Peak-Valley Segmentation Algorithm for Fatigue Time Series Data

Z. M. NOPIAH[1], M. I. KHAIRIR[2], S. ABDULLAH AND C. K. E. NIZWAN
Department of Mechanical and Materials Engineering
Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor
MALAYSIA
[1]zmn@vlsi.eng.ukm.my
[2]mihsankk@vlsi.eng.ukm.my

Abstract: - This paper presents the peak-valley (PV) segmentation algorithm for the purpose of producing a reliable method of fatigue time series segmentation and statistical segment-by-segment analysis of fatigue damage. The time series were segmented using a piecewise linear representation (PLR) based segmentation algorithm and consecutively the peak-valley (PV) segmentation algorithm. Statistical analysis and fatigue damage calculations were made on each segment and scatter plots were produced based on the relationship between segmental damage and its corresponding kurtosis value. Observations were made on the scatter plots produced by the PV segmentation algorithm to determine the reliability of the data scattering for fatigue data clustering prospects.

Key-Words: - Time series, segmentation, peak-valley, data scattering, kurtosis, fatigue damage.

1 Introduction

It has been established over the years that proper evaluation of statistical properties will give reasonable diagnostic indication of damage in critical automotive components [1]. Although there are a large number of such statistical attributes such as root mean square value, crest factor, skewness, kurtosis, and so on, kurtosis has emerged as one of the good indicators of damage of automotive components such as gears.

This paper discusses on the segmentation of fatigue data (represented as time series), the statistical evaluation of each segment of the data, and the resulting data scatter. A peak-valley (PV) segmentation algorithm was introduced in order to produce reliable data scatters. Scatter plots of PV segmented data were made to see if the method resulted in reliable data scatters based on the statistical property of segmental kurtosis. It is hypothesized that by using PV segmentation on the fatigue time series data, the resulting data scatter produced would be more reliable and suitable for fatigue data editing.

2 Literature Background

2.1 Time series behavior

Since all the data that were measured from this experiment are recorded as strain time histories, it is also important to have a better understanding about the behavior of this data before applying the detection of abrupt changes algorithm. The identification of fatigue data behavior is based on the existence of time series component which involves with the identification of trend, cyclical, seasonal and irregular component in time period *t.* The method used in the identification of this component is called the classical decomposition of time series. This process is used to segregate and to analyse the existence components in a systematic manner. The trend component represents the long-run growth or decline over time. On the other hand, the cyclical component refers to the rises and falls of the series over unspecified period of time. The seasonal component also known as seasonal variation refers to the characterization of regular fluctuations occurring within a specific period of time. Although this component is individually identified, in fact it also related to each other in a certain mathematical functional form. The type of relationship that these components have is divided into two which are multiplicative effect and additive effect. Multiplicative effect means that the components are interacted to each other such that the sizes of the seasonal variation increase in accordance with in the level of data. On the other part, additive effect involve with the assumption that the components of the series are interacted in additive manner [8].

2.2 Segmentation

For the purpose of this study, a time series segmentation algorithm that inputs a time series and returns a piecewise linear representation (PLR) was used for the initial segmentation of the time series data. Based on the studies by Keogh et al. [2], a segmentation algorithm that has a global perspective of the data produces the best PLR with the least amount of error. Such algorithms are called batch algorithms, and of the two segmentation methods that fall under this category, Bottom-up

segmentation algorithm has proven to be the best at performing batch segmentation with the least amount of error [2].

2.2.1 Piecewise Linear Representation

By definition, a piecewise linear representation (PLR) refers to the approximation of a time series *T*, of length *n*, with *K* straight lines [2]. The Bottom-up algorithm first creates the finest approximation of the data, which contains at most *n*/2 segments. Then it recursively calculates the cost of merging each pair of adjacent segments and proceeds to merge the segments beginning with the lowest cost pair. The number of segments in the PLR will gradually be reduced until a stopping criterion is met.

Since producing a PLR requires approximation of the time series using straight lines, linear regression was used as the approximation method for the Bottom-up algorithm. This is because the approximating lines produced using the linear regression approach is superior to the method of linear interpolation in terms of Euclidean distance [2].

2.2.2 Peaks and Valleys

A PLR with many non-parallel lines contains a significant number of local optima, which can be either classified as peaks or valleys. A peak is defined to be associated with change in the slope from positive to negative, while a valley is associated with a change in the slope from negative to positive [3]. Peaks in a PLR are essentially the local maxima and valleys are the local minima. Depending on the resulting PLR, some points can be classified as neither peaks nor valleys.

Peak-valley (PV) identification can be used to segment signals so that each segment may contain certain numbers of peaks and/or valleys, according to the needs of the study. This is particularly useful for fatigue time series data, since peaks and valleys feature predominantly in rainflow counting algorithms for fatigue damage calculations [4].

2.3 Kurtosis

In real applications, mechanical signals can be classified as having a stationary or a non-stationary behaviour. Stationary signals exhibit the statistical properties remain unchanged with the changes in time and the statistics of non-stationary signal is dependent on the time of measurement [5]. The most commonly used statistical parameters are the mean value, the root-mean-square (r.m.s.) value and the kurtosis [6].

