Technical Note
Effect of Autocorrelation on Applying Central Limit Theorem to Air Pollutant Concentration Time Series

Chung-Kung Lee*, Ding-Shun Ho, Chung-Chin Yu, Cheng-Cai Wang and Yun-Hua Hsiao

Department of Environmental Engineering, Van-Nung Institute of Technology
Chungli 320, Taiwan, ROC

One-year of hourly average air pollutant concentration (APC) observations, including CO, NO, NO₂, O₃, PM₁₀, and SO₂, was used to examine the effects of autocorrelation on the assumption of Central Limit Theorem (CLT) by calculating the confidence intervals of the data that were known to be dependent. Monte Carlo sampling was used to draw random samples of various sizes from the population (1000 groups of each size), and the sample means and standard deviation of these observed means were then evaluated. Even with small sample sizes, the average of all the means in each group and the observed standard deviation of the means were found to closely approximate the means of the overall population and the standard deviation predicted by CLT, respectively. Moreover, the above consistency was closely related to the coefficient of variation of the population rather than to the degree of long-range-dependence. These results were used to interpret why the right-skewed frequency distribution observed in the mutually dependent air quality data could be accurately described using the lognormal model derived from the CLT. The link between the lognormality and multifractal characteristics in APC time series was also discussed.

Keywords: Long-range dependence, Central Limit Theorem, coefficient of variation, lognormality, multifractality

1. Introduction

Statistical analysis of the APC data from each air quality monitoring station routinely reveals a right-skewed frequency distribution (Georgopoulos and Seinfeld, 1982) and long term memory (Horowitz and Barakat, 1979). For the right-skewness, some probability models have been used to fit APC frequency distribution, and the lognormal is selected most often (Georgopoulos and Seinfeld, 1982; Bencala and Seinfeld, 1976; Ott, 1995). Recently, the so-called successive random dilution theory has been proposed as a physical explanation for the occurrence of the lognormal frequency distribution (Ott, 1995). On the other hand, long term memory may be closely related to the self-similar or scale invariant properties in the APC time series. Our previous investigation (Lee, 2001) found scale invariance in APC time series using monodimensional fractal analysis. Furthermore, moment scaling analysis confirmed the existence of multifractal characteristics in APC time series. The origin of

*Corresponding author:
Tel.:+886-3-4342379 ext.64
Fax:+886-3-4622232
E-mail address: anthony@cc.vit.edu.tw
both the pronounced right-skewness and multifractal phenomena in APC time series was concluded to be explainable using the successive random dilution theory proposed by Ott (1995), and the dynamics of APC distribution process can be described as a random multiplicative process. The link between the multifractal characteristics in APC time series and successive random dilution theory, then the lognormality, is encouraging. However, some commentators argue that the lognormality is inferred directly from the CLT, namely, the observations are independent, and this may be violated for multifractal cascade processes that are closely related to the autocorrelation in the APC time series. Accordingly, it is important to examine the effects of autocorrelation on the predictions made using techniques that are based on the CLT. If the autocorrelation does not significantly influence the results, then this provides further evidence for the connection between the multifractal characteristics and lognormality in APC time series.

The assumption that data are mutually independent is frequently employed to calculate the confidence intervals from a set of observations with the aid of CLT. Since the APC observed at a given time is not independent of the concentration observed at the following observation time (i.e., the serial autocorrelation), however, the convergence toward the true mean usually tends to be more gradual than the CLT predicts. This study first confirmed the long-range dependence of APC time series using the variance method, and then compared the predicted (from CLT) and observed (from Monte Carlo sampling) confidence intervals for these APC data to examine the effects of autocorrelation on the accuracy of CLT based techniques. Finally, these comparative results will be used to provide some explanations for the coexistence of multifractal characteristics and lognormality in APC time series. Overall, the indirect approach used in this study represented an application of Ott's method (1995) to a different data set. This work does not attempt to make a comprehensive study of the effects of autocorrelation on the statistical estimates but instead simply aims to explain the mutual conflict existing in the APC time series, namely, the multifractal characteristics (closely related to the long-range dependence) and lognormality (based on the independence property). These examinations may give us some confidence in adopting multifractal cascade models with a lognormal distribution as an input to model and predict the APC data.

