Environmetrics 00, 1–21

DOI: 10.1002/env.XXXX

Analysis of air quality time series of Hong Kong with graphical modeling

F. Hu[a], Z. Lu[b], H. Wong[c*] and T. P. Yuen[c]

Summary: Identifying the interaction of air pollutants in each monitoring station and the inter-relationship between pollutants at different stations is an important issue in the management and control of air quality and pollutants. In this paper, a graphical model is utilized to analyze the air pollution in Hong Kong using the daily air pollution data from the three monitoring stations located at Tsuen Wan, Tap Mun and Tung Chung in Hong Kong. The model follows broadly the spectral analytic method proposed by Dahlhaus (2000), for determining the edges of a graph. The method extends a graphical model for analyzing multivariate data to multivariate time series, and in our case it is applied to time series over different spatial locations. We adopt the graphical model for time series to investigate the inter-relationship of air pollutants at the three stations in Hong Kong and the interaction between pollutants among the stations. The results obtained have good interpretations in terms of both geographical locations and chemistry.

Keywords: Air pollutants in Hong Kong; Air quality monitoring stations; Graphical model; Partial coherence; VAR model

1. INTRODUCTION

Graphical modeling in data science (Hastie, Tibshirani and Friedman, 2013, Ch. 17) has gained popularity since its introduction in statistics (Lauritzen, 1996; Edwards, 2000). Brillinger (1996), Dahlhaus, Eichler and Sandkühler (1997), Dahlhaus (2000), and Eichler

[a] *Department of Mathematics, Tianjin University, Tianjin 300072, China, and Department of Management Sciences, City University of Hong Kong, Kowloon, Hong Kong*

[b] *Southampton Statistical Sciences Research Institute, & School of Mathematical Sciences, University of Southampton, SO17 1BJ, U.K., and School of Mathematical Sciences, The University of Adelaide, SA 5005, Australia*

[c] *Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China*

[*] *Correspondence to: H. Wong, Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China. E-mail: heung.wong@polyu.edu.hk*

(2012) attempt to extend the graphical methodology to time series data. More recently, Jung *et al.* (2015), Khare *et al.* (2015), Qiu *et al.* (2015), and Wolstenholme and Walden (2015) proposed other methods in determining a graphical model. Tunnicliffe Wilson *et al.* (2015) consider graphical modeling by structural vector autoregressive processes. Due to the ubiquity of time series data, the extension is worthwhile and necessary. An important issue in these works is the problem of estimation and testing. Dahlhaus (2000) proposes a spectral analytic method for testing the existence of an edge between two vertices of the graph. Each vertex actually corresponds to a time series. In this paper, *our objective is to utilize a graphical time series model to analyze the air pollution in Hong Kong* using the daily air pollution data from the three monitoring stations located at Tsuen Wan, Tap Mun and Tung Chung in Hong Kong.

Since 1995, the Environmental Protection Department (EPD) of Hong Kong has collected data on the pollutants sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3) and respirable suspended particulates (RSP). In 2005, the EPD and the Guangdong Provincial Environmental Monitoring Centre (GDEMC) agreed to make an effort to monitor the air pollution status in Hong Kong and the Pearl River Delta region (PRDR), and data on the 4 pollutants were also collected in the PRDR. It covers the most populous region of Guangdong Province and has a population of more than 100,000,000 people. Sixteen stations were established throughout the region, and the 3 stations in Hong Kong are Tap Mun(TM), Tsuen Wan(TW) and Tung Chung(TC). The GDEMC, however, only put the data on the web on a monthly basis. That means the publicly available data from the EPD and the GDEMC are different, with much less information from Guangdong available. The data sets should be very useful for the public to analyze the status of pollution in Hong Kong and PRDR. To our knowledge, very few researches have been done on them.

In this paper, we use *daily* time series data from Hong Kong to analyze the air pollution of Hong Kong. With sufficient data, we are able to apply the test of Dahlhaus (2000). As a comparison, the vector autoregressive model (VAR) is also applied to analyze the data. Our goal is to study the inter-relationship between the three stations and between the pollutants. This will help identify the hot spot of pollutants for management and control of air quality. The results obtained have good interpretations in terms of both geographical locations and chemistry. In a companion paper, we are working on the analysis of air pollution using the available *monthly* data from more stations in the region, based on the Generalized Dynamic Factor Model of Forni *et al.* (2000), which is good for the analysis of short panel time series data.

The rest of this paper is organized as follows: Section 2 gives a description of the data, including geographical and chemical background. Section 3 explains the graphical models and testing method used, while Section 4 gives the analysis, results and interpretations. Section 5 concludes. Additional details including figures and tables are available in an online supporting information file. In our subsequent analysis the main software used is MATLAB. The program codes and the dataset can be obtained from the online supplements.

