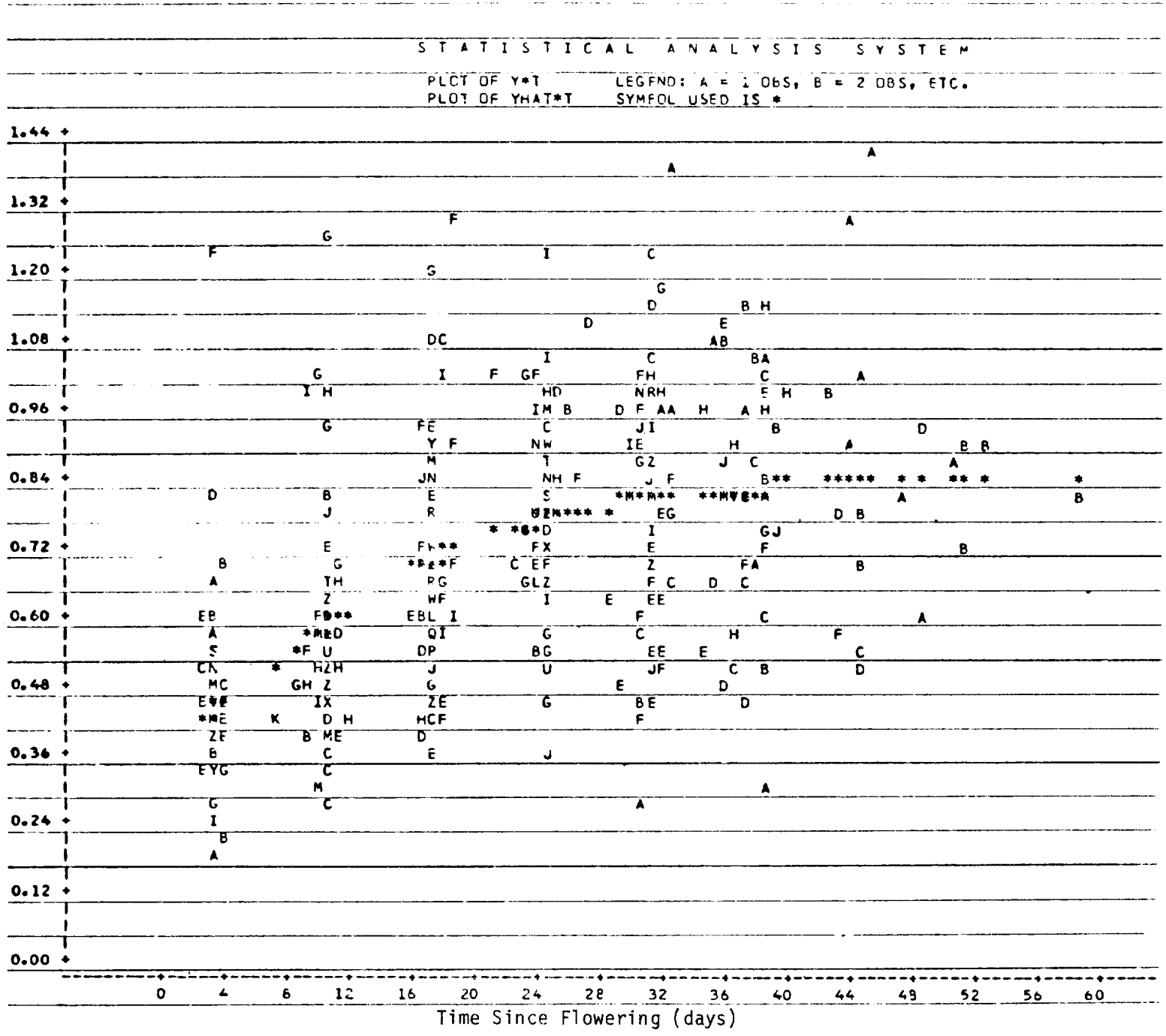

STATISTICAL ANALYSIS SYSTEM
PLOT OF Y*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF YHAT*T SYMBOL USED IS *
1.44
1.32
1.20
1.08
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0.12
0.00
Dried Head Weight (grams)
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 12