2. Materials and Methods

2.1 Data

To facilitate comparison, this study used the same hourly average ambient APC data as our previous investigation (Lee, 2001). The data were supplied by the Environmental Protection Administration of ROC, and were collected at the Chung-Shan air quality monitoring station, Taipei (Taiwan) from January 1998 to December 1998. The selected air pollutants included CO, NO, NO2, O3, PM10, and SO2. The data are described in detail elsewhere (Lee, 2001).
2.2 Methods

The above APC time series is first used to evaluate some standard statistical parameters such as coefficient of variation, skewness, and kurtosis and a variance-type estimator is then applied to all time series to examine the correlation between pollutant concentrations. The long-range dependency of the time series is examined using the Hurst parameter, $H$, with a higher $H$ indicating stronger long-range dependence. The standard method of estimating $H$ is employed; namely the aggregated series of order $m$ are obtained by dividing the original series $X$ into blocks of size $m$ and averaging the values of $X$ over each block, and the sample variance of the aggregated series $\bar{X}^{(m)}$ is then calculated. The same procedures are repeated for a number of $m$ values, and the variances are plotted as a function of $m$ on a log-log scale. If the graph appears to be a straight line, the slope can be used to estimate $H$ with the aid of the equation $\text{var} \, \bar{X}^{(m)} \approx \sigma_0^2 m^{2H-2}$ (see Fig. 1). A detailed description of the procedures may be found in (Teverovsky and Taqqu, 1997; Lee, 2001).

Since the year of observations is considered as the original population, both the population mean $\mu_0$ and standard deviation $\sigma_0$ can be evaluated. Monte Carlo sampling is then applied to draw random samples of various sizes from this population. Using this approach, the APC reading for each hour of the year is equally likely to be
Table 1. Statistical parameters of the examined APC time series.

<table>
<thead>
<tr>
<th>air pollutant</th>
<th>coefficient of variation</th>
<th>coefficient of skewness</th>
<th>coefficient of kurtosis</th>
<th>Hurst Parameter ($H$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO</td>
<td>0.56</td>
<td>1.7</td>
<td>8.2</td>
<td>0.80</td>
</tr>
<tr>
<td>NO</td>
<td>1.18</td>
<td>2.8</td>
<td>14.4</td>
<td>0.76</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>0.50</td>
<td>0.8</td>
<td>5.1</td>
<td>0.87</td>
</tr>
<tr>
<td>O$_3$</td>
<td>1.05</td>
<td>2.1</td>
<td>9.7</td>
<td>0.61</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>0.86</td>
<td>2.0</td>
<td>8.7</td>
<td>0.82</td>
</tr>
<tr>
<td>PM$_{10}$</td>
<td>0.81</td>
<td>2.7</td>
<td>15.5</td>
<td>0.82</td>
</tr>
</tbody>
</table>

selected. For each group, a particular hour cannot be reselected. This work tests six different sample sizes: $n = 9$, $n = 18$, $n = 36$, $n = 72$, $n = 144$, and $n = 576$, obtaining 1000 groups of each size. The arithmetic mean is calculated for each group of samples, thus producing 1000 means for the groups of size 9; 1000 means for groups of size 18; and so on. Furthermore, the standard deviation of these observed means is also calculated. This approach reveals how close the mean of these samples lies to the true mean $\mu_0$. Furthermore, because $\sigma_0$ is also known, it is also possible to see how well the standard deviation of the observed means agrees with $\sigma = \frac{\sigma_0}{\sqrt{n}}$ ($n$ represents the sample sizes), predicted by the confidence interval formula based on the assumption of independence.

3. Results and Discussion

Table 1 lists the standard statistical parameters calculated from the APC time series. Among them, the coefficient of variation indicates the variability of APC history, while the coefficient of skewness measures the relative skewness of APC frequency distribution with a positive value meaning that the distribution has a long tail extending to the right. As Table 1 lists, the variability of concentration history and degree of right-skewness increase with the order of NO$_2$ < CO < PM$_{10}$ < SO$_2$ < O$_3$ < NO and NO$_2$ < CO < SO$_2$ < O$_3$ < PM$_{10}$ < NO, respectively. On the other hand, the autocorrelation exists in all examined APC time series (Hurst parameter $H > 0.5$) and the degree of long-range-dependence increases with the order of O$_3$ < NO < CO < SO$_2$ = PM$_{10}$ < NO$_2$. The autocorrelation function presented in Fig. 2 is also consistent with the above results. Moreover, the autocorrelation is found to decrease slowly in a manner that is clearly not an exponential decay. This slow decay in the autocorrelation function indicates a temporal persistence that may be related to self-similar properties of the time series. Notably however, the autocorrelation plots can be roughly divided into three different groups, with CO, SO$_2$, PM$_{10}$, and NO belonging to the same group, while O$_3$, which indicates a clear diurnal nonstationarity (periodicity) and NO$_2$, which possesses a longer persistence, form their own individual groups, respectively.

