UDC 621-7

A review of fault detection methods for large systems

STEPHEN SHIELDS,
B.Sc., M.Sc., C.Eng., M.I.Mech.E.*

Based on a paper presented at the IERE Conference on Advances in Automatic Testing Technology held at Birmingham from 15th to 17th April 1975

SUMMARY

A critique is made of current research into practical fault finding procedures for the maintenance of complex engineering systems. The half split and other methods currently in practice are analysed and their main weakness shown to be that no account is taken of the various costs involved. Also analysed are cost conscious methods which are useful in diagnosis training or in designing fault detection guides. A brief look is taken at advanced diagnostic techniques which are aided by an on-line computer in selecting the next test to be made.

* *Formerly in the Department of Operational Research, University of Strathclyde, Glasgow; now with the South of Scotland Electricity Board, Forward Planning Division, Cathcart House, Inverlair Avenue, Glasgow G44 4BE.*

1 Introduction

With the ever increasing complexity of engineering equipment of all types and disciplines the engineer and technician are faced with acute maintenance problems. One of the most incomprehensible aspects is that of diagnosing a fault in an inoperable system. The problem is compounded by the increasing obsolescence rate of equipment which allows less time for familiarization. The cost of downtime is often very high, especially in integrated production plants.

There is a real need therefore to develop methods which will enable faults to be found and rectified in the shortest possible time and at the minimum total cost. In addition to direct savings in lost production and consumed maintenance facilities, capital cost savings may be expected since fewer standby systems will be needed.

The problems of fault detection are considered against a maintenance background but the techniques discussed are also applicable in the quality control and assurance field. Similarly, the logical techniques are appropriate in any complex engineering system be it mechanical, electronic, electrical, hydraulic, etc. In fact the author has used some of the techniques in the successful diagnosis of human systems, e.g. in the diagnosis of patients suffering from thyroid disorders[1].

2 Approaches to the Problem

Basically three distinct approaches have been made to improve the efficiency of fault detection.

Approach 1: Personnel training in functional analysis/fault finding guides. This method has the advantage of producing instantly experienced technicians for a very low (or zero) capital outlay. The only cost involved is that of training. This approach is very useful when the frequency of fault occurrence is low and/or the cost of downtime is not particularly high.

Approach 2: The introduction of fully automated testing and switching devices similar to those used on computer, missile and aircraft checkouts and for some makes of motor cars (e.g., Volkswagen). These are really checking procedures in that a complete battery of tests is usually run independent of early indication of a fault. This approach is useful and justifiable when the downtime cost is very high (or the consequences of fault development are serious) and/or the frequency of fault development is high.

Approach 3: This is the middle ground between the two previous approaches and is appropriate when fully automated checking procedures cannot be justified but some automation is necessary to assist in diagnosis. Here the purpose is to locate the fault by using the minimum number of tests or at the least cost (which is often not the same as minimizing the number of tests). This approach is also suitable when automatic testing devices are included but because the elapsed time for a test is significant then the order of testing and the number of tests to be made are important.

This paper concentrates on techniques which are consistent with Approach 3.

3 Formalizing the Approach

The previous casual statement of purpose: '. . . to locate the fault by using the minimum number of tests or at the least cost . . .', requires analysis and enlargement.

3.1 Objective

The objective of the approach is to determine the best sequence of applying tests so that the diagnosis is made in the optimal manner as specified by the criterion or criteria.

3.2 Criteria

The criterion depends on the physical situation. Often it is stated as the minimum number of tests to reach the correct diagnosis or when the tests have unequal durations then it is often stated as the minimum time to reach the correct diagnosis. Although in a few situations one of these criteria may indeed be correct, in most cases the criteria are inadequate since no account is taken of the cost of the various testing actions or the financial and other consequences of incorrect diagnosis. In yet other situations there is danger of incurring further faults by applying particular tests to the system and this danger must be accounted for.

The most widely applicable criterion is to minimize the *total* cost of fault finding and rectification. This is particularly suitable in production and other commercial environments where the downtime results in actual or potential loss of profit or income. This is generally a more appropriate criterion than minimizing time since the consequences of time and the cost of testing/diagnosis/rectification are included.

The most notable exception to the adoption of this criterion is in cases where the *cost* of downtime is not directly applicable, e.g., military operations or where human safety is involved. Other useful criteria are:

maximize the probability of correct diagnosis subject to a fixed testing cost (time) and

minimize testing cost for a given probability level of acceptance.

Decision theorists will be familiar with the methods and the appropriateness of the methods of Wald, Savage, and others, in this area.

