Aerosol and Air Quality Research, 16: 2302–2313, 2016

ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2016.04.0139

Technical Note

Long-Range Correlations in Air Quality Time Series: Effect of Differencing and Shuffling

Asha B. Chelani

Air Pollution Control Division, National Environmental Engineering Research Institute (CSIR-NEERI), Nagpur 440020, India

ABSTRACT

Long-range correlations in the air quality index (AQI) are analysed using rescaled range analysis (*R/S*), detrended fluctuation analysis (*DFA*) and power spectral density analysis. Air quality index in five cities of India is considered for this purpose. Statistical transformations such as differencing and shuffling have been carried out to examine the effect of temporal correlations on long-range correlation property of the time series. All three methods indicated the presence of persistence in original AQI time series. After differencing, long-range correlation property is, however, observed to be distorted. *R/S* analysis did not show the similar results as *DFA* and power spectral density analysis. Shuffled time series is shown to possess persistence as in the original one by using *R/S* analysis, whereas other two methods showed random behaviour at most of the locations. This suggests that the persistence property is largely influenced by short-range correlations in the AQI time series. The incorporation of this information can enhance the performance of the models to forecast the air quality. The similarity in the results of *DFA* and power spectral density analysis suggests that both methods can be relied more than *R/S* analysis in studying the persistence property of the time series.

Keywords: Long-range correlations; Differencing; Shuffling; Air quality index.

INTRODUCTION

Long-range correlation or temporal persistence in air pollutant concentrations has been observed in several studies. A detailed literature review is provided in Chelani (2016). A persistent time series is governed by the patterns which have correlation with the past patterns. Short-range correlations in time series suggest that the temporal correlations among its data decrease exponentially. Long-range correlations imply the temporal correlations decrease slowly and usually follow power-law distribution (Chelani, 2016). For long-memory processes, correlation analysis based traditional methods such as autocorrelation function are not useful. For the process that is not independent, Mandelbrot has provided fractional Brownian motion to model it as persistent or anti-persistent (Kim *et al.*, 2014). Hurst exponent that represents the correlation properties of the time series based on its scaling behaviour is more useful for long-range correlated time series (Hurst, 1951). Hurst effect has been modelled by Mandelbrot (1968) as fractional Brownian motion (fBm) with associated properties as persistent, anti-persistent and random. The existence of persistence or long-range correlations in time series is helpful in making further projections as the measurements rely on the correlations with historical observations (Varotsos *et al.*, 2005; Chelani, 2009). Linear or nonlinear models can be developed and used to obtain forecasts based on the temporal correlations with past observations.

The persistence in the air pollutant concentrations observed over time may either be governed by external influence or internal dynamics (Hu *et al.*, 2001). The temporal correlations may generally be due to the presence of background concentrations and intrinsic evolution of the system. Periodic or seasonal influence in addition to the presence of trend, which is external in nature, may also govern the persistence property. Trend component is usually associated with short-range temporal correlations and may influence the persistence property characterised by Hurst exponent. Recognizing and filtering out short-term correlations associated with trend and/or seasonal component is important to correctly quantify the Hurst exponent.

With the well-established persistent behaviour of air pollutant concentrations time series, this study is an attempt to identify the possible reasons for persistence. For this, two data transformation methods are used. First order differencing is used to nullify the effect of first order linear temporal correlations and shuffling is used to remove the effect of temporal correlations. Analysis is carried out on air pollutant concentration time series observed at various locations in

* Corresponding author.
E-mail address: ap_lalwani@neeri.res.in

India using three methods, detrended fluctuation analysis (*DFA*), rescaled-range (*R/S*) analysis and power spectral density analysis.

STUDY AREA AND DATA USED

Various state and national agencies monitor air pollution data over several locations in India. Central Pollution Control Board (CPCB) and State Pollution Control Boards are the main sponsors for regulatory monitoring of parameters such as particulate matter of size less than 10 micron (PM_{10}), nitrogen dioxide (NO_2) and sulphur dioxide (SO_2). The measurements are with a frequency of twice a week and hence 104 measurements in a year at a particular location. PM_{10}, SO_2 and NO_2 data observed in Nagpur, Amravati, Chandrapur, Delhi and Mumbai during 2012–2014 are used to study the persistence characteristics. The location of the cities is given in Fig. 1. Delhi and Mumbai are metro cities with huge vehicular pollution and other anthropogenic emissions. Nagpur is a major development urban area where recently lots of infrastructural development activities have taken place. Amravati is rural cum urban area where few industries are located. Chandrapur has a major super thermal power plant of around 2000 MW. In each city two sites with different land-use activities are considered.

Fig .1. Location of selected five cities of India.

Ngp1, chp1 and del1 are mainly residential locations, whereas ngp2, chp2 and del2 are industrial sites. Amra1, amra2, mum1 and mum2 are urban sites with residential and commercial activities. The data is available online at the respective websites of the agencies. The data for Delhi is obtained from www.cpcb.nic.in and for other locations the data is obtained from www.mpcb.gov.in. Instead of carrying out the analysis on individual time series of each parameter, air quality index is computed. Recently CPCB has developed air quality index based on health indices which can be a combination of eight parameters. The sub-indices are computed for each parameter and the worst of them represent the AQI. The minimum number of parameters to be used are at least 3. The index is computed using PM_{10}, SO_2 and NO_2. The index has various categories which are health break points as given in Table 1. The details of the index are given in http://www.cpcb.nic.in/AQI_new.php.

RESCALED-RANGE ANALYSIS

Hurst exponent (H) is computed by using rescaled range analysis of the time series. For this the following set of equations based on Hurst (1951) are used;

$$z(k) = \sum_{i=1}^{k} \left[y(i) - \langle y \rangle_{\tau} \right] \tag{1}$$

$$R(\tau) = \max_{1 \le k \le \tau} z(k) - \min_{1 \le k \le \tau} z(k) \tag{2}$$

$$S(\tau) = \sqrt{1/\tau \sum_{i=1}^{k} \left(y(i) - \langle y \rangle_{\tau} \right)^2} \tag{3}$$

where $\langle y \rangle$ is the mean of the time series $y(i)$, τ is the time lag and k is the discrete time, R is the difference of maximum and minimum of the cumulative sum of deviations of time series from its mean, S is the standard deviation. The ratio R/S is used to describe the scaling properties of the time series. Hurst (1951) has shown that if a series of random variables has finite standard deviation and are independent, then R/S statistic increases in proportion to τ^H for large values of τ. Hence

$$R(\tau)/S(\tau) \quad \alpha \quad \tau^H \tag{4}$$

Hurst exponent, H is the slope of R/S against time lag τ on log-log scale. $0.5 < H < 1.0$ for the time series implies persistence, $0 < H < 0.5$ implies anti-persistence, whereas large $H > 0.5$ indicates strong persistence. $H = 0.5$ indicates random time series i.e., the time series is independent and uncorrelated.