Table 2 compares predicted and observed parameters for all APC time series, and lists some key features. The averages of the means in each group very closely approximate the overall population means, as predicted by the CLT equation for averaging processes. The coefficient of skewness of these observed means does not reveal obvious right-skewness. It is also found that the distributions become narrower as the sample size increases (see Fig. 3). Moreover, comparing the standard deviation of the means predicted by the CLT ($\sigma = \frac{\sigma_0}{\sqrt{n}}$) with the observed standard deviation of the means reveals an extremely good agreement. These results indicate that the formulas derived from CLT are good approximation despite these air quality data not being independent, and also provides a solid basis for calculating confidence intervals about the true mean of population. Finally, the difference between predicted and observed parameters is closely related to the magnitude of coefficients of variation, implying that the variability of APC data distribution rather than the long-range-dependence
Table 2. Comparison of predicted and observed parameters for a year of air pollutant concentrations observed at Taipei.

<table>
<thead>
<tr>
<th>Air pollutant</th>
<th>n</th>
<th>mean μ</th>
<th>population observed</th>
<th>obs.-popul./popul. (100%)</th>
<th>standard deviation σ observed</th>
<th>obs.-pred./pred. (100%)</th>
<th>skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>CO (ppm)</td>
<td></td>
<td>1.252</td>
<td>1.242</td>
<td>0.80</td>
<td>0.234</td>
<td>0.239</td>
<td>2.14</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1.252</td>
<td>1.249</td>
<td>0.24</td>
<td>0.166</td>
<td>0.167</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>1.252</td>
<td>1.248</td>
<td>0.32</td>
<td>0.117</td>
<td>0.114</td>
<td>2.56</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>1.252</td>
<td>1.252</td>
<td>0</td>
<td>0.083</td>
<td>0.082</td>
<td>1.20</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>1.252</td>
<td>1.249</td>
<td>0.24</td>
<td>0.059</td>
<td>0.059</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>1.252</td>
<td>1.252</td>
<td>0</td>
<td>0.029</td>
<td>0.028</td>
<td>3.45</td>
</tr>
<tr>
<td>NO (ppb)</td>
<td>9</td>
<td>24.24</td>
<td>24.228</td>
<td>0.05</td>
<td>9.537</td>
<td>9.416</td>
<td>1.27</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>24.24</td>
<td>24.071</td>
<td>0.70</td>
<td>6.744</td>
<td>6.850</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>24.24</td>
<td>24.318</td>
<td>0.32</td>
<td>4.768</td>
<td>4.917</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>24.24</td>
<td>24.36</td>
<td>0.49</td>
<td>3.372</td>
<td>3.255</td>
<td>3.47</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>24.24</td>
<td>24.35</td>
<td>0.45</td>
<td>2.384</td>
<td>2.329</td>
<td>2.31</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>24.24</td>
<td>24.235</td>
<td>0.02</td>
<td>1.192</td>
<td>1.142</td>
<td>4.19</td>
</tr>
<tr>
<td>NO₂ (ppb)</td>
<td>9</td>
<td>32.27</td>
<td>32.377</td>
<td>0.33</td>
<td>5.338</td>
<td>5.378</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>32.27</td>
<td>32.216</td>
<td>0.17</td>
<td>3.775</td>
<td>3.779</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>32.27</td>
<td>32.191</td>
<td>0.24</td>
<td>2.677</td>
<td>2.625</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>32.27</td>
<td>32.282</td>
<td>0.04</td>
<td>1.887</td>
<td>1.835</td>
<td>2.76</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>32.27</td>
<td>32.292</td>
<td>0.07</td>
<td>1.334</td>
<td>1.301</td>
<td>2.47</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>32.27</td>
<td>32.261</td>
<td>0.03</td>
<td>0.667</td>
<td>0.635</td>
<td>4.80</td>
</tr>
<tr>
<td>O₃ (ppb)</td>
<td>9</td>
<td>16.18</td>
<td>16.452</td>
<td>1.64</td>
<td>5.665</td>
<td>5.662</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>16.18</td>
<td>16.038</td>
<td>0.09</td>
<td>4.006</td>
<td>4.237</td>
<td>5.77</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>16.18</td>
<td>16.237</td>
<td>0.31</td>
<td>2.833</td>
<td>2.807</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>16.18</td>
<td>16.166</td>
<td>0.13</td>
<td>2.003</td>
<td>2.068</td>
<td>3.24</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>16.18</td>
<td>16.264</td>
<td>0.48</td>
<td>1.416</td>
<td>1.383</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>16.18</td>
<td>16.206</td>
<td>0.12</td>
<td>0.708</td>
<td>0.681</td>
<td>3.81</td>
</tr>
<tr>
<td>SO₂ (ppb)</td>
<td>9</td>
<td>4.793</td>
<td>4.769</td>
<td>0.50</td>
<td>1.368</td>
<td>1.349</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>4.793</td>
<td>4.815</td>
<td>0.46</td>
<td>0.967</td>
<td>0.981</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>4.793</td>
<td>4.793</td>
<td>0</td>
<td>0.684</td>
<td>0.663</td>
<td>3.07</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>4.793</td>
<td>4.781</td>
<td>0.25</td>
<td>0.484</td>
<td>0.464</td>
<td>4.13</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>4.793</td>
<td>4.789</td>
<td>0.08</td>
<td>0.342</td>
<td>0.346</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>4.793</td>
<td>4.793</td>
<td>0</td>
<td>0.171</td>
<td>0.168</td>
<td>1.75</td>
</tr>
<tr>
<td>PM₁₀ (µg/m³)</td>
<td>9</td>
<td>51.46</td>
<td>51.282</td>
<td>0.36</td>
<td>13.897</td>
<td>13.936</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>51.46</td>
<td>51.622</td>
<td>0.31</td>
<td>9.826</td>
<td>9.898</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>36</td>
<td>51.46</td>
<td>51.326</td>
<td>0.27</td>
<td>6.948</td>
<td>6.879</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>72</td>
<td>51.46</td>
<td>51.113</td>
<td>0.68</td>
<td>4.913</td>
<td>4.802</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td>144</td>
<td>51.46</td>
<td>51.404</td>
<td>0.12</td>
<td>3.474</td>
<td>3.281</td>
<td>5.56</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>51.46</td>
<td>51.483</td>
<td>0.03</td>
<td>1.737</td>
<td>1.687</td>
<td>2.88</td>
</tr>
</tbody>
</table>

