SIMPLE STATISTICS TO MAKE PREDICTIONS OF THE NUMBER AND DURATION OF EXCEEDANCES OF AIR QUALITY THRESHOLDS

G Mahoney, Sefton MBC and the Dept. of Civil Engng., University of Liverpool, (mahoney@garymahoney.ukfreedom.com)
R G Tickell, Dept. of Civil Engng., University of Liverpool, (r.g.tickell@liv.ac.uk)

ABSTRACT
A simple model, proposed by Guigliano et. al. [1] is tested against air pollution data as part of an investigation into the frequency and duration of exceedances of concentration thresholds. Data from the Merseyside region in the NW of England appears to follow the model and encourages the extension of the approach to other applications and sites.

INTRODUCTION
There are many factors that determine how air pollution affects human health; two of which are the duration of the incident and the peak value that pollution levels reach during the incident. The duration is the time that pollution levels remain above a selected threshold and the peak value is the maximum value reached during this time. Figure 1 illustrates these concepts and shows the practical issue of defining exceedances. Duration 1 is reasonably obvious. However small changes in threshold level would change what is recorded as Duration 2. High levels of threshold tend to simplify the issue but reduce the sample size of observations. In the same way, time series of running averages will be smoothed and the multiple peaks are attenuated.

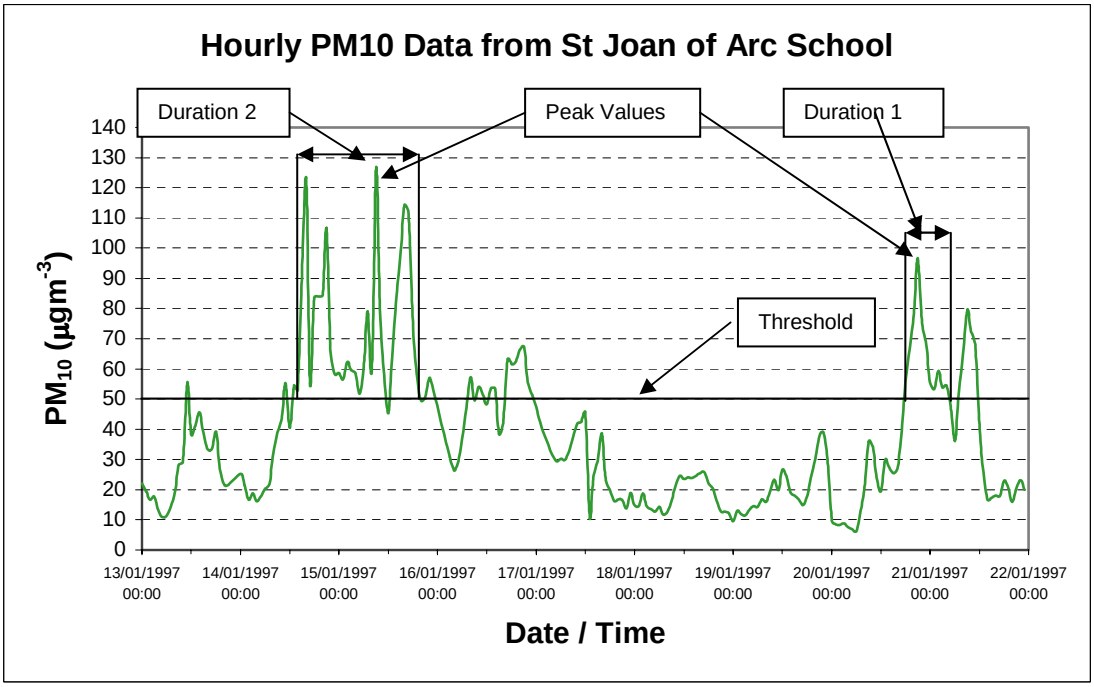

Fig.1 Hourly PM_{10} Data - Duration and Peak Values.

Empirical distributions can be fitted to the data for extrapolation to more extreme events, where large volumes of data have been monitored.
There is theory which models peaks and durations, usually subject to a series of restrictive assumptions of Normality, Independence of Events etc [2]. But where

these assumptions breakdown, as with the current time series, there is a need for an approach to make predictions of duration and extent of exceedances, based on the simple statistics that are readily available.
Sefton MBC collects air pollution data to the same standards and using similar equipment as the UK Government's Automatic Urban and Rural Network.

