PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Outlier Detection in the Design of Fault-Tolerant Biosensor Arrays

M. T. Flanagan, K. A. K. Taha and N. J. Carroll

Phil. Trans. R. Soc. Lond. B 1987 **316**, 61-72
doi: 10.1098/rstb.1987.0017

Phil. Trans. R. Soc. Lond. B **316**, 61–72 (1987) [61]
Printed in Great Britain

Outlier detection in the design of fault-tolerant biosensor arrays

By M. T. Flanagan, K. A. K. Taha and N. J. Carroll

Department of Electronic and Electrical Engineering, University College London, Torrington Place, London WC1E 7JE, U.K.

The possibility of fabricating a digital voter that will detect and eliminate a faulty sensor in an array of identical biosensors is examined. Eleven statistical outlier-detection procedures are applied to the responses of an array of antimony–antimony oxide penicillinase electrodes and to an extensive computer simulation of small array responses. A Dixon excess-over-range test and a maximum normalized residual test are shown to be safe outlier-detection procedures that will detect a faulty sensor and offer an algorithm that may be economically implemented in hardware. The Iglewicz & Martinez test, which can be implemented more conveniently in software, is shown to be very efficient when applied to the real data. However, its poorer performance when applied to the simulated data suggests that further examination of this test is required.

1. Introduction

Several aspects of chemical process control could be significantly enhanced by the use of biosensors (Clarke *et al.* 1982, Turner 1985). However, it is in this demanding environment that the limitations of biosensors are most apparent. Clarke *et al.* (1985) have, for instance, effectively dismissed their use as *in situ* monitors of fermentation. In the long term, this view may be over-pessimistic and, in the short term, their use in offline monitors and in the online monitoring of downstream processing looks increasingly promising (Cleland & Enfors 1983; Decristoforo & Danielsson 1984). None the less, it must be admitted that, in the harsh environment of process control, the standard biosensor defects (e.g. drift, noise, cross-reactivity, and lability of the biomolecule) are often greatly accentuated and compounded by fouling; this leads to crucially limited and unpredictable sensor lifetimes. Thus it is unlikely that the promise of biosensors in process control will ever be fulfilled if the advances in biosensor design that are leading to this promise are not matched by appropriate advances in signal processing, which will allow these defects to be either corrected or sufficiently mitigated. In this paper we will address the problem of the limited and unpredictable lifetimes.

Now that biosensors can be made conveniently and cheaply as multiple arrays, such as enzyme-linked ISFETS (Kuriyama *et al.* 1985) and antimony–antimony oxide enzyme electrodes (Flanagan & Carroll 1986), the possibility of electronically detecting and eliminating an individual faulty sensor in such an array arises. Superficially, this is an analogous problem to that of designing majority voters, e.g. triple-modular redundancy (TMR) systems, in the fault-tolerant design of digital systems (Lala 1984). At its simplest, a TMR system compares three nominally identical digital inputs and rejects an input that differs from the other two. The resemblance is superficial in that the classical voter compares digital signals whose noise component is within the tolerance of the relevant digital logic. In contrast, we wish to compare digitized noisy analogue signals. Identifying a faulty sensor within such an array is more akin to identifying an outlier in statistical analysis. Thus our voter must become a digital processor capable of performing the equivalent of an outlier-rejection procedure. There is a large number of outlier-rejection procedures (Barnett & Lewis 1984) but little documented work on their performance on the small sample numbers that will characterize even an extensive sensor array. In this paper, we compare the performance of eleven outlier procedures, used in this manner, applied to an array of real biosensors and to computer simulations of biosensor responses, with the aim of chosing the most appropriate procedures for conversion to digital voters to be implemented as hardware or as software.

2. Generation of sensor array data

2.1. *Antimony–antimony oxide penicillinase electrodes*

To obtain an 'accelerated lifetime' array of biosensors (i.e. an array that shows, over a conveniently short period in the laboratory, the noise, drift and sensor dropout that will characterize a biosensor array in a harsh environment) we have modified our fabrication of evaporated thin film antimony–antimony oxide enzyme electrodes (Flanagan & Carroll 1986) to give a somewhat less reliable sensor. We have replaced our antimony evaporation procedure with the antimony pH microelectrode fabrication procedure of Vieira & Malnic (1968).

Molten antimony was drawn, under vacuum suction, into 60 mm pyrex tubes (inner diameter 2 mm, outer diameter 4 mm) each containing a tinned copper wire to within 2 mm of the end placed in the molten antimony. This end of the cooled antimony-filled tube was ground flat on a glass plate with carborundum paste. The electrodes were immersed in distilled water overnight. Penicillin monitoring electrodes were then made by immobilizing, on the tip of the electrode, penicillinase (from *Bacillus cereus*) by the gluteraldehyde cross-linking method of Mascini & Guilbault (1977). Penicillin monitoring was performed in 1 mM phosphate buffer, pH 7.0, at a temperature of 23 °C, with a calomel reference electrode as previously described (Flanagan & Carroll 1986).

