CONTINUOUS DATA ANALYSIS WITH ANALOG COMPUTERS USING STATISTICAL AND REGRESSION TECHNIQUES

ABSTRACT: This paper shows how certain fundamental statistical parameters of a given population--the mean, the variance, the auto-correlation function, the cross correlation function, the Fourier transform, or the power spectrum--can be estimated continuously through calculations employing simple analog techniques. In addition, analog methods are developed for continuous regression analysis wherein two or more populations or variables are compared statistically to find significant relationships.

Printed in U.S.A. 264

Bulletin No. ALAC 62023

CONTINUOUS DATA ANALYSIS WITH ANALOG COMPUTERS USING STATISTICAL AND REGRESSION TECHNIQUES

GENERAL

The need for statistical data analysis in the process industries is well known. Measurements of process variables or parameters are subject to random disturbances such as the presence of impurities in varying amounts, environmental changes, weather, etc. Very often it becomes necessary to obtain the "best estimate" of a variable over some prior time interval for purposes of control. It is this concept of "estimate" that introduces statistics.

With the availability of small, rugged and reliable analog computing components especially designed for plant environments, it becomes feasible both economically and technically to apply statistical techniques to the analysis of continuous data for either measurement or control. Of special importance is the fact that an analog device can do a simple or complex calculation task while remaining a small package in terms of physical dimensions and cost. Other computational approaches almost always imply the purchase of a relatively large "minimum" amount of hardware. Thus, one is encouraged to explore the "simple" applications--situations which pay their own way while providing experience in the use and testing of the analog approach.

Such computations can be performed Off-Line, On-Line Open Loop, or On-Line Closed Loop on continuous-signal inputs; digitizing of the analog signal is unnecessary. Noisy signals, unavoidable in many pilot or plant operations, whose deviation traces serve as the basis for subsequent calculations can be "reduced" to more meaningful form by relatively simple and economical computer circuits. The mean and standard deviation of a noisy signal can be recorded continuously and on-line, so that all subsequent problems in interpreting the data can be reduced markedly. More complex but still relatively inexpensive circuits can be used to record continuously, either on-line or off-line, the Fourier Series Coefficients of a signal. Or, in the dynamic testing of systems, the transformation from impulse response to frequency response can be accomplished, thus permitting determination of the best combination of simple input and easily-interpreted output.

THE MEAN

One of the fundamental statistical estimates is that of the "mean" or the "arithmetic average" of a variable or a parameter. When dealing with discrete information, the mean is defined by the summation

$$\bar{f} = \frac{\sum_{i=1}^{N} f_i}{N} \qquad (1)$$

which is recognized easily as the familiar arithmetic average. For data analysis, this statistical property is important for two reasons: 1) it is fundamental to the definition of other statistical parameters, and 2) it applies equally well to normal population distributions and to those that are not distributed normally.

It would appear desirable to utilize this statistical property in the analysis of continuous data or for the measurement of continuous process variables for purposes of control. In order to do so, it becomes necessary to obtain the "estimate" of the mean as a continuous and changing function of time. Specifically, one must be able to define and compute the average or mean value, $\bar{f}$, of a signal, f(t), varying with time over the interval $T_1 \leq t \leq T_2$.

As an example, assume that a steel mill is producing a continuous metal strip which, ideally, should be uniform but actually is fluctuating in thickness (as shown in Figure 1) because of inevitable random disturbances in the process. Over any time interval of reasonable length, the mean thickness should be equal to the nominal value although small deviations are allowable. A sensing device is monitoring the thickness as the strip emerges from the mill, and a transducer is generating a signal, f(t), which is proportional to the instantaneous thickness. We would like to compute the average value of this signal so that it can be compared with the desired or nominal value in order to see if the process is under control.