DETRENDED FLUCTUATION ANALYSIS

The presence of long-range correlations or persistence in

Table 1. Air quality index categories.

AQI	AQI Category	Health Impact
0–50	Good	Minimal impact
51–100	Satisfactory	Minor breathing discomfort to sensitive people
101–200	Moderately Polluted	Breathing discomfort to the people with lung, heart disease, children and older adults
201–300	Poor	Breathing discomfort to people on prolonged exposure
301–400	Very Poor	Respiratory illness to the people on prolonged exposure
401–500	Severe	Respiratory effects even on healthy people

the time series can also be detected by using detrended fluctuation analysis (Peng *et al.*, 1994). The total length of the time series $y(i)$, $i = 1, 2, \ldots, k$ is first integrated as;

$$z(k) = \sum_{i=1}^{k} \left[y(i) - \langle y \rangle_{\tau} \right] \quad (5)$$

where $y(i)$ is the time series with mean $\langle y \rangle$ of all the samples, τ is the time lag, $k = 1, 2, \ldots, N$ and N is the length of the time series. New time series $z(k)$ is divided into segments of equal length n and the least-squares line is fitted to the data in each segment. The y-coordinate of the straight-line segments is denoted by $Z_n(k)$, which is used to detrend the time series $z(k)$ as $z(k) - Z_n(k)$ in each segment. The root mean square fluctuation of integrated and detrended time series is calculated as;

$$F(n) = \sqrt{1/N \sum_{k=1}^{N} \left[z(k) - z_n(k) \right]^2} \quad (6)$$

A linear relationship of average fluctuation $F(n)$ and segment size n on a log-log graph indicates the presence of scaling, i.e., $F(n)$–n^{α}, where α is the scaling exponent, can be obtained as a slope of the line for all the segment sizes. The scaling exponent gives an indication of the nature of the time series. For $0 < \alpha < 0.5$, it indicates the presence of power-law anti-correlations in the time series, $0.5 < \alpha < 1$ suggests the persistence property. If $\alpha = 0.5$, time series corresponds to white noise.

POWER SPECTRAL DENSITY ANALYSIS

The spectral analysis can be carried out by using fast Fourier transform of the time series. Power spectral density $h(f)$, where f is the frequency, is measured as given below;

$$h(f) = \frac{1}{2\pi} \sum_{\tau=-\infty}^{\infty} \rho(\tau) e^{i\tau f} \quad (7)$$

where $\rho(\tau)$ is the auto-covariance function of the time series at different lags τ. The spectral coefficient β can be obtained as slope of the fitted straight line (Bigger *et al.*, 1996). A process can be defined as long-range persistent if $h(f)$ scales asymptotically as a power law as $h(f) \sim 1/f$ (Bigger *et al.*, 1996; Hu *et al.*, 2001). The slope of β, H and α has the following relationship (Hu *et al.*, 2001);

$$H = (1 - \beta)/2; \quad \beta = 2\alpha - 1 \quad (8)$$

Like *DFA* and *R/S* analysis, β gives an indication of the behaviour of the time series. For any time series, β ranges between 0 and 2. If $\beta = 2$, time series is white noise and if $\beta = 1$, time series is pink noise, whereas long-range correlations are present in the time series if $0 < \beta < 1$ (Sprott and Rowlands, 1995).

STATISTICAL TRANSFORMATION OF DATA

In order to ascertain the origin of persistence in AQI time series at various locations, further statistical transformations are carried out to assess the impact of these changes in the time series. Two approaches based on differencing and shuffling of the time series are considered for this purpose. For differencing, the first order temporal correlations are removed by using $y_t = x_{t+1} - x_t$, where x_t is the AQI time series with $t = 1\text{---}n$ and y_t is the differenced time series. Differencing can help stabilize the mean of the time series by removing the changes in the level and thus eliminate trend and seasonality. Shuffling of time series is also suggested as one of the methods to filter out temporal correlations (Shi *et al.*, 2015). The procedure however preserves the distribution of the data. It is carried out by generating the time series of random numbers with length n equivalent to the length of AQI time series. The original AQI time series is then shuffled with respect to the random number time series by arranging the values in the corresponding order of random numbers.

RESULTS AND DISCUSSION

Air quality index observed during 2012–2014 at the study locations is given in Fig. 2. It can be observed that the AQI time series in Nagpur, Amravati and Chandrapur is lower than 200. The frequency distribution of the AQI categories for different locations is given in Table 2. It can be seen that in all the above three cities, AQI falls dominantly in Satisfactory and Moderately Polluted category. At Delhi, the most predominant category is Moderately Polluted and Poor. As AQI is computed by using the sub-indices of each pollutant and the worst represent AQI, it is observed that AQI is mainly dominated by PM_{10}.

DFA and *R/S* analysis is then carried out on AQI time series over different locations. Fig. 3 shows *logx* (i.e., logarithm of τ) vs. *logy* (i.e., logarithm of of *R/S*) of *R/S* analysis for two sites in five cities. Fig. 4 shows *logn* vs. *log* $F(n)$ for *DFA*. Hurst exponent, H is obtained as the slope of the straight line fitted to the plot of *logx* vs. *logy* and

