Editing Very Large Databases

Edward Hill, Jr. and Carol L. French
Department of Energy
Energy Information Administration
Office of Coal and Electric Power Statistics
Division of Electric Power Statistics
Washington, D.C. 20585

Abstract

This paper presents a database edit algorithm that is defined on a two component file. The first component changes over time and the second component is static over time. The algorithm presented is defined to minimize errors in the database with minimum computation time and has been successfully applied to a large database consisting of information on electric power plants each month for one year.

1. Introduction

The basic ob1ectives of this paper are two-fold. The first is to show how data can be analyzed in such a way that errors can be explicitly defined using mathematics, statistics, and heuristics. The second is to develop an algorithm to find the defined errors.

The database to which these methods were developed is the Federal Power commission (FPC) Form Monthly Power Plant Report. It contains monthly data from all electric utilities (approximately 834) or net generation, consumption of coal, oil, and natural gas, and end-of-the-month stocks of coal and oil for each plant (approximately 3069) by prime mover by fuel type combination. These data have many attributes that change from month to month. Ten years of data are available.

We present a database edit algorithm that learns over time as data are loaded into the database. This algorithm is defined over a two component file. The first component is static over time and the second component changes over time.

Data entering the database are affected by many types of errors. The errors considered in this paper are measuring, conversion, and submittal errors introduced by incorrect measurement of various parameters that are elements of the database; i.e., the measuring of electricity generation at the power plant. Conversion errors are errors introduced by submitting data in one form and converting it to another form suitable for computing; i.e., the process of proportioning generation at a power plant according to prime movers. Submittal errors occur during the process of transforming data from one medium to another; i.e., the process of keypunching data into a data entry device. These three types of errors are detected using statistically based rules and computer algorithms. The use of statistical rules to insure data base quality is called database quality assurance.

2. Generalized Database Quality Assurance Algorithm

The basic unit that is processed is a data processing system is called an item or record. An item is made up of two parts: the key and the data. A key is that

part of a record that distinguishes the record from all other records. The data are the parts of records which are not keys and usually comprise multiple fields. A set of records is a file, and a set of files is called a database. A database with keys and data stored on a direct access storage device is called a very large database.

A generalized database quality assurance algorithm involves some test on the functional relationships between various elements in the database. The tests involving functional relationships between various elements in the database are classified as deterministic. Heuristics tests are general rules of thumb applied by the average person to decide if the data are good or bad. A generalized algorithm must have a set of well-defined attributes. Some attributes are:

- The algorithm must automatically check errors using methods based on historical data (i.e. data that have been accepted by the algorithm).
- The algorithm is modular to minimize update time.
- One person is allowed to by-pass the comparator-analyzer in the algorithm.
- The algorithm learns from previous data over time.
- The constrained verification part of the algorithm (that part of the algorithm that allows data to pass as good without entering the comparator-analyzer a second time) implements the management quality assurance policy.

These attributes are implemented in an algorithm flow chart in Figure 1. We discuss first the components of the algorithm and then the mathematics of the feature extractor, including both the initial evaluation of historical data and the development of statistical tests for error detection. The parameters of these tests are updated as each new data point is accepted.

