STATISTICAL PROPERTIES OF TIME SERIES OF AIR POLLUTANTS AND METEOROLOGICAL DATA

SILVANA MIÇO, KLAUDIO PEQINI, MARGARITA IFTI,

POLIKRON DHOQINA, DODË PRENGA

Faculty of Natural Sciences, University of Tirana, Albania

e-mail: silvana.mico@fshn.edu.al

Abstract

Fractal method and descriptive analysis were used in this study to examine the degree to which a time series is scale invariant. One year hourly average air pollution (PM_{10} and $PM_{2.5}$) series as well as other meteorological variables (temperature, wind speed, relative humidity, atmospheric pressure and radiation) time series were studied. The data were obtained from air monitoring station and the MeteoAlb station at Tirana. The series consisted of 6552 hourly average values during the period January 2013 to September 2013. Internal correlation within time series was examined for each variable, identifying the long–range dependency of the time series and annual periodicity. Probability distribution functions and standard statistical parameters (mean, coefficient of variation, skewness and kurtosis) are evaluated to examine the structure of time series. Hurst exponent was estimated using Rescaled Range Analysis method (R/S) and these series exhibited self-similarity on certain time scale. For PM10 and PM2.5 we found that Hurst exponents were 0.82 and 0.83 respectively, showing strong long-term correlations. The time correlation between air pollutants and meteorological variables was discussed based on correlation matrix.

Key words: Time series, fractal method, Hurst exponent, autocorrelation, scale invariance.

Introduction

Time series data are a collection of observations on the values that a variable takes at different times. There is a great variety of examples where data are recorded as time series ranging from financial to physical sciences (Salcedo *et al*., 1999). Air pollution concentrations and meteorological data are often structured as time series and are characterized by large fluctuations. Physical understanding of the complex temporal structure of the air pollutants has attracted increased attention. Recently, numerous statistical analyses were suggested to extract useful information from air pollution concentration and weather variables time series (Lee, 2002; Dai and Zhou, 2017; Yuval and Broday, 2010). Yuval and Broday (2010), found that temperature, wind speed, radiation and precipitation behave in a predicable way at a certain time scale range, whereas air pollutants at the longer time scales exhibit longer memory than meteorological variables. PM_{10} and $PM_{2.5}$

concentrations are significantly affected by the meteorological variables. Żyromski *et al*. (2015), investigated relationships between air pollutants and meteorological variables and found that air pollution is very sensitive to the individual meteorological parameters.

There are many statistical analyses conducted for particulate matter mainly focused on PM_{10} and $PM_{2.5}$ concentrations (Lee 2002, Yuval and Broday, 2010, Galindo *et al.*, 2010). Since PM10 and PM2.5 originate from different sources, the values of $PM_{2.5}/PM_{10}$ ratio can be used to characterize which source dominates. Higher ratios of $PM_{2.5}/PM_{10}$ prove the contribution of anthropogenic sources to particle pollution, whereas the smaller ratios indicate the predominance of coarse particles, which is mainly related to natural sources (Xu *et al*., 2019).

Previous studies conducted on PM10 and PM2.5 sources in Tirana have shown that major anthropogenic sources are diesel vehicle emissions, traffic generated fugitive dust from unpaved roads, construction activities and waste incineration (Mico et al., 2015, Totoni *et al.*, 2012, Civici 2003).

In this work we aim to investigate the statistical properties of $PM_{2.5}/PM_{10}$ ratios in addition to PM_{10} and $PM_{2.5}$ concentrations and meteorological data series to evaluate the correlation between them. A suggested approach starts with the statistical analyses of the collected data and evaluates correlations to the atmosphere conditions. Autocorrelation analysis and fractal methods are considered efficient tools to characterize the scale invariance of the time series. Fractal analyses methods provide a measure of the scale invariance or self-similarity related to the long range dependence in the time series.

Methodology

Data

Air pollutants in this study include PM_{10} and $PM_{2.5}$. The hourly averages pollutants PM_{10} and $PM_{2.5}$ data were supplied by the Institute of Public Health, Tirana, Albania. These data were collected at the Tirana air quality monitoring station. The station was located at the center of Tirana, a heavily populated and traffic related area. The temperature, wind speed, relative humidity, radiation and atmospheric pressure are the measured meteorological variables used in this study. These measurements were provided by MeteoAlb station located at the center of Tirana.