3.3 Depth and Extent of Diagnosis

It has been shown in Refs. 2 and 3 that the three cases:

(i) search (detection) for a fault when it is known that the system has failed,

(ii) checking operability and searching for a fault,

(iii) checking operability,

are all equivalent to diagnosis in different depths. Similarly whether the system has definitely only one fault or many faults, or indeed if it is faulty at all, may be considered to be problems of inspection rather than diagnosis. Finally, whether diagnosis is made when the faulty unit (component) or module is identified depends solely on the definition of a 'unit'.

In subsequent analysis of various methods the 'system' may be a radar-missile system where the 'units' are the surveillance radar, tracking radar, communications system, computer, missile launcher and missile, or modules of these. Alternatively the 'system' may be a domestic hi-fi set and the modules or components represent the 'units'. The terms 'system' and 'units' are thus general.

4 Logical Fault Finding Techniques

To detect a fault in a system one normally carries out various tests and the physical tests used in any situation depends on the nature of the system under investigation. Depending on the physical situation then responsive tests, elemental probe tests, signal tracing and substitution, stress methods, replacements, etc. may be used. The choice of suitable tests is very much a physical engineering matter; sometimes a straightforward and obvious choice, in other cases the subject of much analysis.

Independent of the physical tests possible, one is faced with a logical decision problem. That is, in what sequence should the tests be applied to identify the fault. An intuitive approach does not guarantee an optimal solution. It will be found convenient to consider the system as falling into one of three categories. Strictly these categories are non-overlapping, mutually exclusive and exhaustive. In practice, however, the boundaries are not quite so rigid.

A number of the techniques which will be considered require involved calculations to be made so that the best sequence of testing may be identified. These calculations can be made via a Post Office modem to a general-purpose digital computer which may be some distance from the system under investigation, or by a small special-purpose computer on site. To enable these techniques to be used in conjunction with diagnostic training methods or in preparing fault finding guides, it is possible for the calculations (in categories 1 and 2 only) to be made during the preparation of the guides. This overcomes the need to make calculations during actual fault finding.

4.1 Category 1 : System composed of *N* identical elements—perfect information

Although systems composed of identical items are unusual in practice this category may be broadened to include systems where the *a priori* probabilities of failure of the units are approximately equal and the testing costs are also approximately equal. Hence it will be seen that the system may be composed of many physically different elements so long as the testing costs and failure characteristics of the various elements are approximately the same. Perfect information implies that testing errors do not occur.

A simple procedure which does not immediately identify the faulty element but does rectify the fault is to replace the units of the system with spare units which are known to be in good condition. These replacements may be made sequentially and the system operability checked after each replacement. Alternatively the units may be replaced en bloc. Depending on the relative costs of downtime and unit replacement, this strategy may be desirable. The replacement methods are attractive when

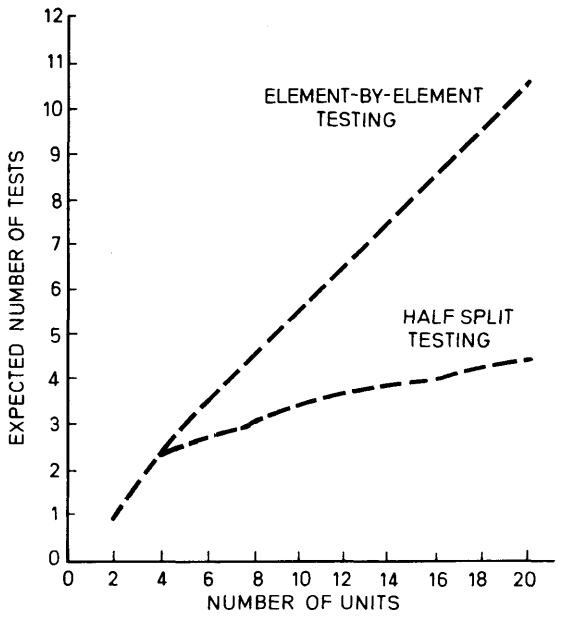


Fig. 1. Comparison of element-by-element and half split testing methods.

the cost of downtime is high and/or the cost of unit replacement is low and/or faults in units may be rectified without system downtime. Obviously this strategy is not limited to systems in this first category but is applicable (often with greater benefit) to more complex systems. The value of this approach is easily determined for any system.

Element-by-element testing is often possible. Since the various elements have the same probability of failure and all testing costs are the same then the order of testing is quite unimportant and a random (non-overlapping) strategy may be chosen.

