THE STUDY OF FLOOD FREQUECNY ESTIMATES USING CAUCHY VARIABLE KERNEL

Young-Il Moon[1], Young-Il Cha[2], and Ashish Sharma[3]

[1] Assistant Professor, Dept. of Civil Engineering, University of Seoul, Korea
[2] Graduate Student, Dept. of Civil Engineering, University of Seoul, Korea
[3] Assistant Professor, School of Civil and Envir. Eng., University of New South Wales, Sydney, Australia

Abstract: The frequency analyses for the precipitation data in Korea were performed. We used daily maximum series, monthly maximum series, and annual series. For nonparametric frequency analyses, variable kernel estimators were used. Nonparametric methods do not require assumptions about the underlying populations from which the data are obtained. Therefore, they are better suited for multimodal distributions with the advantage of not requiring a distributional assumption. In order to compare their performance with parametric distributions, we considered several probability density functions. They are Gamma, Gumbel, Log-normal, Log-Pearson type Ⅲ, Exponential, Generalized logistic, Generalized Pareto, and Wakeby distributions. The variable kernel estimates are comparable and are in the middle of the range of the parametric estimates. The variable kernel estimates show a very small probability in extrapolation beyond the largest observed data in the sample. However, the log-variable kernel estimates remedied these defects with the log-transformed data.

Key Words: frequency analysis, nonparametric, kernel density function, probability precipitation

1. INTRODUCTION

When designing a hydrosystem to control and use of water resources, a frequency analysis based on hydrological data is one of the most important element for designing and planning an economical hydrosystem. Generally, the rainfall data can be easily observed than the flood data in Korea. Therefore, in this paper a comparison of parametric and nonparametric techniques for probability precipitation in Korea is presented.

A currently used approach to frequency analysis is based on the concept of parametric statistical inference. In these analyses, the assumption is made that the distribution function describing precipitation data is known. Distributions that are often used are Log-normal, Pearson Type Ⅲ, Gumbel, extreme value distributions, Gamma, and others by using the method of moment (MOM), maximum likelihood (ML), probability weighted moment (PWM), or L-moment. However, such an assumption is not always justified. Some difficulties associated with parametric estimation are (1) the objective selection of a distribution, (2) the reliability of distributional parameters (especially for skewed data with a short record length), (3) the inability to analyze multimodal distributions, and (4) the treatment of outliers. The probability weight moment (PWM) method (Greenwood et al., 1979; Hosking, 1989, etc.), L-moment method (Hosking, 1990), or others have been complemented the problems of skewed data with a short record length. Nevertheless, in the process of parametric frequency analysis, the choice of best fitted distribution among the other distributions which are passed the goodness-of-fit tests (χ^2 test, Kolmogorov-Smirnov test, Cramer von Mises test, etc) is still not a easy task. Also, the analysis of bimodal probability density function has many complicated problems when the data has a mixed distribution. The assumption of a pre-chosen distribution, which is based on goodness-of-fit tests and selected as the most appropriate distribution, is no longer valid if the size of the data available is increased. Therefore, parametric techniques may be inadequate for reliable frequency estimates.

The problem is to estimate the probability density function. Since the true distribution is unknown, we have to resort to use a nonparametric approach. The histogram has historically been the choice in estimating the probability density function of a sample in a nonparametric fashion. This method requires that a suitable bin width and starting position be chosen to obtain a decent result. However, estimation using this approach can be subject to large errors. If the number of bins chosen are too small, the resulting probability density function is oversmoothed thereby obscuring potentially important details. On the other hand, the use of too many bins may result in a ragged looking distribution.

In recent years, nonparametric kernel density estimation methods have been introduced as viable and flexible alternatives to parametric methods for flood frequency analysis or probability precipitation estimation. Several nonparametric approaches have been introduced by Adamowski (1985, 1989, and 1996), Adamowski and Feluch (1990), Adamowski and Labatiuk (1987), Lall et al. (1993), Moon et al. (1993), Moon and Lall (1994 and 1995), and Moon (2000). Nonparametric methods do not require assumptions about the underlying populations from which the data are obtained. Also, they are better suited for multimodal distributions. Usually, nonparametric kernel density estimator was relatively consistent across the estimation situations considered in terms of bias and root mean squares error (RMSE) with the advantage of not requiring a distributional assumption while providing a uniform procedure (Lall et al., 1993; Moon et al., 1993).