The r.m.s. value, which is the 2nd statistical moment, is used to quantify the overall energy content of the signal and is defined by the following equation:

$$r.m.s = \left\{ \frac{1}{n} \sum_{j=1}^{n} x_j^2 \right\}^{1/2} \quad (1)$$

where x_j is the j^{th} data and n is the number of data in the signal.

The kurtosis, which is the signal's 4th statistical moment, is a global signal statistic which is highly sensitive to the spikiness of the data. It is defined by the following equation:

$$K = \frac{1}{n(r.m.s)^4} \sum_{j=1}^{n} (x_j - \bar{x})^4 \quad (2)$$

where *r.m.s* is the root mean square as calculated in Equation 1 and $\bar{x}$ is the mean value of the signal data.

For a Gaussian distribution the kurtosis value is approximately 3.0. In some definitions of kurtosis, a deduction of 3 is added to the definition in order to maintain the kurtosis of a Gaussian distribution to be equal to zero. For clarity and convenience, in this study the former definition of kurtosis (where the Gaussian distribution has a kurtosis value of 3) was used since the kurtosis function in MATLAB® uses this definition. Therefore kurtosis values which are higher than 3.0 indicate the presence of more extreme values than should be found in a Gaussian distribution. Kurtosis is used in engineering for detection of fault symptoms because of its sensitivity to high amplitude events [7].

2.4 Fatigue damage

It is common that the service loadings caused by machines and vehicles is evaluated using a strain-life fatigue damage approach [3]. The strain-life approach considers the plastic deformation that occurs in the localised region where fatigue cracks begin under the influence of a mean stress.

The total strain amplitude, ε_a is produced by the combination of elastic and plastic amplitudes

$$\varepsilon_a = \varepsilon_{ea} + \varepsilon_{pa} \quad (3)$$

where ε_{ea} is the elastic strain amplitude and ε_{pa} is the plastic strain amplitude. The elastic strain amplitude is defined by

$$\varepsilon_{ea} = \frac{\sigma_a}{E} = \frac{\sigma'_f}{E}\left(2N_f\right)^b \quad (4)$$

while the plastic strain amplitude is given as

$$\varepsilon_{pa} = \varepsilon'_f \left(2N_f\right)^c \quad (5)$$

where σ_a is the stress amplitude, N_f is the number of cycles to failure, σ'_f is the fatigue strength coefficient, b is the fatigue strength exponent, ε'_f is the fatigue

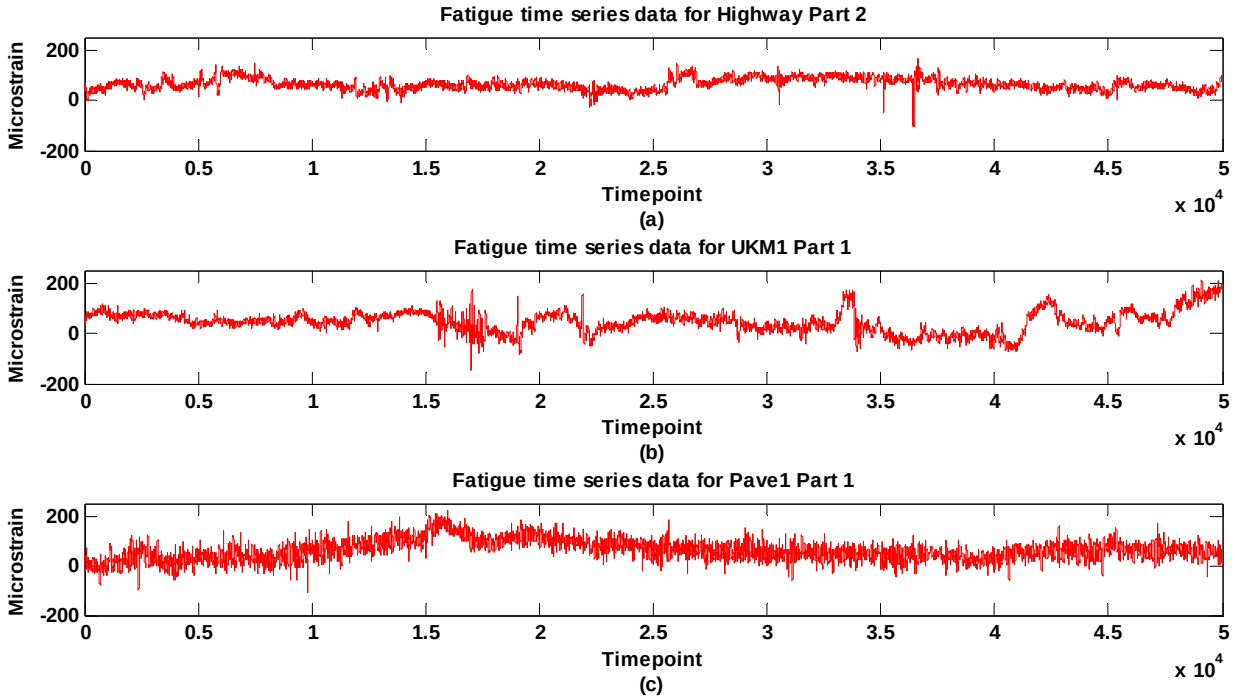


Fig. 1: Fatigue time series data for (a) highway road, (b) in-campus road, (c) pavé road

