USE OF GENERALISED PARETO MODELS TO SET ON-BOARD DIAGNOSTIC FAILURE THRESHOLDS

Paul J. King [1] and Keith J. Burnham[2]

[1]*Powertrain Control Systems and Calibration, Jaguar Land Rover Limited, Coventry, CV3 4BJ, U.K.*
[2]*Control Theory and Applications Centre, Coventry University, Coventry, CV1 5FB, U.K.*

Vehicles sold in the US and Europe have to be equipped with a Diagnostics, called On-Board Diagnostics (OBD), which monitor the performance of various elements of the emission control system. The driver is informed as to any failures by the use of a Check Engine Light on the dashboard of the vehicle and then should return the vehicle to the dealership for rectification. The vehicle manufacturer's aim is to ensure that the Check Engine Light is only illuminated for legitimate failures. For the calibration of an On Board Diagnostics there needs to be sufficient separation between the response of a good sensor and that of a failed sensor the setting of this threshold should be based upon a statistical model of the data so that the predicted failures rate can then be determined. In practice generating a statistical model which is appropriate for the setting of fault thresholds can prove difficult. In this paper this issue is resolved by making use of an extreme statistics method, Generalised Pareto Distribution model to allow a threshold for an Airflow Sensor diagnostic to be set.

Keywords: Automobiles, Detection, Diagnosis, Engines.

1. INTRODUCTION

Vehicles sold throughout the world are subject to an increasingly stringent set of emission thresholds. To achieve certification, all sensors and vehicle sub-systems that may affect vehicle exhaust emissions have to be monitored by an On-Board Diagnostic (OBD) system that is part of the Engine Management System (EMS) or any other embedded controller. This requirement was first introduced in the US in 1988 for OBD1, for open and short circuit faults, and in 1994 for OBD2, for changes in sensor and actuator responses. For Europe this legislation, denoted EOBD, has been introduced for all vehicles built after January 2000. Both sets of legislation link the performance of the different diagnostics to emission thresholds. In the event of component or sub-system failure, a ‘check engine’ light must be illuminated as an indication to a driver that there is a problem, so corrective action can be taken to minimise the pollution caused by such a fault. As the emissions thresholds are continually reduced, more sophisticated techniques are required to be employed to meet these increasingly tightening thresholds.

OBD Diagnostic Calibration engineers develop test plans that invoke the worse conditions for the diagnostics, by introduction of variety of test conditions. These are typically different fuel specifications, operations at different ambient conditions (hot, cold and altitude), with different driving styles and tolerance sensors. Each diagnostic will have it own set of worse case test conditions which have been developed through the experience of the engineer and lessons learnt. The approach used by the calibration engineer is to collect data for these conditions and then fit a Gaussian distribution to the data to determine the standard deviation and by use this information to set threshold to ensure that 'normal' systems does not false flags and that 'failed' system flag in a timely manner. If a Gaussian distribution is in appropriate then other distributions could be used to obtain a better fit. The issue with fitting a parametric model to the data is that it will fit accurately to the sections of which contain the majority of the data as a consequence the information that is contained in the tails of the distribution is never modelled accurately. Unfortunately it is the information and the data in the tails of the distribution which will lead to false flagging of the diagnostic.

The paper is organised as follows: Section 2 introduces the specific Problem Formulation, Section 3 introduces the use of Extreme Value Statistics and Generalised Pareto Distribution, Section 4 details the implementation and analysis and finally, Section 5 details further work and concludes the paper.

2. Problem Formulation

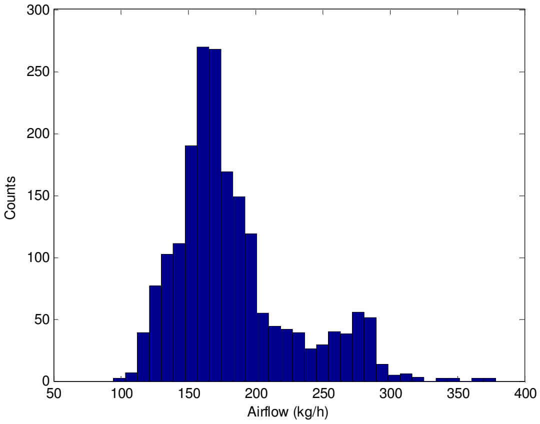
In this paper the calibration of an airflow meter diagnostic is considered. This makes use of a simple calibratable map with Engine Load and Engine Speed as the x and y axes and z axis being the response of the airflow sensor. A threshold can then be set at an appropriate airflow level which will then correctly pass 'normal' systems and ensure that there is a suitable level of detection of 'failed' sensor. In this paper only data collected from a 'normal' system has been considered, but in practice to set the final threshold level the data for both the 'normal' and 'failed' sensor would need to be analysed.