Original hourly averages used in this study include data values for the period from January 1, 2013 to September 30, 2013. Ideally, a specific time series should consist of 6552 hourly values during this period. There were no missing or negative values of meteorological data time series. However, due to instant errors our time series contain less than 6552 hourly values. PM_{10} series had 81 (or 1.2%) missing hourly values. $PM_{2.5}$ contained 81 missing hourly values and 38 negative values (or 1.8%). There are three options to

treat the missing values: to ignore them, to treat as zero values or to interpolate them. Although the missing values of PM_{10} and $PM_{2.5}$ time series consist a very small portion and they are well distributed throughout the period of study we preferred to replace them. Replacing missing or negative values by neighbor averaging is considered the best way in a stable series (Honaker, and King, 2010). For comparison purpose, 270 daily averages for the same months of the year 2012 were also used.

Usually time series statistical analyses require huge amount of data. To examine the similarity of two years time series data variability, firstly PM_{10} and $PM_{2.5}$ hourly data were converted to daily averages. Then, autocorrelation functions (ACFs) were calculated for each time series. The corresponding results presented in the Figure 1 shown that ACFs exhibit an easily observable periodicity for the period under study. So, we can use the hourly data collected over a period of 9 months to perform the statistical analyses.

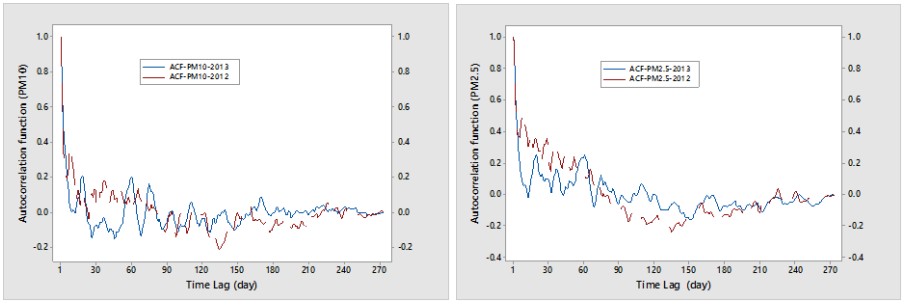


Figure 1. Autocorrelation function graphs for a) PM_{10} and b) $PM_{2.5}$ daily average time series at 2012 and 2013

Statistical Analysis

Statistical analysis of time series were performed using a two stage approach. We combined the statistical analysis used to evaluate autocorrelation functions with fractal method known as rescaled range analysis used to estimate Hurst parameter, H.

In the primary stage we used Rescaled Hurst Analysis to evaluate the degree of persistence in the data series. We evaluate the Hurst exponent, H, to quantify whether a time series exhibits (long) memory that enables the predictabilities. The next stage was based on the statistical analysis of autocorrelation functions and correlation matrices used to evaluate cross-correlation between particulate matter concentrations and meteorological variables.

Hurst Exponent: Rescaled Range Analysis

How consistent is a time series? Hurst ("Father of the Nile") developed a method known as Rescaled Range Analysis (R/S) studying the temporal

structure of the times series of Nile river level (Hurst, 1951). Hurst exponent, H**,** distinguishes a random series from one that is not (Mandelbrot, 2009). Hurst exponent is considered a test to qualify whether a time series exhibits long memory. If a time series is self-similar or not can be determined from the goodness of the fits of the log-log plots (Breslin and Belward, 1999). The exponent H is a real number that takes value in]0, 1 If $0.5 < H < 1$, the time series is persistent, that is, high present values are more probable to be followed by high future values, or the trend continues long into the future (long-term memory). If $H \cong 1$ the series is deterministic and appears in time series generated by long-term cyclic processes (Cadenas *et al*, 2019).

If $H = 0.5$, the time series is random and uncorrelated, with identical properties as a Brownian motion. In the case of $0 < H < 0.5$ anti-persistence can be shown, that is, low values are followed by high future values and vice versa. For each time series we applied the standard procedure of Hurst calculation (Sánchez Granero *et al.*, 2009). This procedure is based on the calculation of the expected value of the rescaled range for the time series, $E[R/S]_n$. As $N \to \infty$ the asymptotic relation holds: $E[R/S]_n = CN^H$ where N is the number of hours to calculate the mean value. Hurst can be calculated by using the linear regression of the function:

$$log\ (R/S)n = H\ log\ N + log\ C \qquad [1]$$

where H is Hurst exponent.