Figure 13

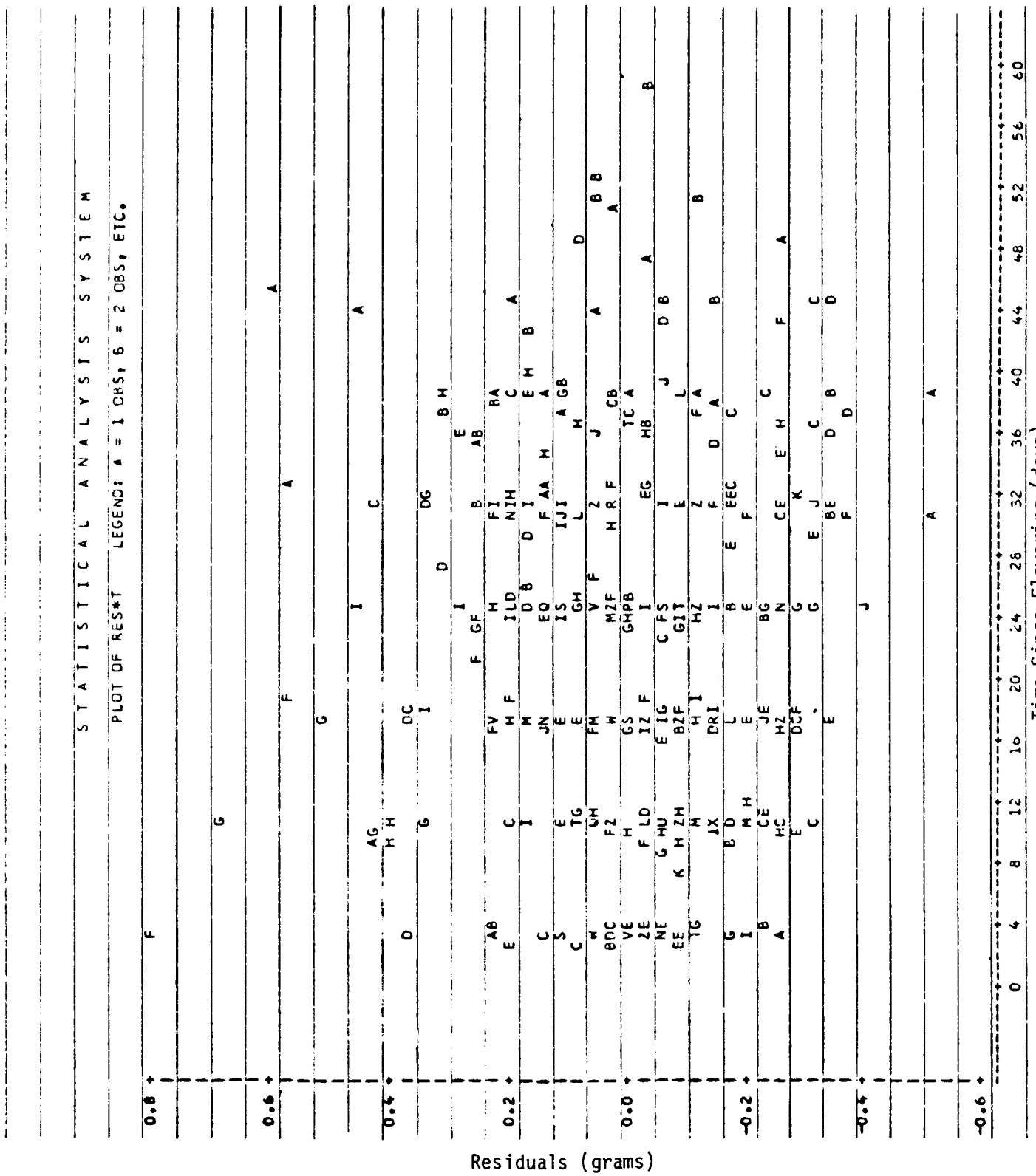