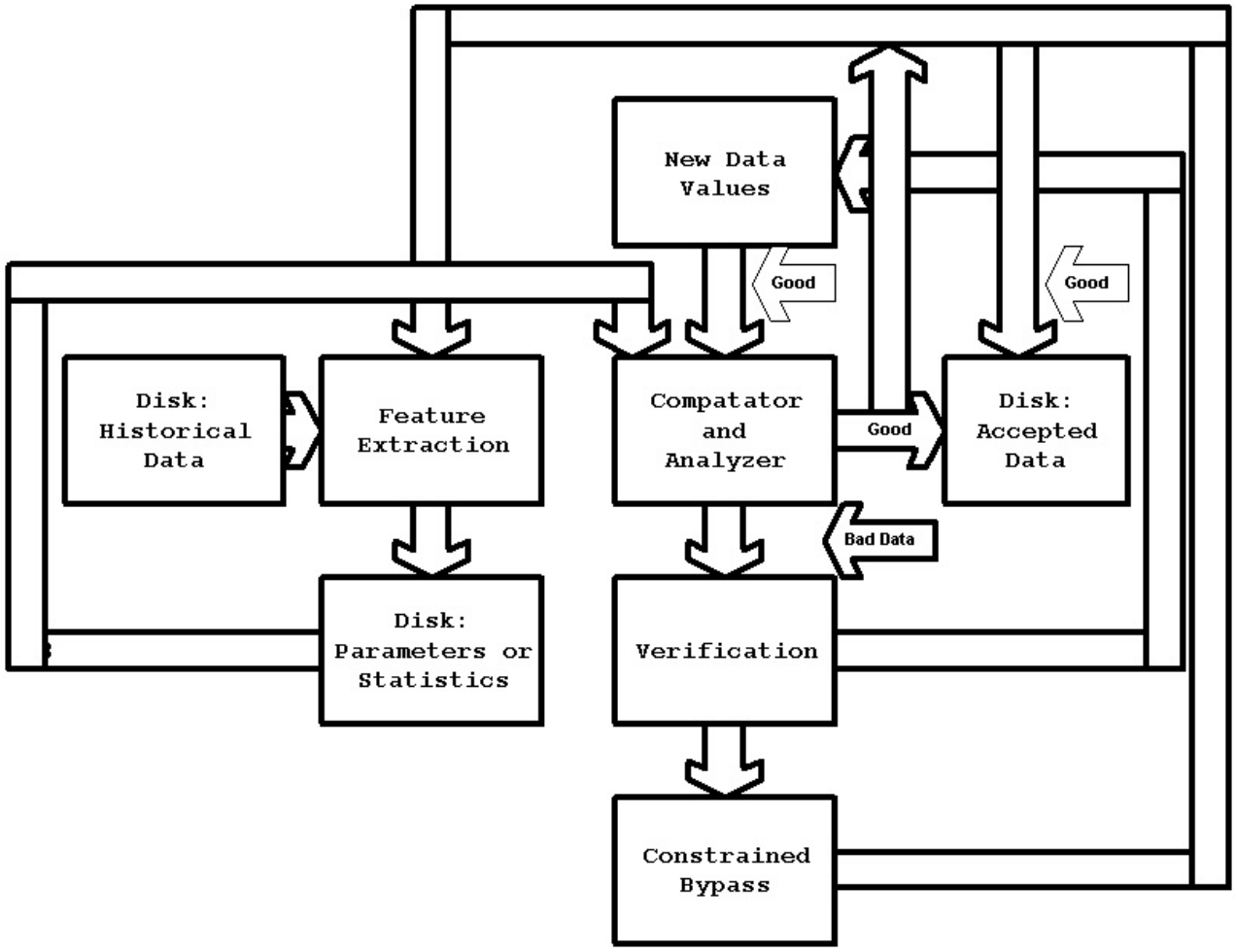

Figure 1. Generalized Database Quality Assurance Algorithm

2.1. Components of the Generalized Quality Assurance Algorithm (QAA)

There are two components of the generalized QAA presented in Figure 1. These components are the feature extractor and the comparator-analyzer.

The feature extractor has two parts. The first part is defined on those features that are extracted overlong time periods. Data from previous months are used to extract features that do not change for many months.
For example, the use of spectral analysis to classify data values as having different structures: no usable structure, but exhibiting white noise: data values with an overall increase of decrease, usually a linear trend; and data values with periodic variation, usually seasonal. These features are static over a fixed time period and they are not updated each month. The second part of the feature extractor is defined on those features that change of must be updated after each data element is accepted in the data base. Features that change for every data value are called dynamic features, and include such statistics as the mean, standard deviation or variance.

Dynamic features raise important implementation issues. Two basic issues must be considered in selecting and processing features. First, we must minimize

computation time. Second, we must minimize the number of disk accesses required to find or change data.

The comparator analyzer must have the capability of operating on both the static and dynamic features. Decisions are made about data that affect the overall quality of the data and data base policy. This part of the algorithm must be designed and analyzed to accomplish the implementer goals, such as sensitivity or specificity criteria or merely error hounds.

In Figure 1, we see that learning occurs after a data element has been accepted as good, or it has a constrained bypass label of good. Automatic learning is accomplished by updating the dynamic components using the feature extractor. The components are changed and written on the disk. When data arrives the components associated with the data are retrieved and used by the comparator-analyzer to check the data.