Autocorrelation Function

Correlation of a given value in a time series with past and future values is called autocorrelation coefficient. When this coefficient is calculated for different time lag values we obtain the Autocorrelation Function (ACF). This function is a way to measure and explain internal association between observations within a time series. If we have a sample X_t (t = 0, 1, 2,…, n), ACF can be calculated by the formula (Box and Jenkins, 1976):

$$Corr(X_t, X_{t-k}) = \frac{\sum_{t=1}^{n+k} (X_t - \bar{X})(X_{t-k} - \bar{X})}{\left(\sum_{t=1}^{n} (X_t - \bar{X})^2\right)^{1/2} \left(\sum_{t=1}^{n} (X_{t-k} - \bar{X})^2\right)^{1/2}} \qquad [2]$$

where k is the time gap being considered and is called time lag. ACF is a simple graphical method to test whether the adjacent values are correlated (for example #1 to #5 and #2 to #6, and so on for a lag k=4). ACF shows the internal association of a given time series. If the internal association within the time series of a variable at a certain period of time is very strong (positive or negative) an accurate predictability of this variable is possible (Salcedo *et al*., 1999, Dai and Zhou, 2017). If it is weak then there is no identifiable association and predictability is limited or impossible. In the Figure 2 are presented the time series constructed from the dataset of hourly averages of air pollutants and meteorological variables.

