



AEROSPACE RECOMMENDED PRACTICE

ARP5783™

Issued 2008-02
Reaffirmed 2018-05

Health and Usage Monitoring Metrics
Monitoring the Monitor

RATIONALE

ARP5783 has been reaffirmed to comply with the SAE five-year review policy.

TABLE OF CONTENTS

1.	SCOPE.....	3
1.1	Purpose.....	3
2.	REFERENCES.....	3
2.1	Definitions, Acronyms, Abbreviations	6
2.2	Goals	8
2.3	Organization of the Report.....	8
3.	METRICS	9
3.1	Introduction	9
3.1.1	1 - Test Conditions.....	9
3.1.2	2 - Algorithm Inputs	10
3.1.3	3 - Algorithm Output.....	11
3.1.4	4 - Non-Diagnostic Performance	12
4.	ACCEPTANCE CRITERIA.....	14
4.1	Introduction	14
4.2	Acceptance Criterion for False Alarm Rates for Flight Critical Use	14
4.3	Acceptance Criterion for Probability of Detection	14
4.4	Interaction between Probabilities of False Alarm and Detection	17
4.5	Acceptance Criterion for Correct Fault Classification (for Flight Critical use)	18
5.	RECOMMENDED METRICS.....	18
6.	NOTES.....	18
APPENDIX A	HYPOTHESIS TESTING	19
APPENDIX B	EXPLANATION OF DECISION MATRIX.....	20
APPENDIX C	DETAILS ON SELECTED METRICS	25
APPENDIX D	IMPACT OF REDUNDANCY	30
APPENDIX E	χ^2 GOODNESS OF BINOMIAL FIT TEST	34
APPENDIX F	METRICS FMEA	36
APPENDIX G	READING ROC CURVES.....	38
APPENDIX H	FAILURE RATES.....	41
APPENDIX I	RELATION BETWEEN EXPONENTIAL AND BINOMIAL FAILURE MODELS	45

SAE Technical Standards Board Rules provide that: "This report is published by SAE to advance the state of technical and engineering sciences. The use of this report is entirely voluntary, and its applicability and suitability for any particular use, including any patent infringement arising therefrom, is the sole responsibility of the user."

SAE reviews each technical report at least every five years at which time it may be revised, reaffirmed, stabilized, or cancelled. SAE invites your written comments and suggestions.

Copyright © 2018 SAE International

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of SAE.

TO PLACE A DOCUMENT ORDER: Tel: 877-606-7323 (inside USA and Canada)
Tel: +1 724-776-4970 (outside USA)
Fax: 724-776-0790
Email: CustomerService@sae.org
http://www.sae.org

SAE values your input. To provide feedback on this Technical Report, please visit <http://standards.sae.org/ARP5783>

APPENDIX J	MULTIVARIATE METRICS.....	47
APPENDIX K	QUALIFYING AN ALGORITHM FOR FAILURE DETECTION.....	49
APPENDIX L	QUALIFYING AN ALGORITHM FOR TRENDING.....	58
APPENDIX M	QUALIFYING AN ALGORITHM FOR FAULT CLASSIFICATION.....	60

INTRODUCTION

This recommended practice investigates metrics that may be useful in evaluating vibration based diagnostic algorithms as health monitors in vehicle health and usage management systems.

There is a twofold need for such monitors. First, is to improve safety of operation, and second, to reduce the cost incurred by the operators enabled by high performance diagnostic algorithms.

To enhance the airworthiness of aircraft, algorithms are being developed for the purpose of detecting on-set of failure. These algorithms show considerable promise in enhancing aircraft safety. It is important to develop tools to evaluate these algorithms, to insure they are able to detect faults while maintaining an acceptably low false alarm rate.

This practice reviews and recommends metrics that can be used to:

1. Measure the effectiveness of candidate algorithms in detecting faults.
2. Measure the algorithms' immunity to false alarms
3. Determine the sensitivity of candidate algorithms to the presence of multiple faults, electromagnetic transients, various flight conditions, sensor placement, different types of gearboxes, state of lubrication and other environmental variables.
4. Establish criteria for accepting algorithms for use in aircraft
5. Evaluate classifications performed by algorithms

The majority of this document is based on the work done by the Rotorcraft Industry Technology Association, RITA, under the Health and Usage Monitoring Systems, HUMS, research program and is used with permission.

1. SCOPE

This recommended practice applies to vibration monitoring systems for rotorcraft and fixed-wing drive trains, airframes, propulsion systems, electric power generators, and flight control systems. It addresses all aspects of metrics, including what to measure, how to measure, and how to evaluate the results.

1.1 Purpose

The objective of this document is to provide recommendations for evaluating the performance of vibration based diagnostic algorithms in vehicle Health and Usage Monitoring Systems.

2. REFERENCES

1. Alban, Lester E. Systematic Analysis of Gear Failures. Metals Park, Ohio: American Society for Metals, 1985.
2. American National Standard. IEEE Guide for General Principles of Reliability Analysis of Nuclear Power Generating Station Safety Systems. ANSI-IEEE Std. 352-1987.
3. American National Standard. Sampling Procedures and Tables for Inspection by Attributes. ANSI/ASQC Z1.4-1993.
4. Arnell, P.P., "The Canadian Forces CH146 Griffon Helicopter Health and Usage Monitoring System," 44th CASI Flight Operations and Flight Test Sections Symposium, pp. 1-10, Toronto, April 28-30, 1997.
5. Becker, Lee A. Crosstabs: Kappa. Univ. of Colorado at Colorado Springs, <http://www.uccs.edu/~lbecker/ctabk.htm>, March 2000.
6. Bruel & Kjaer. Notes on the use of Vibration Measurements for Machinery Condition Monitoring. Application Note 14-222, USA 1980.