2. THE DATA SET FOR GRAPHICAL MODEL

The time series data can be found on the website of the EPD of Hong Kong[†]. These are also included as a csv file in the online supplements of this paper. The time series data plots of the daily average for the four pollutants, SO_2, NO_2, O_3 and RSP, from September 2010 to September 2014 over the 3 monitoring stations at Tsuen Wan, Tap Mun and Tung Chung are shown in Figure 1. In each figure, lines constructed from the LOESS smoother (Cleveland, 1979) are also shown. These plots show that the data have clear seasonal pattern. Each series has a length of 1491. It is observed from the boxplot, Figure 2, that the concentration of SO_2 and NO_2 in Tsuen Wan are higher than that in Tap Mun and the concentration

[†] `http://epic.epd.gov.hk/EPICDI/air/station/?lang=en`

of O_3 in Tap Mun is higher than Tung Chung and Tsuen Wan. O_3 is produced through photochemical reactions between nitrogen oxides (NO_x, all mono-nitrogen oxides including NO_2) and volatile organic compounds (VOCs) under sunlight where NO_x and VOCs are emitted from vehicles. O_3 also reacts with NO_x, producing NO_2 and oxygen. Tap Mun is a rural area with light traffic and Tsuen Wan is an urban area with heavy traffic. Tung Chung is interestingly a semi-rural area. That explains the fact that SO_2 and NO_2 concentration is higher in Tsuen Wan than that in Tap Mun. Given the abundance of its precursors (VOCs and NO_x) in Tsuen Wan which react with and remove ozone in the air, the O_3 level measured at Tsuen Wan station is lower than the Tap Mun station.

[Figure 1 about here.]

[Figure 2 about here.]

In our subsequent analysis we need the time series to be weakly stationary. The series plots, Figure 1, show that all the series have a clear annual cycle of about 365 days. Besides seasonality, the marginal variances of the time series also seem to change over time. We first adopt the Box-Cox transformation (Box and Cox, 1964) on each time series. The Box-Cox transformation has the form

$$T(y_t) = \begin{cases} (y_t^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \log(y_t), & \lambda = 0. \end{cases}$$

The selected parameters are shown in Table S1 of the Supporting Information.

The transformed series are then deseasonalized using the method described in McLeod and Gweon (2012) with the R package "deseasonalize". The purpose is to remove the annual effect in the series. After deseasonalization, the SO_2 series at Tap Mun still shows some evidence of non-stationarity by observing the time series plot and ACF of the deseasonalized series. We further apply first-order differencing on the series. We check the correlograms of the 12 time series after these processing, which show that the time series are weakly stationary. The interested readers may look at these correlograms in Figures S3 – S6.

3. SPECTRAL ANALYSIS AND GRAPHICAL MODELS FOR MULTIVARIATE TIME SERIES

3.1. Frequency domain, and linear filters

Following Priestley (1981), let $X(t)$ be a zero-mean weakly stationary time series. Let the auto-covariance function of $X(t)$ be $\gamma_X(\cdot)$ such that $\sum_{u=-\infty}^{\infty} |\gamma_X(u)| < \infty$.

The spectral density of $X(t)$ is the function $f_X(\cdot)$ defined by

$$f_X(\lambda) = \frac{1}{2\pi} \sum_{u=-\infty}^{\infty} e^{-i\lambda u} \gamma_X(u), \quad -\pi \leq \lambda \leq \pi.$$

Further, $X(t)$ admits the representation

$$X(t) = \int_{-\pi}^{\pi} e^{i\lambda t} dZ^X(\lambda), \tag{1}$$

where $Z^X(\lambda), -\pi < \lambda \leq \pi$ is a complex-valued process with uncorrelated increments. Equation (1) is called the spectral representation of the process $X(t)$.

Let $Y(t)$ be another zero-mean weakly stationary time series. Similarly we define $f_Y(\lambda)$. Suppose now $Y(t)$ and $X(t)$ can be regarded as input and output processes such that

$$X(t) = \sum_{u=-\infty}^{\infty} g(u) Y(t-u), \tag{2}$$

where $g(u)$ is a deterministic sequence. There should be a noise term in (2) for our later work. But to focus on the exposition of the main concept, we drop the noise term here. Then

$\Gamma(\lambda) = \sum_{u=-\infty}^{\infty} g(u)e^{-i\lambda u}$ is called the transfer function, and we have the relationship

$f_X(\lambda) = f_Y(\lambda)|\Gamma(\lambda)|^2.$

Suppose we now have p inputs $Y_1(t), \cdots, Y_p(t)$. We generalize (2) by

$$X(t) = \sum_{u=-\infty}^{\infty} g_1(u)Y_1(t-u) + \cdots + \sum_{u=-\infty}^{\infty} g_j(u)Y_j(t-u) + \cdots + \sum_{u=-\infty}^{\infty} g_p(u)Y_p(t-u), \quad j = 1, \cdots, p. \tag{3}$$