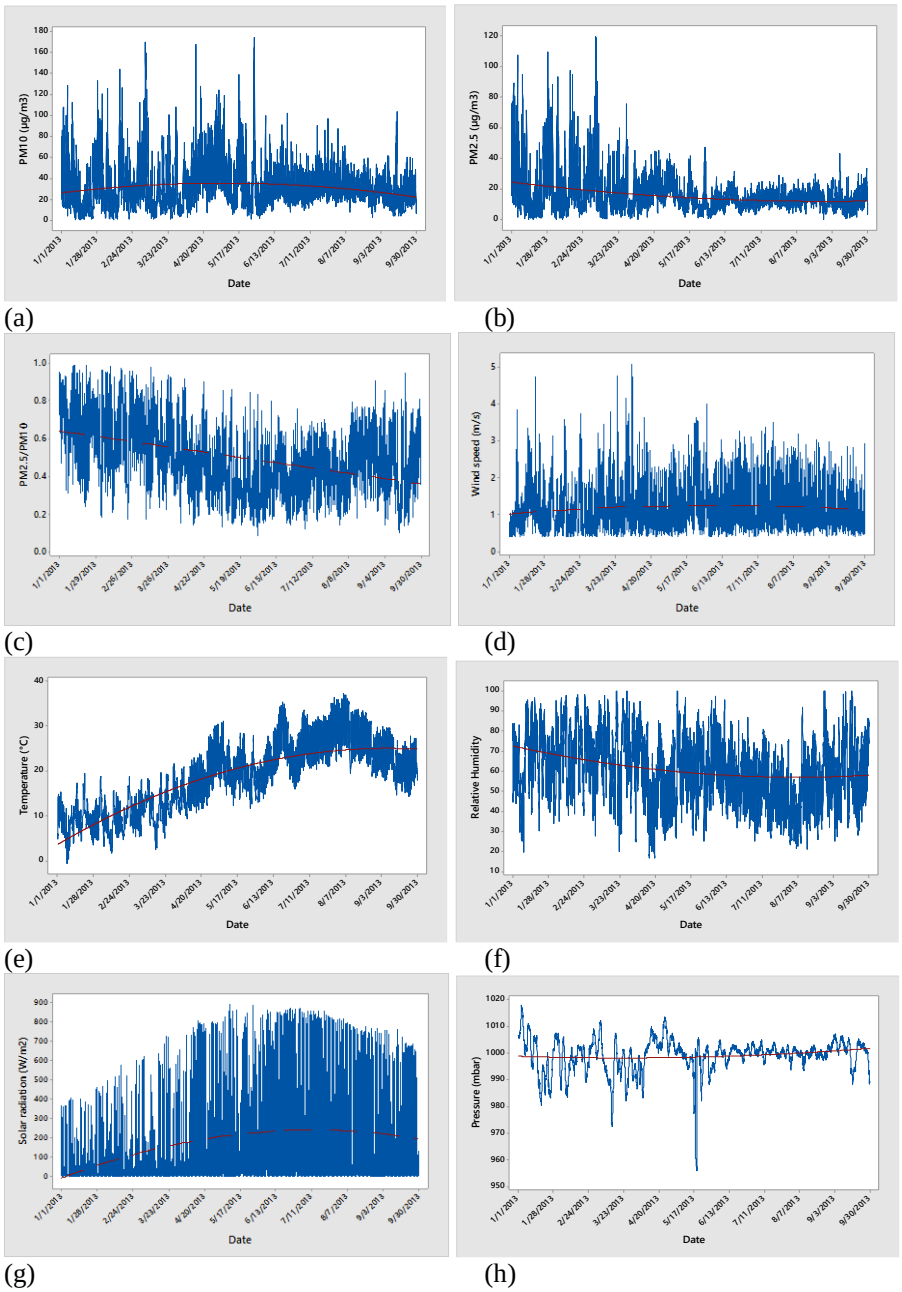


Figure 2. Time series constructed from the dataset of hourly averages of: (a) PM_{10}, (b) $PM_{2.5}$, (c) $PM_{2.5}$ / PM_{10} ratio, (d) wind speed -WS, (e) temperature -T, (f) relative humidity –RH, (g) solar radiation-SR, and (h) pressure –P values.

Results and discussions

Descriptive Analysis

To describe above time series we first evaluated the standard statistical parameters (mean, standard deviation, coefficient of variation and skewness).

Table 1. Descriptive Statistics of variables: PM_{10}, $PM_{2.5}$, Relative Humidity (RH), Temperature (T), Solar Radiation (SR), Pressure, (P) Wind Speed (WS).

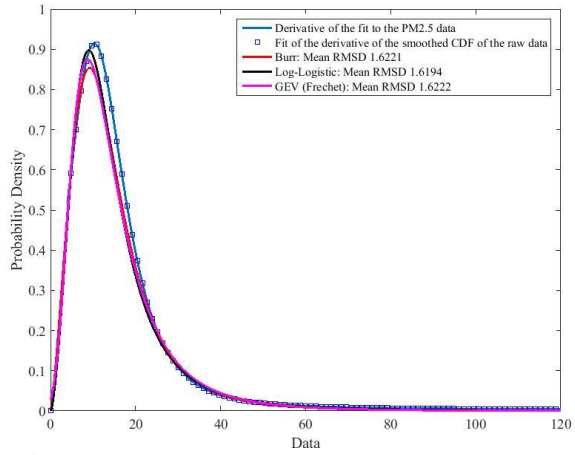
Variable	Mean	Standard deviation	Coefficient of variation (cv) %	Skewness (sc)
PM10 (µg/m3)	32.38	20.31	62.72	1.75
PM2.5 (µg/m3)	15.69	12.20	77.79	2.73
PM2.5/PM10	**0.50**	0.19	**38.92**	0.31
RH (%)	60.90	16.85	27.66	-0.08
T (°C)	18.68	7.77	41.58	-0.03
SR (W/m2)	179.47	251.18	139.96	1.25
P (kPa)	999.23	6.24	0.62	-1.56
WS (m/s)	1.19	0.63	53.45	1.31

The coefficient of variation indicates the degree of dispersion around the mean value. Skewness is a statistical parameter that evaluates the symmetry of distribution pattern, the degree and direction of departure from the symmetry. As shown in Table 1 the large difference between low and high concentrations results in the largest variation and skewness of $PM_{2.5}$ and PM_{10}. The mean value of $PM_{2.5}/PM_{10}$ ratio equal to 0.5 is an indication of the dominance of fine particle pollution (particle size with diameter diameter of 2.5 µm or less) over coarse particles (particles with diameter ranging from from 2.5 to 10µm), proving the high contribution of anthropogenic sources (Seinfeld and Pandis, 2006, Zhao *et al*, 2019). The smaller variability of ratios $PM_{2.5}/PM_{10}$ than that of both $PM_{2.5}$ and PM_{10} indicates the low variety of related sources and factors that influence the ratio. Since values of coefficients of skewness of PM series are positive they are all skewed to the right ($PM_{2.5} > PM_{10} > PM_{2.5}/PM_{10}$). The skewness values of meteorological variables are positive, skewed to the right (WS > SR), and negative, skewed to the left (P > RH > T).