Even though many people have shown that the nonparametric method provides a better fit to the data than the parametric method and gives more reliable flood or precipitation estimates, the nonparametric method implies a very small probability in extrapolation beyond the highest observed data in the sample. In this paper, we tried to show a remedy for these inadequacies by introducing a log-estimator which is a probability density function for log-transformed data, $\ln x_i$ if x_1, x_2, ..., x_n are random variable. The extrapolation is based on the shape of the kernel density function assumed and on the value of bandwidth h. Thus, only a few observations contained in the bandwidth h influence the extrapolation to the tail of the distribution. However, it is possible to remedy these defects by applying the nonparametric kernel estimator to log-transformed data.

2. PARAMETRIC/NONPARAMETRIC FREQUENCY ANALYSES

Which distribution is best for the precipitation data in Korea based on parametric frequency analyses? The question of which distribution gives the best fit may be decided by using the chi- squared statistic and Kolmogorov-Smirnov tests. As shown in Fig. 1, the parametric frequency analyses were considered all the distributions available. The parameters of selected distributions were estimated from the method of moments, maximum likelihood, probability-weighted method, and L-moments. Then, we assumed a particular distribution which was selected from goodness-of-fit tests among the other several distributions to regard it as a population's distribution and progress to analyze. This paper applied Normal distribution, two parameter Log-Normal distribution, three parameter Log-Normal distribution, three parameter Gamma distribution, Log-Pearson Type Ⅲ distribution and Generalized Extreme Value distribution to precipitation data in Korea. The goodness of fit tests was applied to χ^2 test for probability density functions and Kolmogorov-Smirnov test for cumulative distribution functions with 5% significance level.

Rosenblatt (1956) introduced the nonparametric kernel density estimator, defined for all real x by

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x - X_i}{h}\right) \tag{1}$$

where $x_1, \ldots, x_n$ are independent identically distributed real observations, $K(\cdot)$ is a kernel function, and h is a positive smoothing factor assumed to tend to zero as n tends to infinity. Silverman (1986) explained the basic concept of the nonparametric kernel density estimator. From the definition of a probability density, if the random variable x has density f(x), then

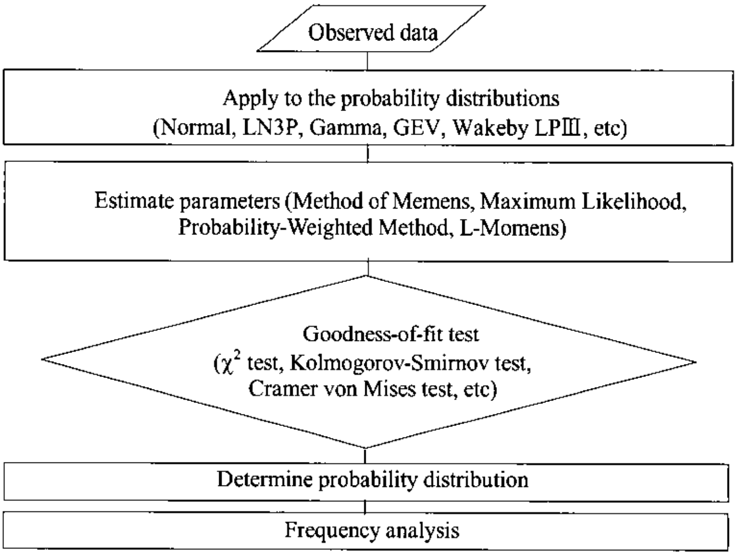
Observed data
Apply to the probability distributions
(Normal, LN3P, Gamma, GEV, Wakeby LPⅢ, etc)
Estimate parameters (Method of Memens, Maximum Likelihood,
Probability-Weighted Method, L-Momens)
Goodness-of-fit test
(χ² test, Kolmogorov-Smirnov test,
Cramer von Mises test, etc)
Determine probability distribution
Frequency analysis