Now consider the spectral representations,

$$X(t) = \int_{-\pi}^{\pi} e^{i\lambda t} dZ^X(\lambda), \text{ and } Y_j(t) = \int_{-\pi}^{\pi} e^{i\lambda t} dZ_j^Y(\lambda), \quad j = 1, \cdots, p.$$

The *jth* term on the right hand side of (3) can be written as

$$\int_{-\pi}^{\pi} e^{i\lambda t} \Gamma_j(\lambda) dZ_j^Y(\lambda),$$

where

$$\Gamma_j(\lambda) = \sum_{u=-\infty}^{\infty} g_j(u)e^{-i\lambda u}$$

represents the transfer function between the *jth* input and the output. Thus, (3) gives for each frequency λ

$$dZ^X(\lambda) = \Gamma_1(\lambda)dZ_1^Y(\lambda) + \cdots + \Gamma_p(\lambda)dZ_p^Y(\lambda). \tag{4}$$

Equation (4) has an important interpretation. At each frequency λ, it can be regarded as a simple multiple regression of output on inputs. Equation (3) is a relationship that involves past, present and future observations. *By transforming it into the frequency domain, in a* *way we have replaced the numerous 'lagged' relations into a sequence of simple regressions.* This is a basic concept for our approach in the next section.

3.2. Graphic model for multivariate time series

Let $G = (V, E)$ denote a graph, where $V = \{1, 2, ..., n\}$ is the set of vertices and $E = \{(a, b) \in V \times V\}$ is the set of edges. Suppose $\boldsymbol{X}(t) = (X_1(t), X_2(t), ..., X_n(t))', t \in \mathbb{Z}$, is a multivariate weakly stationary time series and $Y_{ab}(\cdot) = (X_j(\cdot),\ j \neq a, b)$. We then have $V = 1, 2, \cdots, n$ and each vertex corresponds to one of the time series in $\boldsymbol{X}(t)$.

Without loss of generality, let $a = 1$ and $b = n$. Following Dahlhaus (2000) and the notations of Section 3.1, we in theory remove the effect of $Y_{1n}(t)$ on $X_1(t)$ by determining the optimal filters $g_j(u)$ such that it minimizes $E(X_1(t) - \sum_{j=2}^{n-1} \sum_{u=-\infty}^{\infty} g_j(u) X_j(t-u))^2$. Let the optimal estimates be $\hat{g}_j(u)$. Denote the remainder by $\varepsilon_1(t)$, i.e.

$$\varepsilon_1(t) = X_1(t) - \sum_{j=2}^{n-1} \sum_{u=-\infty}^{\infty} \hat{g}_j(u) X_j(t-u).$$

Similarly define

$$\varepsilon_n(t) = X_n(t) - \sum_{j=2}^{n-1} \sum_{u=-\infty}^{\infty} \hat{g}_j(u) X_j(t-u).$$

Now for general values of a and b, suppose $\mathscr{X}_a = (X_a(t); t \in \mathbb{Z})$ and $\mathscr{Y}_{ab} = (Y_{ab}(t); t \in \mathbb{Z})$ and define

$$\mathscr{X}_a \perp\!\!\!\perp \mathscr{X}_b | \mathscr{Y}_{ab} \iff \operatorname{cov}(\varepsilon_a(t), \varepsilon_b(t+u)) = 0 \quad, \forall u \in \mathbb{Z}. \tag{5}$$

Let $(a, b) \notin E \iff \mathscr{X}_a \perp\!\!\!\perp \mathscr{X}_b | \mathscr{Y}_{ab},\ V = 1, 2, ..., n$. Then $G = (V, E)$ is a partial correlation graph for the time series under study. Thus, (5) is a defining statement of a partial correlation graph. The edges of the graph can be determined from the partial spectral coherence. Let

$$C_{ab}(u) = \operatorname{cov}(X_a(t+u), X_b(t)) \quad \text{for all } u \in \mathbb{Z} \tag{6}$$

be the cross-covariance function of $X_a(t)$ and $X_b(t)$. The cross-spectral density is defined by

$$f_{X_aX_b}(\lambda) = \frac{1}{2\pi}\sum_{u=-\infty}^{\infty} C_{ab}(u)e^{-i\lambda u} \ . \tag{7}$$

This expression can be inverted to yield

$$C_{ab}(u) = \int_{-\pi}^{\pi} f_{X_aX_b}(\lambda)e^{i\lambda u}d\lambda \ . \tag{8}$$

The cross-spectral density and cross-covariance function in (7) and (8) can be easily extended to partial cross-spectrum and partial covariance function. Thus, the cross-spectral density $f_{X_aX_b}$ measures the degree of linear association between all of the variables in frequency domain, and partial cross-spectral density $f_{X_aX_b|Y_{ab}}$ is the cross-spectral density of the residual processes $\varepsilon_a(t)$ and $\varepsilon_b(t)$, measuring the degree of linear association of $X_a(t)$ and $X_b(t)$, after removing the influence of the remaining components. This is because

$$\begin{aligned} (a,b) \notin E &\iff \mathscr{X}_a \perp\!\!\!\perp \mathscr{X}_b | \mathscr{Y}_{ab} \\ &\iff \operatorname{cov}(\varepsilon_a(t), \varepsilon_b(t+u)) = 0 \quad \forall u \in \mathbb{Z} \\ &\iff f_{X_aX_b|Y_{ab}}(\cdot) = 0 \ . \end{aligned} \tag{9}$$