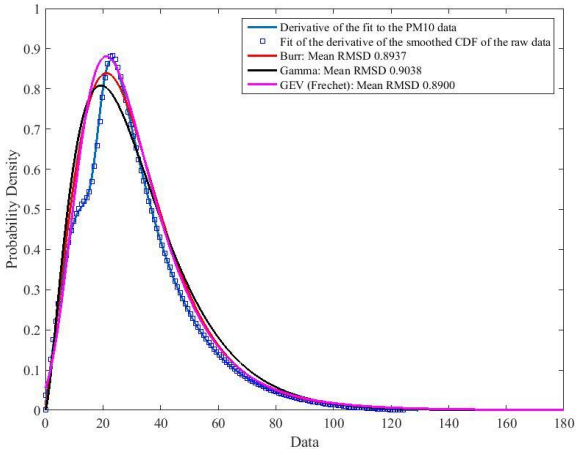
Probability Density Functions

The study of the data distributions is very important in statistics of time series. Probability distribution function (PDF) and cumulative distribution function (CDF) are important in statistics and closely associated. It is well known that CDF represent the cumulative values of PDF and PDF is simply the derivative of CDF. To avoid binning sensitivity we firstly proceeded with CDF of the series and then obtained PDF from calculating the derivative of CDF. We have fitted the series of particulate matter

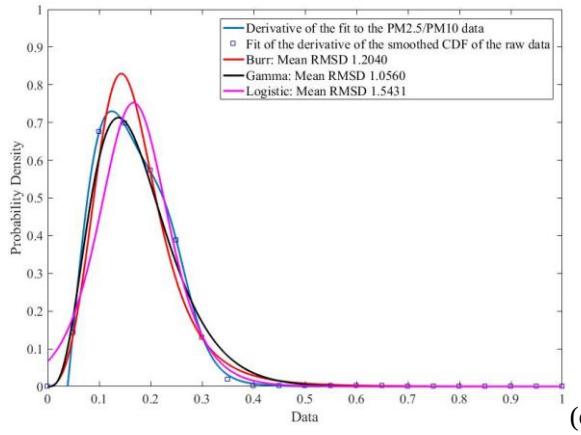
concentrations and the ratios with several theoretical distributions that are shown in the Figure 3.



(a)



(b)



(c)

Figure 3: The plot of PDFs of time series of: (a) PM_{10} (b) $PM_{2.5}$ (c) $PM_{2.5}/PM_{10}$ ratios.

No accurately fitting of our empirical distributions with a particular theoretical distribution was observed. Positively skewed to the right distributions (PM2.5> PM10> $PM_{2.5}/PM_{10}$ ratio) were obtained. We found that the most suitable compared to the other probability distributions were: $PM_{2.5}$– Log-logistic distribution, PM_{10}– Fréchet distribution and $PM_{2.5}/PM_{10}$ ratio– Gamma distribution.

Hurst Exponent Calculation

We calculated H for all time series by dividing them into equal subseries and estimating the statistical R/S for each segment. Linear regression resulting from the rescaled range analysis performed for all time series is shown in the Figure 4. Hurst exponent values are presented in Table 2.

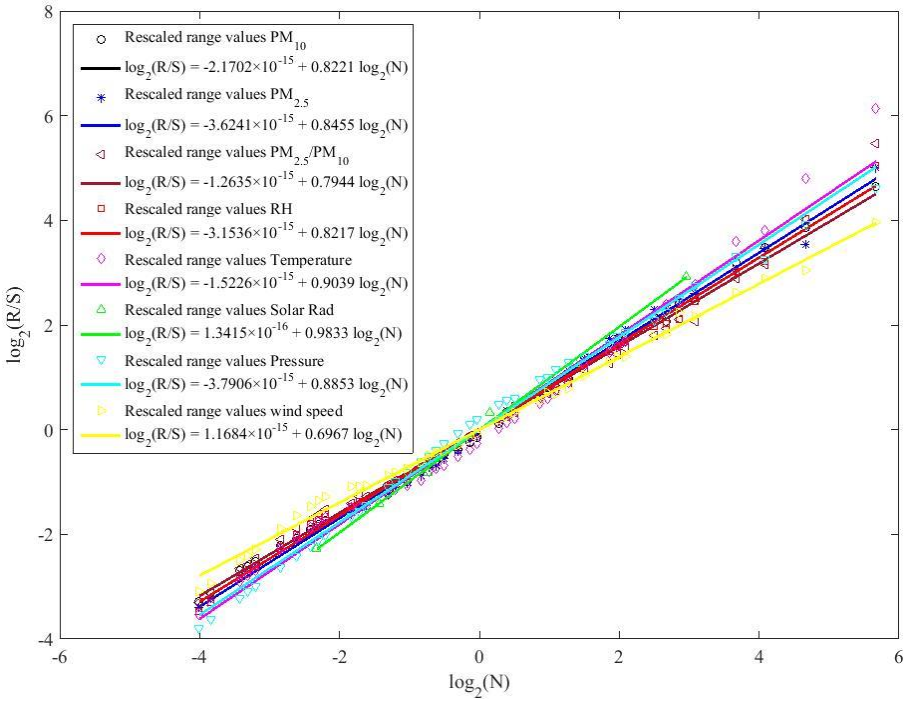


Figure 4. Calculation of Hurst exponent of PM and meteorological variable time series by the R/S method. N is the number of hours in each of the subseries.