Fig. 1. The procedure of parametric frequency analysis

$$f(x) = \lim_{h \to 0} \frac{1}{2h} p(x - h < X < x + h) \tag{2}$$

For any given h, P(x-h < x < x+h) can be estimated by the proportion of the sample falling in the interval (x-h, x+h). Thus, a natural estimator is given by choosing a small number h and setting

$$\hat{f}(x) = \frac{1}{2hn} \; [\#\,\text{of } X_i, \ldots\ldots, X_n \text{ falling in } (x - h, x + h)] \tag{3}$$

To express the estimator more transparently, define the weight function w(x) by

$$w(x) = \begin{cases} \frac{1}{2}\,, & \text{if abs}(x) < 1 \\ 0\,, & \text{otherwise} \end{cases} \tag{4}$$

Then it is easy to see that the estimator can be written as

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - X_i}{h}\right) \tag{5}$$

It follows from equation (5) that the estimator is constructed by placing a box of width 2h and height $(2nh)^{-1}$ on each observation and then summing to obtain the estimator. This weight function is the kernel function which satisfies the condition

$$\int K(t)\,dt = 1, \quad \text{where } t = \frac{x - X_i}{h} \tag{6}$$

The kernel function is usually required to be unimodal with peak at x = 0, smoothness, and a symmetric function, that is, a density $(\int K(t)\,dt = 1)$ with expectation 0 $(\int tK(t)\,dt = 0)$ and finite variance $(\int t^2 K(t)\,dt = \text{constant})$.

When applying the method in practice, it is necessary to choose a kernel function and a smoothing parameter. Some useful kernel functions are given in Table 1 and Fig. 2. Usually, different kernels should be examined depending the objective. For example, if continuity and differentiability of the density is needed, one may choose a kernel with infinite support rather than one with finite support.

While the choice of kernel does not seem to be critical, the choice of smoothing factor is quite a different matter. The value of h is critical

Table 1. Typical kernel functions

Kernel	K(t)
Rectangular	1/2 for $\lvert t \rvert < 1$, 0 otherwise
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)$
Epanechnikov	$\frac{3}{4}\left(1 - \frac{1}{5}t^2\right)/\sqrt{5}$ for $\lvert t \rvert < \sqrt{5}$
Rajagopalan	$\frac{3h}{1 - 4h^2}(1 - t^2)$, where $\lvert t \rvert \le 1$
Cauchy	$\frac{1}{\pi(1 - t^2)}$

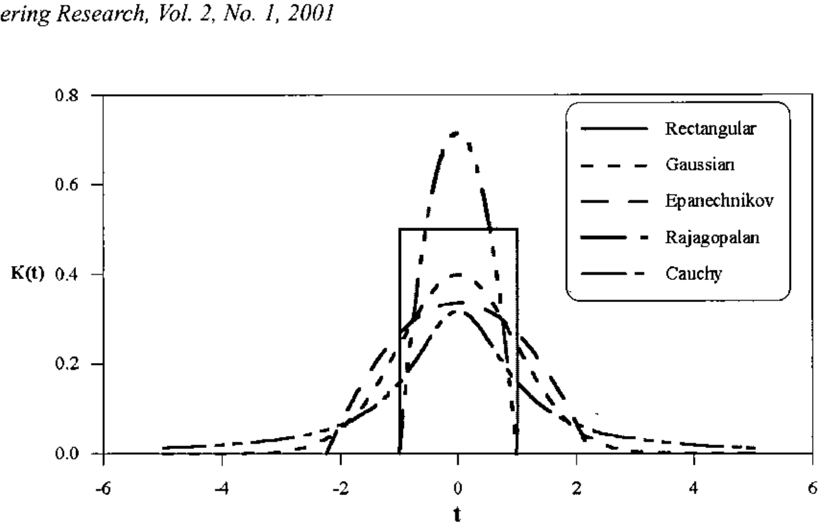

Fig. 2. The Shape of Kernel Functions

and, in practice, not obvious. Too large an h implies large bias, an oversmoothed estimate, and consequent loss of information. Too small an h implies large variance and too rough an estimate (Adamowski and Labatiuk, 1987). Since all error measures depend on the unknown density, generally they cannot be used in deriving analytical expressions for selecting the smoothing factor h. Several measures of performance for using the data to produce suitable values for the smoothing parameter h have been proposed. The smoothing parameter h can be obtained by Maximum Likelihood Cross-Validation (Habbema et al., 1974; Duin, 1976), Least Squares Cross-Validation (Hall, 1983; Hall and Marron, 1987; Stone, 1984), Breiman et al. Method (Breiman et al., 1977), and Adamowski Cross-Validations (Adamowski, 1985).