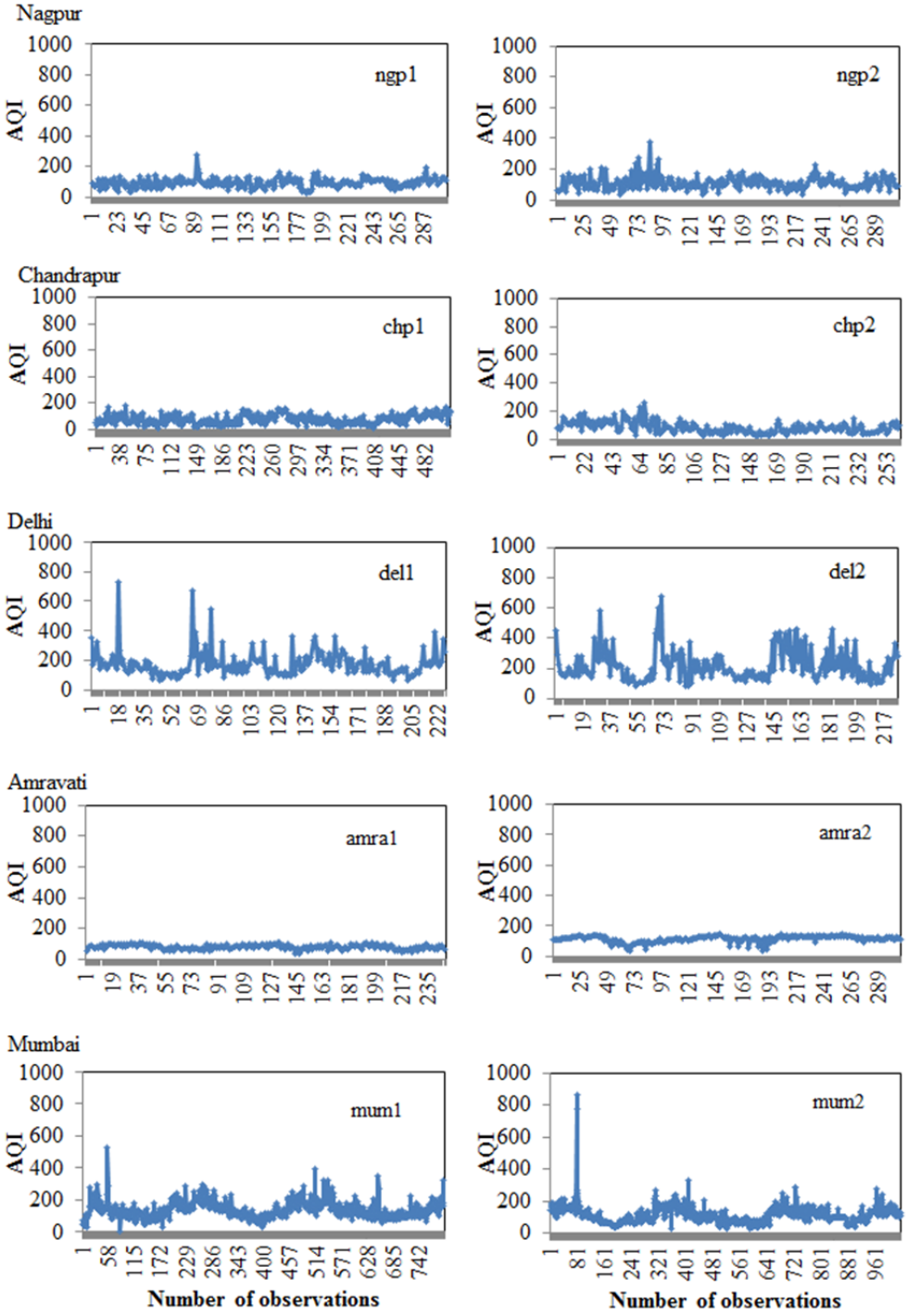

Fig. 2. Air quality index time series during 2012–2014 in five cities of India.

Table 2. Percentage distribution of AQI in different categories.

Category	Bin	Nagpur		Chandrapur		Delhi		Amravati		Mumbai	
		ngp1	ngp2	chp1	chp2	del1	del2	amra1	amra2	mum1	mum2
Good	0–50	7.5	2.9	27.0	22.1	0.0	0.0	1.6	1.6	1.6	5.9
Satisfactory	51–100	51.0	39.8	45.0	45.6	8.4	2.1	89.8	19.5	22.8	36.8
Moderately Polluted	101–200	41.2	54.0	28.0	31.2	68.6	52.6	8.5	78.9	65.1	54.1
Poor	201–300	0.3	2.9	0.0	1.1	15.5	26.9	0.0	0.0	9.7	3.0
Very Poor	301–400	0.0	0.3	0.0	0.0	6.2	9.8	0.0	0.0	0.6	0.1
Severe	401–500	0.0	0.0	0.0	0.0	0.0	6.4	0.0	0.0	0.0	0.0
	> 500	0.0	0.0	0.0	0.0	1.3	2.1	0.0	0.0	0.1	0.2

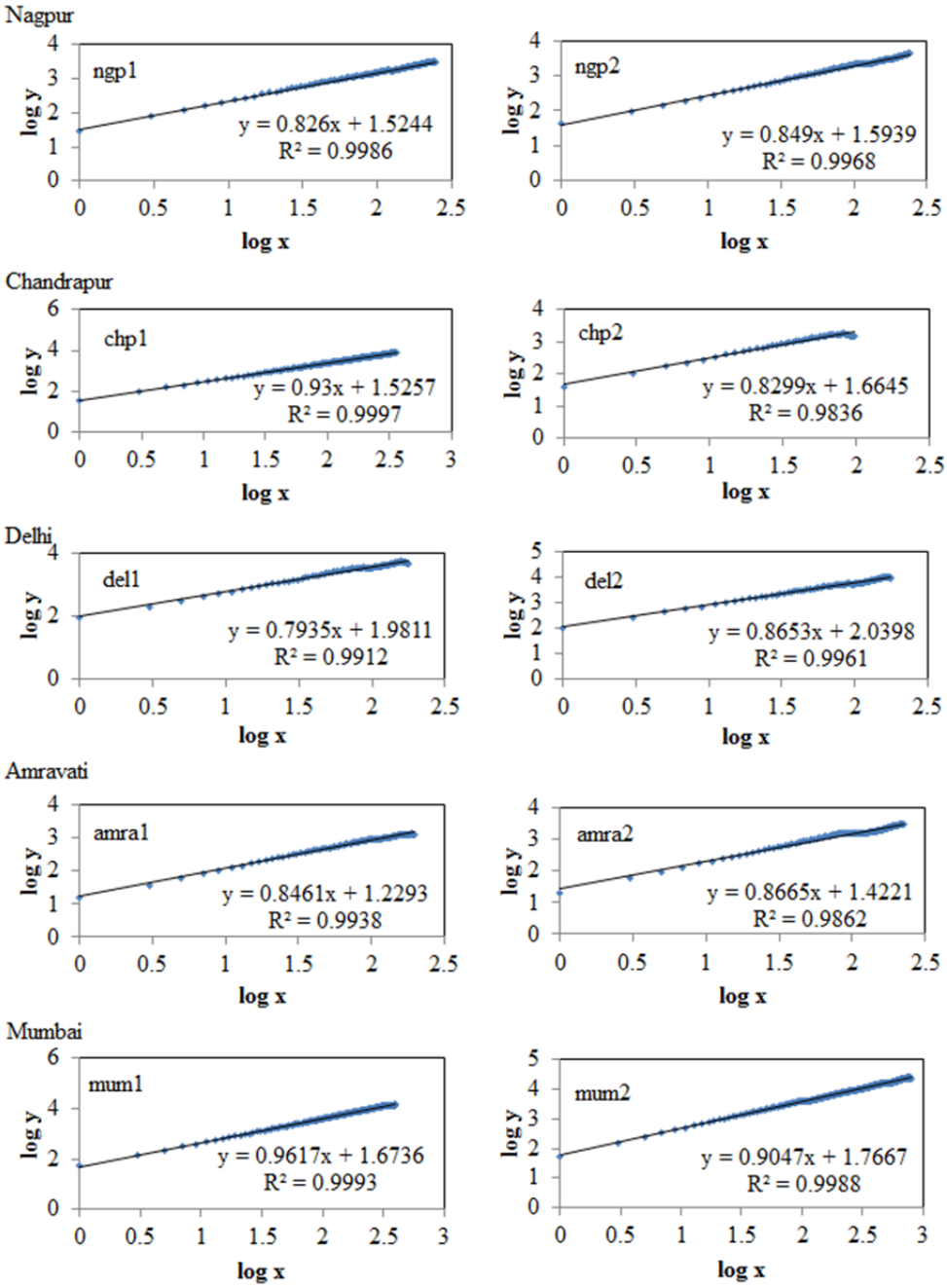

Fig. 3. *R/S* analysis of AQI time series for five cities. Logx indicates logarithm of τ and *logy* indicates logarithm of *R/S*.