Hurst exponent values are almost sufficiently greater than 0.5, showing a persistent behavior of our time series. Time series are not independent from time. PM10 series has more fracture than PM2.5. High values of Hurst for T, SR and P are an indication of long-term memory in the time series. Otherwise, low value of Hurst for wind speed time series presents a tendency toward randomness.

Table 2: Hurst values calculated using rescaled range analysis for time series.

Series	**PM_{10}**	**$PM_{2.5}$**	**$PM_{2.5}/PM_{10}$**	**RH**	**SR**	**T**	**WS**	**P**
H	0.79	0.84	0.84	0.82	0.98	0.90	0.69	0.88

Autocorrelation Function

Figure 5 shows the results of autocorrelation analysis. The time lag was made to run from 1 hour to 200 hours. All time series except from P series seem to have periodic behaviour. Behaviour of ACF is clearly identical for PM_{10} and $PM_{2.5}$, sinusoidal with 12 h period. $PM_{2.5}/PM_{10}$ ratio has similar behaviour as RH, T, WS, SD (sinusoidal with 24h period, due to alteration between day and night).

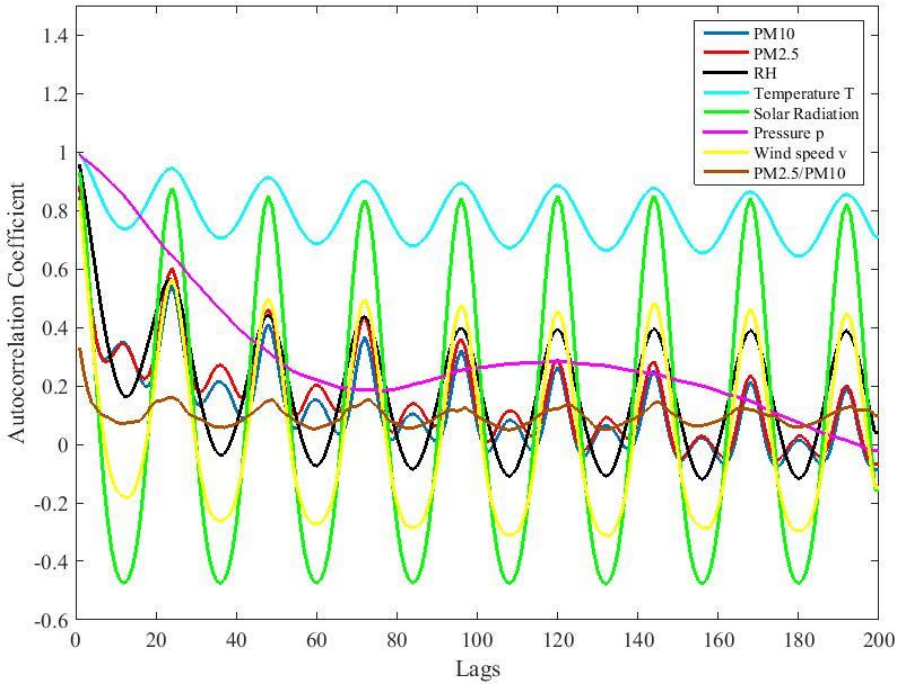


Figure 5. ACFs as the function of lags (in hours) for all series.

Solar radiation and wind speed varied more frequently than other variables. Pressure has different behaviour, presenting a slower exponential decay than other series. Pressure series is less independent in all time legs of 70 hours. This is in accordance with time series graphs presented in the Figure 2,

where pressure varies less frequently than others. For any value of time lag ACF of temperature is highest.

Cross-Correlation

The concentrations of PM_{10} and $PM_{2.5}$ were positively correlated with each other, where correlation coefficient is 0.76. Poor associations of $PM_{2.5}$ and PM_{10} concentrations with meteorological variables were observed by cross correlation analysis. It may be related by the performance of the cross-correlation function applied in the case of short time series (Vio and Wamsteker, 2001). Another important factor that can influence cross-correlations is the multifractal structure of meteorological time series (Baranowski *et al*, 2015).