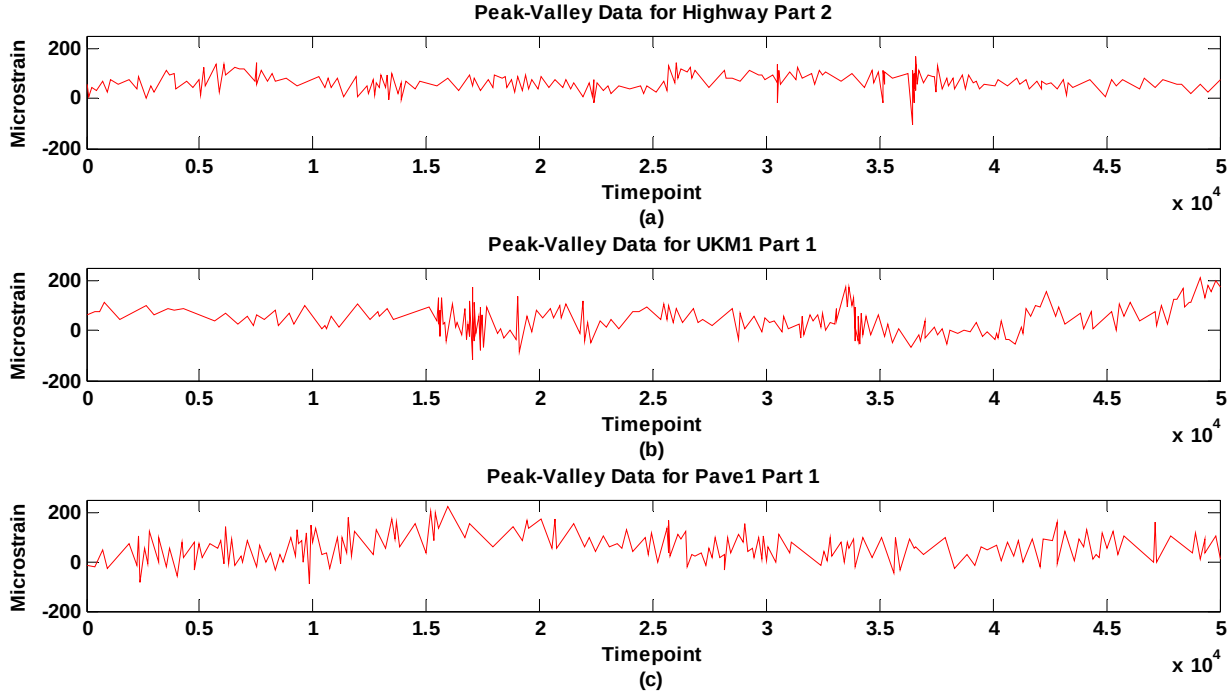


Fig. 2: Peak-Valley data for (a) highway road, (b) in-campus road, (c) pavé road

ductility coefficient, c is the fatigue ductility component and E is the modulus of elasticity.

Combining Equations 4 and 5 gives the Coffin-Manson relationship, which is mathematically defined as

$$\varepsilon_a = \frac{\sigma'_f}{E}(2N_f)^b + \varepsilon'_f(2N_f)^c \quad (6)$$

which is essentially Equation 3 above and is the foundation of the strain-life approach.

Some realistic service loads involve nonzero mean stresses. One common mean stress effect model is the Smith-Watson-Topper (SWT) strain-life model, which is defined by

$$\sigma_{\max}\varepsilon_a = \frac{\sigma'^2_f}{E}(2N_f)^{2b} + \sigma'_f\varepsilon'_f(2N_f)^{b+c} \quad (7)$$

and the damage parameter is taken to be the product of the maximum stress and the strain amplitude of a cycle. In our study the strain-life approach and the Smith-Watson-Topper strain-life model for mean stress effect were used in all fatigue damage calculations. Fatigue damage is derived from the number of cycles to failure where the relationship is

$$Damage = \frac{1}{N_f} \quad (8)$$

and therefore fatigue damage have values in the range (0, 1] where zero denotes no damage (extremely high or infinite number of cycles to failure) and 1 means total failure (one cycle to failure).

2.1 Methodologies

2.1.1 Data acquisition

The fatigue data for this study was obtained from field tests conducted on the lower suspension arm of a mid-sized sedan car using strain gauges and data logging instrumentation. The fatigue data were measured on the car's front lower suspension arm as it was subjected to a variety of road load services. All data were recorded as strain time histories and Figure 3 shows the fatigue data measurement set-up that was used during the tests. The strain value was measured using a strain gauge that was connected to a SoMat eDAQ® data logger for data acquisition. Experimental parameters such as the sampling frequency and type of output data being measured were specified in TCE eDAQ V3.9.0 software.