Given the spectral coherence, $R_{X_aX_b}(\cdot)$, and partial spectral coherence, $R_{X_aX_b|Y_{ab}}(\cdot)$, are defined by

$$R_{X_aX_b}(\lambda) = \frac{f_{X_aX_b}(\lambda)}{\left[f_{X_aX_a}(\lambda)f_{X_bX_b}(\lambda)\right]^{1/2}} \text{ and } R_{X_aX_b|Y_{ab}}(\lambda) = \frac{f_{X_aX_b|Y_{ab}}(\lambda)}{\left[f_{X_aX_a|Y_{ab}}(\lambda)f_{X_bX_b|Y_{ab}}(\lambda)\right]^{1/2}}$$

respectively, then

$$(a,b) \notin E \iff R_{X_a X_b|Y_{ab}}(\cdot) = 0. \tag{10}$$

Thus, the edges in the graph can be characterized using partial spectral coherence. See Brillinger (1981), Brillinger (1996), or Dahlhaus (2000) for more details. For brevity, sometimes we just refer to spectral coherence and partial spectral coherence as coherence and partial coherence respectively. Next, the estimation of spectral density is introduced.

3.3. Spectral density estimation

Consider the spectral density matrix

$$\mathbf{f}(\lambda) = \begin{pmatrix} f_{1,1}(\lambda) & \cdots & f_{1,n}(\lambda) \\ \vdots & \ddots & \vdots \\ f_{n,1}(\lambda) & \cdots & f_{n,n}(\lambda) \end{pmatrix}.$$

The spectral density matrix $\mathbf{f}(\lambda)$ with elements $f_{X_a X_b}(\lambda) = f_{a,b}(\lambda), a, b = 1, 2, ..., n$ is estimated entry-wise by

$$\hat{f}_{X_a X_b}(\lambda) = \frac{1}{2\pi} \sum_{k=-M}^{M} w_M(k) \hat{C}(k) e^{-i\lambda k}, \tag{11}$$

where $w_M(k)$ is the lag window, and the integer M is the lag number. We choose the Hanning window as the lag window and it is given by

$$w_M(k) = \begin{cases} \frac{1}{2}(1 + \cos(\frac{\pi k}{M})), & |k| \leq M, \\ 0, & |k| > M. \end{cases}$$

For M, after trying with the 3 values $M = \sqrt[3]{T}, 0.5\sqrt{T},$ and $\sqrt{T}$, we choose $0.5\sqrt{T}$ as it gives a reasonable balance between resolution and variance.

The cross-covariance estimator is

$$\hat{C}(k) = \begin{cases} \frac{1}{\sum_t h^2(\frac{t}{T})} \sum_t h(\frac{t+k}{T})(X_a(t+k) - c_a^T) h(\frac{t}{T})(X_b(t) - c_b^T), & k \geq 0, \\ \hat{C}(-k), & k < 0. \end{cases} \tag{12}$$

Here

$$c_a^T = \frac{\sum_{t=1}^{T} h(\frac{t}{T}) X_a(t)}{\sum_{t=1}^{T} h(\frac{t}{T})} \quad \text{and} \quad c_b^T = \frac{\sum_{t=1}^{T} h(\frac{t}{T}) X_b(t)}{\sum_{t=1}^{T} h(\frac{t}{T})} \ ,$$

(Brillinger, 1981). The function $h(x)$ in (12) is called the data window or taper with the properties stated in Section 3.3 of Brillinger, 1981. The cosine taper was used in our estimations. That is,

$$h(\frac{t}{T}) = \begin{cases} \frac{1}{2}(1 - \cos(\frac{2\pi t}{T})), & 1 \leq t \leq T, \\ 0, & elsewhere. \end{cases}$$

3.4. Spectral coherence and partial spectral coherence

Given the estimated spectral densities, the estimates of spectral coherences are obtained by

$$\hat{R}_{X_a X_b}(\lambda) = \frac{\hat{f}_{X_a X_b}(\lambda)}{\left[\hat{f}_{X_a X_a}(\lambda) \hat{f}_{X_b X_b}(\lambda)\right]^{1/2}}. \tag{13}$$