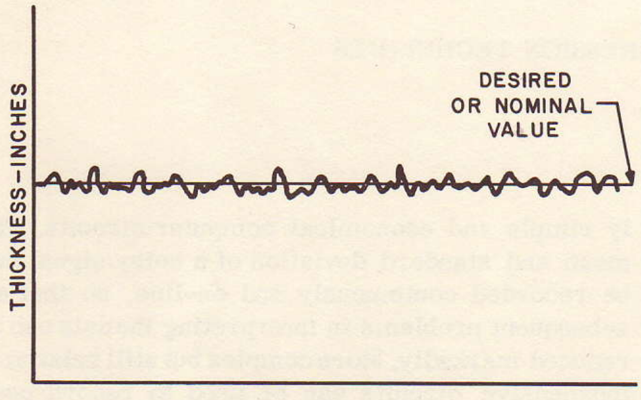

Figure 1. Thickness of Steel Plate from a Rolling Mill as a Function of Time

The most obvious definition of the mean or average value for f(t) over the interval $T_1 \leq t \leq T_2$ is

$$\bar{f}\,(t) = \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} f(t)\, dt \qquad (2)$$

This value can be computed with the simple circuit of Figure 2. At time T_1 the integrator is placed in the COMPUTE mode, and at time T_2 its output is observed. The integrator then can be reset and another average taken.

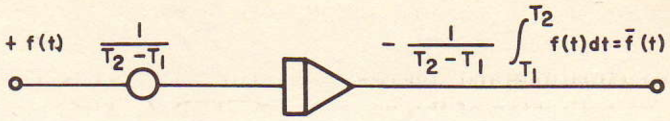

Figure 2. Analog Circuit for Calculation of Estimate of the Mean for a Fixed Time Interval

A refinement to this circuit would be to generate $\bar{f}(t)$ as a continuously varying function of time as shown in Figure 3. The time interval then must be considered as a variable so that the average is computed continuously from time T_1. In Figure 3, T_1 is considered to be zero computer time and T_2 has been replaced by t since the upper limit of the integral is a variable.

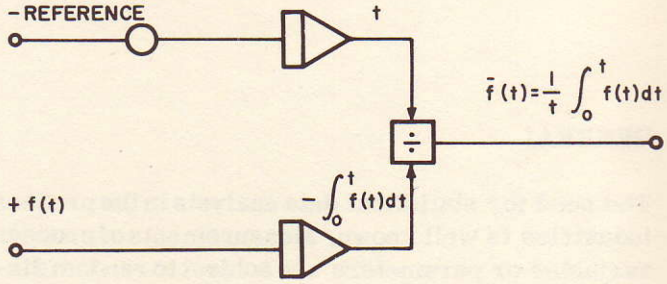

Figure 3. Analog Circuit for Calculation of a Continuous Estimate of the Mean for a Fixed Time Interval

The circuit of Figure 3, although theoretically an improvement over that of Figure 2, has two rather obvious limitations: 1) the uncertainty of the division when t = 0, and 2) the need to select maximum running time in advance, since the integrator will eventually overload. This latter difficulty also occurs with the circuit of Figure 2. In each circuit, the integration can continue only over a certain range of time, and the circuit then must be reset. However, the past values of $\bar{f}(t)$ are lost in the resetting; the average computed during the second "run" are independent of the values of f(t) obtained during the first run. If a succession of runs of length T are made and the circuits are reset each time, it is clear that the last computed average depends only on the behavior of f(t) in the last T units of time. In other words, from the point of view of the most recent average, information older than T units of time is obsolete.

The resetting necessary with the previous two circuits can be avoided, and a much simpler circuit obtained without worry about overloads and division circuits. The clue to the method is the fact that past values of $\bar{f}(t)$ become obsolete. Since the basic signal, f(t), is continuous, it seems advantageous to let past information become obsolete gradually rather than abruptly. This means that f(t) has to be defined in such a way that recent values count much more heavily than earlier values and the behavior of f(t) in the remote past has very little effect. This suggests that a weighted average be used.

The weighted-average-$\bar{f}(t)$ of a function, f(t), over $T_1 \leq t \leq T_2$ with weight function ϕ (t) is defined by (1)* where $\phi(t) \geq 0$ in the interval $T_1 \leq t \leq T_2$. Thus

$$\bar{f}(t) = \frac{\int_{T_1}^{T_2} f(t)\ \phi(t)\ dt}{\int_{T_1}^{T_2} \phi(t)\ dt} \tag{3}$$

The integral in the denominator serves to "normalize" the expression. The function ϕ (t) can be chosen arbitrarily to emphasize or de-emphasize various parts of the interval from T_1 to T_2.