Seven penicillin electrodes were prepared as a batch and were monitored daily for three weeks, with a 1 mM penicillin G solution as the test sample. In a finally developed, digitally controlled sensor array, several measurements will be made over a time period much shorter than that characterizing the change of analyte concentration in the monitored process. This will allow signal averaging of the individual sensor responses before such signal processing as drift compensation, e.g. by adaptive filtering (Thijssen *et al.* 1985), and faulty-sensor detection. As we have used a constant test sample, to concentrate on the latter problem, we can use a simple averaging procedure. The data was smoothed, in real time, by taking the following running average over three points:

$$x_i''(j) = \tfrac{1}{3}(x_i'(j-2) + x_i'(j-1) + x_i'(j)),$$

where $x_i''(j)$ is the smoothed response of the ith sensor at the jth measurement point and $x_i'(j)$ is given by $f_i\, x_i(j)$, where $x_i(j)$ is the observed response and f_i is simply a scaling factor, $m_1/x_i(1)$, introduced to reduce the data spread to facilitate the signal processing. Here m_1 is the mean, calculated at the first measurement point, of the sensor response, $x_i(1)$.

The smoothed data are shown in figure 1, where it can be seen that we have suitably noisy,

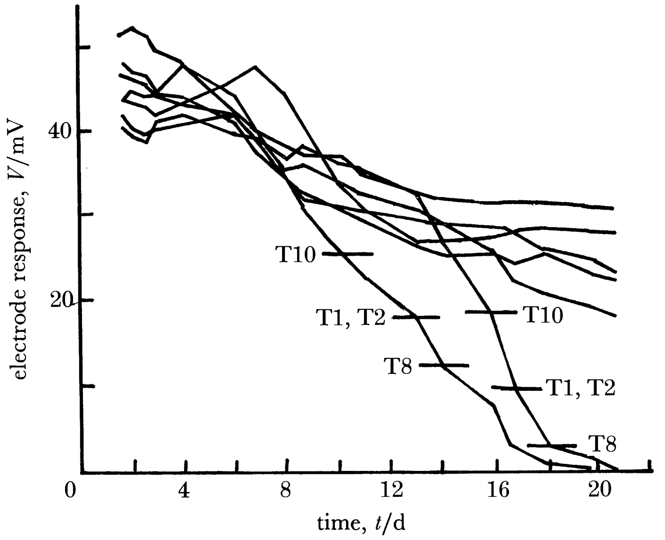

FIGURE 1. The smoothed responses of the seven batch-produced electrodes to a standard sample of 1 mM penicillin G in 1 mM phosphate buffer, pH 7.0. The electrode response value at which tests T1, T2, T8 and T10 identified the two outliers are marked.

drifting data with two clear sensor failures, these failures will be referred to as the first and second real outliers throughout this paper. These data are used as the real data set in the comparison of the outlier-rejection procedures.

2.2. *Computer simulations*

Any experimental data set is necessarily a small number of selected populations. Consequently, it was felt necessary to carry out an extensive computer simulation, based on small samples drawn from a normal distribution of zero mean and unity standard deviation, in testing the outlier-rejection procedures. The Numerical Algorithms Group's (NAG 1984) FORTRAN 77 programs, GO5DDF and GO5CCF, were used to generate these samples on the SERC Prime 550 minicomputer at U.C.L. These programs use the algorithm of Brent (1974). Samples (1000) of size n were generated and used in evaluating the efficiency of any statistical procedure applied to an array of size n. The appearance of a faulty sensor was mimicked by decreasing the value of the lowest generated response in each of the 1000 samples, by increments of one tenth of the population standard deviation, until an outlier was detected or until a distance of one hundred standard deviations had been reached. In the latter case the test was deemed to have failed on that sample. The use of the normal distribution is justified below.

3. DETECTION OF OUTLIERS

3.1. *Normality of the data*

An outlier in a set of data is an observation that appears inconsistent with the remainder of the data (Barnett & Lewis 1984). Statistical procedures have been developed which either allow the accommodation of the outlier within the data or confirm its inconsistency, thus allowing it to be rejected. In transferring these procedures to the elimination of faulty sensors, we are clearly only concerned with the latter approach. The confirmation is obtained by applying a statistical test to examine whether the extreme observation is not only an extreme but is also 'statistically unreasonable'. This implies that we know, or can make a reasonable assumption about, the distribution of responses of an array of correctly functioning sensors. The several production process and operating errors, in a batch of nominally identical devices, are likely to combine to give an overall normal distribution of responses (Steiner 1967; Box *et al.* 1978) and thus we have chosen the normal distribution as our most probable general model for an array of sensor outputs. Furthermore, statistical outlier detection procedures, based on theoretical normality, are often robust to non-normality (Box *et al.* 1978).