The material for the lower suspension arm is SAE1045 steel, and this material's specifications were used in all fatigue damage calculations. The road load conditions were from a stretch of highway road to represent mostly consistent load features (Figure 1a), an in-campus road to represent load features that might include turning and braking, rough road surfaces and speed bumps (Figure 1b), and a stretch of brick-paved (pavé) road to represent noisy but mostly consistent load features (Figure 1c). All data was recorded at a constant sampling rate of 500 Hz. Each set of data contains 50000 discrete points, giving a total signal length of 100 seconds per signal.

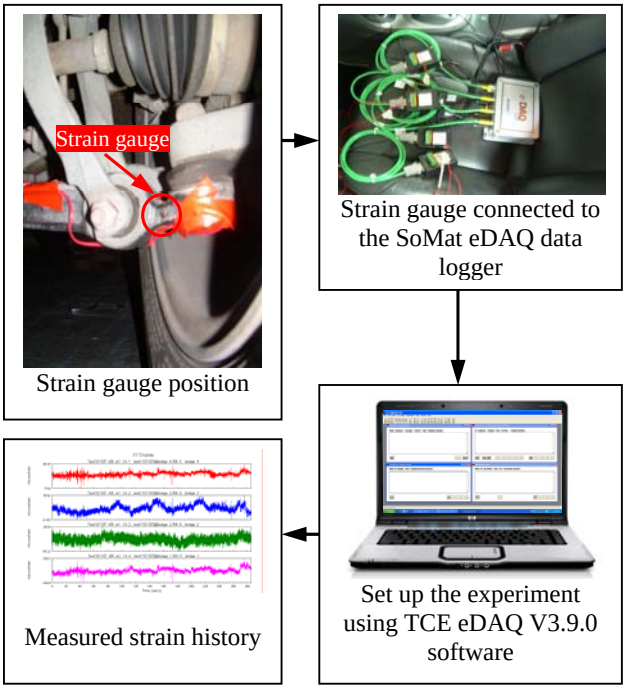


Fig. 3: Diagrammatic flow chart for fatigue data collection process

2.1.2 Statistical analysis

For the purpose of this study, kurtosis, mean, and root mean square were chosen as the global statistical parameters evaluated for each signal. The global statistical values for each signal are presented in Table 1.

Table 1: Global statistical parameters for the measured fatigue data

Data	Kurtosis	Mean	R.M.S.
Highway	3.41	63.73	67.72
Campus	3.70	45.91	62.78
Pave	3.48	61.37	73.00

We can see that the global kurtosis values of the signals are quite close to 3.0 which is the kurtosis of a Gaussian distribution. To further clarify this evaluation, normal distribution fitting was done on all three sets of signal data.

We can see from Figure 4 that the distribution for all three sets of data roughly approach that of a Gaussian distribution. These distributions mirror the

characteristics of the averaged sum of a large number of independent random variables, as explained by the Central Limit Theorem. This supports the earlier assumption that the fatigue loads in the signals are independent and random in nature.

2.1.2 Time series behavior analysis

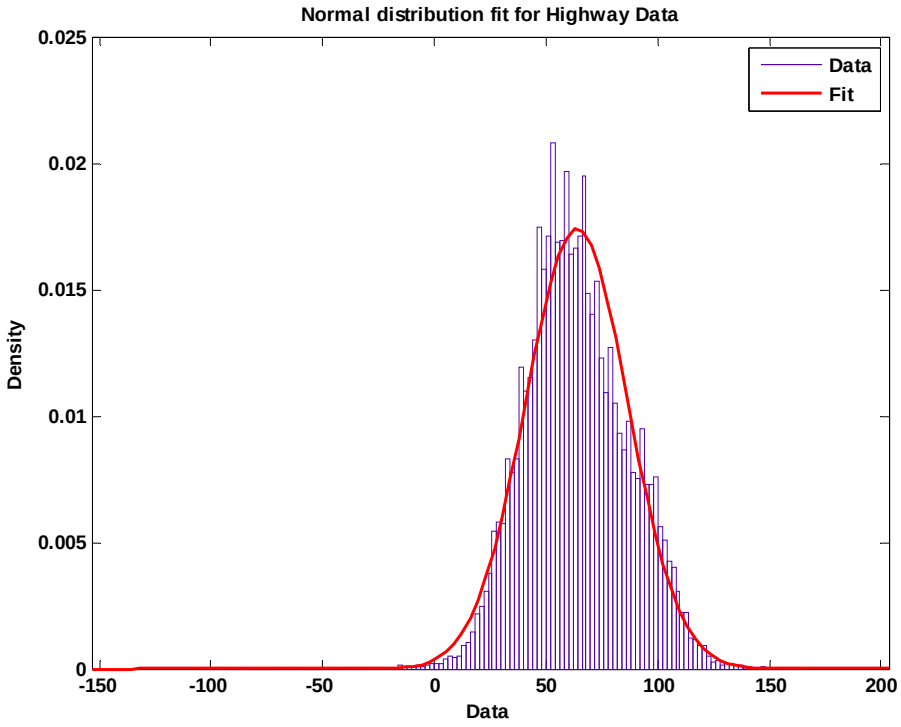
In order to further understand the characteristics of the fatigue time series data, time series behavior analysis was performed based on the classical decomposition of time series.