Remembering the requirement that the recent past must be emphasized and the remote past de-emphasized, it follows that we should choose a weighting function, ϕ (t), which is increasing and such that $\lim_{t \rightarrow -\infty} \phi(t) = 0$. Many functions have this property but the exponential function is a natural one and leads to a simple computer circuit. Picking an exponential weighting function, $e^{\alpha t}$ ($\alpha > 0$), Equation 3 becomes

$$\bar{f}(t) = \frac{\int_{T_1}^{T_2} e^{\alpha t}\ f(t)\ dt}{\int_{T_1}^{T_2} e^{\alpha t}\ dt} \tag{4}$$

$$\bar{f}(t) = \alpha \frac{\int_{T_1}^{T_2} e^{\alpha t}\ f(t)\ dt}{e^{\alpha T_2} - e^{\alpha T_1}} \tag{5}$$

This can be simplified by letting $T_1 \rightarrow -\infty$, or

$$\bar{f}(T) = \alpha\, e^{-\alpha T_2} \int_{-\infty}^{T_2} e^{\alpha t}\ f(t)\ dt \tag{6}$$

The minus infinity in the lower limit serves to indicate that the average has been generated for such a long time that the effect of what happened before T_1 is negligible. In other words, since the exponential weighting function, $e^{\alpha t}$, approaches zero as $t \rightarrow -\infty$, the importance of events prior to T_1 is negligible if T_1 is suitably chosen.

Dropping the subscripts, Equation 6 can be written as

$$\bar{f}(T) = \alpha\, e^{-\alpha T} \int_{-\infty}^{T} f(t)\ e^{\alpha t}\ dt \tag{7}$$

Re-arranging

$$\bar{f}(T) = \alpha \int_{-\infty}^{T} f(t)\ e^{-\alpha(T - t)}\ dt \tag{8}$$

Otterman (2) defines this to be the "Exponentially Mapped Past" or EMP of f(t) over a time interval defined by α †.

Implementation of the analog circuit for solving this equation is reasonably straightforward. Differentiating Equation 7 with respect to machine time, T, (t is a dummy variable) gives

$$\frac{d\,\bar{f}(T)}{dT} = \alpha \left\{ (-\alpha e^{-\alpha T}) \int_{-\infty}^{T} e^{\alpha t}\ f(t)\ dt + e^{-\alpha T} \left[e^{\alpha T}\ f(T) \right] \right\} \tag{9}$$

$$\frac{d\,\bar{f}(T)}{dT} = \alpha \left[-\bar{f}(T) + f(T) \right] = \alpha f(T) - \alpha \bar{f}(T) \tag{10}$$

*Numbers in parentheses in the body of the text refer to references listed in APPENDIX 1.

†Those familiar with linear analysis and, in particular, convolution integrals, will recognize Equation 8 as the output of a filter whose impulse response is $\alpha e^{-\alpha t}$; that is, a first-order filter with time constant $1/\alpha$.

Equation 10 is implemented by the simple circuit of Figure 4, which is recognized easily as the circuit for a simple filter or first order lag. Note that the input and output signals have been written in terms of the more familiar notation for time, t, which is not to be confused with the dummy variable of Equation 7.

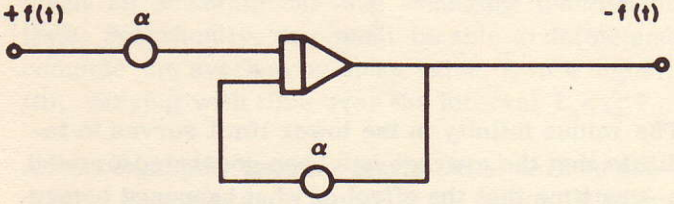

Figure 4. Analog Circuit for Obtaining the EMP Estimate of the Mean

The value of the constant, α, determines how fast past information becomes obsolete. It is chosen arbitrarily to be large enough to filter out nonessential random fluctuations and small enough not to obscure long term trends. A useful rule of thumb can be developed by examining the response of the circuit of Figure 4. If f(t) changes abruptly (step input), $\bar{f}(t)$ will follow gradually, making 95% of the change in 3 time constants or a time interval of $3/\alpha$. In other words, as shown in Figure 5, after three time constants, the integrator has forgotten 95% of the information it had before the step change. Consequently, the EMP average defined by Equation 8 is an estimate* of the mean over a time interval approximately equal to $3/\alpha$.