3. Mathematics of the Feature Extractor

As the information is entered into the computer, it begins flowing through the generalized data editing algorithm. The information is subjected to a variety of consistency checks. The checks involve tests on static and dynamic feature components. This first level of classification of the data is defined on the set of static features. The size of the FPC Form 4 database with approximately 17,280,000 items prohibits the use of manual classification. The correlation of these data in most cases makes it impossible to classify them according to their properties, without the use of computing machines. Various mathematical or statistical techniques are used to characterize power generation, fuel consumption, and stocks data according to the properties of the historic data. Classifying data according to the properties of the data assigning the properties to a fixed and dynamic class is called feature extraction. (See Table 1.)

Table 1. Feature Extraction Property Classification

	Classification	
Method	Fixed	Dynamic
Spectral Analysis	X	
Dixon Test	X	X
Influence Function Test	X	X
T-Test	X	X
Exponential Lag Time Series Test	X	X
Regression Analysis		
L-One Norm	X	
L-Two Norm	X	
L-Infinity Norm	X	

The tests can also be classified into categories of univariate and multivariate. (See Table 2.)

Table 2. Test Classification

	Univariate			Multivariate
Available Test	White Noise	Trend	Seasonal	
Spectral Analysis	X	X	X	X
Dixon			X	
T-Test	X	X	X	
Influence Function			X	X
Exponential Lag Time Series	X	X	X	
Regression Analysis				
L-One	X	X	X	
L-Two	X	X	X	X
L-Infinity	X	X	X	

The techniques to extract features from data are presented in more detail in sections 3.1 – 3.6, of this paper.

3.1. Spectral Analysis Test

We are interested in classifying variables that change over time into classes based on the historic behavior of the data. These classes are used to effectively evaluate whether new data value are correct. This is done by estimating the spectral and cross-spectral densities of a multivariate time series.

Spectral analysis is used to find cyclical patterns or periodicities in the data. The variables are classified into x categories with associated parameters and bounds. Various methods are used to decide whether a new datum value falls within the appropriate bounds, in which case we call it good. If it lies outside the boundaries we call it bad and take appropriate steps to verify the data, such as telephoning the response to check the data.

A finite Fourier transform is used to decompose the historic values of a specified variable into a sum of sine and cosine waves of different amplitudes and frequencies. The finite Fourier transform is

$$f(t) = \frac{a_0}{2} \sum_{k=1}^{m} (a_k \cos w_k t + b_k \sin w_k t) \qquad (1)$$

where

$$m = \begin{cases} 1 + \dfrac{n}{2} & , \text{if } n \text{ is even} \\ 1 + \dfrac{n+1}{2} & , \text{if } n \text{ is odd} \end{cases}$$

n is the number of observations,

a_k are the cosine coefficients,

b_k are the sine coefficients, and

$$w_k = \frac{2\pi k}{n}.$$

The periodogram is defined as

$$P(k) = \frac{n}{2}(a_k^2 + b_k^2), \qquad (2)$$

where k = 1, …, n, and m ≥ (n -1) / 2 .

P (k) represents a sum-of-squares in an analysis of variance sense for each decomposition of the process into 2 degrees-of-freedom components for each of the frequencies.

To classify the data using spectral analysis for periodicities we hypothesize that a time series is represented by

$$f(t) = \mu + A\cos wt + B\sin wt + e_t, \qquad (3)$$

where the e_t are normal independent random variable with A and B fixed.
The test for seasonality is a test of the hypothesis A = B = 0 against the alternative A ≠ 0 or B ≠ 0. This is accomplished by using

$$F_{2m-1}^2 = \frac{\sum_{k=1}^{m-1} P_n(w_k)}{\sum_{\substack{j=1 \\ j\neq k}}^{m} P_n(w_j)}$$

where F_{2m-2}^2 has the F-distribution with 2 and 2m-2 degrees of freedom.