DURATION AND PEAK VALUES OF EVENTS

Fine particulate matter PM_{10} and ozone are the only air pollutants for which there are breaches of the National Air Quality Strategy (NAQS) Standard in Sefton. It was decided that the PM_{10} time series was more appropriate for study in the first instance.
The NAQS Standard for PM_{10} was 50 μgm^{-3} as a 24hr running average. The time series chosen was 24 hour running average PM_{10} levels measured throughout 1997 at St. Joan of Arc School, Bootle. This station is 50m from a heavily trafficked road, on the other side of which lies the Port of Liverpool. The histogram of the distribution of these is shown in Figure 2.

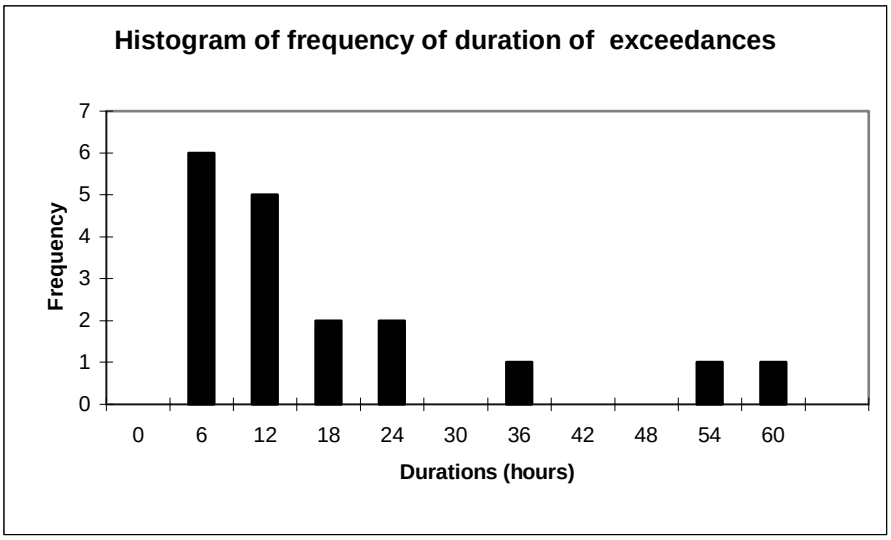

Fig. 2 Duration of Exceedances - St. Joan of Arc School (PM_{10} – 1997 – Threshold level 50 μgm^{-3})

Weibull (3 parameter) and Log Normal distributions were fitted to the data. The Kolmogrov- Smirnov test was used to assess the goodness of fit of each distribution. Both distributions satisfied the KS test at 99.99% significance giving very similar results. [KS value when alpha = 0.01and n = 18 is 0.37. The largest modular difference between calculated and observed exceedance was: 0.138 for Weibull and 0.164 for Log Normal]. The Weibull distribution was chosen for ease of computation and flexibility, in the absence of any significant difference in the goodness of fit.
The largest peak values occurring in each period of exceedance of the threshold were then determined and a Weibull distribution fitted, as for the durations.
Purely from an empirical point of view it is tempting to extrapolate fitted distributions to predict longer durations or higher peaks with a smaller probability of occurrence. However this always carries the dangers of extrapolating beyond observed events (e.g. physical truncation or mixed distributions) and it is little use at sites of interest where no data is available. The alternative is to seek for a reasonably robust method based on prediction of Weibull parameters and other characteristics of the site; hence the interest in investigating the procedure suggested by Guigliano et. al. [1].

PREDICTION OF EVENTS USING SIMPLE CROSSING STATISTICS

Giugliano's work, in 1998, used simple statistics such as the annual mean value, to make predictions of the number and duration of exceedances of air quality standards. Because the key statistic in Guigliano's work is the annual mean value (Ma), the number and duration of exceedances cannot be reported on less than an annual basis. The use of seasonal means has not been considered in the present discussion, though the extension is obvious.

Sefton's 1997 data for CO, NO_2 and PM_{10} was used initially to develop the functions, which were then tested against data from 1998. The NO_2 and CO data was in the form of hourly averages, while the PM_{10} data was in the form of hourly values of a 24 hour running average.