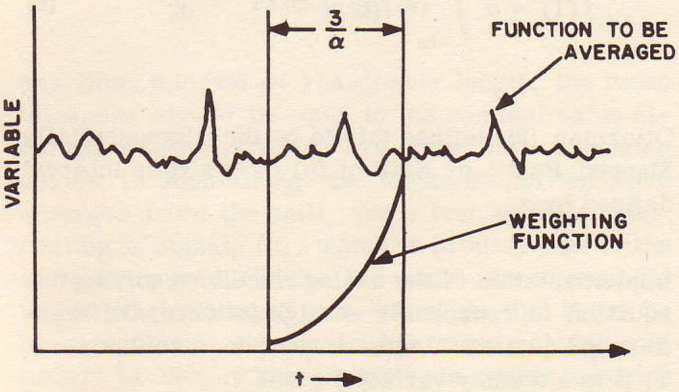

Figure 5. The EMP Mean of a Continuous Variable Provides a Measure of the Average of the Variable for a Continuously Updated Fixed Time Interval. Note the 95% decrease in the value of the weighting function over a period of length $3/\alpha$. This means that the weighted average at time, t, is virtually independent of values that occurred prior to time $t - 3/\alpha$.

In the circuit of Figure 4 it is obvious that an initial condition applied to the integrator will improve the computed average at the beginning. This value should represent a good guess as to the nominal or expected mean value of f(t). One normally would have such an estimate available. If it is a good estimate, the computed average will be reasonable from the start; if it is a bad one, it will not make any difference after about three to five time constants.

THE VARIANCE

A second important statistical parameter is the variance which is used to give a basic measure of the distribution of a population. It is defined as the square of the standard deviation and is equal to the mean-squared deviation of the variable from its mean. For discrete data, an estimate of the variance is obtained with the summation

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left[f_i - \bar{f}\right]^2 \qquad (11)$$

The term N - 1 corresponds to the number of degrees of freedom involved in the calculation of the estimates of the variance (3); practical considerations dictate that the number of samples, N, will be larger than one.

In a manner similiar to the definition of the mean, Otterman (2) defines the EMP variance as

$$\sigma^2(T) = \alpha \int_{-\infty}^{T} \left[f(t) - \bar{f}(t)\right]^2 e^{-(T-t)} dt \qquad (12)$$

which, based on the preceding development of the EMP mean, will be recognized as the weighted average of the square of the deviation of the variable from its mean.

The computer circuit for calculating an estimate of the EMP variance is developed easily without recourse to mathematical manipulations. From the definition, and remembering that averaging is accomplished by the first-order filter circuit, the following operations are required:

1) form the mean with the first-order filter circuit.

*If a 99% criterion were used, the time interval would be approximately $5/a$.

2) subtract the mean from the current value of f(t).

3) square the difference of the mean from the current value of f(t).

4) average the square with a second filter circuit.

From these requirements the circuit of Figure 6 is derived easily.

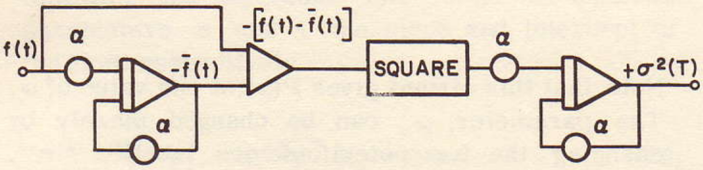

Figure 6. Analog Circuit for Calculatior of the EMP Estimate of the Variance

Example: The LD Steel Process can be used as an example of the use of the EMP mean and variance for the control of a process. This is the oxygen steel making process wherein it is possible to control bath temperatures without an external fuel supply by charging the vessel with materials that are thermally balanced. The charge materials consist of hot metal (iron), scrap, and lime.

The hot metal temperature can range from 2200°F to 2600°F, and, hence, it is necessary to measure the temperature of the iron to obtain a correct thermal balance.