The test statistic for white noise using equation (3) to test the null model is $f(t) = \mu + e_t$, and the alternative model is $f(t) = \mu + A\cos wt + B\sin wt + e_t$.

$$\varepsilon = \frac{mP_n(L)}{\sum_{k=1}^{m} P_n(w_k)}, \qquad (4)$$

where P_n (L) is the largest periodogram ordinate in a sample of a periodogram ordinates each with two degrees of freedom. This is a multiple of a random variable with a chi-square distribution with two degrees of freedom.
For $P_n(L) \geq P_n(w_k)$ for every k we have white noise.

The most efficient way to compute the finite Fourier coefficients is presented by Cooley and Tukey [7]. Details of theoretical results of this method are

presented by Fuller [11], and Box and Jenkins [4, 5]. The method used here for spectral analysis is presented by the SAS Guide [20].

The results of the spectral analysis used to classify variables. The classes that are of interest here are data values with no usable structure, data values with an over-all increase or decrease, and data values with periodic variation.

The type of test applied to the data values is determined by the category. Tests over all data and difference tests are applied effectively to white noise data. Difference tests and ratio tests are effective on a class of linear trend data. The effective tests for the seasonal class are tests over successive seasonal values within classes.

3.2. Dixon Test

The data we are analyzing are time series, and it is possible that the mean or variance, or both, change over time (a possible seasonal effect). The Dixon test is applied to the data in seasonal groups. The historic series are also used to estimate parameters needed to test a specific new data value (used to determine the best critical value of the statistic). The Dixon criteria, based entirely on ratios of differences between the observations may be used in cases where it is desirable to avoid calculation of the standard deviation or where quick judgment is required. For this test, the sample criterion or statistic changes with sample size. This univariate outlier test is simple and can be applied to indicate whether or not the largest or smallest observation is significantly far removed from the main body of the data.

Consider a sample observations $x_1, x_2, \ldots, x_n$.

Let x_1 denote the smallest value and x_n the largest value. The Dixon statistic is

$$r_{10} = (x_n - x_{n-1}) / (x_n - x_1). \qquad (5)$$

The symmetric population statistic is

$$r_{01} = (x_2 - x_1) / (x_n - x_1).$$

Details on this method and other outlier detection methods are presented in various sources [1, 7, 10, 12, 15].

3.3. T-Test

This test is suitable for white noise. The mean and standard deviation are updated as dynamic variables in the feature extractor. After the variables are updated we say that the algorithm has learned from the current accepted data. Consider a sample of n observations $x_1 \leq , \leq x_2, \ldots, \leq x_n$.

Let x_n be a suspected outlier. The sample mean of all values is

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

Using the mean in (6) we calculate the standard deviation as

$$s = \left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}\right)^{1/2}. \qquad (7)$$

The test for a single outlier is written using equations (6) and (7) as

$$T_n = \frac{(x_n - \bar{x})}{s}. \qquad (8)$$

Equation (8) is used to test for any doubtful value. The value T_n is used to test the hypothesis that all observations in the sample come from the same normal population. The acceptance region is defined by the significance level. Details about this method are presented in [8, 13, 15].

3.4. Influence Function Test

A paper written by Hampel in [14] treats the first derivative of an estimator viewed as functional and the ways in which it can be used to study local robustness properties. To understand this approach, consider a sample of real-valued observations $x_1, \ldots, x_n$ and their sample mean

$$\bar{x}_n = \frac{\left(\sum_{i=1}^{n+1} x_i\right)}{n},$$

as estimator for the population mean. The basic idea involved here is: How does the mean change if we introduce a new observation at some point x? The new mean is defined as x

$$\bar{x}_{n+1} = \frac{\left(\sum_{i=1}^{n+1} x_i\right)}{n+1}.$$

The change is

$$\bar{x}_{n+1} - \bar{x}_n = \frac{(n\bar{x}_n + x)}{n+1} - \bar{x}_n$$

$$= \frac{(x - \bar{x}_n)}{n+1}. \qquad (9)$$

Studying equation (9) we find the influence of a single observation with value x. This shows the relationship between the original sample size and the new observation. The influence of the single observation x, is inversely

proportional to the original sample size. When the sample size is fixed, the error increases linearly with the differences between x and the original mean over all bound.