scaling exponent, α is obtained as the slope of straight line fitted to the plot of *logn* vs. *logF*(*n*). For all the locations, *H* and $\alpha > 0.75$ is observed indicating strong persistence in AQI time series. Further, the nature of area does not influence the value of Hurst or scaling exponent using both methods. Scaling exponent β of power spectral density analysis is also computed as slope of power spectral density vs. frequency. It can be observed from Fig. 5 that the spectral density decreases with frequency. The slope β suggests the persistent behaviour in AQI time series for all the locations.

Maraun *et al.* (2004) and Varotsos and Efstathiou (2015) argued that power-law scaling may not be assumed a priori but needs to be established and proposed estimating the local slopes from *DFA*. The local slopes computed over an optimal moving window then need to be evaluated for constancy of scaling exponent. Further details on the choice of window size are given in Maraun *et al.* (2004). Considering the number of divisors on *x*-axis, local slopes are computed with moving window length of 3 for each location. The constancy of the slopes is evaluated by 95% confidence interval. The results are plotted in Fig. 6 for original time series. It can be seen that the scaling exponent is quite within the limits and maintain the constancy.

Scaling exponents of *R/S*, *DFA* and power spectral density analysis are also computed for the differenced time series at all the locations. Plot of *logx* vs. *logy* and *logn* vs.

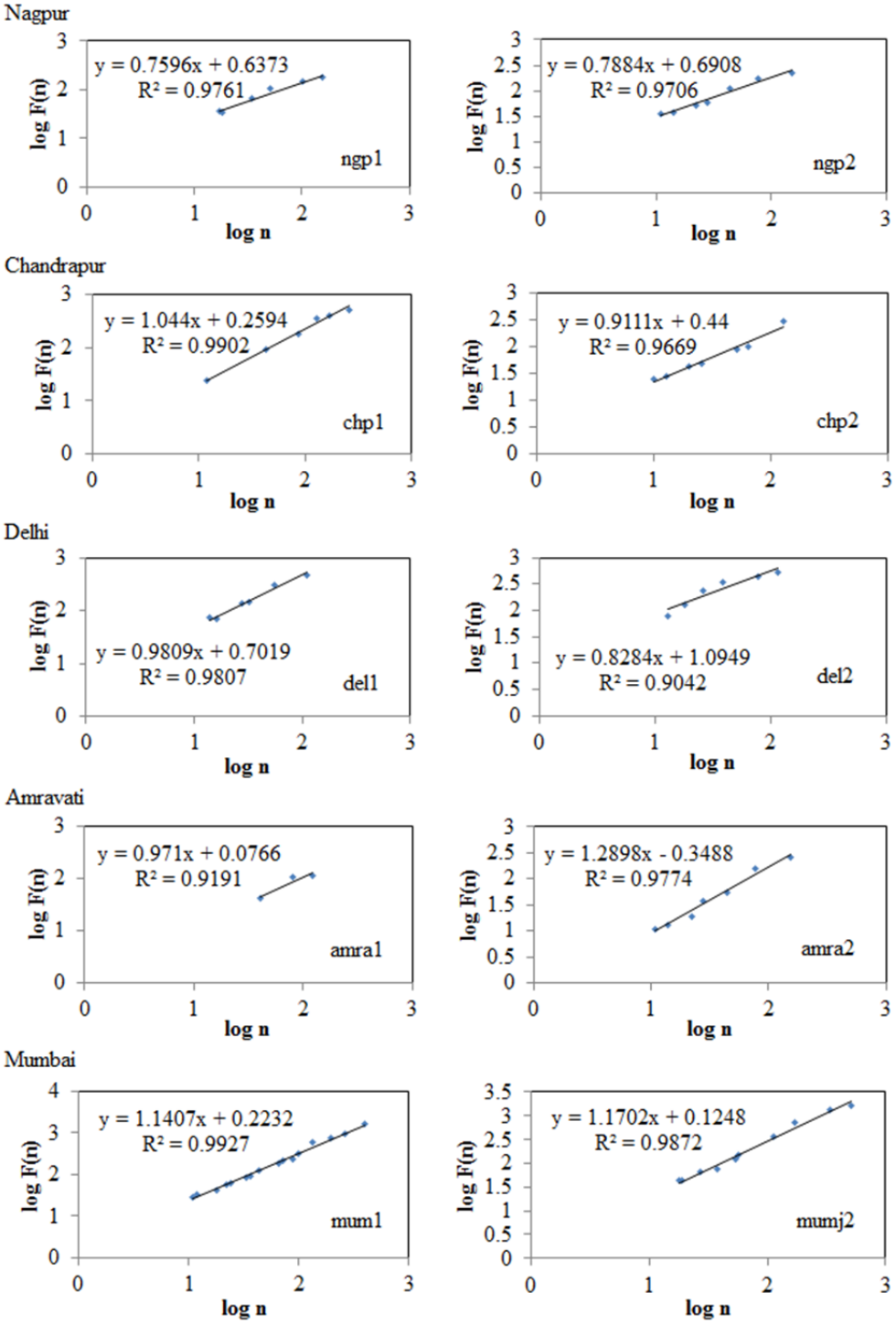

Fig. 4. Detrended fluctuation analysis of AQI time series for five cities.

log F(*n*) is given in Fig. 7 and Fig. 8. The power spectral density analysis results are given in Fig. 9. It can be seen that H is ≤ 0.5 at all the locations suggesting either random or anti-persistent behaviour of AQI time series after differencing. Scaling exponent, α of *DFA* on the other hand is observed to be < 0.3 at all the locations indicating anti-persistence in differenced AQI time series. Scaling exponent β of power spectral density analysis is observed to be > 1, which too suggests anti-persistence in differenced AQI time series. The results of *DFA* and power spectral density analysis are quite similar whereas *R/S* analysis shows random behaviour for few locations. The analysis also suggests that the persistence is certainly due to short-term temporal correlations in AQI time series. The random or anti-persistent behaviour after removing first order linear correlations show the significance of short memory. Short memory indicates that time series is correlated for only a short duration, which can be accounted for by the traditional techniques such as