In the classical decomposition of time series, a few identification methods were selected for all time series components. The methods used in detecting the existence of trend, cyclical, seasonal and irregular behaviors were the linear trend line, the residual method, and the method of seasonal differencing.

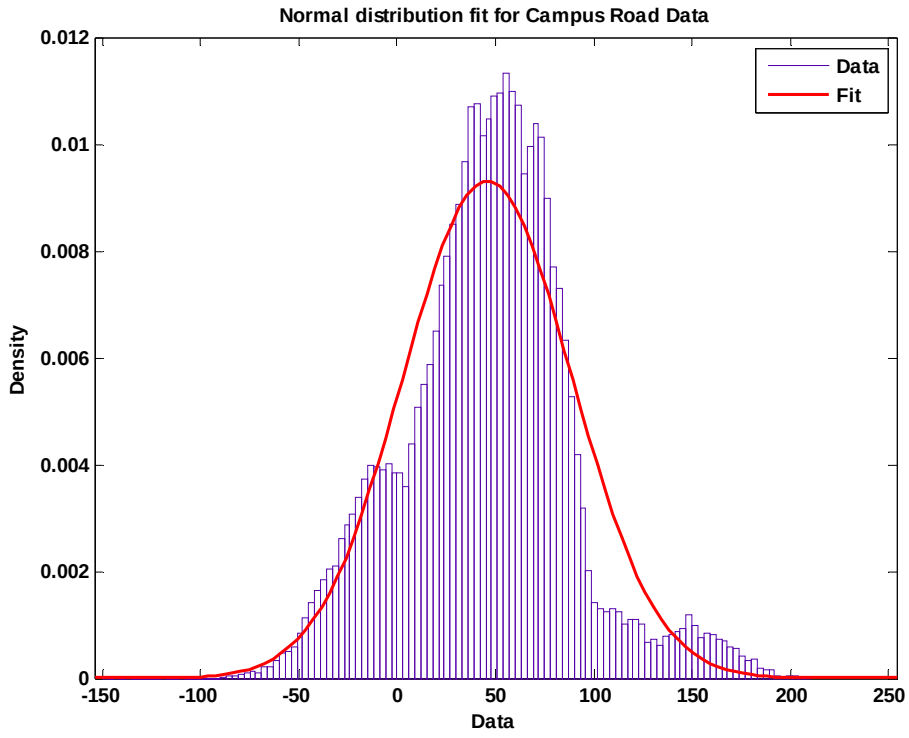
Table 2: Summary of the fatigue time series behavior

Data	Trend	Cyclical	Irregular
Highway	Positive	Yes	Random
Campus	Positive	Yes	Random
Pavé	Negative	Yes	Random

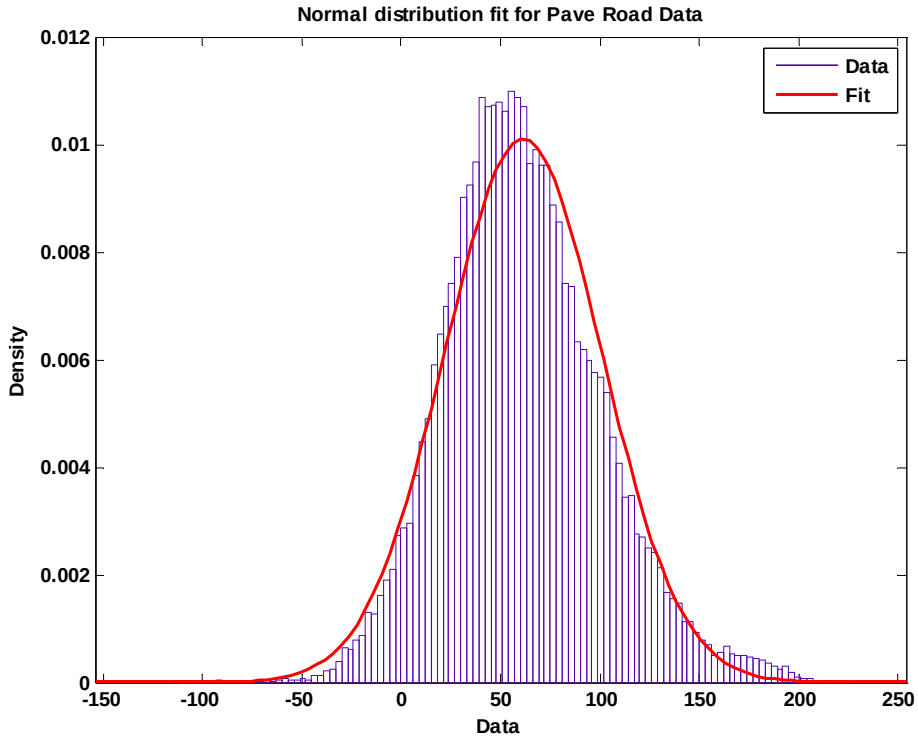
Table 2 shows that all three sets of data have similar time series component behavior except for the Pavé (brick-paved road) data which has a negative trend. In this time series behavior analysis, the seasonal component behavior was not considered since events that occur in the fatigue data were determined to be random and independent and therefore not regular fluctuations occurring within specific periods of time.