determines the applicability of CLT to air quality data, and that care is needed when applying CLT to data sets with high coefficients of variation. An important practical implication of above analyses is that the formulas derived for independent observations can be used to calculate statistical characteristics for air quality data that are not independent. The above may provide a partial explanation for why the right-skewed frequency distribution in APC time series can be accurately represented using the lognormal model despite the APC data being mutually dependent. This also implies that the link between the multifractal characteristics and lognormality may be reasonable. Notably however, many studies have modeled the concentration probability density function using distributions other than lognormal and good reasons exist for believing that the probability density function is not lognormal (Hanna, 1984; Yee and Chan, 1997; Lewis and Chatwin, 1997;
Chatwin et al., 1995). On the other hand, the lognormality is well known to be merely one of possible methods of generating multiplicative cascades. Accordingly, to adopt the multifractal cascade model to simulate and predict the APC data with greater confidence, further comparison of the different multiplicative cascade models is needed to identify which is best, using different probability density functions as inputs (Lee et al., 2002; Schertzer et al., 1997).

Acknowledgments

The authors would like to thank the National Science Council of the Republic of China, Taiwan for financially supporting this research under Contract No. NSC89-2211-E238-001.

References

Ott W. R. (1995), Environmental Statistics and Data Analysis. Lewis, Boca Raton, USA.

Received for review, August 10, 2001
Accepted, November 12, 2001
AAQR-2002-09
Instructions to Authors

These instructions are provided for helping authors with the preparation of their manuscripts or papers. Submitted papers should report on new work or ideas. Full-length papers, critical reviews, technical notes, comments on published papers and short communications are acceptable for submission.

Manuscript
1. Papers must be written in English. Spelling and grammar should be checked by an English speaker if English is not the first language.
2. An original manuscript and three good-quality copies should be submitted to the Editors-in-Chief.
3. The manuscript must be typed in 12-point, double spaced on one side of A4 or 8.5” x 11” papers, and kept in the maximum length of 20 pages including text, illustrations, tables and references. A good-quality laser printer (or equivalent) is always used. All pages should be numbered on the bottom of each page.
4. Manuscripts should be submitted in the sequence: title, name of author, abstract, keywords, introduction of paper, text, conclusions, acknowledgements, appendices, references, tables and figures.

Title
1. Use informative and specific, but not exceed 20-word title.
2. Avoid chemical formula in title.