In analysing our real data for normality we have applied the *W* test of Shapiro & Wilk (1965). This test has been applied to small samples, a difficult area of normality testing, and a comparative study of its performance has shown that it is quite sensitive to a wide range of non-normality (Shapiro *et al.* 1968). Table 1 shows the results of the *W* test applied to our real data with those points removed at which an outlier was later identified. The data are normal, at the 5% significance level, if the value of the observed *W* statistic (second column) is greater than that tabulated for the 5% level (third column). The initial responses are normally distributed, as are all but two of the 18 subsequent responses. One of the two exceptions was a sensor whose response was higher than might have been expected.

Table 1. *W* test for normality applied to the real data

time/h	calculated *W* statistic	*W* statistic at the 5% level of significance	time/h	calculated *W* statistic	*W* statistic at the 5% level of significance
38.6	0.964	0.803	261.8	0.914	0.803
48.7	0.964	0.803	312.3	0.897	0.788
62.1	0.972	0.803	335.0	0.954	0.788
72.0	0.917	0.803	354.9	0.917	0.788
95.8	0.882	0.803	402.8	0.803	0.762
141.5	0.981	0.803	433.8	0.975	0.762
162.0	0.751	0.803	477.5	0.990	0.762
191.1	0.791	0.803	497.0	0.969	0.762
207.8	0.940	0.803	520.0	0.913	0.762
241.3	0.945	0.803			

3.2. *Rejection procedures*

Barnett & Lewis (1984) have extensively reviewed the procedures for detecting outliers in statistical data. Table 2 shows the eleven outlier rejection procedures we have chosen to examine in the context of small sensor arrays. All require normal distributions. In making this choice, we have been guided by Barnett & Lewis's (1984) definition and choice of a useful test: one that performs reasonably well, even if not optimally, with respect to some meaningful alternative hypothesis and one for which there is some information on the percentage points. Additionally, we require a test that at best can be effectively and economically implemented in digital hardware, and at worst, be efficiently implemented in software. Our choice includes the standard discordancy tests, e.g. skewness test, *W* test for normality, maximum normalized residual tests, Dixon's tests and the more recent test of Iglewicz & Martinez (1982). All but the skewness and *W* tests follow a general philosophy of examining some measure of the distance of the putative outlier from the sample with respect to the sample spread and have been

TABLE 2. OUTLIER-REJECTION TESTS EXAMINED.

(Here x_1 is the response of lowest value and x_n that of the highest value; $\bar{x}$ is the mean value of the sample and s^2 is the estimate of the variance; n is the sample size.)

test	description	statistic	available tables for the statistic	reference
T1	maximum normalized residual test for lower outlier	$(\bar{x}-x_1)/s$	0.1, 0.5, 1, 2.5, 5 and 10% significance level for $n = 1$ to 147	Grubbs & Beck (1972)
T2	excess/range test for lower outlier (Dixon-type test)	$\frac{x_2-x_1}{x_n-x_1}$	0.5, 1, 2, 5, 10, 20, 30, 40 and 50% level for $n = 3$ to 30	Dixon (1951)
T3	Dixon-type test for lower outlier	$\frac{x_2-x_1}{x_{n-1}-x_1}$	as T2	Dixon (1951)
T4	Dixon-type test for lower outlier	$\frac{x_2-x_1}{x_{n-2}-x_1}$	as T2	Dixon (1951)
T5	Dixon-type test for lower outlier	$\frac{x_3-x_1}{x_n-x_1}$	as T2	Dixon (1951)
T6	Dixon-type test for lower outlier	$\frac{x_3-x_1}{x_{n-1}-x_1}$	as T2	Dixon (1951)
T7	Dixon-type test for lower outlier	$\frac{x_3-x_1}{x_{n-2}-x_1}$	as T2	Dixon (1951)
T8	skewness test	$\Sigma\frac{(x_j-\bar{x})^3}{n\,s^3}$	1, 5 and 10% level for $n = 3$ to 25	Pearson & Hartley (1966)
T9	W test for normality	W statistic	1, 2, 5, 10 and 50% level for $n = 2$ to 50	Shapiro & Wilk (1965)
T10	test for extreme outlier with a robust estimate of variance, s_b^2	max $((x_n-\bar{x})/s_b)$, $(\bar{x}-x_1)/s_b)$	generating equation	Iglewicz & Martinez (1982)
T11	maximum normalized residual test for lower outlier	$(\bar{x}-x_1)/s$	tests against a constant, C	Anscombe (1960)

specifically designed to detect outliers. The skewness and W tests have not been so designed but have, none the less, proved useful in such detection. These tests may be used to identify either upper or lower outliers. We have limited the tests to lower-outlier identification, as the failure mode of our biosensors is loss of output signal.