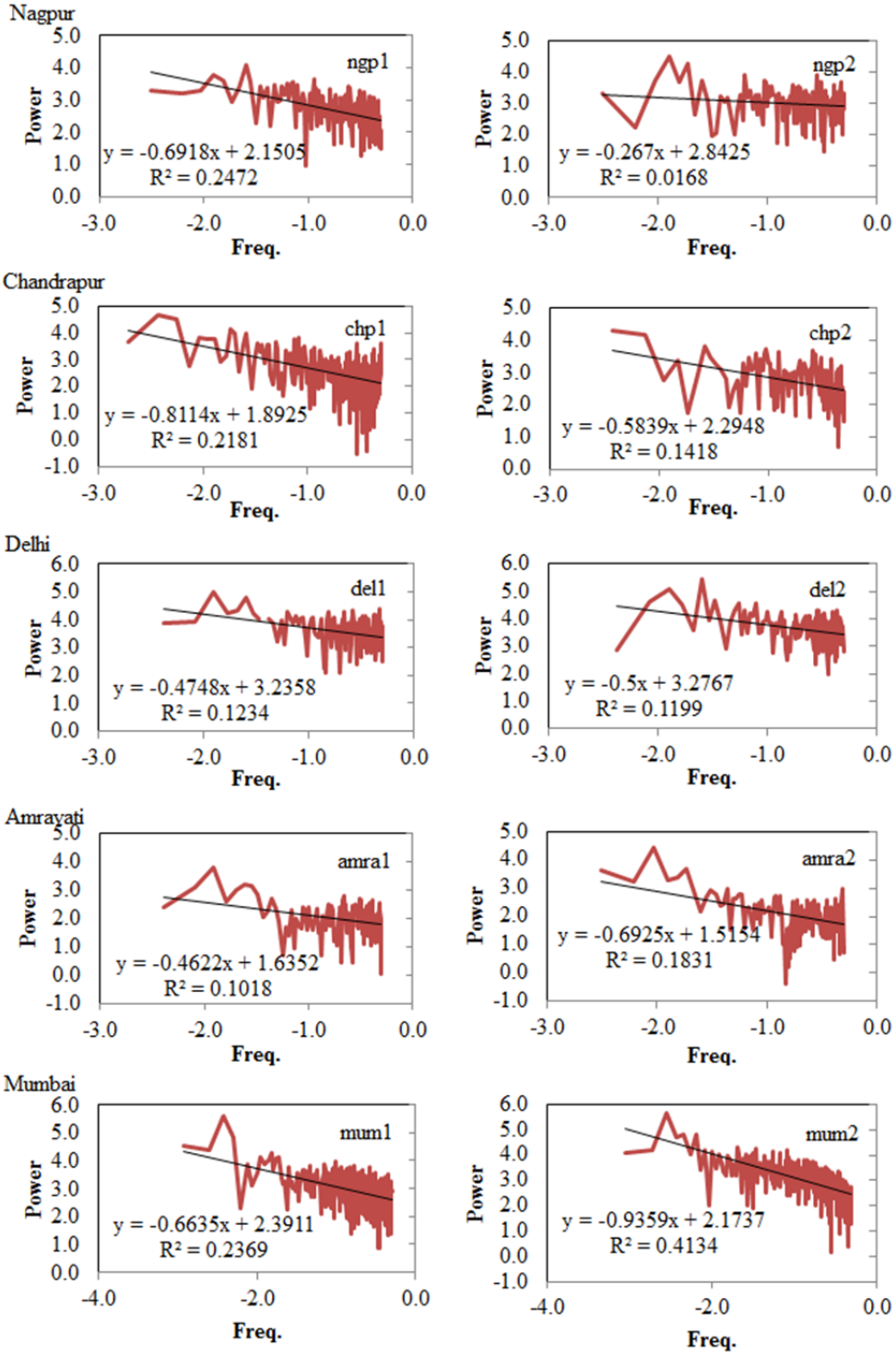

Fig. 5. Power spectral density analysis of AQI time series for five cities.

autocorrelation function. The effect of eliminating first order temporal correlations on Hurst exponent component is also studied in Hu *et al.* (2001) and Chelani (2009). Hu *et al.* (2001) studied the effect of trend superposed over correlated noise on scaling behaviour of noise. Different types of trends were considered and it was observed that presence of trend hinders the accuracy of understanding the correlation properties of the noisy data using *DFA* and *R/S* analysis. It was suggested to carry out trend removal exercise before applying any statistical technique related to persistence study.

For the shuffled AQI time series, scaling exponent H is observed to be > 0.7 for all the locations except for ngp1 and ngp2 in Nagpur. Scaling exponent α is observed to be approximately 0.5 except for the del1 site in Delhi where it is 0.98. Power spectral density analysis also shows the similar