(a)



(b)



(c)

Fig. 4: Normal distribution fit for (a) Highway data, (b) Campus road data, (c) Pavé road data

2.1.3 Peak-Valley Segmentation

Because the Bottom-Up segmentation method produces the best PLR with the least amount of error, for the purpose of this study, the Bottom-Up segmentation algorithm which was developed by Keogh et al. [2] was used to segmentise the time series signals. As the algorithm was run, the number of segments in the PLR will gradually be reduced until a stopping criterion is met. The stopping criterion for the algorithm was set to be the number of segments in the resulting PLR, which for the purpose of simplicity and statistical acceptability, was decided to be 300 segments. This procedure is important so that we can simplify the original signal into a PLR with a workable number of critical points. Peak-valley identification was then used to classify these points on the PLR into peaks and valleys. Datapoints and timepoints on the PLR are grouped into sets and a set of ordered pairs is obtained using the Cartesian product of two sets.

$$X = \{x_j : j = 1,2,...,n; x_j \neq x_{j+1}\} \quad (9)$$

$$T = \{t_j : j = 1,2,...,n\} \quad (10)$$

$$D = X \times T \quad (11)$$

```
algorithm varargout = peakvalley(c, T)

%% initialize first point as peak or valley
if (c(1)>c(2))
   peak(1)=c(1);
   else
   valley(1)=c(1);
end
%% assign peaks & valleys and their timepoints
for n=1:(length(c)-2)
   if (c(n+1)>max(c(n),c(n+2)))
      peak(i)=c(n+1); tpeak(i)=T(n+1);
   elseif (c(n+1)<min(c(n),c(n+2)))
      valley(j)=c(n+1); tvalley(j)=T(n+1);
   else % ignore if not peak or valley
   end
end
%% last point designation as peak or valley
if c(length(c))>c(length(c)-1)
   if tpeak(i-1)<tvalley(j-1)
      peak(i)=c(length(c));
      tpeak(i)=T(length(c));
   else
   end
else
   if tpeak(i-1)>tvalley(j-1)
   valley(j)=c(length(c));
   tvalley(i)=T(length(c));
   else
   end
end
```

Fig. 5: The peak-valley identification algorithm

The ordered set X contains elements that denote the datapoints of the PLR, the ordered set T contains the corresponding timepoints, n is the number of points and D is the ordered pair obtained by pairing each datapoint in X with a timepoint in T.

$$P = \{x_j : x_{j-1} < x_j > x_{j+1}\} \cup \{x_1 : x_1 > x_2\} \cup \{x_n : x_n > x_{n-1}\} \quad (12)$$

$$V = \{x_j : x_{j-1} > x_j < x_{j+1}\} \cup \{x_1 : x_1 < x_2\} \cup \{x_n : x_n < x_{n-1}\} \quad (13)$$

$$P \cap V = \phi \quad (14)$$

$$(P \cup V) \subset X \quad (15)$$

The ordered set P contains elements of X that are classified as peaks in the signal, V is an ordered set of datapoints that are classified as valleys, and P and V are non-intersecting and their union is a proper subset of X.

As in Equation 12, a datapoint x_j is classified as a peak if it is strictly greater than x_{j-1} (the datapoint before it) and x_{j+1} (the datapoint after it); conversely Equation 13 states that a datapoint is classified as a valley if it is strictly smaller than the points before and after it. These conditions are relaxed for the endpoints of the dataset where a comparison of only two adjacent datapoints are needed. Additionally there may be points that are not classified as neither peaks nor valleys because they do not fit the conditions for both sets. These datapoints are therefore ignored and not included in the peak-valley algorithm.

The timepoints that correspond to valley datapoints are then identified

$$T_V = \{t_j : x_j \in V\} \quad (16)$$

and these timepoints were then used as segmentation points for the purpose of segmenting the original data (Figure 1). The algorithm for the peak-valley identification process is shown in Figure 5 and the resulting peak-valley data can be seen in Figure 2. A summary of the complete peak-valley segmentation procedure is represented as a flowchart in Figure 6.

The PV segmented data was then analysed using the GlyphWorks® software package, where the fatigue damage for each segment of the time series was calculated. The segmented data was also run through a MATLAB® algorithm that calculates the kurtosis values of each segment. Another MATLAB® algorithm generates comparison scatter plots of fatigue damage against segmental kurtosis values. Based on these

scatter plots, patterns of data scattering, if any, were identified and noted.

3 Results and Discussion

By introducing PV identification in the segmentation process, segmental kurtosis analysis can be made accurately since every segment contains one overall peak. It is reasonable and practical to perform kurtosis analysis on data segments that each contains an overall peak so that the kurtosis measurement is a better representative of the segmental peakedness of the time series.

The resulting data scatter from this segmentation method also better fits the characteristics expected of a kurtosis versus fatigue damage scatter plot. Kurtosis shows the presence of significantly high amplitudes or peaks in each segment, which supposedly translates into a higher fatigue damage value for the particular segment. Therefore it is expected that scatter plots of kurtosis versus fatigue damage should reflect this trend in some manner. In Figures 7a, 7b and 7c we can see the high kurtosis points are only present in the high fatigue damage range. This means that these scatter plots truthfully reflect the hypothesized relationship between kurtosis and fatigue damage.