3.3. *Results at the 5% significance level*

The ten tests T1–T10 were applied both to the real data and to computer simulations of sample sizes 3, 4, 6, 8 and 10 (1000 simulations each). Initially, the customary significance level of 5% was chosen in testing the statistics as to whether an extreme response indicated an outlier and hence a faulty sensor. Figure 2 shows, in the form of a bar graph, the distance, measured from the sample mean, at which each test discovered the first real outlier. All tests

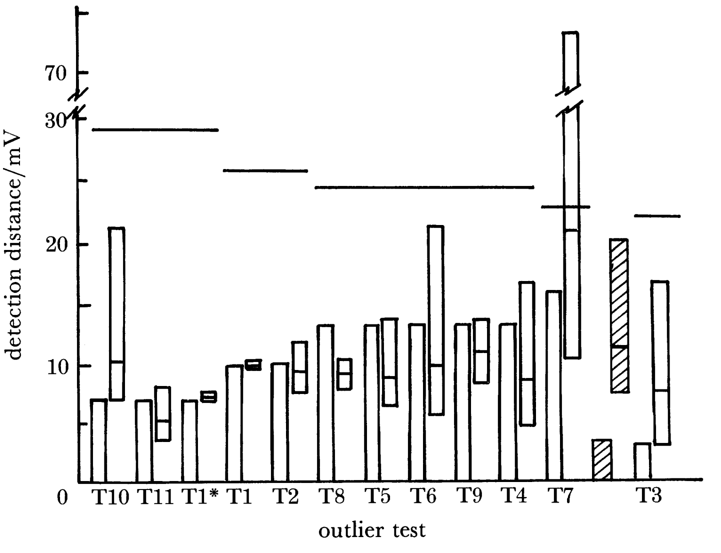

FIGURE 2. A bar graph of the efficiency of the outlier tests in detecting the first outlier. The first bar associated with each test indicates the electrode response at which detection occurred; the second bar indicates the range of detection values found by simulation. The line within the second bar indicates the mean detection value found by simulation. The continuous horizontal lines indicate the positions of zero electrode response. Test T11 was performed with C set at 1.28. T1* was tested at the 15% significance level; the remaining outlier-detection procedures were tested at the 5% significance level. The bar indicating a false detection is hatched.

discovered this outlier with varying degrees of efficiency; only test T3 made an additional false detection. Figure 3 shows the corresponding results for the second real outlier. In this case tests T3, T4 and T9 failed to detect the outlier. Figure 4 illustrates typical distributions, obtained from the computer simulations, of the outlier detection distances for these tests. The range of detection distances given by such distribution curves for each test is also shown, as the second bar associated with each test, in the bar graphs in figures 2 and 3. The mean values of the distribution are also marked on the second bars. These simulation values were converted to equivalent mV by varying the position of the lower extreme point of the real data until its normalised residual corresponded to the appropriate point on the simulation distribution curve. All the real outlier detection distances either lay within the theoretical range or were so close as to be accommodated by the error introduced by the long interval between sampling, i.e. one day. It can also be seen that the tests which failed to detect the second outlier are characterized by a theoretical range that lies predominantly below zero, i.e. detection was improbable before the sensor had totally failed and was generating no response.

Four tests, T1, T2, T8 and T10, performed best when applied to both the real and the simulated data but differed in the order of efficiency betweend the two sets of data. Their ranking is listed in table 3. Unfortunately, the spread between the best and the worst detection distances associated with these four tests is still large, as can be seen in figure 1, where the detection points are indicated on the real data. Thus a difficult choice has to be made within this group of four. The very efficient behaviour of T10, when applied to the real data, makes it very attractive. However, its poor behaviour when applied to the simulated data suggests caution. This is a relatively new test for outliers and there is, as yet, little literature on the test

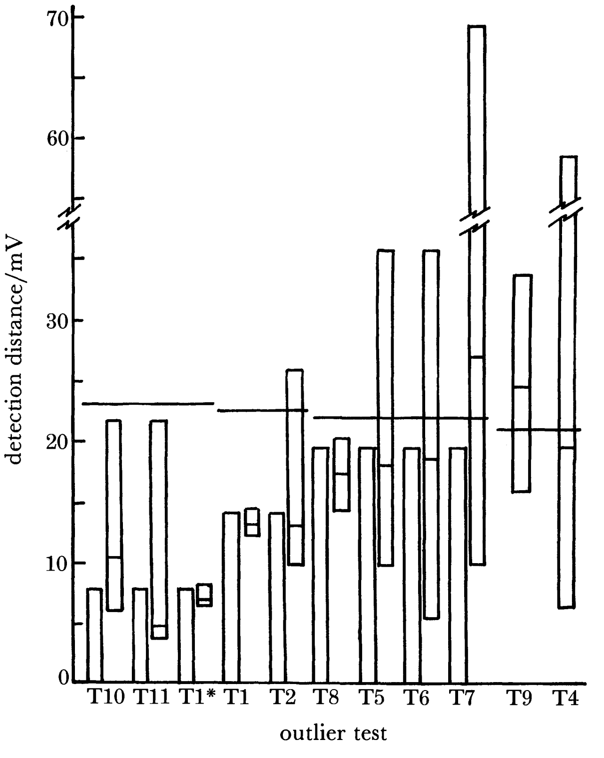

FIGURE 3. A bar graph of the efficiency of the outlier tests in detecting the second outlier. The distance of 0 mV in the case of no detection was taken at the time when both failed sensors ceased to respond. (Symbols as in figure 2.)

