Asian Journal of Water, Environment and Pollution, Vol. 4, No. 1, pp. 195-198.

Research Note

Occurrence of Natural Hazards and Outbreak of Epidemic: A Statistical Scrutiny

Sutapa Chaudhuri* and Surajit Chattopadhyay

Department of Atmospheric Sciences, University of Calcutta
Kolkata – 700 019, India
✉ chaudhuri_surtapa@yahoo.com

Received November 28, 2005; revised and accepted July 17, 2006

Abstract: Present paper delves the impact of natural hazards upon outbreak of epidemics in India. A statistical methodology is adopted to correlate the occurrence of natural hazards with the outbreak of epidemic. The experimentation setup consists of correlation and regression analysis and random walk approach. After a rigorous study, flood is identified as the most responsible natural hazard for the outbreak of epidemic.

Key words: Natural hazards, epidemic, statistical methodology, random walk, correlation, regression.

Introduction

Many of the Asian and Pacific developing countries are situated in the world's hazard belts and are subject to floods, droughts, cyclones, earthquakes, windstorms, tidal waves, land slides, etc. Since the International Decade for Natural Disaster Reduction began in 1990, the total number of deaths due to natural disasters in the region has exceeded 200,000 and the estimated damage to property over this period has been estimated at US$ 100 billion. Vulnerability to disasters has increased due to the increased aggregation of people in urban centres, environmental degradation, and a lack of planning and preparedness. Purpose of the present paper is a bit different with an aroma of newness. This study considers earthquake, drought, and flood as three important natural hazards of India. Their importance in creation of an epidemic is measured through statistical exploration of the data archive of the said hazards along with the application of the concept of fractal dimension and random walk for the period of 1975 to 2001. The said hazards have significant impact upon the human society. Firstly, such severe natural events lead to enormous loss of lives and property. Secondly, these hazards lead to immense water, soil, and air pollution. As a consequence, in various cases, the affected zone experiences an outbreak of epidemic. The present paper delves the relation between outbreak of epidemic and occurrence of natural hazards through statistical procedure.

*Corresponding Author

Methodology

Methodology adopted in this study comprises the following:

- Identification of the pair of events having maximum correlation.
- Preparation of a regressive model for the previously identified pair.
- Introduction of the concept of random walk to see the inherent patterns of the datasets and to give a qualitative support to the previously achieved results.

Pearson Correlation

One very important application of statistical ideas in the study of natural disasters is in making sense of a new set of data. In most of the cases, the data are paired. The nearly universal format for graphically displaying paired data is scatter plot or *x-y* plot. Geometrically, a scatterplot

is simply a collection of points in the plane, whose two Cartesian co-ordinates are the values of each member of the data pair. Often an abbreviated, single-valued measure of association between two variables, say, x and y, is needed. In such situations, the data analysts almost automatically and sometimes fairly uncritically calculate a correlation coefficient (Wilks, 1992, 1995, 1999). Usually, the term "correlation coefficient" is used to mean the "Pearson product-moment coefficient of linear correlation" between two variables x and y. One way to view the Pearson correlation is as the ratio of the sample covariance of the two variables to the product of the two standard deviations:

$$\rho_{xy} = \frac{\text{cov}\,(x,\, y)}{s_x s_y} = \frac{\sum_{i=1}^{n} [x_i' y_i']}{[\sum_{i=1}^{n} (x_i')^2]^{\frac{1}{2}} [\sum_{i=1}^{n} (y_i')^2]^{\frac{1}{2}}} \quad (1)$$

where the primes denote the anomalies or subtraction of the mean values from the original values. The correlation coefficients lie between (–)1 and (+)1, i.e., $-1 \le \rho_{xy} \le 1$.

Regression Model

Regression is most easily understood in the case of "simple" linear regression, which describes the linear relationship between two variables, say x and y. Conventionally, the symbol x is used as predictor and y is used as predictand. The predictor variable is the independent variable and the predictand variable is dependent variable. Very often, more predictor variables are required in practical forecast problems, but the ideas for simple linear regression generalize easily to this more complex case of multiple linear regression.

Fitting a regression equation to a data pair (x_i, y_i) means to find a straight line

$$\hat{y}_i = a + bx_i \quad (2)$$

The constants a and b are to be found out in such a way that the vertical distances between the line and the data points is minimized. The regression constants found these ways are

$$b = \rho_{xy} \frac{\sigma_y}{\sigma_x} \quad (3)$$

$$a = \bar{y} - b\bar{x} \quad (4)$$

where ρ_{xy} = Pearson correlation coefficient, σ_x = Standard deviation of x and σ_y Standard deviation of y.