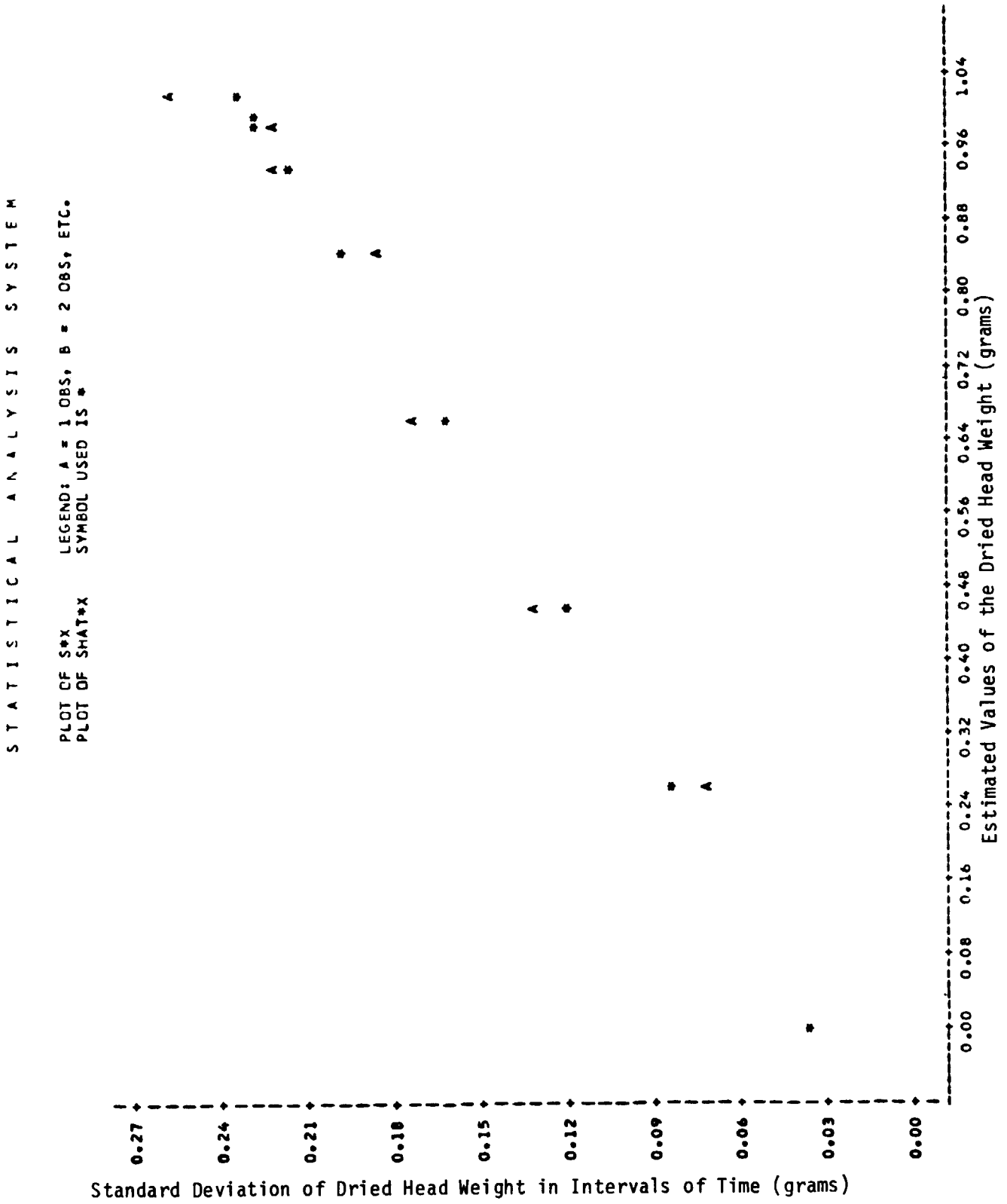