7. Choi, Bernard C.K., "Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test," American Journal of Epidemiology, Vol. 148, no. 11, pp. 1127-1132, Dec. 1, 1998.
8. DeGroot, Morris H. "Conditional Probability," Chapter 2 in Probability and Statistics (2nd Edition). Reading, MA: Addison-Wesley, 1986.
9. Dey, Kieron A. (IIT Research Institute). "Sampling Inspection and Statistical Q.C.," Chapter 5 in Practical Statistical Analysis for the Reliability Engineer (State-of-the-Art Report). Rome Air Development Center, Reliability Analysis Center, a DOD Information Analysis Center. SOAR-2, Spring 1983.
10. DOD. System Safety Program Requirements. MIL-STD-882C, Jan 19, 1993.
11. Dorner, William W. "Using Microsoft Excel for Weibull Analysis," www.qualitydigest.com/jan99/html/body_weibull.html, Jan. 1999.
12. Dowling, M. and K. Trainor. "High Frequency Spectral Analysis Study for Helicopter Bearing and Gear Failure Detection," Report to NADC, Warminster, Pa, Naval Air Development Center Final Report ADC-90033-60 issued January 22, 1990.
13. Dubois, Thomas and Richard R. Pruyn, "The Feasibility of Data Fusion Techniques Applied to Rotorcraft Transmission Diagnostics," Signal Processing, Sensor Fusion, and Target Recognition II; Society of Photo-Optical Instrumentation Engineers, Orlando, FL, Apr. 12-14, 1993, SPIE Proceedings. Vol. 1955, p. 105-116, 1993.
14. Ford, Terry, "Airworthiness technology," Aircraft Engineering, A85-27501, pp. 11-14, February 1985.
15. Gong, Peng. Remote Sensing and Image Analysis. Chapter 7, Information Extraction, <http://nature.Berkeley.EDU/~gong/textbook/chapter7/html/sect73.htm>, University of California at Berkeley, Mar 2000.
16. Hanley, James A. and Barbara J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, Volume 143, No. 1, pp. 29-36, April 1982.
17. Harris, William D., Terry Larchuk, Ed Zanoni and Lew Zion, "Application of probabilistic methodology in the development of retirement lives of critical dynamic components in rotorcraft," AHS 55th Annual Forum, 14 pages, Montreal Quebec, Canada, May 25-27, 1999.
18. Heeger, David. Signal Detection Theory Handout (Psych 30). Stanford University, <http://white.stanford.edu/~heeger/psych30/lecture-notes/sdt/sdt.html>, Dec. 21, 1999
19. HHMAG. The Report of the Flight Deck Health Monitoring Indications Working Group, 87 pages. Submitted for Ratification by the Helicopter Health Monitoring Advisory Group, March 25, 1999.
20. Hopkins, Will G. A New View of Statistics. University of Queensland, www.uq.oz.au/~hmrburge/stats (in book form), 1997.
21. Kanal, L. "Patterns in pattern recognition: 1968-1974," Machine Recognition of Patterns (Ashok K. Agrawala, ed.), NY: IEEE Press, 1977. Reprinted from IEEE Trans on Information Theory, vol. IT-20, pp. 697-722, Nov. 1974.
22. Klein, Barbara D., Dale L. Goodhue and Gordon B. Davis, "Can humans detect errors in data? Impact of Base Rates, Incentives, and Goals," University of Michigan, www-personal.umd.umich.edu/~bdklein, Feb. 24, 2000.
23. Lander, Anthony. Statistical Notes. University of Birmingham, web.bham.ac.uk/landerad/statistics.html, Feb. 24, 2000.
24. Lank, G.W., "Theoretical aspects of importance sampling applied to false alarms," IEEE Transactions on Information Theory, Vol. IT-29, No. 1, pp. 73-82, January 1983.

25. Mark, William D., "Analysis of the vibratory excitation of gear systems: Basic theory," J. Acoust. Soc. Am, Vol. 63, No. 5, pp. 1409-1430, May 1978.
26. Mark, William D., "Analysis of the vibratory excitation of gear systems. II: Tooth error representations, approximations, and application," J. Acoust. Soc. Am, Vol. 66, No.6, pp. 1758-1787, Dec. 1979.
27. Montgomery, Douglas C., "Lot-by-Lot Acceptance Sampling for Attributes," Chapter 10 in Introduction to Statistical Quality Control. NY: John Wiley & Sons, 1985.
28. Norman, Donald A., "A comparison of data obtained with different false-alarm rates," Psychological Review, Vol. 71, no. 3, pp. 243-246, May 1964.
29. Okure, Mackay A.E. and Michael A. Peshkin, "Quantitative Evaluation of Neural networks for NDE Applications Using the ROC Curve," Northwestern University, <http://apo.mech.nwu.edu/publications/hokure/ROC.html>, Dec. 22, 1999
30. Parsons, Thomas, "Recognition: features and distances," Chapter 7 in Voice and Speech Processing. NY: McGraw-Hill, 1987.
31. Pollack, Irwin and Donald A. Norman, "A non-parametric analysis of recognition experiments," Psychonomic Science, Vol. 1, No. 5, pp. 125-126, May 1964.
32. Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. Numerical Recipes in C: The art of scientific computing (2nd ed.) Cambridge University Press, 1992.
33. Real, E. and D. Tufts, Estimation of Prescribed False Alarm Rate Thresholds and ROC Curves from Local Data Using Tolerance Regions, ASAP 1999 (Seventh Annual ASAP Workshop).
34. Riegelman, Richard K., "Diagnostic Discrimination of Tests," Chapter 15 in Studying a Study and Testing a Test: How to read the medical literature, Little Brown, 1981.
35. RIS. An Explanation of Vibration in Rotating Machinery. Rochester Instrument Systems, 1983.
36. Rose, H. John, "What 'HUMS' Means to Boeing Helicopters," Aerospace Technology Seminar, Republic of Singapore, May 18, 1995.
37. Rozak, James N. RITA HUMS Diagnostic Database. RHDDDB 98-06-1.1(4), Draft 2.0.
38. SAE International E-32. A Methodology for Quantifying the Performance of an Engine Monitoring System. Aerospace Information Report (AIR) 4985, January 2005.
39. Savinsky, Yu and G.S. Borisov, "The total fault-and-failure analysis at the design stage of a civil helicopter to ensure safety, availability, and minimized direct operating costs," 21st European Rotorcraft Forum, vol. 3, pp. IX7.1-7.14, St. Petersburg, Russia, Aug. 1995.
40. Shapiro, Jeffrey H., "An extended version of Van Trees' Receiver Operating Characteristic Approximation," IEEE Trans. Aerospace and Electron. Syst., http://www.cis.jhu.edu/wu_publications/s/shapiroj2.html, 1999.
41. Shipley, Eugene E., "Gear Failures," Machine Design, December 7, 1967.
42. SMS. Typical Machinery Problems. SMS 8010-015.
43. Stott, G.F. and E.V. Stanfield, "Measurement of false alarm rate in a synthetic radar environment," Colloquium on specifying and measuring performance of modern radar systems, Colloquium Digest, IEE No. 221, 1998.
44. Tantaratana, Sawasd, "Two-stage CFAR detectors for multiple-range-bin radars," IEEE Trans. On Aerospace and Electronic Systems, Vol. AES-19, No. 5, pp. 688-703, September 1983.