To estimate partial coherence, we have to first find the partial cross-spectrum of $X_a(\cdot)$ and $X_b(\cdot)$ given $Y_{ab}(\cdot)$, which is given by

$$f_{X_a X_b | Y_{ab}}(\lambda) = f_{X_a X_b}(\lambda) - \mathbf{f}_{X_a Y}(\lambda) \mathbf{f}_{YY}(\lambda)^{-1} \mathbf{f}_{X_b Y}(\lambda)^*, \tag{14}$$

where $\mathbf{A}^*$ is the conjugate transpose of matrix $\mathbf{A}$ (Brillinger, 1981). To illustrate the calculation of partial cross-spectrum, suppose we are interested in the chemistry of air pollutants at Tsuen Wan. Let

$$\mathbf{X}(t) = \begin{pmatrix} X_1(t) \\ X_2(t) \\ X_3(t) \\ X_4(t) \end{pmatrix} = \begin{pmatrix} \mathrm{SO}_2(t) \\ \mathrm{NO}_2(t) \\ \mathrm{O}_3(t) \\ \mathrm{RSP}(t) \end{pmatrix}$$

be a multivariate weakly stationary process with 4 components. Then, the spectral density matrix is

$$\mathbf{f}(\lambda) = \begin{pmatrix} f_{1,1}(\lambda) & f_{1,2}(\lambda) & f_{1,3}(\lambda) & f_{1,4}(\lambda) \\ f_{2,1}(\lambda) & f_{2,2}(\lambda) & f_{2,3}(\lambda) & f_{2,4}(\lambda) \\ f_{3,1}(\lambda) & f_{3,2}(\lambda) & f_{3,3}(\lambda) & f_{3,4}(\lambda) \\ f_{4,1}(\lambda) & f_{4,2}(\lambda) & f_{4,3}(\lambda) & f_{4,4}(\lambda) \end{pmatrix}.$$

The partial cross-spectrum of SO_2 with NO_2 after removing the linear effect of O_3 and RSP or $Y_{12}(\cdot)$ is

$$f_{X_1 X_2 | Y_{12}}(\lambda) = f_{1,2}(\lambda) - \mathbf{f}_{X_1 Y}(\lambda)\mathbf{f}_{YY}(\lambda)^{-1}\mathbf{f}_{X_2 Y}(\lambda)^*,$$

where

$$\mathbf{f}_{X_1 Y}(\lambda) = \begin{pmatrix} f_{1,3}(\lambda) & f_{1,4}(\lambda) \end{pmatrix}$$

$$\mathbf{f}_{YY}(\lambda) = \begin{pmatrix} f_{3,3}(\lambda) & f_{3,4}(\lambda) \\ f_{4,3}(\lambda) & f_{4,4}(\lambda) \end{pmatrix}$$

$$\mathbf{f}_{X_2 Y}(\lambda) = \begin{pmatrix} f_{2,3}(\lambda) & f_{2,4}(\lambda) \end{pmatrix}.$$

Normalizing the partial cross-spectrum estimates leads to the estimate of partial coherences of SO_2 and NO_2 given O_3 and RSP, which measures the dependence between SO_2 and NO_2

after removing the linear effect of O_3 and RSP. It is given by

$$\hat{R}_{X_1X_2|Y_{12}}(\lambda) = \frac{\hat{f}_{X_1X_2|Y_{12}}(\lambda)}{\left[\hat{f}_{X_1X_1|Y_{12}}(\lambda)\hat{f}_{X_2X_2|Y_{12}}(\lambda)\right]^{1/2}}. \tag{15}$$

3.5. Tests for coherence and partial coherence

The edges of the partial correlation graph are characterized using partial spectral coherence. An edge is missing if the two components are uncorrelated given the others. Under the hypothesis of $R_{X_aX_b|Y_{ab}}(\cdot) = 0$, the edge in a graph can be determined using the test statistic S, which is equal to

$$\frac{(n-q)\hat{R}^2_{X_aX_b|Y_{ab}}(\lambda)}{1-\hat{R}^2_{X_aX_b|Y_{ab}}(\lambda)} . \tag{16}$$

Here $2n$ is the equivalent degrees of freedom of the spectral density estimator and q is the number of components other than X_a and X_b. S follows the F-distribution with 2 and $2(n-q)$ degrees of freedom. Similarly, a test statistic for a test of zero coherence is given by

$$\frac{(n-1)\hat{R}^2_{X_aX_b}(\lambda)}{1-\hat{R}^2_{X_aX_b}(\lambda)} , \tag{17}$$

following the F-distribution with 2 and $2(n-1)$ degrees of freedom (see Section 8.4 of Koopmans (1995) for more details). When data window $h(\cdot)$ in (12) is implemented, the equivalent degrees of freedom are corrected for the effect of tapering by dividing the original equivalent degrees of freedom by a factor

$$\frac{\int_0^1 h^4(v)dv}{(\int_0^1 h^2(v)dv)^2}.$$

See Section 9.2 of Koopmans (1995). Based on the test statistics, the graphical model can be determined. The two tests are multiple tests on coherence and partial coherence at each frequency λ.