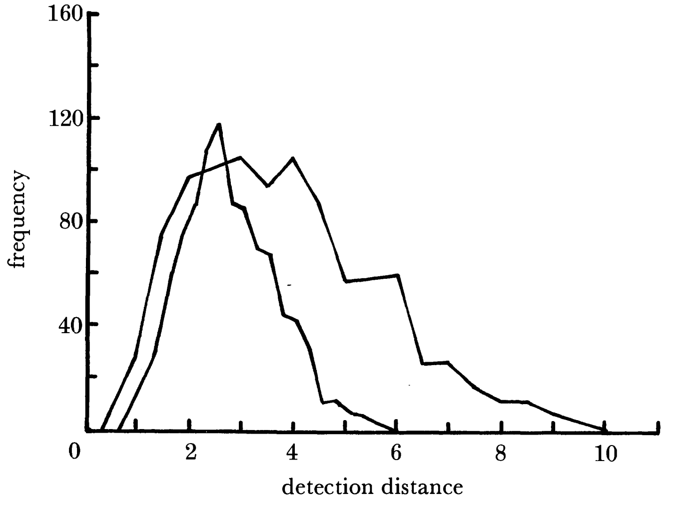

FIGURE 4. A frequency plot of the outlier-detection distances found on applying tests T1 and T7 to a thousand simulations of a sample size of 8. The detection distance is expressed as a normalized residual calculated by using the mean and variance of the population from which the samples were drawn.

TABLE 3. THE BEST FOUR TESTS, AT THE 5% SIGNIFICANCE LEVEL, IN ORDER OF RANK, PLUS TEST T1 AT THE 15% LEVEL (*) AND TEST T11 AT $C = 1.28$ (+)

(The simulated data mean detection distance is defined in figure 6.)

real data: first outlier detection distance / mV	real data: second outlier detection distance / mV	simulated data: mean detection distance		simulated data: range (standard deviation of the detection distance)	
$n = 7$	$n = 6$	$n = 7$	$n = 6$	$n = 7$	$n = 6$
T10 6.81	T10 7.91	T2 1.67	T2 1.89	T10 0.18	T10 0.22
T1 9.90	T1 14.16	T1 1.86	T1 1.92	T2 0.20	T2 0.24
T2 9.90	T2 14.16	T8 2.16	T10 2.45	T1 0.22	T1 0.26
T8 13.13	T8 19.50	T10 2.21	T8 2.54	T8 0.29	T8 0.34
T1 6.81	T1 7.94	T1 1.11	T1 1.16	T1 0.16	T1 0.18(*)
T11 6.81	T11 7.94	T11 0.37	T11 0.37	T11 0.10	T11 0.10(+)

to help us; it is also a laborious test, as will be discussed below. Tests T1 and T2 behaved in a very similar manner and consistently with respect to both the real and simulated data; they clearly represent two safe procedures that are also simple to perform.

3.4. *Complexity of the chosen tests*

The numerically simpler tests are the more attractive to us as the ultimate aim of this work is to make a microelectronic voter. If we restrict ourselves to tests in which a statistic is compared with tabulated values at a chosen significance level (T1–T10) the Dixon-type tests (Dixon 1950, 1951) are the most attractive; hence the presence of six such tests in the original choice. One of the safe choices, T2, is such a test. In T2 the calculated statistic is simply the excess, i.e. the distance between the lower extreme and the next lowest response, divided by the range, i.e. the distance between the lowest and the highest response. This statistic is compared with appropriate tabulated values.

Test T1 is slightly more complex than T2 in that the calculated statistic is the maximum normalized residual, i.e. the distance between the lowest response and the sample mean divided by the standard deviation of the sample. T11, a variation of test T1, in which the look-up table is dispensed with by comparing this statistic with a constant, is discussed below.

Test T10 uses a robust estimator of the variance, which introduces a somewhat longer calculation. The test was derived as a two-sided test (Iglewicz & Martinez 1982). In our one-sided application the statistic reduced to a maximum normalized residual in which the sample variance is replaced by s_b^2, where m is the sample median and

$$s_b^2 = (V/W)\, n^{\frac{1}{2}};$$

$$V = \left[\sum_{|u_i|<1} (x_i - m)\ (1-u_i^2)^4\right]^{\frac{1}{2}};$$

$$W = \max\left\{1,\ -1 + \left|\sum_{|u_i|<1} (1-u_i^2)\ (1-5u_i^2)\right|\right\};$$

$$u_i = (x_i - m)/(9\,\text{median}\,|x_i - m|).$$

The 5% significance level is obtained from $2.064\sqrt{n}/(n-1)^{0.387}$.