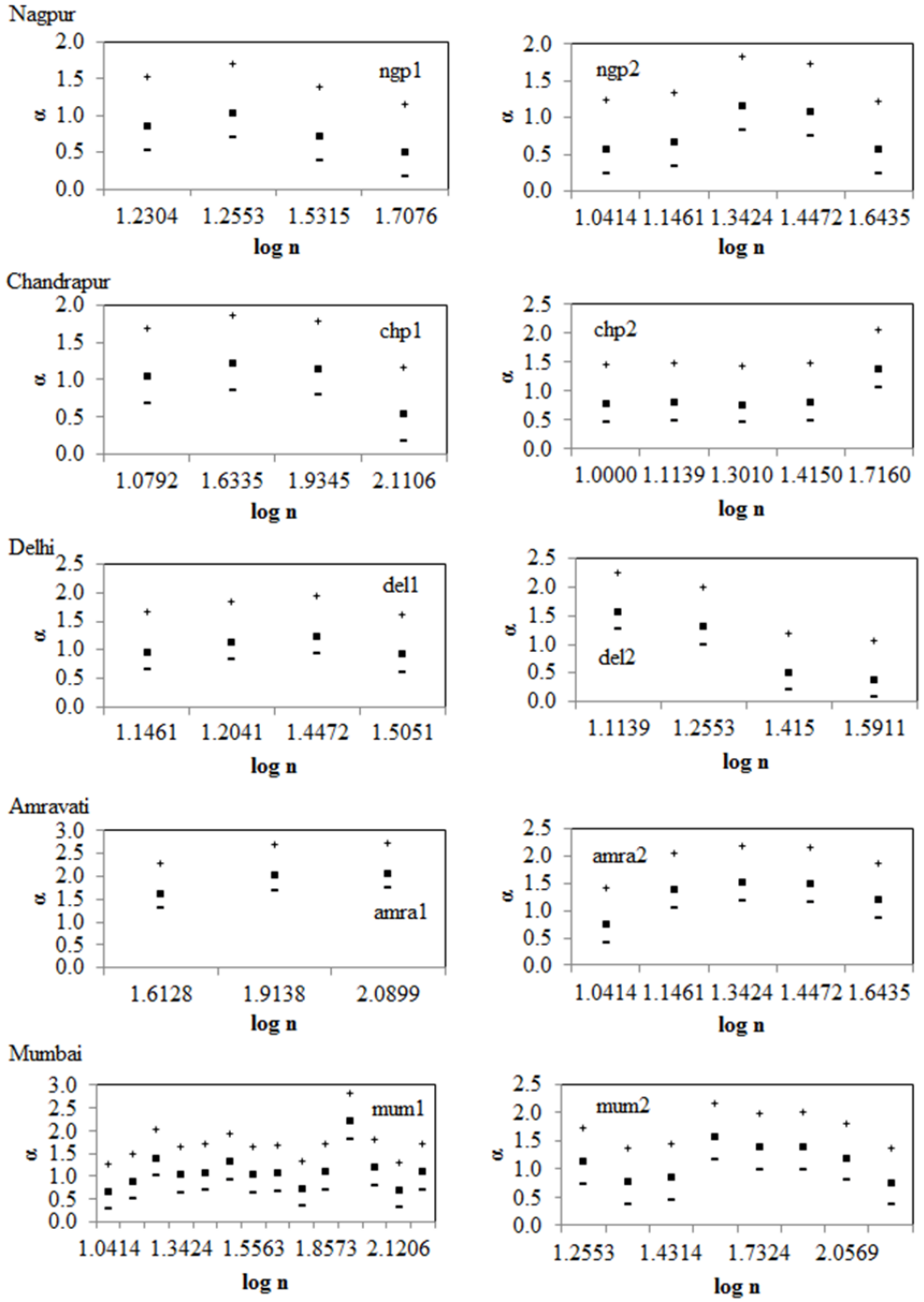

Fig. 6. Local slopes of *DFA* for original AQI time series.

results where $\beta < 0.127$ is observed at all the locations except for del1 site in Delhi which shows persistent behaviour of shuffled AQI time series. At del2 site in Delhi, $\alpha = 0.3$ is observed suggesting anti-persistence in AQI time series. For differenced and shuffled time series, local slopes were computed over the whole scaling region (results not shown here). It is observed that the scaling exponent is well within the confidence limits. The results of power spectral density analysis and *DFA* are somewhat closer, although the values of the scaling exponent are not. Apparently the behaviour of the time series is observed to similar by *DFA* and power spectral density analysis. *R/S* analysis on the other hand shows different behaviour of the AQI time series.

Kim *et al.* (2014) also observed the similar results while comparing various methods of scaling exponent estimation and found that *DFA* outperforms in estimating the appropriate scaling exponent for short and long term memory time series. Further when using differenced time series which removes any significant linear trends, it is observed that *R/S* method shows anti-persistent or even random behaviour of AQI time series for some locations, whereas *DFA* and power spectral density analysis show anti-persistent behaviour. This

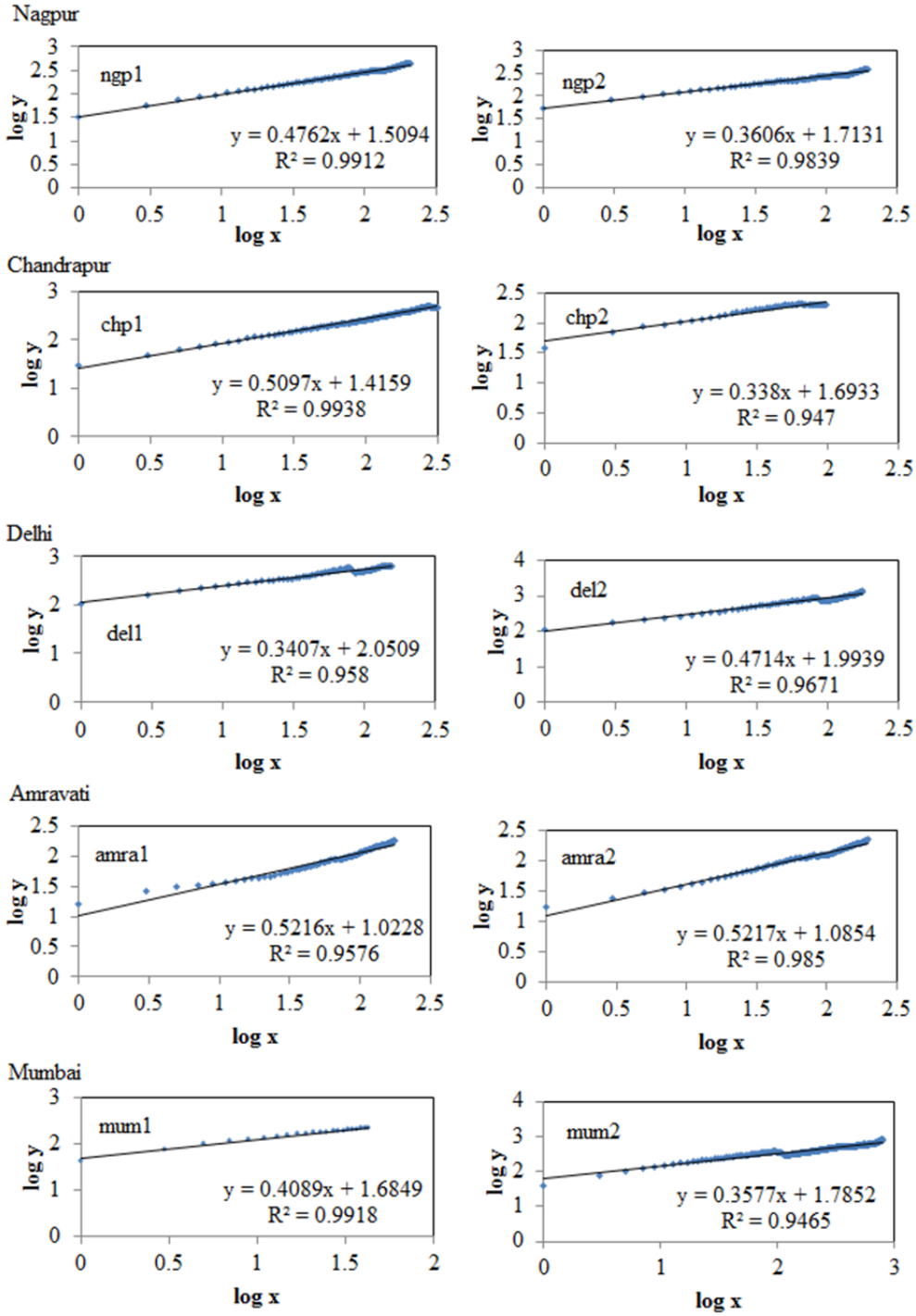

Fig. 7. *R/S* analysis of differenced AQI time series for five cities. Logx indicates logarithm of τ and *logy* indicates logarithm of *R/S*.

change in behaviour from long-term memory to short-term memory or randomness suggests that all methods account for first order temporal correlations and show larger scaling exponent (> 0.6) even if long-term correlations are not present in the time series. Even if the local slopes are computed, which is the criteria for constancy of the scaling exponent of DFA (Maraun *et al.*, 2004), it too shows the similar results as single scaling exponent for full scaling region. Another interpretation may be differencing distorts the time patterns in the original time series. A common logic suggests that differencing result in removing the first order temporal correlations and can thus lead to short-memory time series, which is well taken care of by the three methods.