For purposes of illustration the calibration of only one of the points in the calibration map will undertaken in this paper. The data used is centred at a nominal calibration point of 150 to 225Nm for the Engine Load and 1750 and 2250 rpm for Engine Speed. Figure 1 shows the histogram for the measured airflow sensor response, the resulting distribution is not Gaussian and is bi-modal. This phenomenon is due to the operation of the diagnostic in that it doesn't capture any dynamic behaviour and therefore the bulk of the data centralised around 160 represents the steady-state driving condition with the second smaller distribution centralised around 275 capturing the more transient data. The data for the second distribution is generated when the driver transitions through the particular engine speed and load region. From the distribution in Figure 1 the data is bi-modal so this reinforces the validity of the approach being highlighted in this paper.

3. Extreme Value Statistics

There are two general forms of Extreme Value statistics one is Block Maxima which considers that maximum (or minimum) of a sequence of observations. The distribution of this new set of data can then be approximated to the Generalised Extreme Value (GEV) distribution (Fischer 1928) which takes the form of one of the following distributions Gumbel, Frechet and Wiebull. This can then be used to determine the probability of an event exceeding a minimum or maximum level over given period of time based upon a data set which normally contains information for a lesser period of time. This is useful tool for example in the specification of the height of flood defense walls. Where you might want to specify sea defenses to provide protection for 100 years however you only have data for the past 10 years. Using can then use GEV as a framework which will allow you to set the height level based upon the amount of risk of flooding you are willing to accept.

Figure 1: Airflow Histogram for specific Engine Speed and Load



The second approach is to consider the Peaks Over Threshold (POT), this theorem states that the observations exceeding a high threshold, under very general conditions, are approximately distributed as the Generalised Pareto distribution (GPD) (den Haan 1974 and Pikands 1975). This distribution has three forms Exponential, Pareto and Beta. This has been used widely in the Finance and Insurance Industry to determine risk (Embrechts 1977, McNeil 1999). The potential values of a risk have a probability distribution which can never be observed exactly, although past losses due to similar sets of data, may provide information as to the risk. Extreme events occur when a risk takes values from the tail of its distribution. It is this form of Extreme Value Statistics which is most applicable for the setting of diagnostic thresholds and will therefore be further developed in this paper.

Generalised Pareto Distribution

The GPD is a two parameter distribution with the following function

$$G_{\xi\beta}(x) = \begin{cases} 1-(1+\xi x/\beta)^{-1/\xi} & if \quad \xi \neq 0 \\ 1-e^{-x/\beta} & if \quad \xi = 0 \end{cases}$$

where $\beta > 0$, and where $x \geq 0$ when $\xi \geq 0$ and $0 \leq x \leq -\beta/\xi$ when $\xi < 0$. If ξ is defined as the shaping parameter with $\xi > 0$ indicating that the distribution has a heavy-tail, and a value of $\xi = 0$ indicating that the distribution is exponential (McNeil 1999).

Let $X_1 X_2 \cdots$ be a set of independently and identically distributed random variables representing losses with unknown Cumulative Distribution Function *F(.)* The distribution of (excess) losses, *y*, over a threshold μ is given by

$$F_u(y) = P\{X - \mu + y \mid X > u\} = \frac{F(y+\mu) - F(\mu)}{1-F(\mu)}$$

for $0 < y < x_0 - \mu.$, where $x_0 > \infty$. is the right endpoint of *F*. The excess distribution represents the probability that a loss exceeds the threshold μ by at most an amount *y*, given that it exceeds the threshold. For a large class of underlying distributions *F*, as the threshold μ is progressively raised, the excess distribution F_μ converges to a generalised Pareto distribution.

Assuming that N_u out of total *n* data points exceed the threshold μ, the GPD is fitted to the N_u excess values using maximum likelihood methods. The choice of the threshold is a trade-off between having sufficiently high threshold to ensure that enough data is included with the tail of the distribution and not including too much data from the body of the distribution

The GPD distribution function can also be written as:

$$F(x) = (1 - F(\mu))G_{\xi\beta}(x-\mu) + F(\mu)$$

for $x > \mu$. Where the tail estimator *F(μ)*, can be constructed using an empirical estimator, in other words $(n - N_u)/n$,

Putting together the empirical estimate of *F(μ)* and the maximum likelihood estimates of the parameters of the GPD an estimation of the tail can be obtained

$$\hat{F}(x) = 1 - \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-\mu}{\hat{\beta}}\right)^{-1/\hat{\xi}} \qquad (1)$$