Guigliano suggests the following empirical description of the number of events:

$$Nd = \alpha.Nc.R^{\beta}.\exp(-\gamma.R) \qquad (1)$$

where Nd is the number of events above the threshold, Nc is the total number of values in the time series, R is the ratio of the threshold to the annual mean value and α, β and γ are constants. Guigliano's notation is retained and R should not be confused with the coefficient of determination (R^2) in subsequent regression analysis. Eqn. 1 can be rearranged into a straight-line form:

$$\left[Ln(Nd) - Ln(\alpha.Nc)\right]/R = \beta.Ln(R)/R - \gamma \qquad (2)$$

Using Eqn. 2, the constants α, β and γ were determined for NO_2, PM_{10} and CO, measured at the St Joan of Arc monitoring station in 1997, by iterative line fitting. An example of this is shown in Figure 3.

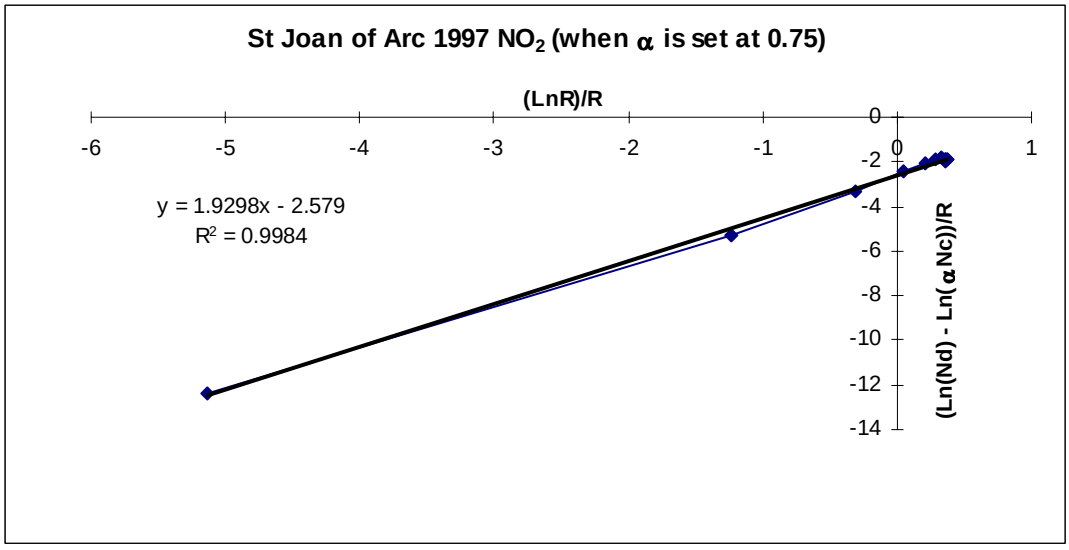

Fig. 3 Example of Determining α, β and γ Constants for Eqn. 2.

The concentration of points between 0 and 1 with only a small number of points on the negative side of the graph does give some cause for concern as to the robustness of the equation derived. However, for each pollutant, the R^2 value obtained was high (>0.9) and the agreement between the observed and the calculated number of events was good, as shown in Figure 4.