A two-color radiation pyrometer method is used to measure the iron temperature while it is being poured into the vessel. A typical trace is shown in Figure 7. (For further details of the process the reader is referred to reference 4.)

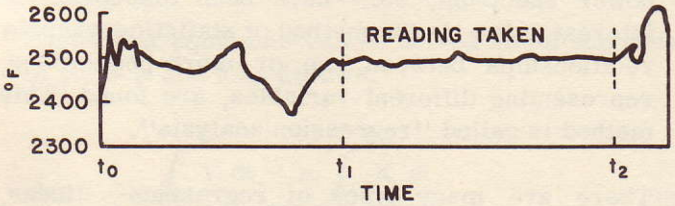

Figure 7. Plot of Hot Metal Temperature vs Time for the LD Steel Process

The initial variations in temperature, $t_0<t<t_1$, are due to the presence of smoke and the formation of voids in the pour. The operator disregards these readings until the smoke has been blown away (by a fan) and smooth pouring is established. Note that even after these conditions have been attained, the temperature measurement, $t_1<t<t_2$, is subject to fluctuations.

At present, the "temperature" reading is inserted manually into the charge balance computer (the mean value of temperature is "guesstimated" by the operator). This could be automated easily with an EMP mean value circuit since the transducer signal is a continuous electrical signal. The condition that the reading of the mean value circuit should not be used at the beginning of the time history, $t<t_1$, for reasons mentioned previously, can be automated by using the standard deviation (variance) as a control criterion, i.e., when $\sigma^2(T)$ is greater than a reference value, do not use $\bar{f}(T)$, when $\sigma^2(T)$ is less than a reference value, use $\bar{f}(T)$. The reference value chosen will depend on the maximum variance expected during smooth pour conditions. This can be mechanized readily on the analog computer by means of a comparator.

With this simple technique a better estimate of the mean temperature could be inserted into the charge computer automatically and economically.

AUTOCORRELATION

The autocorrelation function, defined as an integral between fixed limits, is converted easily to a continuous EMP autocorrelation function, ϕ (T), by the definition

$$\phi(T) = \alpha\int_{-\infty}^{T} f(t)\, f(t-\tau)\, e^{-\alpha(T-t)}\, dt \qquad (13)$$

Cross correlation also could be accomplished by the substitution of a second function, g(t), into the time delay box, τ, shown in Figure 8, so that the output of the delay box is $g(t-\tau)$ and the output of the multiplier becomes $-[f(t)]\,[g(t-\tau)]$.

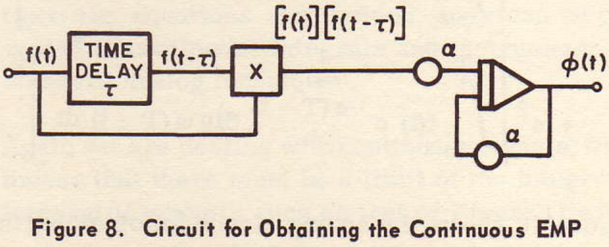

Figure 8. Circuit for Obtaining the Continuous EMP Autocorrelation Function for Time Delay τ

Reasonable time delays are obtained easily by assembling linear analog computing components. Figure 9 shows a fourth-order Padé circuit for generating a time delay, τ. This circuit is accurate to within 1 degree of phase shift for input frequencies in f(t) such that the product of the maximum useful signal frequency, ω_m, with the time delay, τ, shall not exceed 6.5 radians, i.e.,

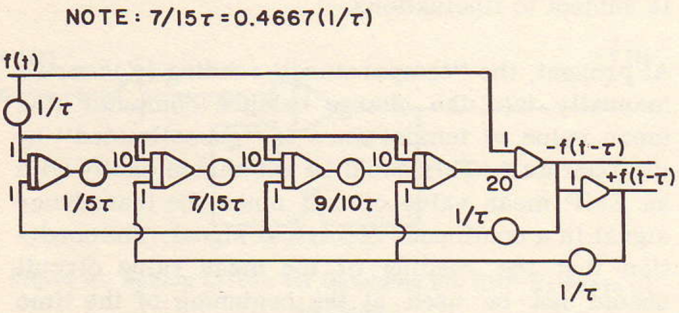

Figure 9. Circuit for Fourth-Order Pade Approximation for Ideal Time Delay of Magnitude τ (5)