For shuffled time series, *R/S* analysis did not show any change in the behaviour of the time series, whereas *DFA* and power spectral density analysis showed significant change. *R/S* analysis showed that the persistence in the original AQI time series as observed by the three methods

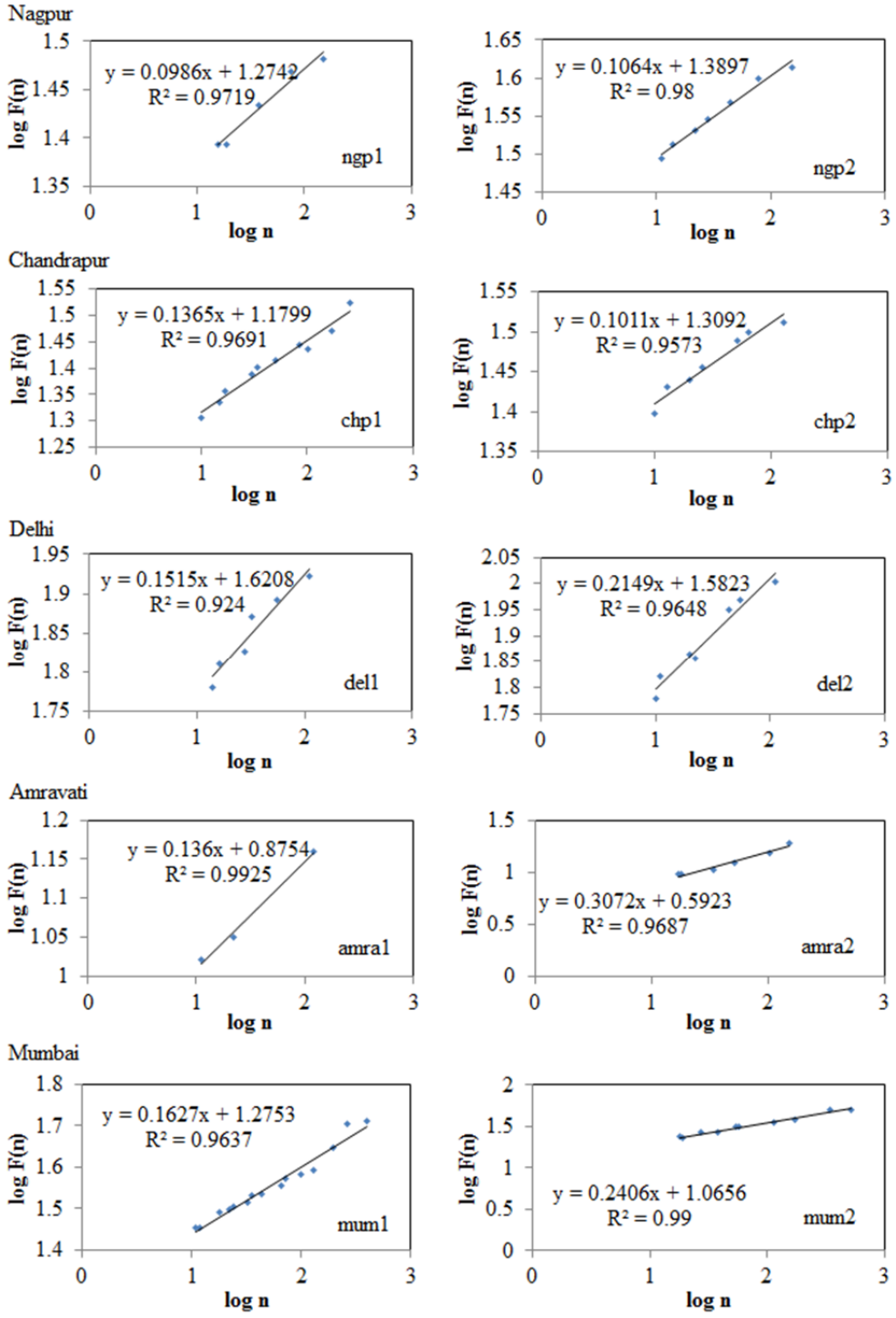

Fig. 8. *DFA* of differenced AQI time series for five cities.

were retained in shuffled time series, whereas other two methods showed that the long-range correlations were in fact not inherent in the time series. A careful investigation is therefore required before using any transformation of the time series while computing the scaling exponent by using any of the three methods. Rangarajan and Ding (2000) suggested that when $H < 0.5$ is observed, one should also perform a power spectral density analysis on the same data set. Verification of Eq. (8) can then be considered as the clue on the consistency of the results of *R/S* or *DFA*. In this study, the results of the *DFA* and power spectral density analysis appear to be closer. In summary, *DFA* and power spectral density analysis appear to be more suitable methods for analysing the persistence or long-range correlations in the time series as these two methods take care of appropriate transformations in the original time series and show quite similar results. *R/S* analysis on the other hand shows different behaviour.