The half split method is a frequently employed and very efficient testing strategy. The method starts by making a group testing of any $N/2$ units (where the system consists of N units.) If the system fault is found to lie in these $N/2$ units then these are further sub-divided into two groups of $N/4$ units and the process continues. If at any point the fault is not found to lie in the tested group then it is assumed to lie in the corresponding untested group and it is this untested group that is subsequently divided. The method continues until a single unit is isolated as the defective item. The method requires modification in some series/parallel circuits. It will be observed in Fig. 1 that for systems in this category the half split method is vastly superior to element-by-element testing.

4.2 Category 2: System composed of N non-identical items—perfect information

This is the general case where the system comprises N items with probabilities of failure p_i $(i = 1 \ldots N)$ and testing costs t_j $(j = 1 \ldots M)$ where M is the total number of possible tests. (Usually $M \geqslant N$ but this need not be so.) Again no testing errors can occur.

In this case the half split method is generally unsuitable since it requires all the probabilities to be equal for it to pinpoint the fault in the minimum time. Moreover, since it does not consider the costs of the various tests, it is unable consciously to minimize these or any other costs. Nevertheless, because of the great power of this method reasonable results can often be obtained in practice.

Kozlov[4] considers a special case where the test costs are equal (though the probabilities need not be the same). He proposes a method which involves calculating the conditional probability of failure for each unit by the formula:

$$p_i^* = \frac{p_i}{1-p_i}\left[\sum_{i=1}^{N}\frac{p_i}{1-p_i}\right]^{-1}$$

Thereafter the method is similar to the half split method except that the groupings are made such that the sum of the conditional probabilities is equal in each split and *not* necessarily the number of units.

To cover the general case of unequal probabilities and unequal testing costs the entropy concept is frequently used. The entropy concept has been used in thermodynamics and physics for a number of years but it was Shannon[5] who first suggested its use as an information measuring technique. Good[6] considered its use in diagnosis. Basically, entropy is a measure of uncertainty and in test selection is defined as:

$$E = -p_i \log p_i$$

where p_i is the probability of item failure. Kozlov[4] proposes the use of entropy concept in engineering system diagnosis. The technique is similar to that used in Ref. 1. Although proposed under this category of system, entropy-based models are also suitable for more complex systems where testing errors can occur.

The methods of dynamic programming[7] have also been used for fault finding in systems of this category and it has been shown[8] that the optimal policy for element-by-element testing—the present extent of this technique in diagnosis—is to select the tests which minimize the ratio p_i/t_i. (p_i and t_i are as previously defined.) Some of the specialized models based on the dynamic programming concept which have been described are:

The case where the system fails when K out of N items fail, i.e., the system has spare or standby facilities.[9]

A special case where dynamic programming is applied to a half split approach. The model, however, is not generally suitable for engineering systems.[10]

The updating of item failure probabilities based on previous testing experience. It is worth noting that the title of this paper may be misleading since the technique itself is not adaptive—only the updating of item failure probabilities is adaptive.[11]

4.3 Category 3: System composed of N non-identical units—imperfect information

This is the ultimate in system complexity. It is a general system as in the previous category but it is further complicated by the possibility of testing errors. This means in effect that a unit which is not faulty may be indicated as faulty, while a unit which is faulty will be occasionally indicated as not faulty. Since testing errors can occur the system is then similar in test outcomes to one where the fault is intermittent. Hence the techniques