Name of author and affiliation
1. List the first, middle and last name of each author.
2. Include the fax, telephone, e-mail numbers and address for the corresponding author.

Abstract
1. Maximum of 250 words should be used in abstract.
2. The purposes, methods, contents, conclusions as well as new significant findings should be described in abstract.

Keywords
1. Use maximum five keywords to enforce literature retrieval of title
2. Keywords should define the topic of the paper and enforce literature retrieval in title.

References
1. All references to literature in the text should be quoted in the form Connell (1997), or Connell et al (1997) if there are more than two authors. The manuscript should be carefully checked to ensure that the information given in the text is exactly the same as that given in the reference list.
2. References should be arranged in alphabetical order (each author’s surname first).
3. Some examples for arranging the references:
   (i) Journals:
   (ii) Books:
   (iii) Edited books:
   (iv) Conference proceedings, symposia, etc.:
4. The title of the cited journal should be abbreviated as in the Word List of Scientific Periodicals. The title of the journal should be given in full if the correct abbreviation is not known.

Tables
1. Tables are submitted on separate pages. Captions to tables are provided on a separated list (with copies) as well with the manuscript.
2. Tables are numbered consecutively with Arabic numerals (Table 1, etc) in order referred to in text. The text should include references to all tables.
3. Excessive tabulation of data or large table should be avoided.
4. Each table should have a brief and self-explanatory title.
5. Tables should be presented in the simplest style. The vertical lines should not be used.
6. Explanations should be given in footnotes with superior italic Roman alphabets at the bottom of
the table, if it is necessary for the understanding of the table.

Figures
1. All figures are provided on separate pages. Each figure should have a caption and put the captions on a separate list (with copies) with the manuscript.
2. Figures should be numbered consecutively with Arabic numerals (Fig. 1, etc), according to their appearance in the text. References to each figure should be included in the text.
3. Figures should be set in text size, but be clear and large enough to allow to a single-column width.
4. Original drawings or good-quality printouts on glossy paper are only acceptable.
5. Color figures are avoided, unless paid for by the author.

Units
1. The use of S.I. units throughout is highly recommended.
2. Other metric units may be used if authors have the discretion of the Editor-in-Chief in advance.

Symbols
1. Define in text or in a list of notation where units or dimensions should be given.
2. Use mathematics type if it is possible.
3. Identify uncommon symbols in the first time they appear in the text, including Greek letters and other non-Latin symbols.
4. The name of chemical compounds should be typewritten in full at the first occurrence, with the abbreviation in parentheses.
5. Take care to clearly differentiate zero (0) and the letter (O), one (1) and the letter (l), times sign (×) and the letter (x).

Equations
1. All equations must be clearly typewritten.
2. Subscripts and superscripts should be clearly legible.
3. The meanings of all symbols must be defined immediately after the equation in which they are first used.
4. Equations should be sequentially numbered in parentheses at the extremely right of the line, according to their appearance in the text.
5. When referring to equations in the text, preface the number with the word “equ” or “eqns” and place the number within the brackets (eqn 1, etc).

Acknowledgements
Written in a separate section briefly before the references.

Electronic manuscripts
1. Authors are required to submit electronic manuscripts on 3.5” disks.
2. The following information must be clearly indicated on the disks:
   a. operating system
   b. word-processing package and version used.
   c. the name of the text files
   d. authors’ names
3. The traditional hard-copy manuscripts will be considered as the final version if there is even one difference between the hard-copy manuscripts and those submitted on the floppy disks.
4. The floppy disks will be sent back to the authors. If revisions are required, the returned disk to the authors should be revised and then resubmitted to the Editor-in-Chief, together with three revised hard-copy manuscripts.

Proofs and Reprints
1. Unless otherwise specified, proofs will be sent back to the authors for proofreading and correction.
2. Authors will be charged for any other substantial corrections.
3. Authors should return the page proofs of their articles within 48 hours of receipt.
4. Reprints can be ordered when proofs are returned.
5. The decision of the Editor-In-Chief is final.

Manuscripts Submission
All manuscripts will be submitted to Editor-in-Chief.

Editor-In-Chief: Prof. Chiu-sen Wang
Address: Department of Public Health,
National Taiwan University,
NO. 1, Jen-Ai Road, Section 1
Taipei 100, Taiwan
Tel.: +886-2-2341-0065
Fax: +886-2-2351-6701
e-mail: cswang@ccms.ntu.edu.tw

© Copyright 2002, All rights reserved.