Transportation Research Part D 16 (2011) 57–60

Contents lists available at ScienceDirect

Transportation Research Part D

journal homepage: www.elsevier.com/locate/trd

Notes and comments

Complexity analysis of CO concentrations at a traffic site in Delhi

A.B. Chelani *

Air Pollution Control Division, National Environmental Engineering Research Institute (NEERI), CSIR, Nehru Marg, Nagpur 440 020, India

ARTICLE INFO

Keywords:
Lempel–Ziv complexity
Hurst exponent
Dispersion analysis
Traffic related CO pollution

ABSTRACT

The Lempel–Ziv complexity measure is used to examine the complexity in carbon monoxide concentration emissions over time. Rescaled range analysis and dispersion analysis is also carried out to confirm the findings. For this, time series of carbon monoxide concentrations observed between 2000 and 2009 at a traffic site in Delhi are used to examine separately the entire time series, and the series by years and months. The presence of complexity is observed in carbon monoxide concentrations over the period. The analysis showed the long-range correlations up to 15 months. The Lempel–Ziv complexity of time series over different years however indicates complexity in all the years except 2000, 2001 and 2006, reflecting the probable impact of control measures in Delhi.

1. Introduction

Carbon monoxide (CO) contributes indirectly to global warming and cause harmful effects to human health in terms of cardiovascular and respiratory disorder. The CO concentration time series in ambient air is shown to be nonlinear and chaotic in recent studies (Chelani and Devotta, 2007). Understanding of complexity in the time series can provide useful information on its nature, which can be used to develop more accurate and reliable predictions.

The complexity of a finite sequence is a measure on the extent to which the sequence resembles a random one (Lempel and Ziv, 1976). To ascertain the complex nature, air pollutant concentration time series are analyzed using correlation integral, detrended fluctuation analysis, rescaled range analysis (Chelani, 2009). Correlation integral provides the lower bound on the number of variables required to describe the system. The need for large data sets, however, often makes the use of this technique impracticable (Tsonis et al., 1993). Rescaled range analysis and detrended fluctuation analysis though allow the study of persistence or the random nature of the time series. These techniques have been used to examine the self-organizing behavior of air pollution time series (Lau et al., 2009); an important characteristics of complexity. Kolmogorov complexity and Lempel and Ziv complexity (LZC) measures have also been used to analyze the complexity in the finite sequences. Based on the symbolic representation of data, LZC is a measure of degree of irregularity in the time series; it neither requires large data sets nor the stationarity of the time series. LZC has not, though, been applied to analyze the complexity in air pollution processes. In this study, the LZC measure is applied to CO concentrations observed between 2000 and 2009 at traffic site in Delhi. To confirm the findings of LZC, rescaled range and dispersion analyses are also carried out.

2. The analysis

2.1. Study area

Studies of air pollution in Delhi regularly report high concentrations of NO_2, PM10, CO and ozone, and a continuous air quality-monitoring programme has been initiated by the Central Pollution Control Board in New Delhi. Twenty-four hourly

* Tel.: +91 07122249886.
E-mail address: ap_lalwani@neeri.res.in

1361-9209/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.trd.2010.08.008

CO data during 2000–2009 at Bahadur Shah Zafar Marg (ITO) site in Delhi is used here. ITO is a traffic intersection site with traffic volumes of between 113,000 and 176,000 vehicles per day. Commercial buildings are located near the site.[1]

2.2. Lempel–Ziv complexity measure

Lempel–Ziv complexity is a nonparametric measure of irregularity or disorder in a sequence of data. Based on the symbolic dynamics it is initially used in data compression. It defines subsets of a sequence and counts their occurrences. The time series x_i with mean $\bar{x}$ is first transformed into a new binary sequence s as,

$$s(i) = \begin{cases} 0 & \text{if } x_i < \bar{x} \\ 1 & \text{if } x_i \geqslant \bar{x} \end{cases} \tag{1}$$