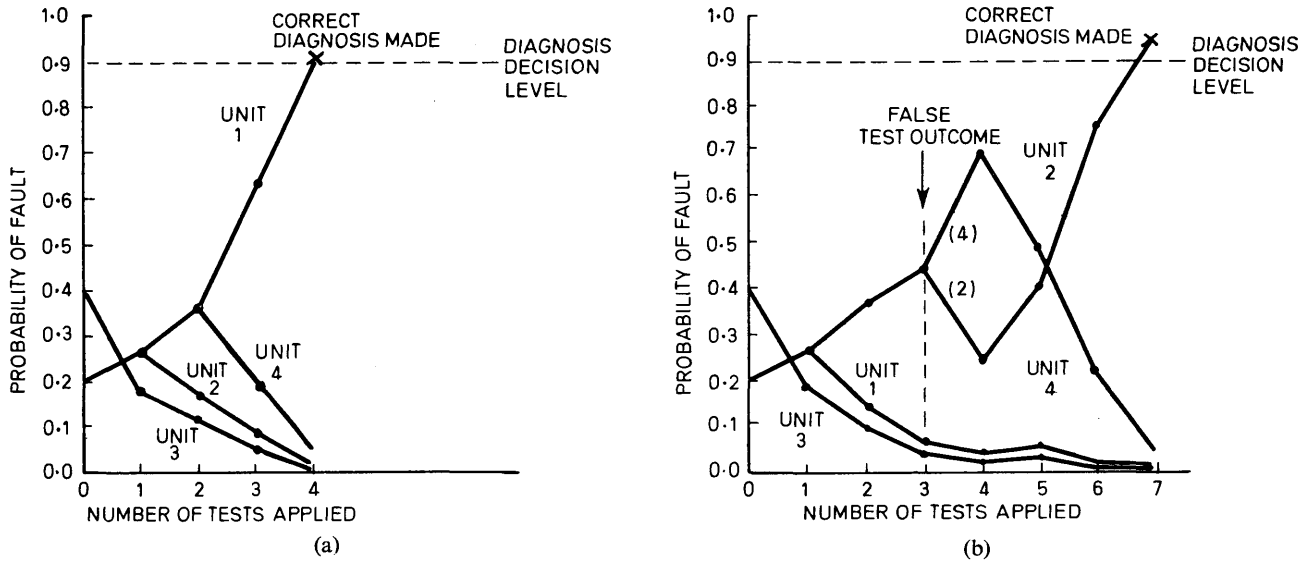


Fig. 2. Examples of changes in probability level when using the entropy method. (Reproduced from Ref. 15 by permission of the Institution of Mechanical Engineers).

developed for systems with testing errors may also be used to diagnose systems with intermittent faults. In systems in this category it is often necessary to apply a test a number of times and the ' balance of evidence ' is accepted rather than the results of any particular tests.

Because of the possibility of testing errors the half split method is generally quite unsuitable since this method places absolute confidence in every test outcome. Consequently if even one error is made in testing then an incorrect diagnosis is guaranteed. Hence frequent incorrect diagnoses may be expected if this method is used for systems in this category.

Methods based on the entropy concept would appear to be of significant value in diagnosing this type of complex system since one or more false test readings may occur and still the correct result obtained, that is, provided sufficient correct test readings are also received. After a test is made the *a priori* probabilities of failure are updated (e.g. Bayesian statistics) and the new *a posteriori* probabilities indicate the extent of present knowledge concerning the relative probabilities of failure. Figure 2(a) shows the case where no false responses are received with consequential swift diagnosis. Figure 2(b) shows the case where a false response is obtained but subsequent correct responses sway the balance in favour of the true fault and the correct system diagnosis is eventually made. In both these cases it will be seen that diagnosis is made when one of the probabilities is greater than or equal to 0·9. But why should diagnosis be made when the probability is 0·9? Why not (say) 0·8 or 0·99 or any other convenient level? This indicates the main problem in using the entropy method, namely selection of the best probability level at which diagnosis should be made. The selection of the probability level is closely related to the number of tests which should be made. The answer lies in analysing the consequences of an incorrect diagnosis. If the 'cost' associated with an incorrect diagnosis is high then in general a large number of tests will be necessary to give a high probability of correct diagnosis and a corresponding low probability (and expected cost) of incorrect diagnosis. If the cost of incorrect diagnosis is low then errors in diagnosis are more tolerable and few tests will be necessary to satisfy the lower diagnosis decision level. In lowering the diagnosis decision level the incidence of incorrect diagnosis must be expected to increase. The selection of the correct probability level is thus a balance between the cost incurred in further testing and the potential cost of an incorrect diagnosis due to insufficient information. The selection of this level is regretfully an arbitrary decision and thus detracts from the usefulness of the method.

An extension of the dynamic programming method[12] attempts to overcome some of the weakness of the entropy method by including in the computational procedure the cost of replacing a good item when it has been incorrectly diagnosed as the faulty item. Also the probability of testing errors occurring is included. These inclusions help to remove some of the subjective element in deciding the diagnosis decision level. It was stated earlier that the method of dynamic programming has only been proved for element-by-element testing and this constraint is immensely restrictive. Were this barrier to be removed then this method is likely to be very useful in the majority of fault finding situations.

Adaptive dynamic programming is a special version of the general technique. Its potential value as a diagnostic technique is enormous and the method is currently the objective of much research. A modification of the general technique of dynamic programming has been applied to the sequential testing of marine boiler tubes.[13] In this particular use of the general technique the very desirable attribute of adaptive dynamic programming was experienced, namely the technique included its own stopping rule (equivalent to the specification of a diagnosis decision level). Hence a totally quantitative approach to fault finding was achieved.