3.5. *Variation of the level of significance*

The choice of significance level is dependent on the application: 5% was chosen in the first part of this paper as a customary level. Increasing the significance level will increase the efficiency of outlier detection but at the expense of false identification of outliers, i.e. of declaring correctly operating sensors to be faulty. We have applied test T1 to both the real and the simulated data with the tabulated values at the 15% significance level in assessing outliers. It can be seen from figures 2 and 3 and from table 3 that T1 behaves as well as T10 at this level.

A commonly used test for outlier identification in curve fitting is the Anscombe (1960) maximum normalized residual test (see, for example, Reich *et al.* (1972)). T11 is such a test. Its simplicity arises from the fact that no attempt is made to test the statistic at a chosen constant significance level. Instead the maximized normalized residual is compared with a constant, C; if it is greater than C, the extreme point is declared an outlier. We have determined the minimum value of C for which T11 behaves as well as T10 when applied to the real data and applied it to the simulated data at this value of C. The results can be seen in figures 2 and 3 and in table 3. The value of C was found to be 1.28; this value corresponds to a 56% and to a 65% significance level for the first and second outliers respectively. Figure 5 shows graphs of the probability of false detection, in this case simply the significance level (Hays 1981), against the value of C to be used in test T12 for several sample sizes. Figure 5 allows an easy choice of the value of C. If T12 is used, care must be taken not to use too large a value of C. There

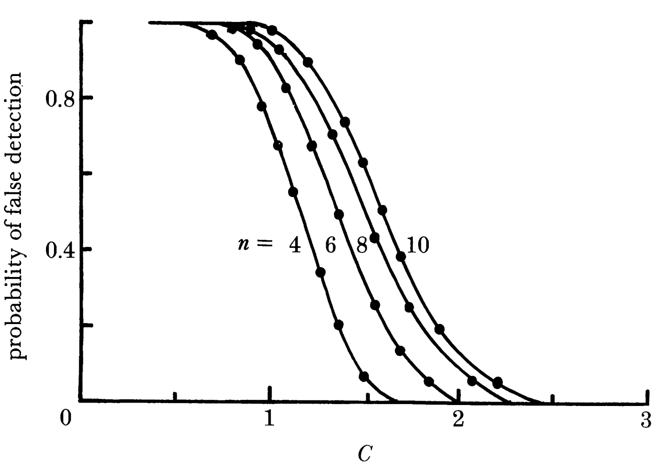

FIGURE 5. The probability of false detection of outliers by test T11 for sample sizes (n) of 4, 6, 8 and 10 as a function of the value at which C is set.

is an upper bound to the value of C, which is given by $(n-l)/\sqrt{n}$ (see Appendix 1). Figure 6 shows the relation between C and the mean distance at which an outlier is detected; figure 5 illustrates one of the consequences of applying Anscombe's tests to small sample sizes. For small samples the level of significance, for a fixed value of C, depends strongly on the sample size and thus, when an outlier is rejected, can change markedly.

3.5. *Sample size*

The simulated data have been used to examine the effect of sample size on the efficiency of the outlier tests. The efficiency increases rapidly on increasing the sample size from 3 to 6 in all cases but then increases only slowly. A typical result is shown in figure 7 where the variation, with sample size, of the mean and range of the outlier detection distances for test T1 are plotted. This result is encouraging in that it suggests that the analysis presented in this paper is a reasonable guide for the small array sizes that are likely to characterize most biosensor arrays.

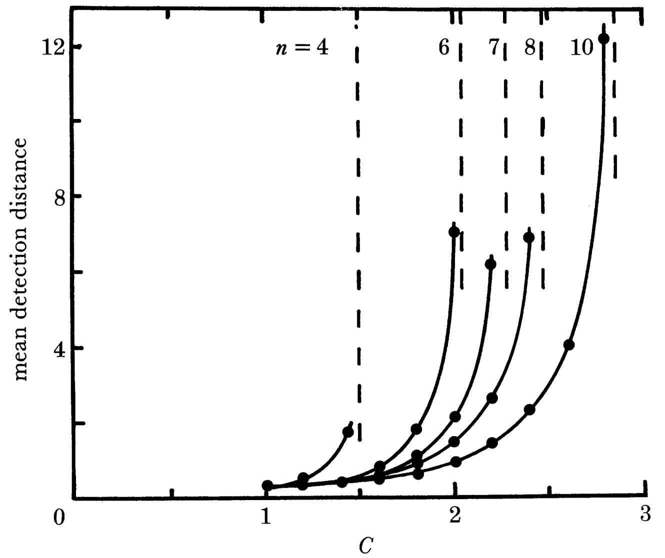

FIGURE 6. A plot of the value at which C must be set in test T11 to detect an outlier whose normalized residual is given by the ordinate. The vertical dotted lines indicate the upper bound to C given by equation (6) (Appendix 1); n is the sample size. The detection distances were measured as the distances between the original positions of the lowest points and their positions when detected as outliers after successive decrements. They are normalized to the variance of the population from which the samples were drawn.