In this context, the sample variance of a observation is affected by adding a single observation in x as

$$S_{n+1}^2 - S_n^2 = \frac{(\chi^2 - S_n^2)}{n+1}. \qquad (10)$$

When | x | < s_n, the sample variance decreases by as much as

$$\frac{S_n^2}{n+1}.$$

As | x | increase without bound, the variance grows quadratically over all bounds.

The approach described above illustrates the characteristics of influence function test in general. A particular influence function test depends on the parameter being estimated, the observation vector whose influence is being measured, and the distribution function of a random observation of a random observation vector. The influence function defined mathematically is

$$\theta = T(F), \qquad (11)$$

a functional of the distribution F. The influence function presented in [6, 9] is T (y : θ) at y.

$$I(y:\theta) = \lim_{\varepsilon \longrightarrow 0} \left(\frac{\bar{\theta} - \theta}{\varepsilon}\right), \qquad (12)$$

where

$\bar{\theta} = T(\bar{F})$ and $F = (1-\bar{\varepsilon})F + \varepsilon\delta y$ is a perturbation of F by δy, the distribution function for a point mass of one at y.

The influence function test over a11 data in the historical sample may be considered. The new observation is used with the influence function test to decide whether the new observation is an outlier or a valid observation. We reject an observation with too great an influence.

3.5. Exponential Lag Time Series Test

The exponential lag test is suitable for between-variable comparisons and for the detection of white noise. We wish to forecast some data values for at least one period based on historical data. It is desirable to be able to make these forecasts quickly, cheap1y, and easily. The number of pieces of information required to do the forecast must be kept at a minimum. The method must be adaptive and capable of learning from past experiences. The method must have

predictive capability for at least one period. A model that incorporates all of these attributes is the exponential model.

This model varies with the type of data being forecasted. Predicting data with no definite seasonal pattern and no long run trend is accomplished using the following procedure. Take the weighted average of all past observations and use it as a forecast of the present mean of the distribution.

$$\bar{E}_t = \alpha E_t + (1-\alpha)\bar{E}_{t-1} \qquad (13)$$

where E_t = data values during the t'th period,

$\bar{E}_t$ = forecast of the expected data values in the t'th period, and $0 \le \alpha \le 1$.

This model applied to seasonal data are defined to adjust for the seasonal trends.

$$\bar{E}_t = \alpha\left(\frac{E_L}{F_{t-L}}\right) + (1-\alpha)\bar{E}_{t-1} \quad , \qquad (14)$$

where $0 \le \alpha \le 1$, and L is a month period usually 12 months. The estimate of the expected deseasonalized data values in period t is

$$F_t = \bar{\beta}\left(\frac{E_L}{\bar{E}_t}\right) + (1-\beta)F_{t-L} \quad ,$$

where $0 \le \beta \le 1$, for the estimate of the seasonal factor for period t. The value of E_t from equation (13) is used in forming a new estimate of the seasonal factor in equation (14). The new estimate F_t, is a weighted sum of the current estimates, $\frac{E_t}{\bar{E}_t}$, and the previous estimate, E_{t-L+1}. A forecast of the expected data values in the following period is

$$E_t^* = \bar{E}_t F_{t-L+1} \, . \qquad (15)$$

For more details on this method see [2, 12, 18, 24]. The exponential lag time Series test may be used over all data values to calculate the necessary information for the exponential lag time series model. Equations (13), (14), and (15) are used to forecast the data values for the following period. Selective data values are used to take into account the seasonal trends. This method is used since the exponential lag data values as a function of α. In this

case, the predicted value would be a better approximation of the value in the following period.

3.6. Regression Analysis

Regression analysis methods are suitable for trend, white noise, and seasonal data. These methods are sometimes limited because they assume knowledge of some functional relationship between the data values. The regression methods discussed here are those where the functional relationship is known and we desire to find the associated coefficients to minimize the error in some norm [3, 16, 21, 22].

3.6.1. L-One Norm

In this norm we desire to determine the coefficients of F(t) to satisfy the L-One norm criterion. The L-One criterion requires that the absolute sum of the deviations be minimized as

$$\sum_{i=1}^{n} | f(t_i) - F(t_i) | . \qquad (16)$$

A modified Levenberg-Marquardt algorithm using linear programming is used to solve the L-one approximation problem [16, 22]. This method works best when the errors are independent and follow a double exponential distribution.