Figure 2: Raw Histogram data

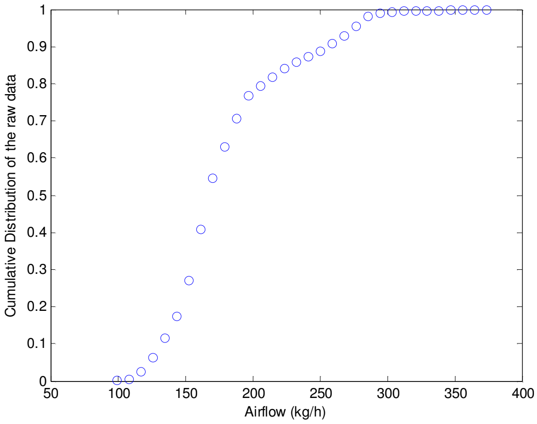
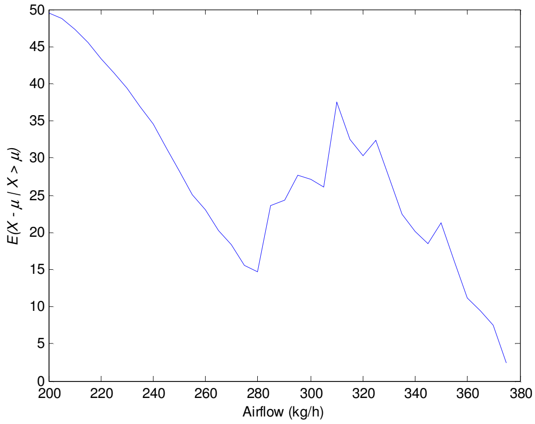


Figure 3 : Mean Excess Function results



For given $q > F(\mu)$, the estimate can be calculated by inverting the tail estimate (1) to get:

$$\hat{x}(q) = \mu + \frac{\hat{\beta}}{\hat{\xi}}\left(\left(\frac{n}{N_u}(1-q)\right)^{-\hat{\xi}} - 1\right) \qquad (2)$$

Another important property of the GPD is the mean excess function over a level μ, for this type of distribution is a linear function of μ given by

$$E(X - \mu \mid X > \mu) = \frac{\beta + \xi\mu}{1-\beta} \qquad (3)$$

4. Process Monitoring & Analysis

An appropriate value of μ, the point at which data is taken from the tail of the distribution has to be chosen. To help with the determination of start of the tail the cumulative distribution of the raw data is plotted in Figure 2 it is then possible to visually inspect the data. From this it can be seen that the tail of the distribution lies in the region of airflow between 250 to 300kg/h. To determine the point of the tail more accurately use can be made of the linearity of the equation (3). Figure 3 shows the results of this equation, the distribution up to an airflow of 280 kg/h gives a linear slope, as indicated by (3), indicating that this distribution can be modelled using a GPD. The airflow data greater than 280 kg/h shows an initial deviation before once again producing a linear relationship. The data for the second distribution is not as smooth as the amount of data included within the calculation is being continually reduced. Making use of the information an airflow of 285 (kg/h) is selected as the value μ this then allows as much data to be included within the estimation of the distribution of the tail.

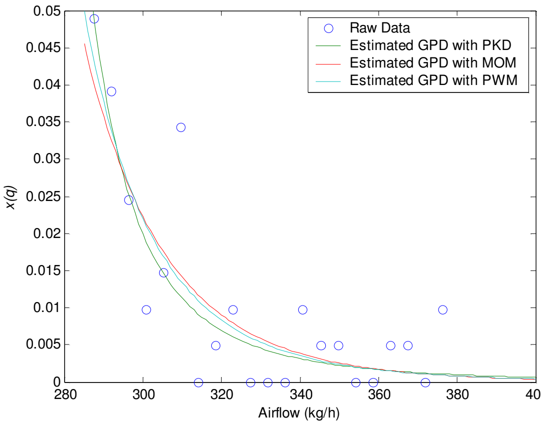
Use has been made in this paper of the WAFO (Wave Analysis for Fatigue and Oceanography) MATLAB toolbox from Centre for Mathematical Sciences Mathematical Statistics at the Lund University (http://www.maths.lth.se/matstat/wafo/index.html). Within the toolbox there are a number of different estimation techniques that are available for the estimation of the parameters of (1). There are three techniques used in the WAFO MATLAB library these are; Pickands' (PKD) estimator, Method of Moments (MOM) and Probability Weighted Moments (PWM) (Pickands 1975, Hoskings 1985, O'Connell 2004). The Pickands estimator gives generally good estimates of the parameter ξ when it is in the following range $-5 \le \xi \le 5$, and $\beta > 0.5$ (WAFO 2000).