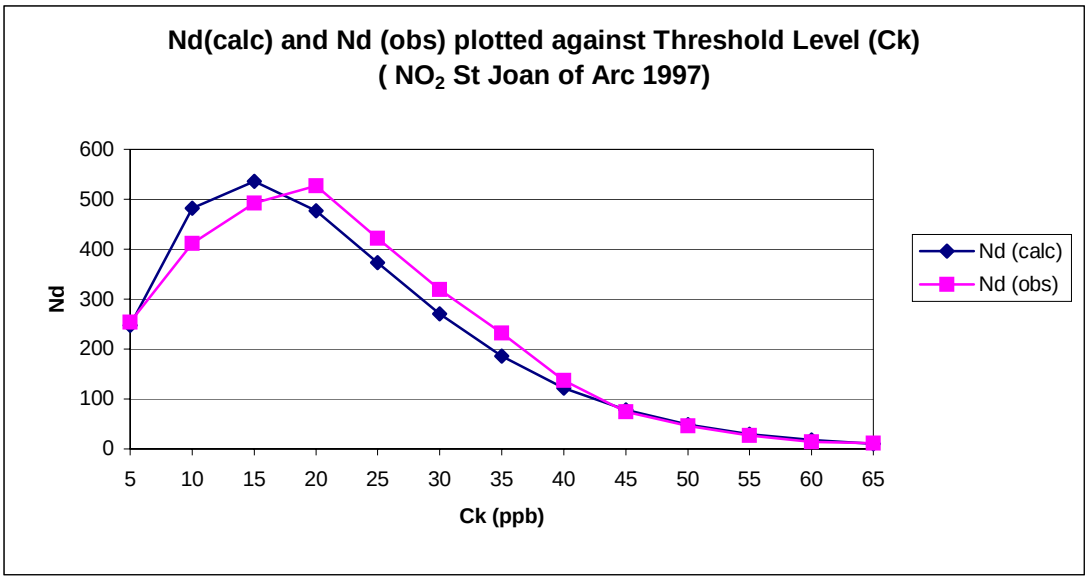

Fig. 4 Observed and Calculated Exceedances (Nd) v Thresholds (Ck).

For the technique to be of real use to air quality practitioners, it would be desirable to predict the number of exceedances at sites other than those where the measurements were made to determine the α, β and γ coefficients or in future years at the same site. To examine the effectiveness of the technique when used in this way, coefficients for St Joan of Arc (1997) were used to calculate threshold exceedances at a nearby monitoring station (Hugh Baird) in 1997 and for the St Joan of Arc site in 1998. Applying the coefficients to data from Hugh Baird gave good comparisons between calculated and observed durations, as can be seen in Figure 5. The results when using coefficients from 1997 with 1998 St Joan of Arc data were less satisfactory tending to underestimate the number of events.

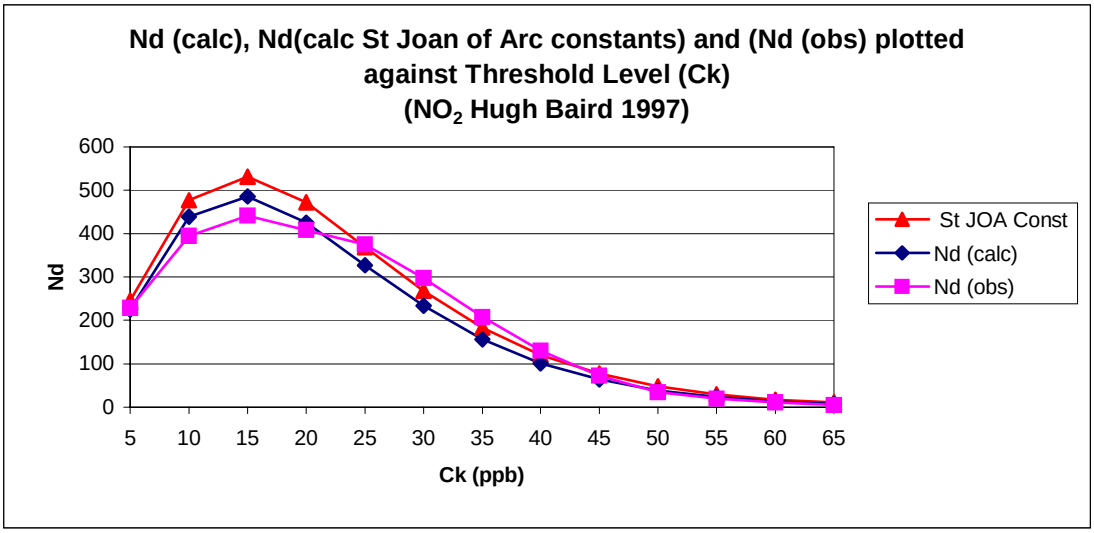

Fig. 5 Number of Exceedances (Nd): Observed, Calculated and Calculated (Using Constants from St. Joan of Arc) v Thresholds (Ck) for Hugh Baird data, NO_2, 1997.

Because of the concerns expressed over the clustering of data points used to determine the β and γ constants, another method of determining these constants using

multiple linear regression was explored. This required Eqn. 1 to be rearranged into the form:

$$Ln(Nd/Nc) = Ln(\alpha) + \beta.Ln(R) - \gamma.R \quad (3)$$

For each pollutant, Minitab regression analysis was used to express Ln(Nd/Nc) in terms of LnR and R with α, β and γ being determined from the constants obtained.

Plots of Nd v Ck were produced in the same way and with much the same degree of agreement, indicating that the first procedure for fixing α, β and γ was viable.