45. Teal, Richard Scott. Sensor Analytic Redundancy for Helicopter Flight Control. MS Thesis, MIT, February 1984.
46. Teal, Rich, Dave Miller, Jeff Evernham, Don Marquith, Dan Deibler, Terry Larchuk, and Fred White, "Regime Recognition for MH-47E Structural Usage Monitoring," American helicopter Society 53rd Annual Forum, VA, 1997.
47. Tsien, Christine L. and Hamish S. F. Fraser, "Classification trees: What they are and an example from a clinical domain," MIT Laboratory for Computer Science, 6.872 Lecture, Feb 19, 1998.
48. United States Department of Defense. Sampling Procedures and Tables for Inspection by Attributes, MIL STD 105D, 1963.
49. United States Army Management Engineering Training Agency (USAMETA). Design and Analysis of Experiments. Rock Island, IL: USAMETA, July 1971.
50. Vinall, P.D. "Airworthiness of Helicopter Transmissions," Civil Aviation Authority, Royal Aeronautical Society Conference: Helicopter Transmissions, Proceedings of the Symposium, London, England, 30 p., Feb. 6, 1980.
51. Western, W.J.H. and N. J. Boulding, "UK Military Experience and Viewpoint," Helicopter Airworthiness in the 1990's: Health and Usage Monitoring Systems – Experience and Applications, Royal Aeronautical Society, pp. 4-1 – 4-15, Nov. 29, 1990.
52. Whalen, Anthony D., "Hypothesis Testing," Chapter 5 in Detection of Signals in Noise. New York: Academic Press, 1971.
53. Zar, Jerrold H., "Rank Correlation," Section 18.6 in Biostatistical Analysis. Englewood Cliffs, NJ, 1974.

2.1 Definitions, Acronyms, Abbreviations

Algorithm	Series of steps taking raw data to a useable state
ANOVA	Analysis of Variance
A posteriori	Probabilities based on acquired data
binomial	Distribution of elements of one of two classes sampled with replacement
CFAR	Constant False Alarm Rate
Chi-square	Distribution of sums of squared values. Useful for determining confidence limits of variances and for goodness of fit tests.
Coherence	Measure of causality between input and output. A spectrum of linear correlation coefficients at every frequency.
Condition index	Numerical output of an algorithm which indicates condition of the mechanical component
Conditional probability	Probability of an event given that another event has occurred
Confusion matrix	Matrix showing numbers or proportion of correct and incorrect classification.
Correlation	A measure of the strength of the relationship between two variables
Critical value	A tabulated value for a distribution
Density function	Derivative of a statistical distribution

d	Difference between ranks of two related variables
D'	Discriminability index (separation / spread)
d ^c	Distance from the intersection of the minor diameter and the ROC curve to the center of the ROC graph
discriminability	Index reporting the separation of peaks divided by their spread
False alarm rate	Proportion of fault-free cases that trigger an alarm
Gold standard	Data of rotorcraft component whose condition is certified in writing
Hold-out	Data reserved for final testing; not used during algorithm design
HUMS	Health and Usage Monitoring System
hypergeometric	Distribution of elements of a class sampled without replacement
measurand	Physical variable being measured
Miss rate	Probability of missing a fault
Neyman-Pearson criterion	Method of determining threshold that maximizes the probability of detection given a fixed false alarm rate.
Null hypothesis	Hypothesis of no fault.
pdf	Probability density function
P(j)	Probability of j successes
POD	Probability of detection (given a fault)
POFA	Probability of false alarm
Posterior probability	(Conditional) probability of a fault (or non-fault) estimated after an observation has been taken.
Prevalence	Proportion of all cases which have a fault
Prior probability	Probability of occurrence of a fault, estimated from historical data
Rank	Position of measurement in data ordered by value
Rayleigh	Distribution of peak values of narrowband Gaussian signals
Receiver operating Characteristic	A plot of the probability of detection versus the probability of false alarm with the threshold as the along-curve parameter and SNR as the between-curve parameter.
Rician	Distribution of the envelope of a sine wave plus narrowband Gaussian noise
RITA	Rotorcraft Industry Technology Association, Inc.
ROC	Receiver Operating Characteristic

Sensitivity	Probability of detection given that a fault occurred
Separation	Distance between noise-only and signal-plus-noise pdf modes or means
SNK	Student–Newman–Kuels
SNR	Signal to Noise Ratio
Spall	Advanced pitting of gear teeth due to contact fatigue
Specificity	Probability of correct classification of no-fault
Spread	Dispersion of data usually measured in standard deviations.
Test statistic	An observation compared to a threshold and used to conclude if the null hypothesis is true or false
Tolerance factor	Multplier to convert sample spread to estimated population spread

2.2 Goals

To accomplish the stated purpose, guidelines are proposed for measuring the probabilities of detection and false alarm rate. When possible, this will be done on single and multiple faults as well as sensor failure modes. Ideally the tests will also be conducted under a variety of flight, lubrication and environmental conditions. These conditions include:

- noise of various distributions
- noise spikes such as electromagnetic transients
- various wind and turbulence conditions
- different flight regimes
- temperature extremes
- good and poor states of lubrication.