STATISTICAL ANALYSIS SYSTEM
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
Residuals (grams)
Time Since Flowering (days)

Figure 14


STATISTICAL ANALYSIS SYSTEM
PLOT OF S*X
PLOT OF SHAT*X
LEGEND: A = 1 OBS, B = 2 OBS, ETC.
SYMBOL USED IS *
0.27
0.24
0.21
0.18
0.15
0.12
0.09
0.06
0.03
0.00
Standard Deviation of Dried Head Weight in Intervals of Time (grams)
0.00
0.08
0.16
0.24
0.32
0.40
0.48
0.56
0.64
0.72
0.80
0.88
0.96
1.04
Estimated Values of the Dried Head Weight (grams)

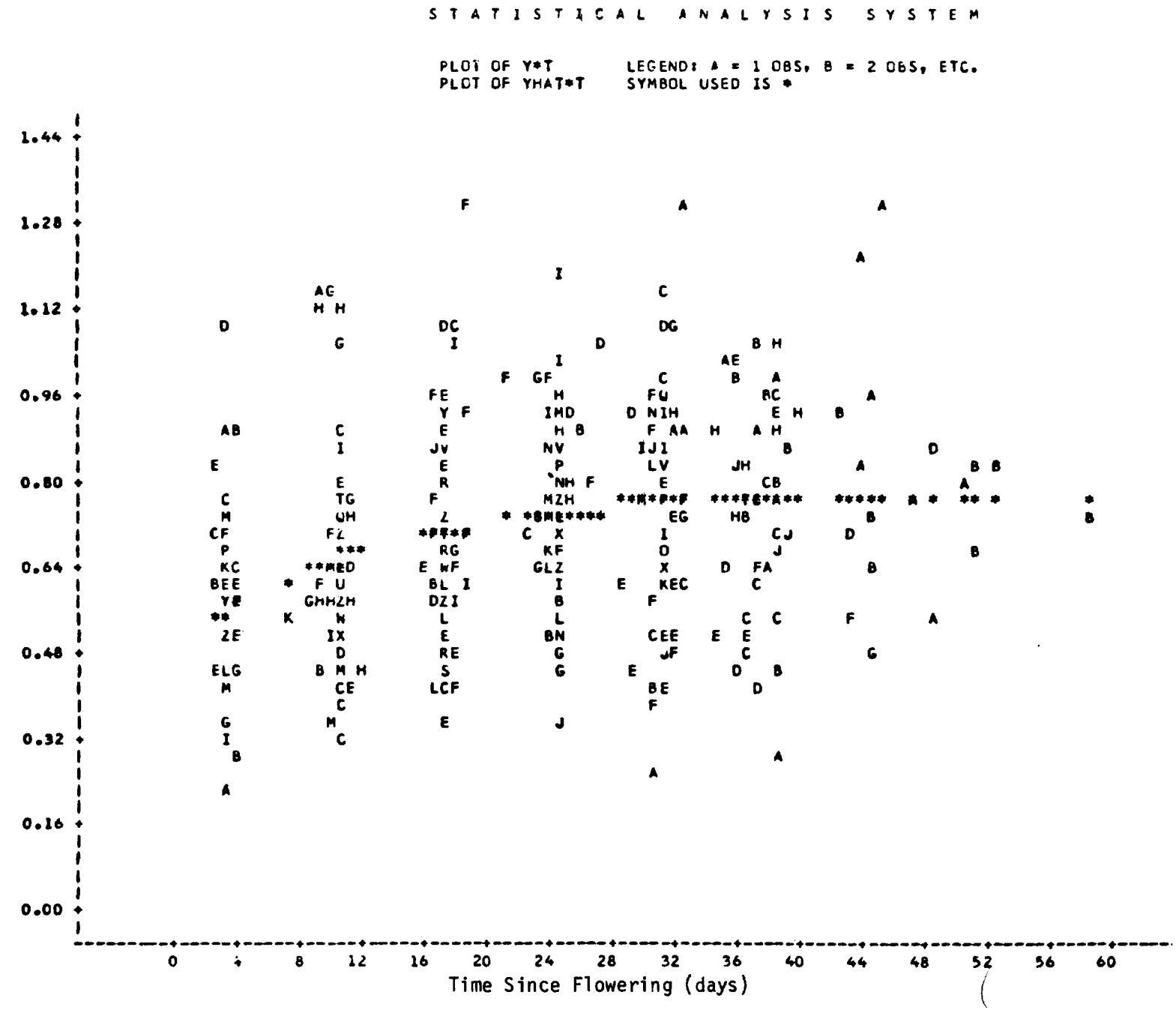

S T A T I S T I C A L A N A L Y S I S S Y S T E M
PLOT OF Y*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
PLOT OF YHAT*T SYMBOL USED IS *
Dried Head Weight (grams)
1.44
1.28
1.12
0.96
0.80
0.64
0.48
0.32
0.16
0.00
0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60
Time Since Flowering (days)