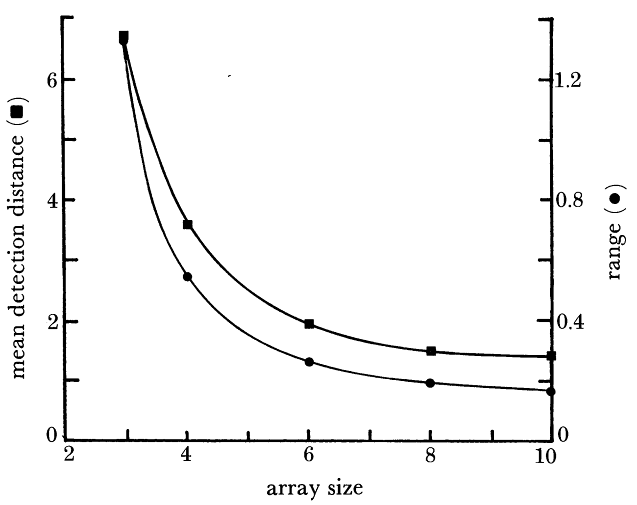

FIGURE 7. The efficiency of test T1, at the 5% significance level, as a function of sample size for 1000 simulations at each sample size. The detection distances (filled squares) are as defined in figure 6. The range of detection distances (circles) was obtained from frequency plots analagous to those presented in figure 4.

4. Conclusions

The results of the analysis of outlier-detection procedures in the detection of faulty sensors, presented above, indicate that it is possible to produce a data processor that is able to recognize a faulty sensor. The Dixon excess-over-range test and the maximum normalized residual outlier test are safe outlier detection tests that may be used to detect faulty sensors; however, the safety is at the expense of rapidity of detection. The more recent test of Iglewicz & Martinez (1982), when applied to the real data, gave a detection rate that was more appropriate to a detection system for faulty biosensors. However, it did not behave consistently when applied to the simulated data. Further examination of this test and of the possibility of deciding on the most appropriate significance level to be used with any test is required. In this paper we have chosen to describe only the distribution of the correctly functioning sensors, i.e. a normal distribution. We have then adopted a very general outlier-rejection philosophy. An approximation of the distribution function describing faulty sensors would allow a formulation of more rigorous alternative hypotheses in the rejection procedure. It is intended to do this to improve the adequate, but less than optimum, choice of procedures that has arisen from this work. An improvement of the computationally simpler tests, or more confidence in their use at higher significance levels is desired, as it will be possible to implement these tests economically in hardware; for example, by a combination of programmable-logic arrays and custom-design techniques.

The final fault-tolerant system will also have to accommodate noise and drift while monitoring a varying analyte concentration. The variation in analyte concentration in many process-control applications will be slow by comparison with the measurement period, thus allowing data smoothing. The drift is a much greater problem, but the advances in adaptive filtering, e.g. Kalman filtering, which has already been applied to sensor drift (Thijssen *et al.* 1985), are encouraging and would be compatible with the above outlier-rejection procedures.

M.T.F. is the Standard Telephone and Cable Research Fellow in Bioelectronics. N.J.C. thanks the Science and Engineering Research Council for support throughout the period of this work.

References

Anscombe, F.J. 1960 Rejection of outliers. *Technometrics* **2**, 123–147.

Barnett, V. & Lewis, T. 1984 *Outliers in statistical data*, 2nd end. Chichester & New York: John Wiley.

Box, G. E., Hunter, W. G. & Hunter, J. S. 1978 *Statistics for experimenters*, pp. 43–46. New York: John Wiley.

Brent, R. P. 1974 Algorithm 488: A Gaussian pseudo-random number generator [G5]. *Communications of the Association for Computing Machinery* **17**, 704–706.

Clarke, D. J., Calder, M. R., Carr, R. J. G., Blake-Colman, B. C., Moody, S. C. & Collinge, T. A. 1985 The development and application of biosensing devices for bioreactor monitoring and control. *Biosensors* **1**, 213–320.

Clarke, D. J., Kell, D. B., Morris, J. G. & Burns, A. 1982 The role of ion-selective electrodes in microbial process control. *Ion-selective Electrode Rev.* **4**, 75–131.