All these variations are required to be within the allowed operating environment.

2.3 Organization of the Report

The overall approach to testing diagnostic algorithms is covered in Section 3. It includes a discussion of the kinds of tests to be conducted and how they will be made as realistic as possible. It further describes test metrics, the yardstick by which we measure the performance of HUMS algorithms. This section lists the metrics, describes them, and indicates why and how they are used. Acceptance criteria are discussed in Section 4. Conclusions and recommendations are presented in Section 5.0.

3. METRICS

3.1 Introduction

Numerous metrics have been reviewed. Those listed in this section have been selected as being most effective in evaluating vibration based diagnostic algorithms.

Four groups of metrics are presented. The most important group is related to an algorithm's output and evaluates how the algorithm performed. The other three groups document the test condition as it pertains to data quality, input to the algorithm, and the non-diagnostic performance of the HUM system. In order not to interrupt the flow of the report, the mathematics behind the algorithm metrics is presented in appendices.

3.1.1 1 - Test Conditions

The following test conditions are recorded to document channel characteristics (all tests) and dynamic conditions (flight tests). The asterisk indicates recommended measurements. This information will be shared with all parties.

TABLE 1

Test Conditions		
Metric	Description	Limitations
Data Collection		
*Sample rate	Places limitations on bandwidth.	
*Bandwidth	Single number giving approximate limit of unity response.	Approximate measure.
*Channel frequency response	Frequency response spectrum. More detailed than bandwidth. Should include the effect of filters.	A vector. Not a single index.
*Channel crosstalk	Low crosstalk is a necessary condition for sensor independence.	
Transducer-transducer correlation	Noise non-correlation is a necessary condition for sensor independence	
*Transducer-transducer coherence	Best way to measure effect of transducer cross-correlation; provides correlation coefficient at every frequency.	
*Transducer sensitivity	Converts electrical to physical units.	Must be tied in with gain.
*Channel gain	Necessary for amplitude measurements.	
Flight Conditions (if a flight test)		
*Flight/hover regime		
Torque		
NR shaft thrust	Rotor Speed	
Environmental Conditions (if a flight test)		
Outside air Temperature		
Indicated airspeed		
Pressure altitude		

3.1.2 2 - Algorithm Inputs

The characteristics of the input signal and noise need also be measured. These are important for determining the probabilities of false alarm (POFA) and detection (POD). The signal and noise characteristics are shown in the following table.

TABLE 2

Input to Diagnostic Algorithms		
Metric	Description	Limitations
Discrete Signal Measurements		
Signal p-p before averaging	Important time domain measurement for discrete signal.	
*Signal p-p after averaging	Signal should remain constant with synchronous averaging.	
Signal peak in spectrum	Cross-check on time domain measurement.	
*Signal frequency	Key measurement.	
First sideband level	First sideband amplitude level	
*% amplitude modulation	AM sideband energy as percent of carrier	
*% frequency modulation	FM sideband energy as percent of carrier	
*Harmonic distortion	harmonic energy as percent of fundamental	
Noise Measurements		
*Mean	First statistical moment	
*Standard deviation	Second statistical moment	
*Skewness	Third statistical moment	
*Kurtosis	Fourth statistical moment	
*Rms level	Measures background noise; captures both AC and DC components	
*Pdf (probability density function)	Amplitude distribution of noise energy	
*PSD (Power spectral density)	Frequency distribution of noise energy	
Measures for Testing Distribution*		
χ^2 Goodness of Fit	Tests any empirical distribution against theoretical.	Requires substantial amount of data.
Shapiro and Wilk's W test	Normality test for small samples.	Limited to normal distribution.
D'Agostino's D	Normality test for small samples. More powerful than W test.	Limited to normal distribution.
Transient Noise Measures		
*Peak amplitude	Single most important measure for transient	
*Crest factor	Ratio of Peak (transient) to RMS (background level)	
*Reps per second	Repetition rate if periodic.	
*Damping factor	Measure of sharpness of transient	
Combined Signal and (non-transient) Noise Measurements		
*SNR (aka discriminability, effect-size)	Signal peak amplitude divided by RMS noise.	

It is recommended that the noise distribution be determined by means of fitting with an appropriate distribution or combination of distributions, paying particular attention to the tails. It is further recommended that a commercial program, be utilized to definitively characterize the noise distribution for record and to allow the HUMS suppliers a baseline for comparing their distribution evaluations.

3.1.3 3 - Algorithm Output

The results from the diagnostic tests must be measured in order to evaluate the algorithms. These are described in the following table. The values are calculated after the Vehicle Health Management System reports the decisions and condition indices produced by its diagnostic algorithms.

TABLE 3

Metrics for Output from Diagnostic Algorithms		
Metric	Description	Comments
<i>Essential Measures for Binary Decisions</i>		
*P(Detection)	$P(D1 F1)$. Aka Sensitivity, or Power of a test	Varies with noise and signal level.
*P(False Alarm)	$P(D1 F0)$. Will not vary when using CFAR. Aka Producer's risk, Size of a test, or Type I error.	Varies with noise unless using CFAR. With CFAR approximately constant.
<i>Additional Measures for Binary Decisions</i>		
*Decision Matrix	Shows hit and miss rates.	Not an indicator per se.
Specificity	$P(D0 F0)$. Aka Correct rejection rate.	Redundant.
Miss Rate	$P(D0 F1)$. Aka Consumer's risk or type II error.	Redundant.
Predictive value of positive test	$P(F1 D1)$. Proportion of detected cases having a fault.	Redundant.
Predictive value of negative test	$P(F0 D0)$. Proportion of negative cases that are fault free.	Redundant.
P(no fault given an alarm)	$P(F0 D1)$	Redundant.
P(fault given no alarm)	$P(F1 D0)$	Redundant.
Prevalence	$P(F1)$. Probability of having a fault	Redundant.
P(no fault)	$P(F0)$. Proportion of all tests which have no fault.	Redundant.
Unconditional P(detection)	$P(D1)$. Proportion of all tests which are positive.	Redundant.
P(no-detection)	$P(D0)$. Proportion of all tests which are negative.	Redundant.
*Accuracy	$P(D1 F1 \& D0 F0)$. Proportion of all test which were diagnosed correctly. Universally understood measure. Can be used to compare two algorithms.	Doesn't distinguish between FA and detection probabilities. Varies with noise.
*ROC Area	Area under ROC curve. Has visual impact. Area can be used to compare two algorithms.	Varies with noise and signal. Biased toward high Pfa end of ROC curve.
*ROC dc	Distance between center of ROC plot and intersection of ROC and minor diameter. Can be used to compare two algorithms.	Varies with noise and signal level. Doesn't distinguish between FA and P(detection).
*P(isolation)	Probability of correct isolation	See Appendix H.