Intuitively the range of kurtosis values for the highway data should generally be lower than the kurtosis values for the in-campus road data. This is because the highway stretch is generally straight and the road surface is mostly consistent, and therefore the service loads on the lower arm of the car's suspension are generally low and consistent (see Figures 1a and 1b). On the other hand, the in-campus road consists of speed bumps, rough road surfaces, curves and intersections (which prompted for turning and braking) and so it is expected that some segments of the in-campus road load data have more significant peaks than the highway data. In Figures 7a and 7b we can see that the maximum kurtosis value for the highway data is smaller than the in-campus road data, which is consistent with what is expected for both signals. A scatter plot for the pavé road data (Figure 7c) is also included here to show that the scatter plot is also reliable for input signals that are noisy but mostly consistent, in this case, the road load data from the brick-paved (pavé) road (Figure 1c).

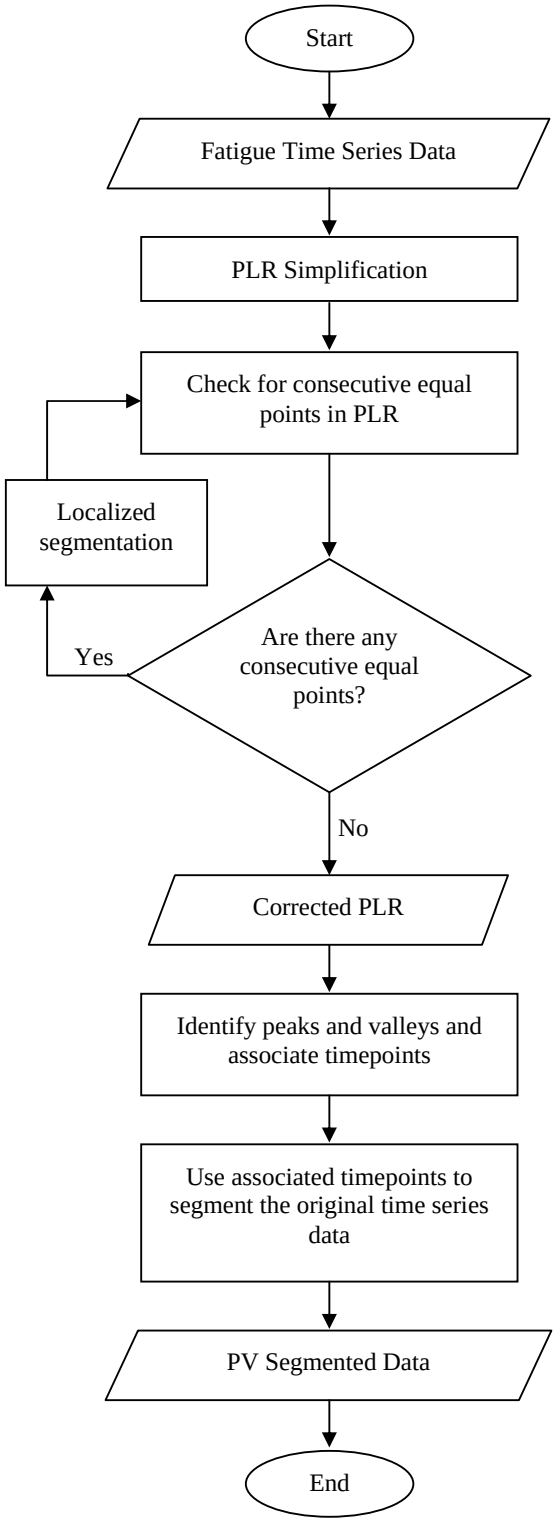
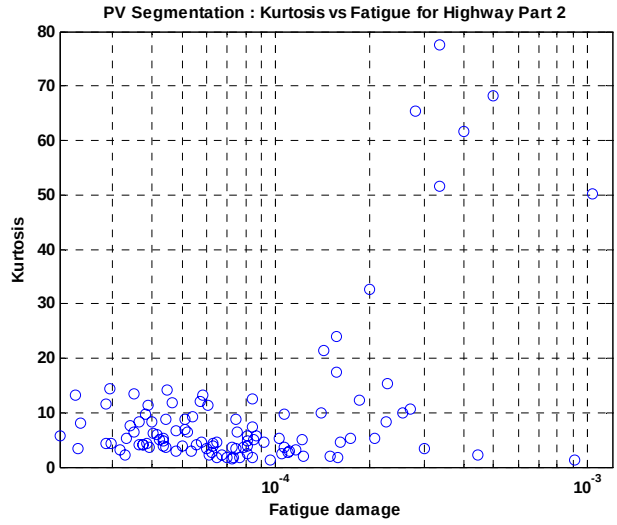
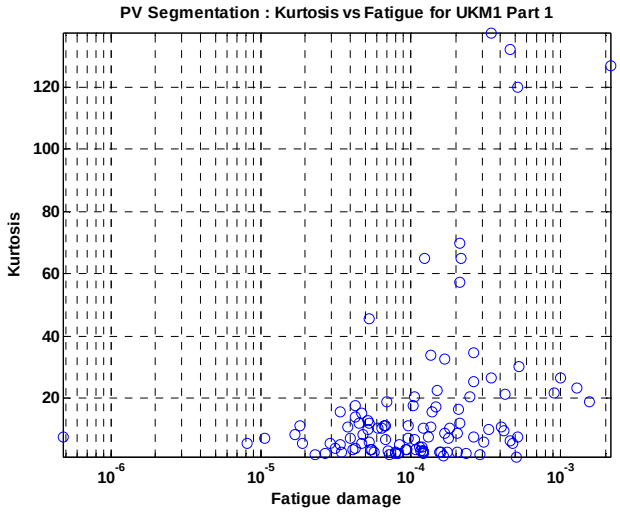