Table 1 shows the results obtained by making use of the three different estimation techniques. Figure 4 shows the comparison of these three estimated models results for equation (2) against the raw data. Visually all of responses for the three different estimation techniques appear to give a reasonable fit to the data. The only marked difference between them being that the response for the PKD model results in a steeper initial curve and a flatter final tail. This can be seen in the estimated model results, in Table 1, with PKD having the highest shaping factor ξ. Note: the value of ξ for MOM estimation method results in a value close to zero indicating that the estimated model has an exponential decay.

Table 1: Comparison of Estimation Techniques

Estimation Method	ξ	β
Pickands (PKD)	0.4613	16.2988
Method of Moments (MOM)	0.0673	21.9689
Probability Weighted Moments (PWM)	0.1538	19.9308

Figure 4: Comparison of the different estimation techniques



Using (2) and the estimated GPD parameters allows a calibration diagnostic threshold to be determined based upon different levels of σ (Pyzdek 2009). A value of 4, 5, and 6 σ will result in a normal sensor being detected as a failed sensor 0.62%, 0.023% and 0.00034% of the total distribution of data. The estimated diagnostic threshold $\hat{x}(q)$ for the different levels of σ is given in Table 2

Table 2: Determination of the Diagnostic Threshold

Estimation Method	4σ	5σ	6σ
PKD	315	548	2335
MOM	315	404	550
PWM	314	419	660

All of the three estimation techniques produce a very similar diagnostic threshold result for 4σ, the resulting airflow threshold lying within the range of the existing set of data. For 5σ and 6σ the estimation techniques have to predict beyond the currently available data and as such their results start to vary quite markedly. The threshold predictions for MOM and PWM for 5σ are only just beyond the range of the data and as such have resulted in similar set of result. The values for 6σ vary quite markedly and as such should not be used for setting the diagnostic threshold.

The results in Table 2 will give an indication to the calibrator what level of risk they can expect for setting a particular diagnostic threshold. There is still, however, a level of judgment that has to be made as the estimation methods give different threshold values for the same level of risk or σ level. Any final failure threshold would have to set based upon, not only the risk that the diagnostic will false flag a potentially good sensor, but also that it can reliably detect a failed sensor. The data collected for a failed sensor would need to analysed in a similar way to determine finally what level of separation will be achieved between the good and the failed sensors

5. Further Work

This analysis needs to be repeated across all of the other sections of data in the calibration map for the airflow diagnostic. The process, as it stands, requires a high level of judgement to determine where to set the tail threshold μ and which, if any, of the available estimation techniques is generating appropriate estimates. Further work is required to turn this into process which can be used by calibration engineers to support their work.

CONCLUSION

The process outlined in the paper shows the estimation of Generalised Pareto Distribution model fitted to the tail experimental data of an airflow meter sensor. Using this model the level of σ, and therefore risk, can be defined by the calibration engineer and used to determine the diagnostic failure threshold.

References

De Haan, L., D. W. Janssen, K. G. Koedjik, and C. G. de Vries (1994). "Safety first portfolio selection, extreme value theory and long run asset risks." In J. Galambos, J. Lechner, and E. Simiu (eds.), Extreme Value Theory and Applications.Dordrecht, Netherlands: Kluwer, pp. 471–488.

Embrechts, P., C. Kluppelberg, and T. Mikosch (1997). Modeling Extremal Events for Insurance and Finance. Berlin: Springer.

EOBD DIRECTIVE 98/69/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 October 1998

Fisher, R. A., and L. H. C. Tippett (1928). "Limiting forms of the frequency distribution of the largest or smallest member of a sample." *Proceedings of the Cambridge Philosophical Society*, 24: 180–190.

O'Connell, J.Shao, Q, (2004). Further Investigation on a New Approach in Analysing Extreme Events, CSIRO Mathematical and Information Sciences, Report Number 04/41

OBD-II Title 13, California code regulations, Section 1968

Pickands, J. (1975). Statistical inference using extreme order statistics. Annals of Statistics, 3: 119–131.

Pyzdek, T and. Keller, P (2009). The Six Sigma Handbook, Third Edition. New York, NY: McGraw-Hill. ISBN 0071623388.

McNeil, A. J. (1999). "Extreme value theory for risk managers." In Internal Modeling and CAD II. London: Risk Books, pp. 93–113.

Hosking, J.R.M., Wallis, J.R. and Wood, E.F. (1985). Estimation of the generalized extremevalue distribution by the method of probability-weighted moments. *Technometrics*, **27**, pp. 251-261.

WAFO, (2000). A MATLAB Toolbox for Analysis of Random Waves and Loads, Tutorial, Version 2.0.02, The WAFO Group, Centre for Mathematical Sciences, Mathematic Statistics, Lund University, August 2000,