TABLE 3 (CONTINUED)

Metrics for Output from Diagnostic Algorithms		
Metric	Description	Comments
Measures for Classification		
*Confusion Matrix	A matrix showing percent hits for each class.	A matrix rather than single parameter.
% Correct classification	Matrix trace in terms of probability.	
*Kappa coefficient	Single parameter showing measure of agreement.	
ANOVA	Tests if condition indices can detect differences in faults	Can't discern which faults are significantly different
SNK Test	Compares individual condition indices	Can't do for subsets
Scheffe's multiple contrast	Compares subsets of condition indices	Low power test (can have significant Type II error)
Fukunaga J measures	Measures separability of classes	
Measures for Comparing Multiple Fault Levels		
*Spearman Rank Correlation	Correlates qualitative fault to quantitative algorithm condition index. Tests monotonicity. Useful for trending as well as correlation.	Not as stable as ordinary linear correlation coefficient r.

The first three metrics in the above table (algorithms output) are the essential measures of the performance of the system. The other metrics provide a deeper understanding of the process.

Confusion matrices and the associated kappa coefficient should be used to indicate relative successes and failures when testing on different kinds of faults.

Note that measurements of signal strength, noise, and the signal to noise ratio can equally be listed under the "Metrics for Output from Diagnostic Algorithms" as for the "Input to Diagnostic Algorithms" table because to some extent the signal is a function of how well the individual algorithm extracts features,. Additionally the "noise" is a function of the extent to which the algorithm reduces the noise by processing the raw signal. It is recommended to record both pre-processing and post processing signal and noise. Part of the post processing "noise" is the variability in the condition index for repeated tests on essentially the same data.

Most of the probabilities in the above table can be described in overall terms, for a particular component, or for a specified failure mode on a component. The various treatments, as well as time rates, are enumerated in Appendix H.

The appendices contain the technical basis for the metrics. Appendix B shows the mathematical details of calculating the POD, POFA, and associated measures. This is related to hypothesis testing discussed in Appendix A. Appendix C provides details on many of the other metrics. Appendix G provides guidance on interpreting ROC (receiver operating characteristic) curves. Appendix J discusses multivariate metrics. Appendix H covers isolation probability and time rates of failure. Appendix H also ties the general probabilities listed in this section with specific component and failure rates addressed in SAE AIR4985, 2005.

3.1.4 4 - Non-Diagnostic Performance

The system non-diagnostic performance can also be evaluated.

TABLE 4

Metrics for System non-Diagnostic Performance		
Metric	Description	Comments
Mean Times to Process		
MTT Detect given failure	Average time required to detect a failure (given the failure).	
MTT Detect prior to failure	Average time required to detect a failure condition given a slow buildup to failure.	
MTT Diagnose given detection	Average time required to isolate or diagnose the cause of a failure after failure has been detected.	
System Requirements		
Minimum Memory	Minimum memory required to operate system in present configuration.	
Actual Memory	Actual memory used in tests.	
Minimum Storage	Minimum storage required to operate system in present configuration.	
Actual Storage	Actual storage used in tests.	
Consecutive measurements	Number of consecutive measurements a system made to decide fault/no-fault. (Can be averaged or calculated on case basis.)	
Redundant Sensors	Number of independent sensors required to determine fault definitively for this system.	
System Performance		
Number of crashes	Number of times system crashed and had to be rebooted	
Number of repeats	Number of times HUMS supplier requested test repeats due to system problem	
Setup time	Amount of time operator required to set up.	
Operator intervention	Degree of operator intervention required to operate: 1-10 little to much	
Margin	$\text{Margin} = \frac{\text{Condition Index} - \text{Threshold}}{\text{Threshold}} \cdot 100$	How close is Condition Index to threshold? Can be averaged or stated on a case basis.
HUMS supplier's rank	Rank algorithms and performance for each component and fault. Include generic algorithm and system.	

4. ACCEPTANCE CRITERIA

4.1 Introduction

This section suggests acceptance limits for key probabilities: probability of detection, probability of false alarm, and probability of fault classification.

4.2 Acceptance Criterion for False Alarm Rates for Flight Critical Use

"Any HUM system fitted must be more reliable than the system it is monitoring, otherwise it loses credibility and can cause rather than prevent accidents." Western, 1990

4.3 Acceptance Criterion for Probability of Detection

While a detection probability of 0.9 or higher is desirable, this may not always be possible. In the first place, POFA and POD are correlated because both are measured on the high side of the threshold. If the threshold is raised to decrease the probability of false alarm, the probability of detection is decreased accordingly. This can be seen in the following illustration:

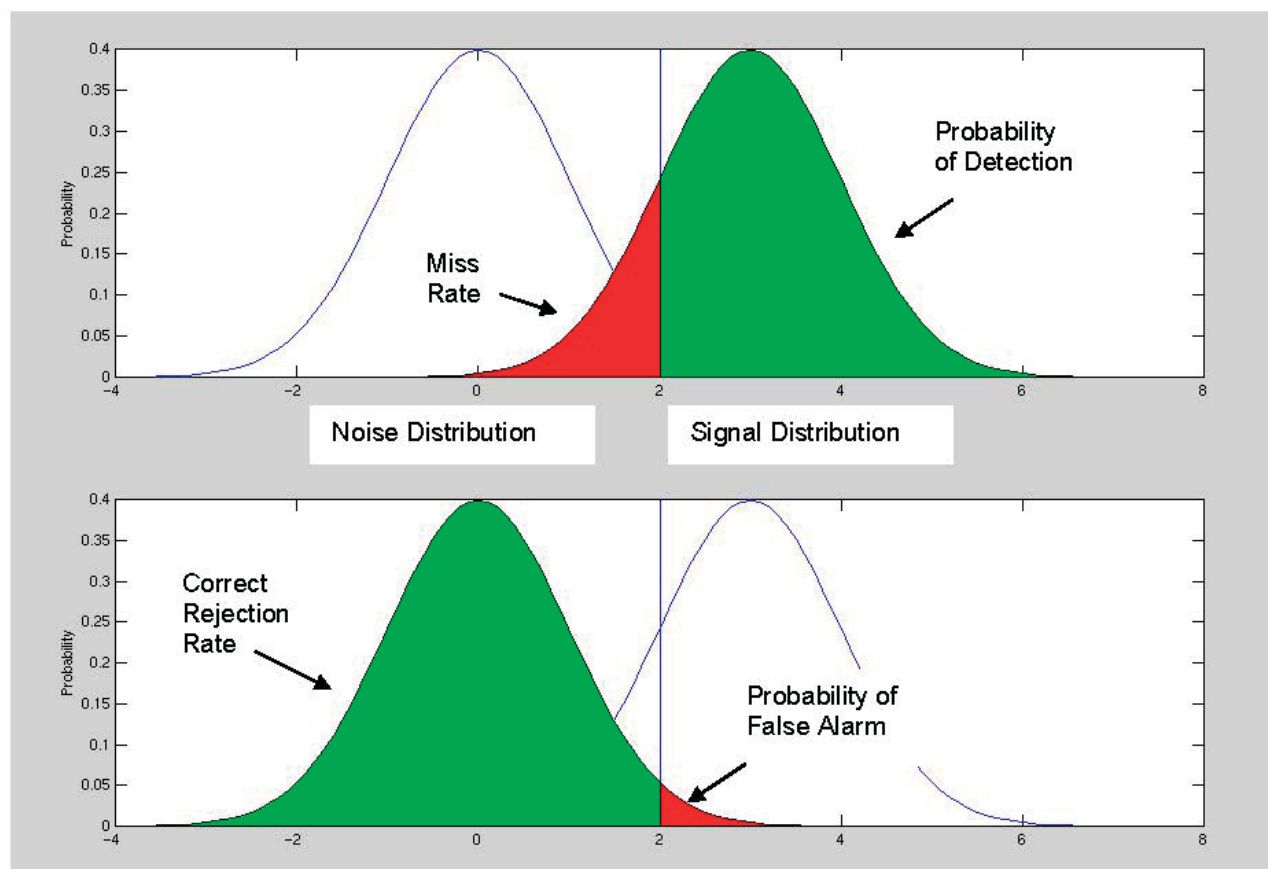


FIGURE 1

The upper figure highlights the signal distribution. The probability of detection can be seen on the right side of the threshold. The lower figure emphasizes the noise distribution. POFA is also on the right side of the threshold. To decrease POFA, we raise the threshold (move it to the right), and this has the unfortunate effect of decreasing the probability of detection. The following figure shows the threshold being moved successively to the right. The probability of false alarm (POFA) is shown in red. But it can be appreciated that as the POFA is decreased, so is the probability of detection (POD). As it is shown if the probability of false alarm is made more stringent, the probability of detection will degrade. A consequence of this is that the warning time between fault detection and catastrophic failure is reduced.

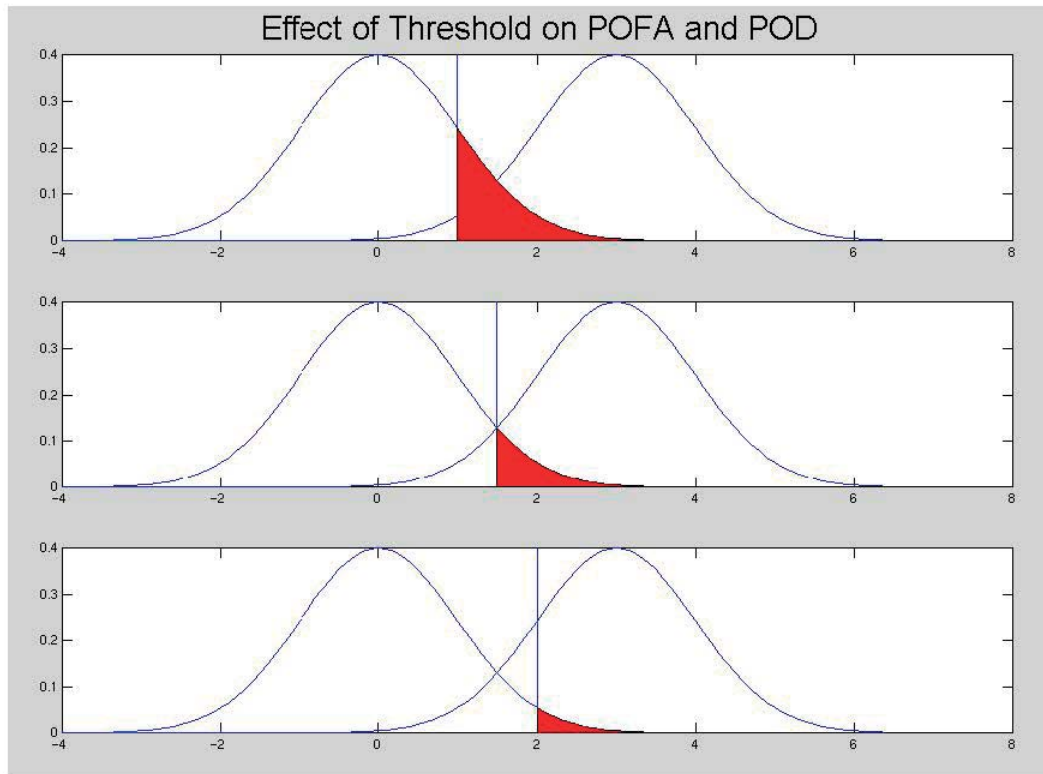


FIGURE 2

A curve known as the receiver operating characteristic shows the effect of a changing threshold on POFA and POD over the full range. See below.