The threshold of mean of the time series is specified in Eq. (1) to account for the concentrations above and below mean as an indicator of complexity (Zhang et al., 2001). The distinct patterns can be determined by arranging the sequence s from left to right and making subsequences by increasing one unit of time. This results in computing the complexity counter $c(n)$. To elaborate more, $c(n)$ can be generated using the comparison, copying and insertion by parsing from left to right of the sequence (Sen, 2009). The set of operations that need to be carried out on set s are given below;

- Let p and q be the two subsets of s and pq be the concatenation of p and q. The set $pq\Pi$ can be derived after removing the last digit of pq.
- Let $\upsilon(pq\Pi)$ be the vocabulary, i.e. the set of all possible words or digits of $pq\Pi$.
- For each copy and insert a new set s^* of p and q is to be created. If an insert is carried out, it is noted clearly by putting a letter + in the appropriate location.
- For the sequence $s = s(1), s(2)—s(r)–s(n)$, one starts with initializing $c(n) = 1$ as first digit is always unknown. For the second digit, i.e. $s = [s(1), s(2)]$, $p = s(1)$, $q = s(2)$, $pq = s(1)s(2)$, $pq\Pi = s(1)$, $\upsilon(pq\Pi) = s(1)$ and $q \notin \upsilon(pq\Pi)$.
- Generalizing the above, $p = s(1)—s(r)$, $q = s(r+1)$, $pq = s(1)—s(r)$, $s(r+1)$, $pq\Pi = s(1)—s(r)$.
- If $q \in \upsilon(pq\Pi)$, then q is a subset of $pq\Pi$ and renew q as $q = s(r+1), s(r+2)$ and again check whether it belongs to $\upsilon(pq\Pi)$. The procedure is repeated till $q \notin \upsilon(pq\Pi)$.
- If $q \notin \upsilon(pq\Pi)$, $q = s(r+1)s(r+2)—s(r+i)$ is a new sequence and not a subset of $\upsilon(pq\Pi)$. Hence an insert of one unit needs to be carried out in the complexity counter $c(n)$. In this case, p needs to be renewed as $s(1), s(2)—s(r+1)–s(r+i)$ and $q = s(r+i+1)$. If $q \in \upsilon(pq\Pi)$, $c(n)$ remains intact. The procedure continues until the last digit.

The different subsequences are counted in complexity measure $c(n)$, which can further be normalized to make it independent of data size,

$$C(n) = \frac{c(n)}{b(n)} \tag{2}$$

where $b(n)$ is the complexity measure of a binary random sequence with uniform probabilities,

$$b(n) = \frac{n}{\log_2 n} \tag{3}$$

2.3. Rescaled range analysis

The time series x_i is normalized by subtracting the mean $\langle x \rangle$ as;

$$z_k = \sum_{i=1}^{k} [x_i - \langle x \rangle_\tau] \tag{4}$$

where τ is the time lag and k is the discrete time. The slope of the curve of R/S against τ on log–log scale gives an estimate of Hurst (1951) exponent H, which determines the persistent or random nature of the time series. The range R and standard deviations S can be obtained;

$$R_\tau = \max_{1 \leqslant k \leqslant \tau} z_k - \min_{1 \leqslant k \leqslant \tau} z_k \tag{5}$$

$$S_\tau = \sqrt{1/\tau \sum_{i=1}^{k} (z_i - \langle z \rangle_\tau)^2} \tag{6}$$

The time series with $0.5 < H < 1.0$ implies a persistent time series, whereas $H > 0.5$ indicates the presence of strong persistence. $0 < H < 0.5$ implies anti-persistence and $H = 0.5$ indicates random time series, i.e. the time series is independent and uncorrelated. The more elaborate description is given in Chelani (2009).

[1] Missing values do not impact much on the LZC measure due to symbolic representation, but they have some impact on the rescaled range analysis. Hence missing data (about 15.6%) are filled out by mean of the whole series.