A final technique worth mentioning is that of pattern recognition.[14] Again the potential is significant but until

the results of current research become more freely available a critical appraisal is not possible.

5 Conclusion

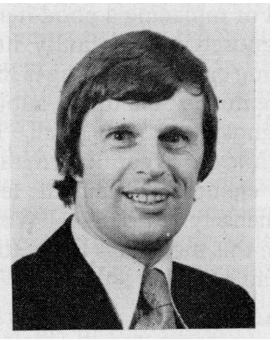
Depending on the complexity of the system under investigation so different groups of testing strategies are possible. Some techniques are universally superior to others but equally a large grey area exists where no one technique has absolute supremacy. For any given system the relative merits of the techniques change as the parameters of the system change, e.g., changes in probability of unit failure with time, changes in probability of testing errors, changes in test costs, etc. Current research is attempting to improve existing methods and develop new methods of fault finding. Additionally it is concerned with the selection of the most appropriate technique in any particular circumstance.

6 References

1. Taylor, T. R., Shields, S., and Black, R., 'Study of cost conscious computer-assisted diagnosis in thyroid disease', *Lancet*, 8th July 1972, pp. 79–83.
2. Rabinovich, V. I., Rozov, M. A., and Timonen, L. S., 'The subject matter and scope of diagnosis in engineering', *Avtometriya*, 1965, No. 1.
3. Kuznetsov, P. I., Pchelintsev, L. A., and Gaydenko, V S., 'Checking and Failure Search in Complex Systems' (Sovetskoye Radio Press, Moscow, 1969).
4. Kozlov, B. A., and Ushakov, I. A., 'Reliability Handbook', pp. 305–10 (Holt, Rinehart and Winston, New York, 1968).
5. Shannon, C. E., 'A mathematical theory of communication', *Bell Syst. Tech. J.*, **27**, No. 3, pp. 379–423, July 1948, and No. 4, pp. 623–56, October 1948.
6. Good, I. J., 'Some statistical methods in machine-intelligence research', *Virginia J. of Sci.*, **19**, pp. 101–10, 1968.
7. Bellman, R., 'Dynamic Programming' (Princeton, New Jersey, 1957).
8. Gluss, B., 'An optimum policy for detecting a fault in a complex system', *Operations Research*, **7**, No. 4, pp. 468–77, August 1959.
9. Butterworth, R., 'Some reliability fault-testing models', *Operations Research*, **20**, No. 1, pp. 335–43, February 1972.
10. Cameron, S. H., and Narayanamurthy, S. G., 'A search problem', *Operations Research*, **12**, No. 4, pp. 623–9, August 1964, (Letter to the Editor).
11. Chu, W. W., 'Adaptive diagnosis of faulty systems', *Operations Research*, **16**, No. 5, pp. 915–27, October 1968.
12. Firstman, S. I., and Gluss, B., 'Optimum search routines for automatic fault location', *Operations Research*, **8**, No. 4, pp. 512–23, August 1960.
13. Devanney, J. W., 'A note on adaptive boiler tube pulling', *Naval Research Logistics Quarterly*, **18**, No. 3, pp. 423–7.
14. Bonch-Bruevich, A. M., Milokhin, N. T., and Shibanov, G. P., 'Applicability of certain algorithms to efficiency checks and equipment inspection', *Kibernetika*, **5**, pp. 95–100, January 1969.
15. Shields, S., 'Engineering diagnosis—the state of the art-science', Proc. I. Mech. E. Conference: 'Terotechnology—does it work in the Process Industries?', I. Mech. E., September 1975.

Manuscript first received by the Institution on 30th January 1975 and in final form on 12th November 1975. (Paper No. 1720/ACS12).

The Author



Mr. Stephen Shields was apprenticed as a Mechanical Fitter with Wellman Engineering Corporation, Belfast, from 1960 to 1966, and studied part-time at the College of Technology, Belfast, for his H.N.C. in Mechanical Engineering. He then attended the University of Strathclyde, obtaining a B.Sc. degree in mechanical engineering in 1968. During the next two years he held engineering appointments with Northern Ireland companies and in 1970-71 he returned to the University of Strathclyde to work for an M.Sc. degree in operational research; following a period as an Operational Research Engineer with Denholm Ship Management Ltd., of Glasgow, he took up an appointment as a Lecturer in the Department of Operational Research at Strathclyde. For the past year he has been with the South of Scotland Electricity Board as a First Engineer in the Forward Planning Division.