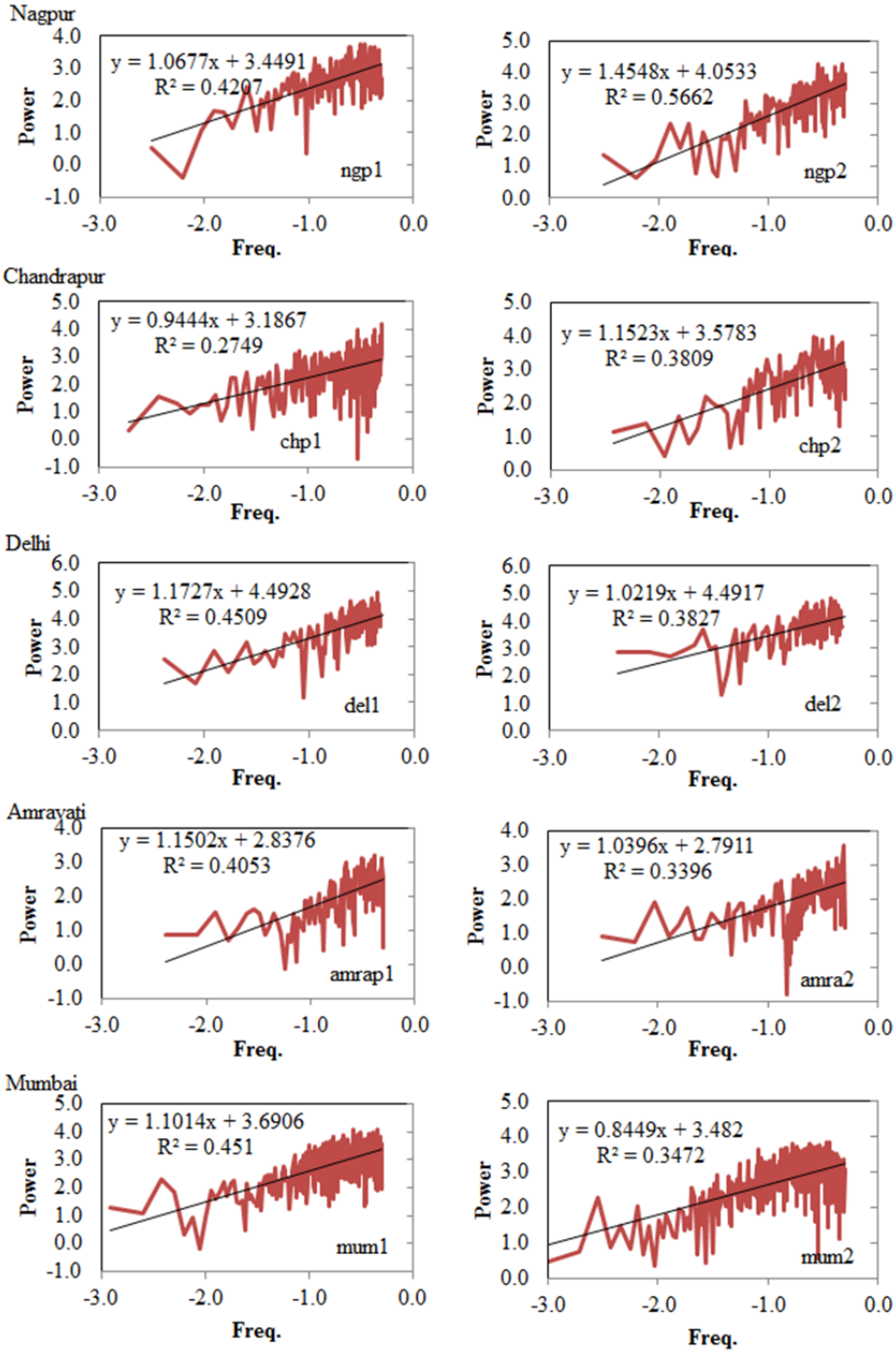

Fig. 9. Power spectral density analysis of differenced AQI time series for five cities.

The above exercise suggests that while developing the predictive model of air pollutant concentrations, if the short-range correlations are taken in to account, it will enhance the performance of the model as it accounts for the major portion of the original time series. It is also interesting to note that differencing results in the time series which is either random or anti-persistent in nature. In order to improve accuracy, the models can utilize this information on the nature of resultant time series. If, for example, after differencing, the persistent time series turns out to be random, only linear model accounting for linear temporal correlations will be sufficient. If after differencing, persistent time series turns out to be anti-persistent, model taking into account the linear temporal correlations and long-memory models accounting for negative correlations can be developed. In any case one can obtain the predictive model utilizing the linear temporal correlations either short or long range in nature with certain degree of accuracy.

CONCLUSION

Air quality index observed in five cities of India is analysed for the presence of long-range correlations using *DFA, R/S* and power spectral density analysis. The three methods indicated the presence of strong persistence in AQI time series. Statistical transformations such as differencing and shuffling are then carried out to examine the influence of linear temporal correlations on persistence or long-range correlation property of AQI time series. *R/S* analysis suggested that differenced AQI time series is either random or anti-persistent, whereas *DFA* and power spectral density analysis indicated anti-persistence in AQI time series. The analysis suggested that long-range correlation property is influenced by the presence of linear first order correlations in the AQI time series. For the shuffled AQI time series, *R/S* analysis showed the similar behaviour of the time series as the original one i.e., the presence of persistence whereas *DFA* and power spectral density analysis showed random behaviour except at few sites. *DFA* and power spectral density analysis take care of appropriate transformations in the original time series and show quite similar results. The persistence property is largely influenced by short-range correlations in the AQI time series. The incorporation of this information can enhance the performance of the models to forecast the air quality. The similarity of the results of *DFA* and power spectral density analysis suggests that both methods can be relied more than *R/S* analysis in studying the persistence property of the time series.

ACKNOWLEDGEMENT

The author is thankful to anonymous reviewers for constructive comments that helped improve the manuscript.

REFERENCES

Bigger, T., Richard, C., Steinman, A. B., Linda, M., Joseph L. and Richard, J. (1996). Power law behavior of RR-interval variability in healthy middle-aged persons, patients with recent acute myocardial infarction, and patients with heart transplants. *Circulation* 93: 2142–2151.

Chelani, A.B. (2009). Statistical persistence analysis of hourly ground level ozone concentrations in Delhi. *Atmos. Res.* 92: 244–250.

Chelani, A.B. (2016). Long memory in air pollutant concentrations. *Atmos. Res.* 171: 1–4.

Hu, K., Ivanov, P.C., Chen, Z., Carpena, P. and Stanley, H.E. (2001). Effect of trends on detrended fluctuation analysis. *Phys. Rev. E.* 64: 011114.

Hurst, H.E. (1951). Long term storage capacity of reservoirs. *Trans. Am. Soc. Civil Eng.* 116: 770–808.

Kim, B.S., Kim, H.S. and Min, S.H. (2014). Hurst's memory for Chaotic, SOI, and Tree-ring series. *Appl. Maths* 5: 175–195.

Mandelbrot, B.B. and Ness, J.W.V. (1968). Fractional brownian motions, fractional noises and applications. *SIAM* 10: 422–437.

Maraun, D., Rust, H.W. and Timmer, J. (2004). Tempting long memory – On the interpretation of DFA results. *Nonlinear Processes Geophys.* 11: 495–503.

Peng, C.K., Buldyrev, S.V., Havlin, S., **S**imons, M, Stanley, H.E. and Goldberger, A.L. (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E* 49: 1685–1689.

Rangarajan, G. and Ding, M. (2000). Integrated approach to the assessment of long range correlation in time series data. *Phys. Rev. E* 61: 4991–5001.

Shi, K., Liu, C. and Huang, Y. (2015). Multifractal processes and self-organized criticality of $PM_{2.5}$ during a typical haze period in Chengdu, China. *Aerosol Air Qual. Res.* 15: 926–934.

Sprott, J.C. and Rowlands, G. (1995). Physics Academic Software, American Institute of Physics.

Varotsos, C., Ondov, J. and Efstathiou, M. (2005). Scaling properties of air pollution in Athens, Greece and Baltimore, Maryland. *Atmos. Environ.* 39: 4041–4047.

Varotsos, C.A. and Efstathiou, M.N. (2015). Symmetric scaling properties in global surface air temperature anomalies. *Theor. Appl. Climatol.* 121: 767–773.

Received for review, April 25, 2016
Revised, June 30, 2016
Accepted, June 30, 2016