$$\tau\omega_m \leq 6.5 \text{ radians} \qquad (14)$$

FOURIER AND POWER SPECTRUM ANALYSIS

The EMP Fourier transform of f(t) is defined as

$$F(\omega) = \alpha\int_{-\infty}^{T} f(t)\, e^{-\alpha(T - t)}\, e^{-j\omega t}\, dt \qquad (15)$$

or

$$F(\omega) = \alpha\, e^{-j\omega T}\int_{-\infty}^{T} f(t)\, e^{-\alpha(T - t)}\, e^{j\omega(T - t)}\, dt \qquad (16)$$

The EMP power spectrum is defined as

$$\begin{aligned} P(\omega) &= \left|F(\omega)\right|^2 \qquad (17)\\ &= a^2\left[\int_{-\infty}^{T} f(t)\, e^{-a(T - t)} \operatorname{Cos}\omega\ (T - t)\, dt\right]^2\\ &\quad + a^2\left[\int_{-\infty}^{T} f(t)\, e^{-a(T - t)} \operatorname{Sin}\omega\, (T - t)\, dt\right]^2 \end{aligned}$$

Figure 10 shows the analog circuit for obtaining the power spectrum, $P(\omega)$, of the Fourier transform, $F(\omega)$. The real component, E_1, is formed from the transfer function

$$\frac{E_1}{f(t)} = -\frac{(P + \alpha)\alpha}{(P + \alpha)^2 + \omega^2} \qquad (18)$$

and the imaginary component, E_2, from

$$\frac{E_2}{f(t)} = \frac{\alpha\omega}{(P + \alpha)^2 + \omega^2} \qquad (19)$$

Note that this circuit gives $P(\omega)$ at one value of ω. The parameter, ω, can be changed merely by changing the two potentiometers labelled "ω".

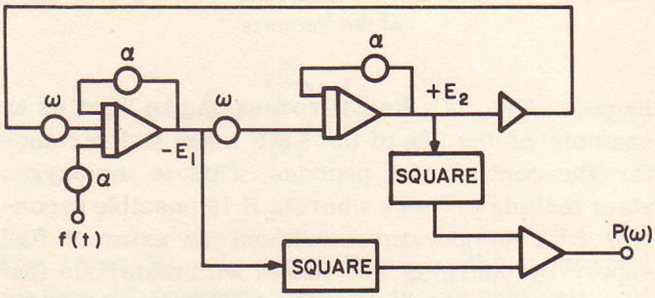

Figure 10. Circuit for Calculation of EMP Fourier Transform and Power Spectrum

An alternative is to build similar circuits in parallel, all having the same input, f(t), and differing only in the setting of ω. This will allow many points of the power spectrum to be obtained *simultaneously.*

REGRESSION ANALYSIS

Until now, only those statistical parameters that describe a single population--mean, variance, power spectrum, etc.--have been discussed. Of interest, also, is the method of statistics whereby relationships between two or more populations, representing different variables, are found. This method is called "regression analysis".

There are many types of regression---linear, quadratic, high order, multivariate, etc. These terms refer to the type of expression used to relate the variables. For example,

$$y = mx + b \qquad \text{(linear regression)} \qquad (20)$$

$$y = ax^2 + bx + c \qquad \text{(quadratic regression)} \qquad (21)$$

Regression consists, essentially, of finding the best "fit" to a set of data using a least-squares criterion. While several authors have hinted at obtaining a "least squares" fit by analog techniques for special cases, none have shown a straightforward solution to the regression problem as defined above for continuous variables.