Otherwise, the association of $PM_{2.5}/PM_{10}$ ratios with meteorological variables remained statistically significant (Table 3).

Table 3. Pearson cross-correlation for air pollutants and meteorological variables

	PM_{10}	**$PM_{2.5}$**	**$PM_{2.5}/PM_{10}$**
PM2.5	0.76		
PM2.5/PM10	-0.14	0.44	
RH	-0.26	0.11	0.61
T	0.09	-0.27	-0.59
SR	0.10	-0.22	-0.55
P	0.17	0.19	0.01
WS	-0.20	-0.35	-0.34

Relative humidity had a significantly positive effect on $PM_{2.5}/PM_{10}$. Since RH had negative influence on PM_{10} greater than positive influence on $PM_{2.5}$, higher RH get worse pollution mainly with $PM_{2.5}$. Temperature, solar radiation and wind speed had negative effect on $PM_{2.5}/PM_{10}$ ratios, while pressure had weak effect. The vertical structure of the atmosphere and vertical air movement are significantly controlled by temperature, solar radiation and wind speed (Seinfeld and Pandis, 2006). WS tends to eliminate air pollutants. WS was more efficient to $PM_{2.5}$ than PM_{10}, resulting in a negative correlation with ratios. SR had stronger effect on $PM_{2.5}$ than PM_{10} resulting in higher values of ratios. Lower values of SR means reduced vertical dispersion of pollutants. Higher values of temperature favour the dispersion of pollutants, affecting fine particles more than coarse particles (Xu *et al*, 2019).

Pressure is inverse proportional with temperature and positively correlated with particulate matter, affecting $PM_{2.5}$ slightly more than PM_{10}, but the other factors can influence, resulting in a poor correlation with $PM_{2.5}/PM_{10}$ (Dai and Zhou, 2017).

A determinant factor that influences the level of PM concentrations is the variation of PBL (Planetary Boundary Layer), which is significantly affected by meteorological factors (Mandija *et al*, 2017). The combination of long-range transport processes with local aerosol emissions also determines the level of PM concentrations. But these two sets of processes are influenced differently by meteorological factors and have different variations.

Conclusions

By using R/S and ACF statistical methods were obtained similar results. All time series showed strong persistence, except for pressure time series were lowest value of Hurst exponent was observed. $PM_{2.5}/PM_{10}$ ratio had similar behavior as meteorological data (sinusoidal with 24 hours period) and high correlation was found between RH, T, SR and WS. 12 hours period observed for PM10 and PM2.5 series is related with meteorological variations and human daily activities. Autocorrelation analysis is very helpful to identify the effect of meteorological variables on the $PM_{2.5}/PM_{10}$ ratios more than on PM_{10} and $PM_{2.5}$ separately. The study of $PM_{2.5}/PM_{10}$ can figure out the contribution of fine particles and coarse particles.

Harmonizing the uncontrolled meteorological variables by the controlled human activities air pollution reduction strategies can be improved. During the hours with higher ratios the measures should be focused on anthropogenic sources. During the hours with lower ratios measures should be focused on the coarse particle pollution. The study provides a resource for environmental agencies to focus reducing emission by integrating PM2.5/PM10 ratios with meteorological factors and air pollutant sources. Further studies taking into account the different factors influencing PM dispersion processes also using multifractal approach are necessary in order to better describe the statistical properties of PM and meteorological data time series.

References

Adães, J., Pires, J.C.M. (2019): Analysis and Modelling of PM2.5 Temporal and Spatial Behaviors in European Cities. Sustainability, 11, 6019.

Baranowski, P., Krzyszczak, J., Slawinski, C., Hoffmann, H., Kozyra, J., Nieróbca, A., Siwek, K., & Gluza, A. (2015). Multifractal analysis of meteorological time series to assess climate impacts. Clim Res. 65:39-52. doi.org/10.3354/cr01321

Breslin, M., C. and Belward, J., A. (1999): Fractal dimensions for rainfall time series. Mathematics and Computers in Simulation 48 437-446.

Box, G. E. P. and Jenkins, G.M. (1976): Time Series Analysis: Forecasting and Control (2nd ed.) Holden Day: San Francisco.