The following example illustrates the concept of a graph between a few time series. Let

$$\begin{aligned}
&X_1(t) = a_1X_1(t-1) + \varepsilon_1(t),\\
&X_2(t) = b_1X_1(t-1) + b_2X_2(t-1) + \varepsilon_2(t),\\
&X_3(t) = c_3X_3(t-1) + c_2X_2(t-1) + \varepsilon_3(t), \text{ and}\\
&X_4(t) = d_4X_4(t-1) + d_2X_2(t-1) + \varepsilon_4(t).
\end{aligned} \tag{18}$$

Here $\varepsilon_i(t), i = 1, 2, 3$ and 4 are mutually independent and identically distributed. Furthermore, $X_j(t), j = 1, 2, 3, 4$ are weakly stationary time series. Then it is clear that they are all correlated, but $(X_1(t), X_3(t))$, $(X_1(t), X_4(t))$ and $(X_3(t), X_4(t))$ are partially uncorrelated. The graphical model of the 4 time series in (18) is illustrated in Figure 3.

[Figure 3 about here.]

3.6. Alternative method in determining the graphical model

From Section 3.2, the partial correlation graph is to find the association between two variables, after removing the linear effect of all other variables. In particular, it is based on a two-sided filter. As a comparison of the relationship between the input and output processes in (3), we consider a bivariate vector autoregressive (VAR) model. With the fitted models, we calculate the cross-correlation of the residuals and hence obtain the partial cross-correlations.

For example, suppose we are interested in analyzing the chemistry of the air pollutants at Tsuen Wan. Let $S(t)$, $N(t)$, $O(t)$ and $R(t)$ be the SO_2, NO_2, O_3 and RSP series at Tsuen Wan respectively. The VAR model to determine the partial cross-correlation of SO_2 with

NO_2 at Tsuen Wan is

$$\begin{bmatrix} S(t) \\ N(t) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \mathbf{F}_0 \begin{bmatrix} O(t) \\ R(t) \end{bmatrix} + \mathbf{\Phi}_1 \begin{bmatrix} S(t-1) \\ N(t-1) \end{bmatrix} + \mathbf{F}_1 \begin{bmatrix} O(t-1) \\ R(t-1) \end{bmatrix} + \cdots + \mathbf{\Phi}_j \begin{bmatrix} S(t-j) \\ N(t-j) \end{bmatrix}$$

$$+ \mathbf{F}_j \begin{bmatrix} O(t-j) \\ R(t-j) \end{bmatrix} + \cdots + \mathbf{\Phi}_p \begin{bmatrix} S(t-p) \\ N(t-p) \end{bmatrix} + \mathbf{F}_p \begin{bmatrix} O(t-p) \\ R(t-p) \end{bmatrix} + \begin{bmatrix} e_{S \cdot O,R}(t) \\ e_{N \cdot O,R}(t) \end{bmatrix},$$

where

$$\mathbf{F}_0 = \begin{bmatrix} f_{11}^{(0)} & f_{12}^{(0)} \\ f_{21}^{(0)} & f_{22}^{(0)} \end{bmatrix}, \mathbf{\Phi}_j = \begin{bmatrix} \phi_{11}^{(j)} & \phi_{12}^{(j)} \\ \phi_{21}^{(j)} & \phi_{22}^{(j)} \end{bmatrix}, \quad \mathbf{F}_j = \begin{bmatrix} f_{11}^{(j)} & f_{12}^{(j)} \\ f_{21}^{(j)} & f_{22}^{(j)} \end{bmatrix}, \quad j = 1, \cdots, p.$$

Note that there is no $\mathbf{\Phi}_0$ term.

Then, the partial cross-correlation of SO_2 with NO_2 given the O_3 and RSP processes is the cross-correlation of $e_{S \cdot O,R}(t)$ and $e_{N \cdot O,R}(t)$. Other partial cross-correlations are computed in the same manner. The lag order p is determined by the Bayesian information criterion (BIC).

4. ANALYSIS AND RESULTS

Hong Kong is a densely populated city having around 7 million people, but with only a small area of about 1000 km^2. With such a small area, it is expected in general the pollution in the three stations will be highly cross-correlated and partially cross-correlated.

For both NO_2 and SO_2, it can be seen from the boxplot (Figure 2) that the level of concentration, in increasing order, is TM, TC and TW. This suitably reflects the background that TW is a city area with some industries, TM is a rural area, and TC is a semi-rural area. For RSP, TW is slightly higher than TC and TM, while the latter two practically have no difference. For O_3, the order in increasing level is TW, TC and TM. This correctly echoes the statement that "As nitric oxide emissions from motor vehicles can react with and remove O_3 in the air, regions with heavy traffic normally have lower O_3 levels than areas with light traffic. Hence, Tap Mun has steadily recorded more than twice the O_3 levels measured in urban areas ..."(Air Science Group, Environmental Protection Department, the Government of the Hong Kong Special Administrative Region, 2012, p. 18). For the 3 regions, TW and TM have the heaviest and lightest traffic, respectively.