Figure 15

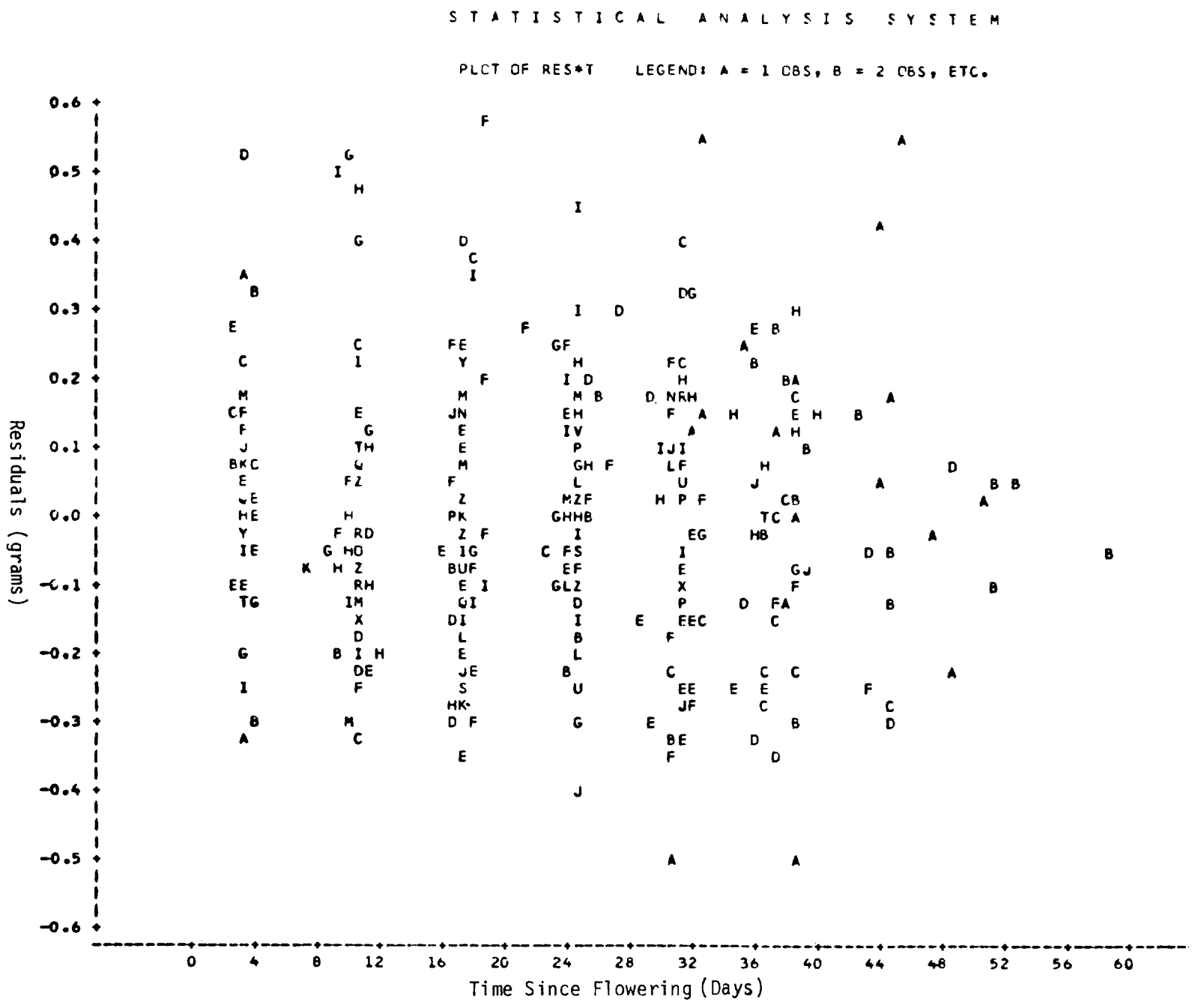

STATISTICAL ANALYSIS SYSTEM
PLOT OF RES*T LEGEND: A = 1 OBS, B = 2 OBS, ETC.
Residuals (grams)
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
0
4
8
12
16
20
24
28
32
36
40
44
48
52
56
60
Time Since Flowering (Days)

Figure 16