Cadenas, E., Campos Amezcua, R., Rivera, W., Espinosa Medina, M.A.,.. (2019): Wind speed variability study based on the Hurst coefficient and fractal dimensional analysis. Energy Science and Engineering 7 (2), 361-378.

Civici, N. (2003): Determination of elemental composition and probable sources of atmospheric aerosol in Tirana by EDXRF analysis. IAEA-CN-103.

Dai, Y. H, Zhou, W. X. (2017): Temporal and spatial correlation patterns of air pollutants in Chinese cities. PLoS ONE 12(8): e0182724.

https://doi.org/10.1371/journal.pone.0182724

Galindo, N., Varea, M., Gil-Moltó, J., Yubero, E., & Nicolás, J. (2010): The Influence of Meteorology on Particulate Matter Concentrations at an Urban Mediterranean Location. Water, Air, & Soil Pollution, 215(1-4), 365–372. doi:10.1007/s11270-010-0484-z

Honaker, J. and King G. (2010): What to do about missing values in time-series cross-section data. American Journal of Political Science, 54(2):561–581.

Hurst, H.E. (1951): Long Term Storage Capacity of Reservoirs. Transactions of the American Society of Civil Engineers, 116, 776-808.

Lee, C. K. (2002): Multifractal Characteristics in Air Pollutant Concentration Time Series, Water, Air, & Soil Pollution 135: 389.

https://doi.org/10.1023/A:1014768632318

Mandelbrot, B. B. 2(009): The Fractal Geometry of Nature, 3rd edn. New York, USA: W.H. Freeman and Company.

Mandija, F., Sicard, M., Comerón, A., Alados-Arboledas, L., ...(2017). Origin and pathways of the mineral dust transport to two Spanish EARLINET sites: Effect on the observed columnar and range-resolved dust optical properties. Atmospheric Research, 187, 69–83. doi:10.1016/j.atmosres.2016.12.002

Mico, S., Tsaousi, E., Deda, A., Pomonis, P. (2015). Electron microscopy characterization of individual aerosol particles nternational Journal of Ecosystems and Ecology Sciences, Vol. 5 (4): 615- 622

Mico, S., Tsaousi, E., Deda, A., Pomonis, P. (2015). Characterization of airborne particles and source identification using SEM/EDS. European Chemical Bulletin, Vol 4, No 4–6, https://doi.org/10.17628/ECB.2015.4.224-229

Salcedo, R. L. R., Alvim , F. M., Alves C., and Martins, F. (1999): Time series analysis of air pollution data. Atmos Environ., 33, 2361-2372.

Sánchez Granero, M. A., Trinidad Segovia, J. E., García Pérez, J. (2008): Some comments on Hurst exponent and the long memory processes on capital markets. Physica A, 387, 5543–5551. doi:10.1016/j.physa.2008.05.053

Seinfeld, J and Pandis, S. (2006): Atmospheric chemistry and physics: From air pollution to climate change. 2nd ed. Ed. John Wiley & Sons, Inc. New Jersey, USA. pp. 22-27

Totoni (Lilo), R., Prifti, L. and Mulla, E. (2012): Particulate Matter pollution, a continuing problem in Tirana air quality: status and trends. Asian Journal of Chemistry. Vol. 24, No.6, 2674-2678.

Xu, G., Jiao, L., Zhang, B., Zhao, S., Yuan, M., Gu, Y., Liu, J., Tang, X. (2019): Spatial and Temporal Variability of the PM2.5/PM10 Ratio in Wuhan, Central China. Aerosol Air Qual. Res. Volume: 2613-2624. doi: 10.4209/aaqr.2016.09.0406

Vio, R., and Wamsteker, W. (2001): Limits of the Cross-Correlation Function in the Analysis of Short Time Series. Publications of the Astronomical Society of the Pacific, 113(779), 86–97. doi:10.1086/317967

Yuval and Broday D. M. (2010): Studying the time scale dependence of environmental variables predictability using fractal analysis. Environ. Sci. Technol. 44, 4629- 4634. doi.org/10.1021/es903495q

Zhao, D., Chen, H., Yu, E. and Luo, T. (2019): PM2.5/PM10 Ratios in Eight Economic Regions and Their Relationship with Meteorology in China, Advances in Meteorology, Volume 2019, Article ID 5295726,

Żyromski A., Biniak-Pieróg M., Burszta-Adamiak E., Zamiar Z. (2014): Evaluation of relationship between air pollutant concentration and meteorological elements in winter months. Journal of Water and Land Development. No. 22 p. 25–32.