Cleland, N. & Enfors, S. O. 1983 Control of glucose-fed batch cultivation of *E. coli* by means of an oxygen stabilised enzyme electrode. *Eur. J. appl. Microbiol. Biotechnol.* **18**, 141–147.

Decristoforo, G. & Danielsson, B. 1984 Flow injection analysis with enzyme thermistor detector for automated determination of bete-lactams. *Analyt. Biochem.* **56**, 263–268

Dixon, W. J. 1950 Analysis of extreme values. *Ann. math. Statist.* **21**, 488–506.

Dixon, W. J. 1951 Ratios involving extreme values. *Ann. math. Statist.* **22**, 488–506.

Flanagan, M. T. & Carroll, N. J. 1986 A thin-film antimony–antimony oxide enzyme electrode for penicillin determination. *Biotechnol. Bioengng* **28**, 1093–1099.

Grubbs, F. E. & Beck, G. 1972 Extension of sample sizes and percentage points for significance tests of outlying observations. *Technometrics* **14**, 847–854.

Hays, W. L. 1981 *Statistics*, 3rd edn, p. 245. New York: Holt, Rinehart & Winston.
Iglewicz, B. & Martinez, J. 1982 Outlier detection using robust measures of scale. *J. statist. Comp. Sim.* **15**, 285–294.
Kuriyama, T., Kimura, J. & Kawana, Y. 1985 Development of biosensors with immobilised enzymes. *Chem. Econ. Engng. Rev.* **17**, 22–27.
Lala, P. K. 1984 *Fault tolerant and fault testable hardware design*, chapter 4, pp. 69–137. New Jersey: Prentice–Hall.
Mascini, M. & Guilbault, G. G. 1977 *Analyt. Chem.* **49**, 795–798.
Numerical Algorithm Group 1984 NAG Fortran Library Manual, mark II, vol. 5. Oxford.
Pearson, E. S. & Hartley, H. O. 1966 *Biometrika tables for statisticians*, 3rd edition, vol. 1. London: Cambridge University Press.
Reich, J. G., Wangermann, G., Falck, M. & Rohde, K. 1972 A general strategy for parameter estimation from isosteric and allosteric–kinetic data and binding measurements. *Eur. J. Biochem.* **26**, 368–379.
Shapiro, S. S. & Wilk, M. B. 1965 An analysis of variance test for normality. *Biometrika* **52**, 591–611.
Shapiro, S. S., Wilk, M. B. & Chen, M. J. 1968 A comparative study of various tests of normality. *J. Am. statist. Ass.* **63**, 1343–1372.
Steiner, E. H. 1967 Statistical methods in quality control. In *Quality control in the food industry* (ed. S. M. Herschdoerfer), vol. 5, pp. 121–229. London: Academic Press.
Thijssen, P. C., Prop, L. T. M. & Kateman, G. 1985 A Kalman filter for calibration, evaluation of unknown samples and quality control in drifting systems. *Analyt. chim. Acta.* **174**, 27–40.
Turner, A. P. F. 1985 Biosensors for process monitoring and control. *Proceedings of Biotech '85 (Europe)*, pp. 181–192. Pinner, U.K.: On-line Publications.
Vieira, F. L. & Malnic, G. 1968 Hydrogen ion secretion by rat renal cortical tubules as studied by an antimony microelectrode. *Am. J. Physiol.* **214**, 710–718.

Appendix 1. The upper limit to C in the test T11

Consider a sample x_1 to x_n. The sample mean, m_1, is given by

$$\sum_{1}^{n} x_i/n = \bar{x}, \tag{1}$$

and the sample variance, s_1^2, is given by

$$\sum_{1}^{n} (x_i - m_1)^2/(n-1). \tag{2}$$

The lower point, x_1, is moved by a decrement, Δx, to give a lower outlier. The new mean, m_2, is given by

$$m_1 + \Delta x/n \tag{3}$$

and the new variance, s_2^2, is given by

$$\frac{1}{n-1}\left[(x_1 + \Delta x - m_2)^2 + \sum_{2}^{n} (x_i - m_2)^2\right]. \tag{4}$$

Equations (1), (3) and (4) lead to

$$s_2^2 = (\Delta x)^2/n + 2\,\Delta x(x_1 - m_1)/(n-1) + s_1^2. \tag{5}$$

The statistic for test T11, T, is $(m_2 - (x_1 + \Delta x))/s_2$. Therefore the upper bound of C, with which T is compared, is given by

$$\lim_{|\Delta x| \to \infty} (T),$$

which can be obtained from (3) and (5) as

$$\lim_{|\Delta x| \to \infty} \left[\frac{(n-1)/n + (m_1 - x_1)/|(\Delta x)|}{(1/n + 2\sqrt{(m_1 - x_1)/((n_1)\ |\Delta x|) + s_1^2/\ |\Delta x|^2}}\right] = \frac{(n-1)}{\sqrt{n}}. \tag{6}$$