Random Walk Approach

The random walk concept is based upon the motion of a particle upon a lattice. At every moment the particle can move to any other point up or down with a probability 0.5. Recording the coordinate of a point as a function of time, that is, $x(t)$, we can get a realization of the random walk process. Since all steps are independent, we can calculate the probability $P(x, t)$ that the system can be found at a distance x at time t. In the continuum approximation, we obtain a Gaussian distribution (Barabasi and Stanley, 1995; Bauchaud, 1990)

$$P(x, t) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right) \quad (5)$$

Discussion

The correlation coefficients for the pairs of disasters mentioned earlier are computed using equation (1) and are made schematically discernible in Figure 1. The figure exhibits that the correlation coefficient attains its maximum for the pair (yearly frequency of flood and yearly frequency of epidemic). The scatterplots corresponding to each pair are displayed in Figures 2 to 4. The following results are apparent from the figures:

- Figure 1 illustrates that yearly frequency of epidemic has the highest linear association with the yearly frequency of occurrence of flood.
- Figure 2 shows that yearly frequency of occurrence of flood and yearly frequency of epidemic maintains a significant degree of linearity. The trend line shows that the data pairs are almost in the same pattern of that of trend line. Thus, there is evidence in support of linear association between occurrence of flood and epidemics. Thus, impact of flood upon the occurrence of epidemic is quite significant.

Since the previous discussion revealed that yearly frequency of epidemic is mostly influenced by the

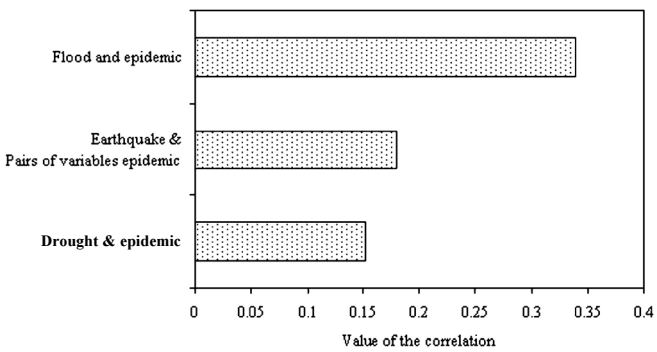

Figure 1: Schematic showing the Pearson correlation coefficients between outbreak of epidemic and occurrence of natural hazards.

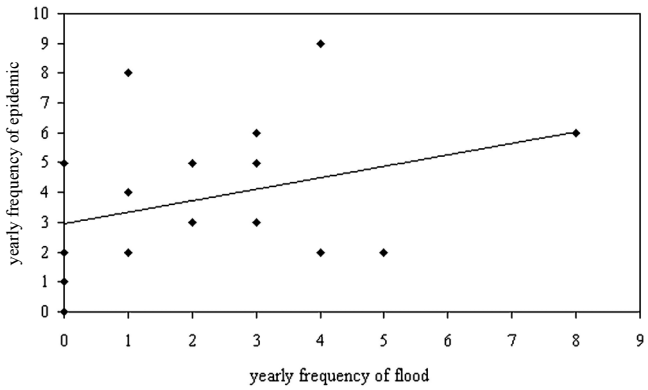

Figure 2: Schematic showing the scatter plot for the yearly frequency of flood and epidemic.

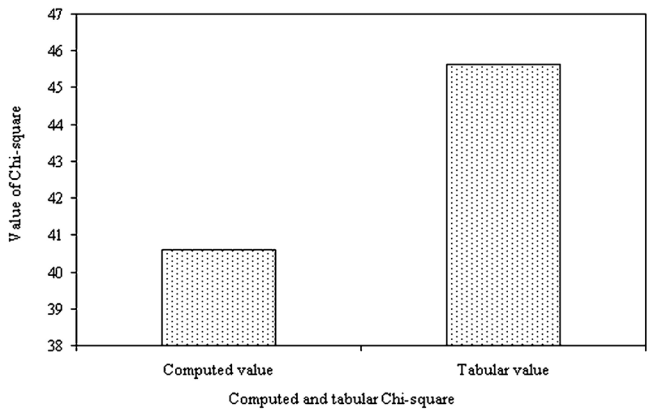

Figure 3: Schematic showing the computed and tabular values of Chi-square at 1% level of significance with 26 degrees of freedom to find the goodness of fit of regression equation with flood as predictor and epidemic as predictand.

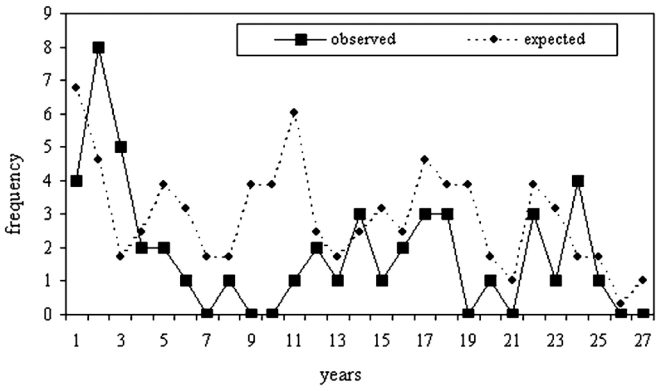

Figure 4: Schematic showing the observed yearly frequency of epidemic and expected frequency on the basis of regression equation with flood as predictor.

occurrence of flood, a predictive equation for epidemic outbreak is developed using flood as predictor. Equations (2), (3) and (4) are implemented to fit a regression equation to the (flood, epidemic) data points using the yearly frequencies of flood as predictor and the frequencies of epidemic as predictand variable.