3.6.2. L-Two Norm

We assume that the error is normally distributed. Given a function F(t) that approximates a function f(t), the L-Two norm is

Minimize

$$(\sum_{i=1}^{n} (f(t_i) - F(t_i))^2)^{1/2}. \qquad (17)$$

In this case the function F(t) may be linear or nonlinear with coefficients to be determined.

In some cases the function F(t) is approximated over all the data values. This assumes you can find a suitable F(t)to approximate f(t). The coefficients are obtained using the methods defined by Shrager [21]. Differences are taken over successive data values to take into account possible seasonal behavior. The function F(t) is determined to minimize the square root of the square of the differences.

3.6.3. L-Infinity Norm

The coefficients of the approximating function F(t) are determined to satisfy the L-infinity norm criterion. The L-infinity criterion requires that the maximum absolute deviation of F(t) from f(t) be minimum: that is,

Minimize

$$\max_{1\le i\le n} | f(t_i) - F(t_i) | . \qquad (18)$$

This nonlinear function approximation is made using a modified Levenberg-Marquardt method and linear programming. An algorithm is used that approximates functions in the L-infinity norm [16, 22]. This method should be used when the data has outliers that must be considered with great importance. In an extrapolation using this norm, the maximum error will always be known. The approximating function is determined using all the data. The coefficients of the approximating function are used to forecast new data values.

4. Mathematics of a Comparator Analyzer

The comparative analyzer uses a new data value and extracted features from the data base on past history to decide the validity of the new data value. This usually involves setting up a discriminate region, confidence intervals, tolerance, α in an exponential lag model and significance levels.

An example is presented below to illustrate how the comparator works. In this example, a quadratic discriminate function is applied to data. A simple discriminate region is defined by considering a function

$F(x) - Ax^2 + Bx + C,$

Where A, B, C are unknown coefficients to be determined in some norm.

$F(x^*) = A(x^*)^2 + Bx^* + C,$

for some x^* (i.e. a new data value).

$$| F(x^*) - F(x) | \le R \qquad (19)$$

When (18) is true then x^* is in the region; otherwise, it is outside of the region defined by the inequality. Confidence intervals for the comparator analyzer are defined as follows:

μ - Is the mean of a normal distribution, when the variance, , is unknown.

$\bar{x}$ - Is a random variable, the sample mean, obtained from a random sample of size n.

$$z = \frac{(\bar{x} - \mu\sqrt{n})}{s}$$

Has a t-distribution with (n – 1) degrees of freedom, and

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(n-1)}$$

In this case Y is an estimator of μ whose distribution depends on μ.

The confidence interval is

$$P\{\bar{x} - t_{m/2:n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{m/2:n-1}\frac{s}{\sqrt{n}}\} = 1 - \alpha$$

where $t_{\alpha/2:n-1}$ in the 100α / 2 percentage point of t.

An error tolerance is used in many cases to decide on a certain acceptable level of error. Let F^* (p, x, y) be a feature function value. Let $\tilde{F}$ (p, x, y) be a test or a new function value. The tolerance is defined as

$$|F^*(p,x,y) - \tilde{F}(p,x,y)| \le \tau.$$

5. Conclusion

We have found that a collection of feature extraction methods is required to locate errors in a data base with a varity of parameters having different statistical properties. Two major objectives govern the methods discussed here. The first is minimum computation time and the second is maximum error resolution. To accomplish these objectives we have partitioned the set of applicable methods of mathematical feature extraction into two classes: namely, fixed and dynamic. The partitioning is not unique since some methods of mathematical feature extraction are in both classes. The dynamic class contains those methods that are computationally feasible in accomplishing our goal of minimum computation time.

The spectral analysis test used in this algorithm has been a useful method for looking at the historical data. This method was used to partition the historical data into subclasses. Computationally, this is a fixed type classification. The results of this analysis are used to decide the type of dynamic test to apply for error resolution. The regression analysis methods were hard to apply because they require previous knowledge about the structure of the data to associate a particular function to fit. These methods are classified as fixed by our computation objective.

The generalized data base quality assurance algorithm presented in figure 1 was applied to a test batch of records. A summary of the results is presented in [2], and Table 3).