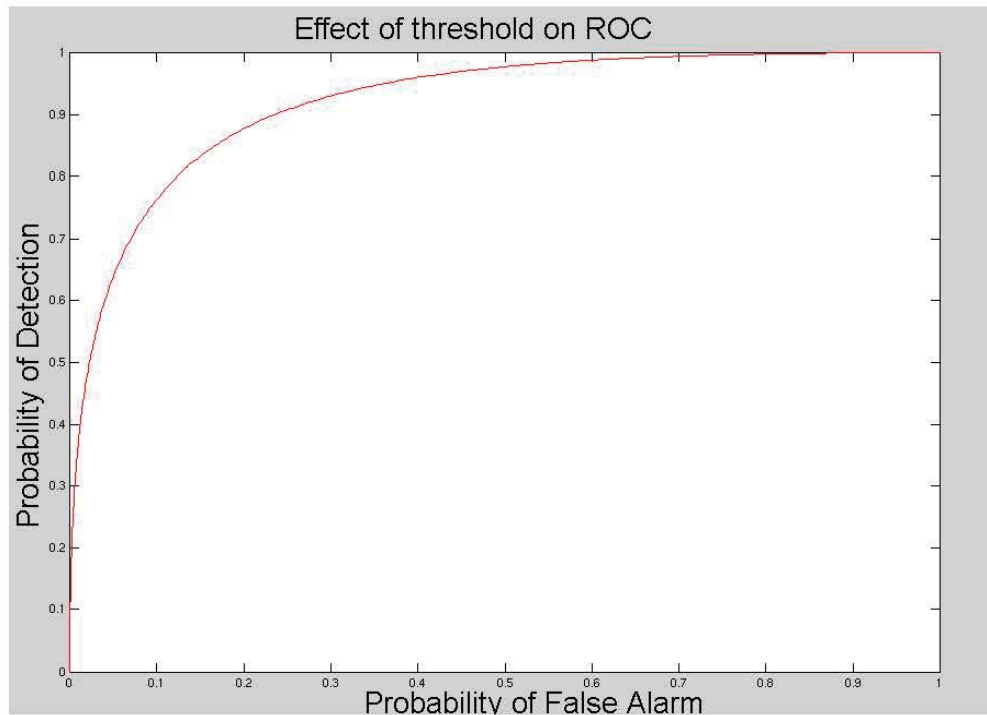


FIGURE 3

The probability of detection is also affected by the signal to noise ratio. For example, if the signal is increased (in other words, the entire signal distribution is shifted to the right) without changing the noise or the threshold, the POD (area in the signal distribution to the right of the threshold) will increase.

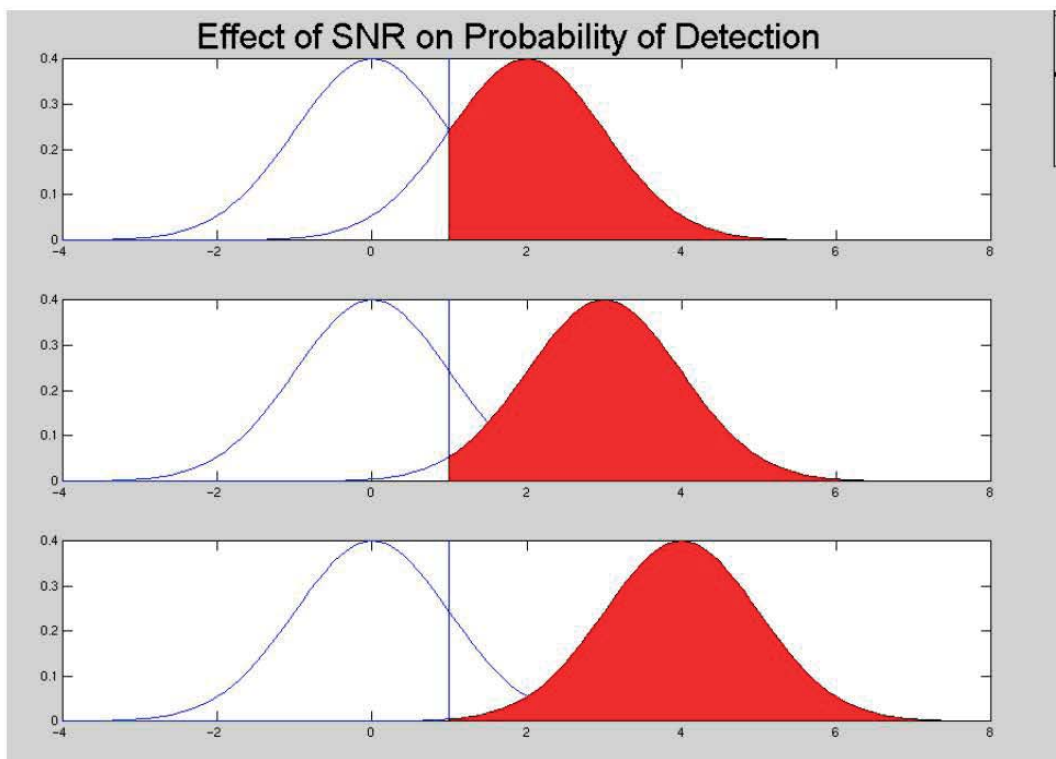


FIGURE 4

This can be appreciated even better in the following figure of the receiver operating characteristic.