DURATION STATISTICS

Having looked at the number of exceedances (Nd), the next stage is to consider the duration statistics. Guigliano et. al. proposed a 2 parameter Weibull distribution for the exceedance duration (d):

$$P(d) = 1 - \exp\left[-\left(\frac{d}{\sigma}\right)^{\lambda}\right] \quad (4)$$

In fitting observed durations, Guigliano's work has a different notation from that used in this study, in which σ is B and λ is identical to F. For flexibility, and because it was found to give a better fit, the present study moved to a 3 parameter Weibull for which:

$$P(d) = 1 - \exp\left[-\left(\frac{d-A}{B}\right)^{F}\right] \quad (5)$$

The durations are ranked in the usual way, such that:

$$P(d_k) = \frac{k}{N+1} \quad (6)$$

Eqns. 5 and 6 can be combined and rearranged such that:

$$Ln\left(-Ln\left(1-\frac{k}{N+1}\right)\right) = F.Ln(d_k - A) - F.Ln(B) \quad (7)$$

allowing B and F to be determined from a LSE straight line plot, iterating on A.

The Weibull equation was then rearranged into the form:

$$d = \left[B.\sqrt[F]{Ln\left(\frac{1}{1-P(d)}\right)}\right] + A \quad (8)$$

Then by substituting the values of A, B, F and the rank probabilities (Eqn 6) the durations of incidents can be back calculated and compared against the observed durations in order to check for trends, wild points etc.

The durations of exceedances of one threshold for each pollutant were examined. The threshold chosen for PM_{10} was 50 μgm^{-3}. 40ppb was chosen for NO_2 because this

represented a significant rise above the normal values, which would be indicative of higher than usual levels of NO_2, but a level that would give a workable number of exceedances. Similarly 0.7 ppm was chosen as the threshold for CO to provide a workable number of exceedances.
The comparison of observed v back calculated durations gave very good results for nitrogen dioxide. In the case of CO and PM_{10}, results were acceptable but there were outlying values at higher durations. The plot comparing observed to calculated durations obtained for PM_{10} is shown as an example.

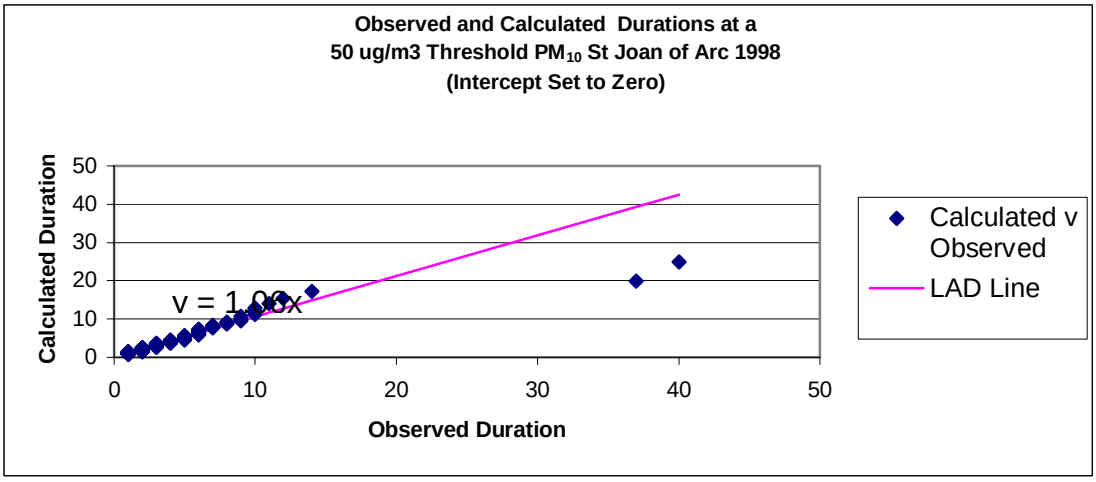

Fig. 6 Comparison of Observed and Calculated Durations at a 50 μgm^{-3} Threshold forPM_{10}, St Joan of Arc 1998 (Intercept Set at 0)