Lall et al. (1993) demonstrated that one should directly focus on kernel distribution function estimates rather than kernel density estimates. The variable kernel estimate $F_n(x)$ of the cumulative distribution function F(x) is defined as :

$$\begin{aligned} F_n(x) &= \int_{-\infty}^{x} \sum_{i=1}^{n} \frac{1}{nhd_{i,k}} K\left(\frac{t-x_i}{hd_{i,k}}\right) dt \\ &= \sum_{i=1}^{n} K*\left(\frac{x-x_i}{nd_{i,k}}\right) \end{aligned} \tag{7}$$

where K(t) is a kernel function, h is a bandwidth, $d_{i,k}$ is the distance from x_i to its k th nearest neighbor, and $K^*(t)=\int_{-\infty}^{t} K(u)du$.

Lall et al. (1993) provide a review of this discussion and compare the performance of different kernels and bandwidth selection methods in the flood frequency context. They found that the variable kernel estimator with heavy-tailed kernel (Cauchy) and bandwidth selection based on Adamowski criteria (VK-C-AC) led to the best tail estimates using kernel methods. The Cauchy kernel is a heavy tailed kernel and may have a better capacity for extrapolation, particularly with heavy tailed densities.

3. RESULTS

The frequency analysis for the precipitation of 26 sites in 5 basin areas (i.e. Han River, Nakdong River, Keum River, Sumjin River, Yeong-

Table 2. Precipitation data of Han River area

Area	Site	Year	Data size	Missing Year
Han River	Seoul	1907～1998	85	1907.1～1097.9, 1950.9～1953.11
	Inchon	1949～1998	50	1950.6～1951.9
	Chungju	1973～1998	26	-

san River in Korea), which are under control of the Korea Meteorological Office, were performed. We used daily maximum series, monthly maximum series, and annual series. In order to select an appropriate distribution, 17 probability density functions were considered for the parametric method. They are Gamma Ⅱ, Gamma Ⅲ, GEV (Generalized Extreme Value), Gumbel (Extreme Value type Ⅰ), Log-Gumbel, Log-normal Ⅱ, Log-normal Ⅲ, Log-Pearson type Ⅲ, Weibull Ⅱ, Weibull Ⅲ,.Exponential, Normal, Pearson type Ⅲ, Generalized logistic, Generalized Pareto, Kappa and Wakeby distributions. The distribution parameters were estimated by method of moments, maximum likelihood, probability weighted moments, and L-moments method. The goodness-of-fit test for the parametric distribution applies Kolmogorov-Smirnov test and χ^2 test with significant level of 5 %.

For nonparametric frequency analyses, variable kernel density function and log-variable kernel density function estimators were used with Cauchy kernel and bandwidth selection Adamowski criteria (VK-C-AC).

The precipitation data of Han River area is shown in Table 2. In this paper, we presented just the data and the results of Han River area to save the space.

The Figs. 3～5 represent the probability density functions (PDF) of annual precipitation amounts, monthly maximum precipitations, and daily maximum precipitations for Seoul, Inchon, and Chungju stations respectively. From the Fig. 3～5, we observed bimodal distributions in Seoul and Chungju stations, and multimodal one for Inchon station. In those cases, a difficulty associated with parametric approach is the in-