Table 3. Algorithm Summary for Test Batch

Processing Time	Number of Errors Found	Number of Variables
22.91 sec	158	13,340

We plan to study the learning capabilities of this algorithm. Many issues must be resolved to improve the quality of a data base. For a given set of data an optimum fixed parameter file update interval must be developed. In the dynamic parameter file an optimum strategy must be found to update the file to insure maximum error resolution, without the necessity of generating too many data points. This involves finding a balance between sensitivity and specificity.

References

1. Anscombe, F. J. Rejection of Outliers, Technometrics, 2, 2 (may 1960), 123-147.

2. Bishop, Y. M. M. Imputation, Revision and Seasonal Adjustment, 1980 Proceedings of the American Statistical Association Meetings. (July 1980).

3. Bishop, Y. M. M. Statistical Methods for Hazards and Health, Environmental Health Perspectives, 20 (October 1977), 149-157.

4. Box, G. E. P. and Jenkins, G. M. Time Series Analysis Forecasting and Control Holden-Day, Inc. San Francisco, 1970.
5. Box. G. E. P. and Jenkins, G. M. Some Recent Advances in Forecasting and Control Part I, European Meeting of Statisticians, (1966), 91-109.

6. Chernick, M. R. The Influence Function and its Application to Data Validation, ORNL Report: TM-6871, (August 1979), 1-18.

7. Cooly, James W. and Tukey, John W. An algorithm for the machine calculation of complex Fourier Series, Math. Of Computation. 19, 90, (April 1965), 297-301.

8. Croxton, F. E. and Cowden, D. J., Applied General Statistics, Prentice-Hall, Inc. 1955.

9. Devlin, S. J., Gnanadesikon, R. and Kettenring, J. R. Robust estimation and outlier detection with correlation coefficients. Biosetrika. 62, 3 (1975) 531-545.

10. Dixon, W. J. and Massey, F. J. Introduction to Statistical Analysis. McGraw-Hill Book Company. 1969.

11. Fuller, W. A. Introduction to Statistics Time Series, John Wiley and Sons, Inc. New York, 1976.

12. Goodwin, G. C., Ramodge, P. J. and Caines, P. E., Recent Results in Stochastic Adaptive Control, 363-367.

13. Grubbs, F. E., Procedures for Detecting Outlying Observations in Samples, Technometrics, 11, 1 (Feb. 1969), 1-21.

14. Hampel, F. R., The Influence Curve and its Role in Robust Estimation. JASA. 69, 346 (June 1974), 383-387.

15. Hawkins, D. M., The Detection of Errors in Multivariate Data Using Principle Components. JASA. 68, 346, 340-344.

16. Hill, E. Discrete Rational Function Approximation in the L-One and L-Infinity Norms Using the Simplex Method. 1978 Proceedings of the ACM, Volume 2, 958-963.

17. Johnson, C.P. Rephrasing the Discrete Model Reference Adaptive Control Problem, 351-355.

18. Montgomery, D. C. and Johnson, I. A. Forecasting and Time Series Analysis, McGrae-Hill, New York, 1976.

19. Pome, P. Design of On-Line Control of Systems From Input-Output Data, 368-371.

20. SAS User's Guide 1979 Edition ISBN 0-917382-06 SAS Institute, Inc. North Carolina.

21. Shrager, R.J. Nonlinear Regression With Linear Constraints: An Extension of the Magnified Diagonal Method, JACM, 17, 3 (July 1970), 446-452.

22. Shrager, R.I. and Hill, E. Nonlinear Curve-Fitting in the L-One and L-Infinity Norms, Mathematics of Computations, 34, 150, (April 1980), 529-541.

23. Van Compenhout, J. M. On The Peaking of the Hughes Mean Recognition Accuracy. The Resolution of an Apparent Paradox, IEEF Transactions on Systems, Mon, and Cybernetics, Vol. SMC-8, No. 5 (May 1978), 390-395.

24. Winters, P. B., Forecasting Sales by Exponentially Weighted Moving Averages. Journal of the Institute of Management Sciences, 6, 3 (April 1960), 324-341.

Acknowledgments. The authors are particularly indebted to Dr. Y. M. M. Bishop for her generous assistance and many valuable suggestions.