In these cases Least Absolute Deviation (LAD), constrained to pass through the origin was used as criterion for straight line fit recognising that the data sample in each case was relatively small and would be particularly susceptible to distortions due to outliers if the usual Least Squares Error was used, see [3].
The next stage of Guigliano's work was to find a function that would link the B and F values calculated for the Weibull distribution to a more easily obtained statistic, the ratio of the threshold level selected and the annual mean value of the pollutant (R).
Guiliano suggested the following empirical power law relationships for the Weibull distribution of duration:

$$B = g.R^h \tag{9}$$

$$F = a.R^b \tag{10}$$

Thus the B and F values in the rearranged Weibull Eqns. (8) can be substituted with the functions from Eqns. 9 and 10.
Theoretical durations were again calculated and compared to the observed durations. In the cases of NO_2 and CO the comparison of observed and calculated durations gave satisfactory results comparable to those obtained using the fitted Weibull coefficients. However the results obtained for PM_{10} were less satisfactory, over predicting PM_{10} by 23% on average. A comparison of observed and calculated durations for CO is shown as an example in Figure 7 using LAD as before.

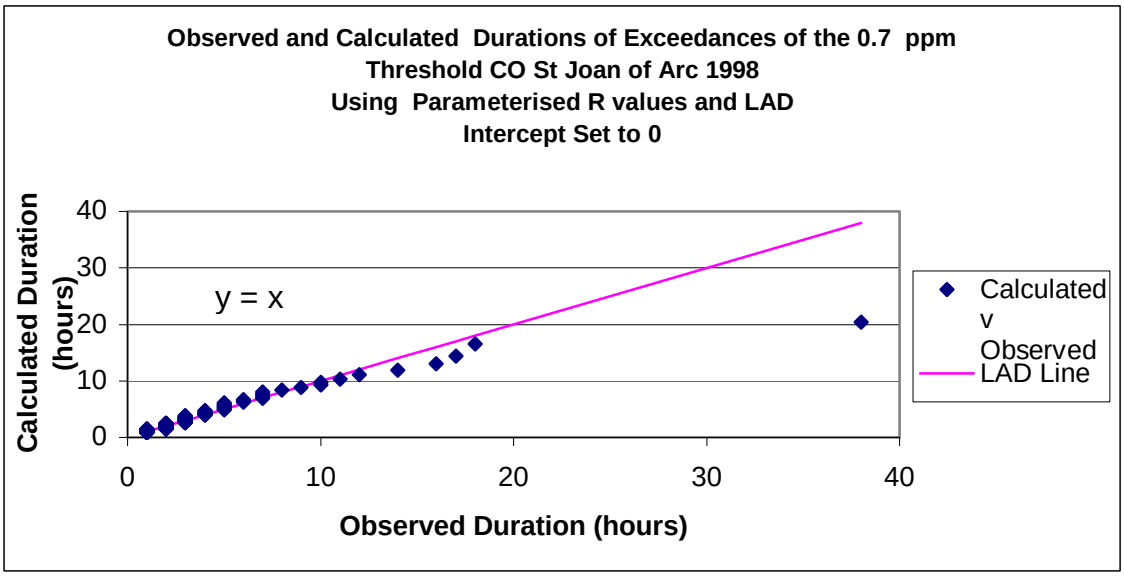

Fig. 7 Comparison of Observed and Calculated Exceedances (Using Parameter Values) of 0.7 ppm Threshold CO St Joan of Arc 1998 Intercept Set at 0

DISCUSSION AND CONCLUSIONS

A sample of observed crossing and duration results have been presented that, to a first order, follow the same patterns found in Guigliano's study. There is sufficient promise in this approach, using simple average statistics, to justify further development in characterising air quality in areas not directly monitored. Furthermore, it supports the use of a more sophisticated assessment of health impacts by consideration of the magnitude and duration of exposure rather than the simple exceedance criteria. Other developments now underway include use with: other pollutants, other seasonal averages and the comparison against a formalised theory for peak and duration statistics.

ACKNOWLEDGEMENTS

The authors are grateful to Sefton MBC for permission to use the air quality data. However the views expressed in this paper are the authors and do not represent those of Sefton Council.

REFERENCES

1. Giugliano M, Cernuschi S, And Marzolo F.(1998) The Duration of High NO_2 and CO Concentrations in an Urban Atmosphere. Atmospheric Environment Vol 32 No.17 pp 2923-2929.
2. Papoulis A (1984) Probability, Random Variables and Stochastic Processes, McGraw-Hill Co., NewYork.
3. Kottegoda N.T. and Rosso R., (1998) Statistics, Probability and Reliability for Civil Engineers, McGraw – Hill Co., New York.