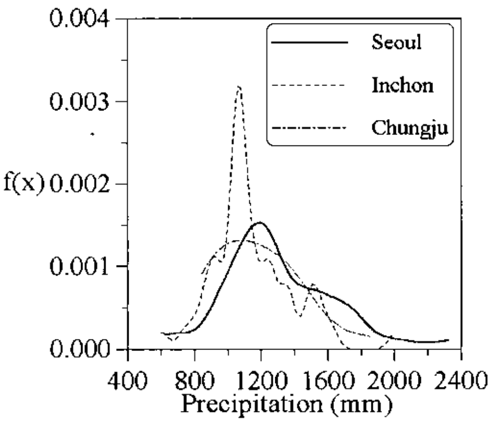

Fig. 3. PDF of annual precipitation of Han River

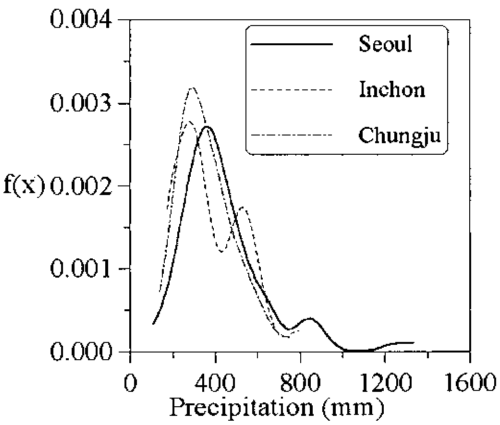

Fig. 4. PDF of monthly maximum precipitation for Han River

Table 3. Probability precipitation of each return period (mm, : maximum or minimum)

Site	Seoul		Inchon		Chungju	
Return Period (year)	100	200	100	200	100	200
Log-Variable Kernel	2365.6	2423.9	2146.8	2610.4	2005.3	2254.7
Variable Kernel	2353.9	2356.9	2023.1	2095.4	1891.0	1898.5
Exponential					2248.2	2453.6
Gamma	2215.9	2335.8	1792.0	1874.8	1879.9	1970.8
GEV	2330.8	2486.6			2014.2	2151.0
G-Logistic	2442.2	2696.1			2099.4	2315.1
Log-Normal	2330.8	2493.4			2009.7	2148.0
G-Parato					1834.3	1872.5
Gumbel	2401.9	2590.2			2039.8	2188.5
Normal			1713.9	1774.5	1790.7	1856.3
Pearson typeIII	2309.0	2456.5			1988.1	2111.1
Wakeby	2330.3	2458.4				

ability to analyze multimodal distributions. Therefore, parametric estimation techniques are inadequate for modelling such an annual maximum process. However, nonparametric methods do not require assumptions about the underlying populations from which the data are obtained. Therefore, they are better suited for multimodal distributions with the advantage of not requiring a distributional assumption.

A comparison among the parametric and the nonparametric distribution estimators of the quantile function for Seoul, Inchon, and Chungju data are shown in Fig. 6~8. Here the method of parametric estimation for parametric methods is considered with L-moments. The maximum and minimum recorded precipitations are presented in Table 3. The parametric estimates of the 100-year precipitation range from 2,216 to 2,442 mm in Seoul, from 1,791 to 2,248 mm in Chungju per year respectively. The log-variable kernel and variable kernel are comparable and are in the middle of the range of the parametric estimates. As shown Fig. 7, only Gamma and Normal distributions were passed the goodness-of-fit test for Inchon station. The variable kernel estimates in Figs. 6~8 shows a very small probability in extrapolation beyond the 50-year return period (i.e., the quantiles beyond the 50-year are almost the same). However, the log-variable kernel estimates (i.e., variable kernel estimator applied to log-transformed data) remedied these defects with the log data.

4. CONCLUSION

The frequency analysis for the precipitation data of 26 sites in 5 basin areas in Korea were performed. We applied nonparametric variable kernel estimators, log-variable kernel estimators, and 17 selected parametric distribution estimators to daily maximum series, monthly maximum series, and annual series. Since the results of the parametric estimators varied according to the distributions and the methods of the parametric estimation, it is not easy to say which parametric estimator is the best. However, for each data set, the nonparametric variable kernel estimator with the Cauchy kernel and Adamowski's bandwidth selection is shown to be competitive with any parametric distribution estimators and has the advantage of not requiring a distributional assumption. In particular, the nonparametric kernel estimators (variable and log-variable kernel estimators) worked better than the parametric estimators for multimodal data. This ability to analyze multimodal density by the nonparametric method is particularly useful in hydrology. Even though only a limited data set was available and estimation outside the range of data was wanted, the log-variable kernel estimator provided good results in the upper tail compared with the variable kernel estimator.

ACKNOWLEDGEMENTS

This work was supported by Korea Research Foundation Grant. (KRF-99-003-E00591)

REFERENCES

Adamowski, K. (1985). "Nonparametric kernel estimation of flood frequency." *Water Resources Research*, Vol. 21, No. 11, pp. 1585-1590.