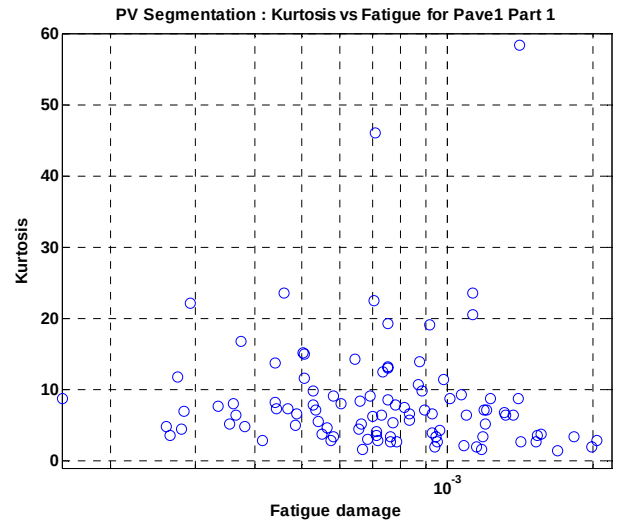
Fig. 6: Flowchart of the Peak-Valley segmentation algorithm



(a)



(b)



(c)

Fig. 7: Scatter plots of kurtosis vs fatigue damage for (a) highway road, (b) in-campus road, (c) pavé road

For further research work, scatter plots of segmental kurtosis versus segmental damage as shown in Figure 4 can be utilized for fatigue data classification and clustering. The utilization of PV segmentation has resulted in the production of reliable data scatters for the described purpose. As seen in Figure 7, significant clusters of data can be visually identified in all of the scatter plots, which opened up the prospect of using data clustering algorithms on the data scatters for future analysis and research works.

4 Conclusion

The study has demonstrated the use of Peak-Valley segmentation of time series data for fatigue analysis. Combining time series segmentation with statistical analysis has produced reliable results. By analysing the data this way, we may identify trends and patterns of data scattering based on critical statistical parameters. From the scattering of data we may acknowledge which parts of the data made significant contribution and which did not. Finally based on our findings we may eliminate or exclude certain parts of the data in order to make further study and analysis of the signal much faster and more efficient without significant loss of data.

As our main focus in this study, we suggested that the implementation of the PV segmentation algorithm will produce significantly reliable scatter plots for fatigue data clustering prospects. Finally, as a possible future work, after identifying and clustering the data in the signal, fatigue data editing through the elimination of certain non-contributory or insignificant segments of the signal may help in reducing the length and complexity of the data and may thus speed up the process of fatigue testing of metal components of mechanical systems or any similar application.

5 Acknowledgements

The authors would like to express their gratitude to Universiti Kebangsaan Malaysia and Ministry of Science, Technology and Innovation, through the fund of UKM-GUP-BTT-07-25-152, for supporting these research activities.

References:

[1] Rao, V. B., Kurtosis as a Metric in the Assessment of Gear Damage: The Shock and Vibration Digest, Vol. 31, No. 6, 1999, pp. 443-448

[2] Keogh, E., S. Chu, D. Hart and M. Pazzani, An Online Algorithm for Segmenting Time Series: Data Mining. ICDM 2001, *Proceedings IEEE International Conference on 29 Nov - 2 Dec 2001*, pp. 289-296.

[3] Abdullah, S., Choi, J.C., Giacomin, J.A., and Yates, J.R., Bump Extraction Algorithm for Variable Amplitude Loading, *International Journal of Fatigue*, Vol 28, 2005, pp. 675-691.

[4] Xiong, J.J. and Shenoi, R.A., A Load History Generation Approach for Full-scale Accelerated Fatigue Tests, *Engineering Fracture Mechanics*, Vol. 75, 2008, pp. 3226-3243.

[5] Bendat, J. S. and Piersol, A. G., *Random Data: Analysis and Measurement Procedures*, 2nd Edition, Wiley-Interscience, New York. 1986.

[6] Hinton, P. R., *Statistics Explained: A Guide for Social Science Students*, Routledge, London. 1995.

[7] Qu, L. and He, Z., *Mechanical Diagnostics,* Shanghai Science and Technology Press, Shanghai, P. R. China. 1986.

[8] Lazim, M. A, Introductory Business Forecasting: A Practical Approach, 2nd edition, 2007,University Publication Centre (UPENA), UiTM 2005, 2007