As an illustration of how least squares fitting would be performed on the analog computer, consider the linear case defined by Equation 20. For this general equation it can be shown (6) that the following two equations will define the unknown parameters *m* and *b*, the slope and intercept of the line, respectively.

$$\sum_{i=1}^{N} Y_i = m \sum_{i=1}^{N} X_i + Nb \tag{22}$$

$$\sum_{i=1}^{N} X_i Y_i = m \sum_{i=1}^{N} X_i^2 + \sum_{i=1}^{N} X_i \tag{23}$$

Since X and Y will be continuous functions of time, the discrete summation from 1 to N can be replaced by a time integral where the total time, t, is proportional to N. Therefore, Equations 22 and 23 become

$$\int_o^t Y\,dt = m\int_o^t X\,dt + bt \tag{24}$$

$$\int_o^t XY\,dt = m\int_o^t X^2\,dt + b\int_o^t X\,dt \tag{25}$$

These two equations can be solved simultaneously to yield *m* and *b* as follows:

$$b = \frac{\int_o^t Y\,dt - m\int_o^t X\,dt}{t} \tag{26}$$

$$m = \frac{\int_o^t XY\,dt - b\int_o^t X\,dt}{\int_o^t X^2\,dt} \tag{27}$$

Figure 11 shows the analog computer circuit for calculating the "least squares" parameters *m* and *b* using the definite integral. One should note that in calculating *m* and *b* from the definite integral we form an "algebraic" loop. This brings up the question of circuit stability. It can be shown* that once the computation is under way, (t>0), the loop will be stable unless X is constant. However, if X is constant it cannot be used as an independent variable in a correlation study. Overloads at the very beginning of the computation (a region of no interest) can be taken care of with feedback limiters on the division amplifiers, or by using a "steepest descent" division circuit (7).

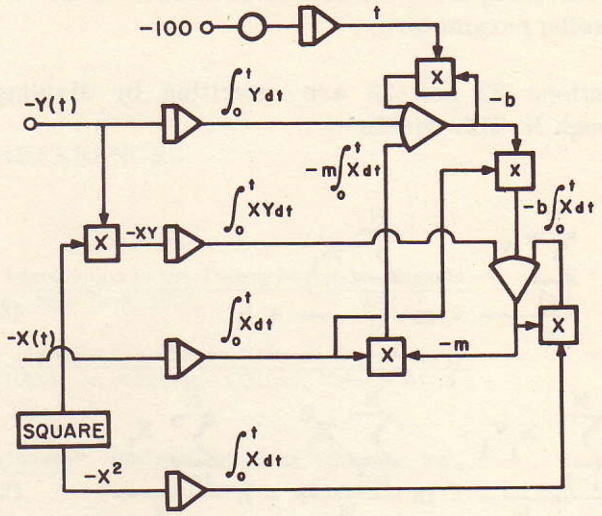

Figure 11. Circuit for Obtaining the "Least Squares" Regression Parameters m and b

It should be observed that *m* and *b* are defined at every instant of time. For small values of t (corresponding to small sample size), the estimates of *m* and *b* will be relatively insignificant and, hence, will be changing rapidly. As the time interval increases, however, the values of *m* and *b* become more significant and actually should reach "steady state" or non-changing values.

The technique for linear regression can be extended to quadratic or higher order regressions, it then being necessary to define a new set of equations-- such as Equations 22 and 23-- for determining the unknown parameters of the regression system. Once the equations are defined, they can be converted to continuous integrals and instrumented by standard analog techniques.

Again we are dealing with continuous signals, which means that there must be a limit of the integration interval if circuits such as that of Figure 11 are to

*See Reference (8)

be used. Just as before, the need for resetting of integrators can be eliminated by converting the equations for *m* and *b* to EMP equations and, thereby, obtaining truly continuous estimates of the regression parameters.

Equations 22 and 23 are rewritten by dividing through N. This yields

$$\frac{\sum_{i=1}^{N} Y_i}{N} = m \frac{\sum_{i=1}^{N} X_i}{N} + b \qquad (28)$$

$$\frac{\sum_{i=1}^{N} X_i Y_i}{N} = m \frac{\sum_{i=1}^{N} X_i^2}{N} + b \frac{\sum_{i=1}^{N} X_i}{N} \qquad (29)$$

Recalling the correspondence between EMP variables and discrete summations, one can transform Equations 28 and 29 immediately into continuous EMP notation which gives

$$\overline{Y} = m\overline{X} + b \qquad (30)$$

$$\overline{XY} = m\overline{X}^2 + b\overline{X} \qquad (31)$$

where *m* and *b* are now the EMP estimates of the regression parameters. The analog circuit required is shown in Figure 12. The statements made with regard to the stability of the circuit shown in Figure 11 apply also to the algebraic loop found in the Figure 12 circuit.