Adamowski, K., and Labatiuk, C. (1987). "Estimation of flood frequencies b a nonparametric density procedure." Hydrologic Frequency Modeling, pp. 97-106.

Adamowski, K. (1989). "A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies." *Journal of Hydrology*, Vol. 108, pp. 295-308.

Adamowski, K., and Feluch, W. (1990). "Nonparametric flood-frequency analysis with historical inforamtion." *Journal of Hydraulic Engineering*, Vol. 116, No. 8, pp. 1035-1047.

Adamowski, K. (1996). "Nonparametric Estimation of Low-Flow Frequencies." *Journal of Hydraulic Engineering*, Vol. 122, No. 1, pp. 46-49.

Breiman, L., Meisel, W., and Purcell, E. (1977). "Variable kernel estimates of multivariate densities." *Technometrics*, Vol. 19, No. 2, pp. 135-144.

Duin, R.P.W. (1976). "On the choice of smoothing parameters for parzen estimators of probability density functions." *IEEE Trans. Comput.*, C-25, pp. 1175-1179.

Greenwood, J.A., Landwehr, J.M., Matalas, N.C., and Wallis, J.R. (1979). "Probability Weighted Moments : Definition and Relation to Parameters of Several Distributions Expressible in Inverse Form." *Water Resources Research*, Vol. 15, No. 5, pp. 1049-1054.

Habbema, J.D.F., Hermans, J., and Broek, V.D. (1974). *A stepwise discrimination program using density estimation.* In G. Bruckman (Ed.). Physical verlag. Compstat Vienna, pp. 100-110.

Hall, P. (1983). "Large sample optimality of least squares cross-validation in density estimation." *Ann. Statist.*, Vol. 11, pp. 1156-1174.

Hall, P., and Marron, J.S. (1987). "Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation." *Probability Theory Rel.*, Fields 74, pp. 567-581.

Hosking, J.R.M. (1989). *The theory of probability weighed moments.* Research Report, RC 12210, IBM Research Division, T.J. Watson Research Center, New York.

Hosking. J.R.M. (1990). "L-moments Analysis and estimation of distribution using linear combinations of order statistics." *Journal* *of Royal Statistical Society*, Vol. 52, No. 1, pp. 105-124.

Lall, U., Young-Il Moon, and Bosworth, K. (1993). "Kernel flood frequency estimators : bandwidth selection and kernel choice." *Water Resources Research*, Vol. 29, No. 4, pp. 1003-1015.

Moon, Young-Il, Lall, U., and Bosworth, K. (1993). "A comparison of tail probability estimators." *Journal of Hydrology*, Vol. 151, pp. 343-363.

Moon, Young-Il, and Lall, U. (1994). "Kernel Quantile Function Estimator for Flood Frequency Analysis." *Water Resources Research*, Vol. 30, No. 11, pp. 3095-3103.

Moon, Young-Il, and Lall, U. (1995). "Nonparametric flood frequency analysis by a kernel quantile function estimator." *European Geology Society*.

Moon, Young-Il. (2000). "The Study of Parametric and Nonparametric Mixture Density Estimator for Flood Frequency Analysis." *Water Engineering Research*, Vol. 1, No. 1, pp.61-73.

Rosenblatt, M. (1956). "Remarks on some nonparametric estimates of a density function." *Ann. Math. Statist.* Vol. 27, pp. 832-837.

Silverman, B.W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, New York.

Stone, M. (1984). "An asymptotically optimal window selection rule for kernel density estimates." *Ann. Statistics*, Vol. 12, pp. 1285-1297.

Young-Il Moon, Assistant Professor, Dept. of Civil Engineering, University of Seoul, Korea
(E-mail:ymoon@uoscc.uos.ac.kr)

Young-Il Cha, Graduate Student, Dept. of Civil Engineering, University of Seoul, Korea
(E-mail:ycha@sidae.uos.ac.kr)

Ashish Sharma, Assistant Professor, School of Civil and Envir. Eng., University of New South Wales, Sydney, Australia
(E-mail:a.sharma@unsw.eclu.au)

(Received May 18, 2000; accepted October 26, 2000)