We next look at the test for partial coherency. The test statistics for partial coherency are plotted separately for Tsuen Wan, and Tsuen Wan and Tap Mum in Figure 4a and Figure 5a. Other related plots are available in Figures 4 and 5. For partial coherence, the error bound is given by the 95% quantile of the $F(2, 88)$ distribution, with a value of 3.1. Thus we see from Figure 4a that in Tsuen Wan all tests are significant, whereas Figure 5a shows that 7 out of 16 cases are insignificant. The insignificant cases are SO_2(TW)/NO_2(TM), NO_2(TW)/SO_2(TM), O_3(TW)/SO_2(TM), O_3(TW)/NO_2(TM), O_3(TW)/RSP(TM), RSP(TW)/SO_2(TM) and RSP(TW)/O_3(TM).

Figure S1 shows the graphs of coherency and partial coherency of the full model, which contains all 12 variables, while Figure S2 shows the cross-correlations and partial cross-correlations. The technical details of the two graphs are described in the Supporting Information. Here we focus on the interpretation from them. It is intriguing to see if the partial coherence and partial cross-correlations give the same results.

[Figure 4 about here.]

[Figure 5 about here.]

Table S2 of the Supporting Information shows the testing results of using partial coherence and partial cross-correlations. A summary of the results and interpretations from Table S2 is as follows:

(1). Following Dahlhaus (2000), if the partial coherence is marginally significant at a few frequencies, the test is regarded as insignificant. For partial coherence, the error bound is given by the 95% quantile of the $F(2, 88)$ distribution, a value of 3.10. For partial correlations, we again use the approximate error bound of $\pm 2/\sqrt{n}$, namely, $\pm 2/\sqrt{1491} = 0.0518$.

(2). A "Yes" in the column "Partial Coherence" or "VAR" means the corresponding test is significant, whereas a "No" means not significant. The last column shows whether the two methods agree or not. Table S2 shows the two methods agree on 57 cases out of a total of 66 or the percentage of matching is 86.3%, which is quite good. Further, even when the two methods do not match in their results of testing, they are in line. That means when one test is not significant, the other test value, though significant, is usually quite small. In summary, the two methods have close agreement in terms of p-values.

The time domain method is inefficient to implement as compared with the frequency domain approach. One has to calculate many partial cross-correlations between 2 variables, and in each case to first decide the lag order of the VAR model by BIC. The frequency domain approach does not have this problem. We tested the two programs, using a computer with a Core i7-4790 3.60 GHz CPU and 16 GB main memory under the Windows 7 64-bit operating system and Matlab 2015a. We find the running time of the frequency domain approach and the VAR method are 93.52 seconds (about 1.56 minutes) and 1332.33 seconds (about 22.21 minutes) respectively. On the other hand the time domain method gives more information. It tells us whether the partial cross-correlation is positive or negative, and also the time lag, which cannot be read from the graphs of partial coherence directly. From the partial cross-correlations, an immediate observation is that the lag-zero one is usually large. In fact, 47 out of the 51 significant values are due to the zero lag. This is not surprising due to the close proximity of the three stations. The flight distance between Tung Chung and Tsuen

Wan, Tsuen Wan and Tap Mun, and Tung Chung and Tap Mun are 25 km, 33 km and 58 km respectively.

For the graphical model, each vertex represents a pollutant from a station. There should be a total of 66 edges if all vertices are connected. From the frequency domain results, there are 50 edges. Figure 6 is the graphical model for our full data. It is not a simple graph due to the large number of vertices. In the figure, only the significant edges are shown. Black line represents the edge between two different pollutants at two different stations, black bold line represents the edge between two pollutants at the same station and dotted line represents the edge between two stations for the same pollutant.

[Figure 6 about here.]

To find the explanation for the interaction of pollutants (chemical substances) in air pollution is in general difficult. It depends on the concentration of the pollutants, chemical reaction between the pollutants, and very often the presence of other substances. Several observations are in order:

(a). For the same pollutant, there is always an edge between two stations. That means the partial coherence and partial cross-correlations between the 3 locations are all significant. For instance, the maximum of the test statistic S for testing partial coherence between SO_2 of TW and TM has a value larger than 10, much greater than 3.08.

(b). For the same location, the partial coherences between different pollutants are all significant except the value between SO_2 of TC, TC(SO_2), and O_3 of TC, TC(O_3). That means only 1 out of 18 partial coherences is not significant. For partial cross-correlations, 3 out of 18 are insignificant.

We next look at the cases of different pollutants more carefully, including results from different stations. Direct counting based on partial coherence gives Table 1.

[Table 1 about here.]

A contingency table analysis shows that the two dimensions concerned are dependent. The chi-square statistic with Yates' continuity correction is 5.87. The 5% critical value for a chi-square distribution with one degree of freedom is 3.84. Thus, the test is significant. Being in the same station or not has a bearing on partial coherence. It is clear from the table that different pollutants have a higher probability of having no edge connected if they are from different stations.