It should be observed that time has, in effect, been taken out of the problem by the conversion to EMP variables. There is no longer any need to ''reset'' the integrators since they are now serving as convolution circuits rather than pure accumulators. It follows that circuits similiar to those shown can be instrumented for continuous higher order and continuous multi-variable regressions. All that is re·quired is more analog computing equipment.

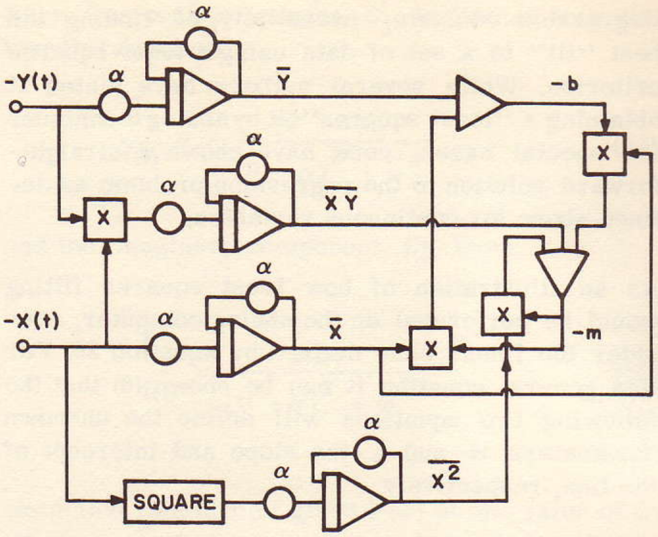

Figure 12. Unscaled Analog Circuit for Calculation of Continuous EMP Values of Regression Parameters m and b

CONCLUSIONS

The conversion of statistical parameters to EMP variables enables data analysis to be performed continuously through the use of relatively simple analog circuits. Limits on the size of the integration interval normally encountered with continuous signals have been eliminated; the need to ''reset'' the integrators is no longer required since they are now serving as convolution circuits rather than pure accumulators. A continuous estimate of statistical parameters can be calculated readily.

The concept of replacing discrete summations with the EMP mean can be a valuable one. In addition to obvious uses for instrumentation and control, for both on-line and off-line systems, this technique also can be used in analog simulation studies. For example, it is sometimes desirable to calculate the rms value of a computed variable. This is accomplished quite simply by 1) squaring the instantaneous value of the variable, 2) taking the mean of the square of the variable with an EMP circuit, and 3) taking the square root of the mean. Other uses arise in simulation work where Gaussian noise is used to disturb a particular parameter.

APPENDIX I: REFERENCES

(1) Davenport, W.B., Jr., and W.L. Root: ''An Introduction to the Theory of Random Signals and Noise'', McGraw-Hill Book Company, Inc., New York, 1958.

(2) Otterman, Joseph: ''The Properties and Methods for Computation of Exponentially Mapped Past Statistical Variables'', IRE TRANS. on Automatic Control, Volume AC-5 Number 1, January 1960, pp. 11-17.

(3) Volk, William: ''Applied Statistics for Engineers'', McGraw-Hill Book Company, Inc., New York, 1958, p. 136.

(4) Slatosky, W.J.: ''End Point Temperature Control in LD Steel Making'', Journal of Metals, Volume 12, March 1960, pp. 226-230.

(5) Brenner, M.M., and J.D. Kennedy: ''Dead Time Simulation for Electronic Analog Computers'', National Simulation Council, December 11, 1957.

(6) Widder, D.V.: Advanced Calculus, Prentice-Hall, 1947, P. 108.

(7) Favreau, R.R.: ''Dividing Circuit Obtained by Applying Method of Steepest Ascent'', Princeton Computation Center Report 132, Electronic Associates, Inc., Princeton, New Jersey.

(8) Hannaner, George: "Algebraic Loops - Some Stability Considerations" Education and Training Memo #22, Electronic Associates, Inc., Princeton, New Jersey.