In this study:

$\sigma_x = 2.13$, $\sigma_y = 1.88$, $\rho_{xy} = 0.34$, $b = 0.30$, $a = 0.72$

Thus, the regression equation with flood as predictor and epidemic as predictand will be $\hat{y} = 0.72 + 0.30x$.

The goodness of fit of this linear equation through the method of hypothesis testing is then examined (Chaudhuri et al., 2001).

In the present problem, the null hypothesis is framed as follows:

H_0: The line is a good fit.

Alternative hypothesis is

H_A: The line is not a good fit.

Under the supposition of the truthness of the null hypothesis, a chi-square statistic (Wilks, 1995) is framed as follows:

$$\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

With 26 degrees of freedom

The computed value of Chi-square is found to be 40.605. Tabular value of Chi-square is found to be 45.642 with 26 degrees of freedom at 1% level of significance. Thus the null hypothesis is accepted. The result is displayed schematically in Figure 3 and 4. Figure 4 shows the actual and predicted yearly frequencies of epidemic. It is apparent from the figure that the actual frequencies of epidemic and predicted values on the basis of flood as predictor are almost of the same pattern. This gives a qualitative support to the goodness of fit of the regression.

Applying the concept of random walk, the probabilities are computed using equation (5). The probabilities are pictorially displayed in Figure 5.

Figure 5 depicts the following:

- Roughness curves pertaining to drought and earthquake are significantly smooth.
- Roughness curves pertaining to flood and epidemic are significantly rough.
- Roughness curves pertaining to flood and epidemic are highly similar with respect to roughness.

Thus, the study through the concept of random walk reveals that outbreak of epidemics is highly possible due to the occurrence of flood. Earthquake and drought are

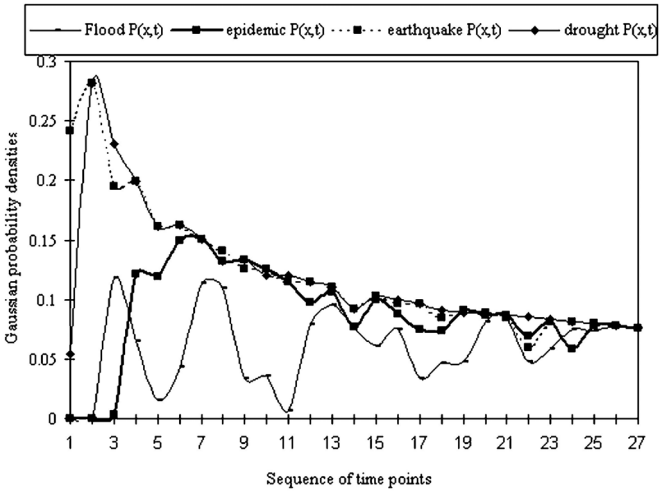

Figure 5: Schematic showing the roughness pertaining to yearly frequencies pertaining to different natural hazards obtained through the Gaussian approximation.

not so significant for the occurrence of epidemics. Moreover, it can further be concluded that predictive model for flood is more important than that of drought and earthquake to reduce the impact of epidemic.

Conclusion

The thorough statistical study leads us to conclude that among different natural hazards, flood is the highest degree of responsibility in resulting an outbreak of epidemic. Thus, a proper prediction model for flood is an absolute necessity in order to save human lives from epidemical death.

References

Barabasi, A.L. and H.E. Stanley (1995). Fractal Concepts in Surface Growth. Cambridge University Press, Cambridge.

Bauchaud, J.P. and A. Georges (1990). Anomalous diffusion in disordered media: Statistical mechanics, models and physical applications. *Phy. Rep*., **195**: 127-193.

Chaudhuri, S. and S. Chattopadhyay (2001). Measure of CINE—A relevant parameter for forecasting pre-monsoon thunderstorms. *Mausam*, **42**: 679-684.

Wilks, D.S. (1995). Statistical Methods in Atmospheric Sciences. Academic Press, USA.

Wilks, D.S. (1992). Adapting stochastic weather generation algorithms for climate change studies. *Climate Change*, **22**: 67-84.

Wilks, D.S. (1999). Inter annual variability of extreme value weather characteristics of several stochastic daily precipitations. *Agricultural and Forest Meteorology*, **93**: 153-169.