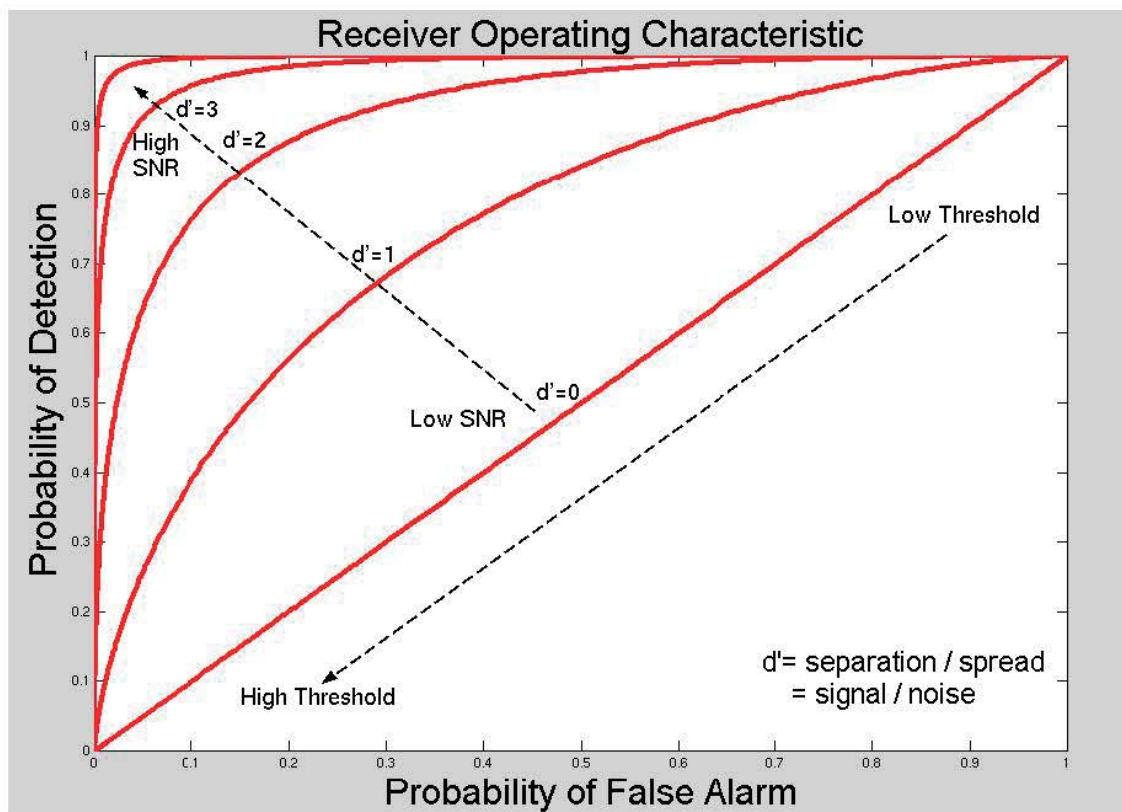
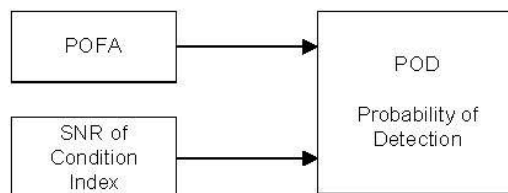


FIGURE 5

It can be seen that as the signal to noise (d') decreases, the POD also decreases for a given POFA. In other words, if the POFA is fixed, the maximum possible POD is limited by the discriminability (or SNR).

4.4 Interaction between Probabilities of False Alarm and Detection

The figure below shows how the POD is influenced by the POFA and SNR:



Prob. of Detection is determined by POFA
and SNR

FIGURE 6

All algorithm suppliers are subject to the same POFA. However, the signal-to-noise ratio in their condition indices may vary, resulting in different probabilities of detection.

4.5 Acceptance Criterion for Correct Fault Classification (for Flight Critical use)

Pilots in Helicopter Health Monitoring Advisory Group (HHMAG) were adamant that an in-flight indicator should not only have an extremely low false alarm rate, but isolate defects and provide quantification as well [HHMAG, 1999]. Similar risks in a nuclear plant environment often accept 95% coverage. Using this reasoning, an algorithm should be able to isolate a fault 95% of the time. To be considered capable of classification, or fault isolation within a major component, it is proposed that an algorithm should be able to correctly isolate a fault within a component 95% of the time.

The probability of isolation given detection should be 0.95.

5. RECOMMENDED METRICS

The following metrics are recommended for evaluating diagnostic algorithms:

Measuring effectiveness

Sensitivity (Probability of Detection)

False Alarm Rate

Decision matrix

Discriminability Index (SNR)

Measuring classification capability

Confusion matrix

Probability of correct classification or isolation

Kappa coefficient

Tending and correlating condition index with fault severity

Spearman's rank correlation

6. NOTES

- 6.1 A change bar (|) located in the left margin is for the convenience of the user in locating areas where technical revisions, not editorial changes, have been made to the previous issue of this document. An (R) symbol to the left of the document title indicates a complete revision of the document, including technical revisions. Change bars and (R) are not used in original publications nor in documents that contain editorial changes only.

APPENDIX A - HYPOTHESIS TESTING

A hypothesis is an assertion about the parameters of a probability distribution. A hypothesis test is an objective evaluation of that assertion.

The null hypothesis is designated by H_0 , the alternative hypothesis by H_1 . Within the context of fault detection, the null hypothesis, H_0 , asserts there is no fault, for the case under consideration, in the gearbox. The alternative, H_1 , claims there is a fault.

D_0 is the case in which no fault is detected. If there is no fault (H_0) and none is detected (D_0), then there is no error. This state is represented in the bottom right cell in the following table.

TABLE A1

The Two Types of Error in Hypothesis Testing		
	H_0 is False (H_1)	H_0 is True (H_0)
H_0 is rejected (D_1)	No error	Type I error Alpha (α) error Error of the First Kind Producer's risk
H_0 is not rejected (D_0)	Type II error Beta (β) error Error of the second kind Consumer's risk	No error

It can be seen that each error has several descriptive names.

Sometimes it is more convenient to work with the power of a test rather than beta, where

$$\text{Power} = 1 - \beta$$