(c). From the graphs of partial coherency (Figure 4 and Figure 5), there are 15 cases of non-significance for different pollutants from the three different stations. Since each case involves two stations, there should be 30 station counts altogether. The distribution of the counts for TW, TM and TC is 10, 12 and 8 respectively.

It is clear that the data do not reject a uniform distribution. Given that there is no edge for the two pollutants coming from different stations, there is no evidence that any of the 3 stations will have a higher chance of getting no edge. A similar analysis for the 36 combinations of different pollutants from different stations is provided by Table 2.

[Table 2 about here.]

But given the frequencies in the table are small, we have to interpret the results with care. The first 3 combinations with more significant test results are NO_2/RSP, NO_2/O_3 and SO_2/RSP. It seems that for pollutants from different stations, NO_2 and RSP are more influential than the other two. From the graphs of partial cross-correlations, we see the partial cross-correlations for different pollutants from the same station are usually much larger than that from different stations. Combining with the results in (b), we believe that being in the same station is a more important factor than the combination of pollutants.

5. CONCLUSIONS

We studied the air pollution of Hong Kong by looking at the data from 4 pollutants and 3 monitoring stations of Hong Kong. Since Hong Kong is small in area, one may think all the pollutants from different stations are highly correlated. This is the case if we just look at correlations and coherences.

By looking at the raw data, we can see the three stations have some local characteristics. We applied the partial coherence method to construct a graphical model for the 12 time series. As a comparison, we also looked at a comparable time domain approach, the VAR approach. The two have good agreement. The same pollutants in different stations are always connected by edges. For different pollutants in the same station, they have very high probability of being partially correlated. For different pollutants between different stations, about 58 percent of them have significant results in the test of partial coherence. But the strength of association is clearly less than that from the same station. There is some evidence that NO_2 and RSP are more influential than O_3 and SO_2. In brief, the intuition that all pollutants are highly associated is not correct. Graphical modeling seems a plausible tool for the investigation of relationships between multivariate spatial time series.

Additional details including figures and tables are available in an online supporting information file. The Matlab program codes and the dataset used can also be obtained from the online supplements.

Acknowledgements

We thank the Editor, the Associate Editor and the Referees for their constructive criticisms and suggestions which have led to improvements of the paper. The research of Heung Wong was supported by a GRF grant of the Hong Kong Research Grant Council PolyU 5029/11P.

REFERENCES

Air Science Group, Environmental Protection Department, the Government of the Hong Kong Special Administrative Region, 2012. Air quality in Hong Kong 2012. `http://www.aqhi.gov.hk/api_history/english/report/files/AQR2012e_final.pdf`, [accessed 16 December 2015].

Box GEP, Cox DR, 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B* **26**(2): 211–252.

Brillinger DR, 1981. *Time series: Data analysis and theory.* San Francisco: Holden-Day, expanded edition.

Brillinger DR, 1996. Remarks concerning graphical models for time series and point processes. *Revista de Econometria* **16**: 1–23.

Cleveland WS, 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**(368): 829–836.

Dahlhaus R, 2000. Graphical interaction models for multivariate time series. *Metrika* **51**(2): 157–172.

Dahlhaus R, Eichler M, Sandkühler J, 1997. Identification of synaptic connections in neural ensembles by graphical models. *Journal of Neuroscience Methods* **77**(1): 93–107.

Edwards D, 2000. *Introduction to Graphical Modelling.* New York: Springer.

Eichler M, 2012. Graphical modelling of multivariate time series. *Probability Theory and Related Fields* **153**: 233–268.

Forni M, Hallin M, Lippi M, Reichlin L, 2000. The generalized dynamic-factor model: Identification and estimation. *The Review of Economics and Statistics* **82**(4): 540–554.

Hastie T, Tibshirani R, Friedman J, 2013. *The Elements of Statistical Learning.* New York : Springer, 3rd edition.

Jung A, Hannak G, Goertz N, 2015. Graphical lasso based model selection for time series. *Signal Processing Letters, IEEE* **22**(10): 1781–1785.

Khare K, Oh SY, Rajaratnam B, 2015. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77**(4): 803–825.

Koopmans L, 1995. *The Spectral Analysis of Time Series.* New York: Academic Press.

Lauritzen S, 1996. *Graphical Models.* Oxford England: Clarendon Press.

McLeod AI, Gweon H, 2012. Optimal deseasonalization for monthly and daily geophysical time series. *Journal of Environmental Statistics* **4**(11).

Priestley M, 1981. *Spectral analysis and time series*, volume 1 and 2. London: Academic Press.

Qiu H, Han F, Liu H, Caffo B, 2015. Joint estimation of multiple graphical models from high dimensional time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* , DOI: 10.1111/rssb.12123.

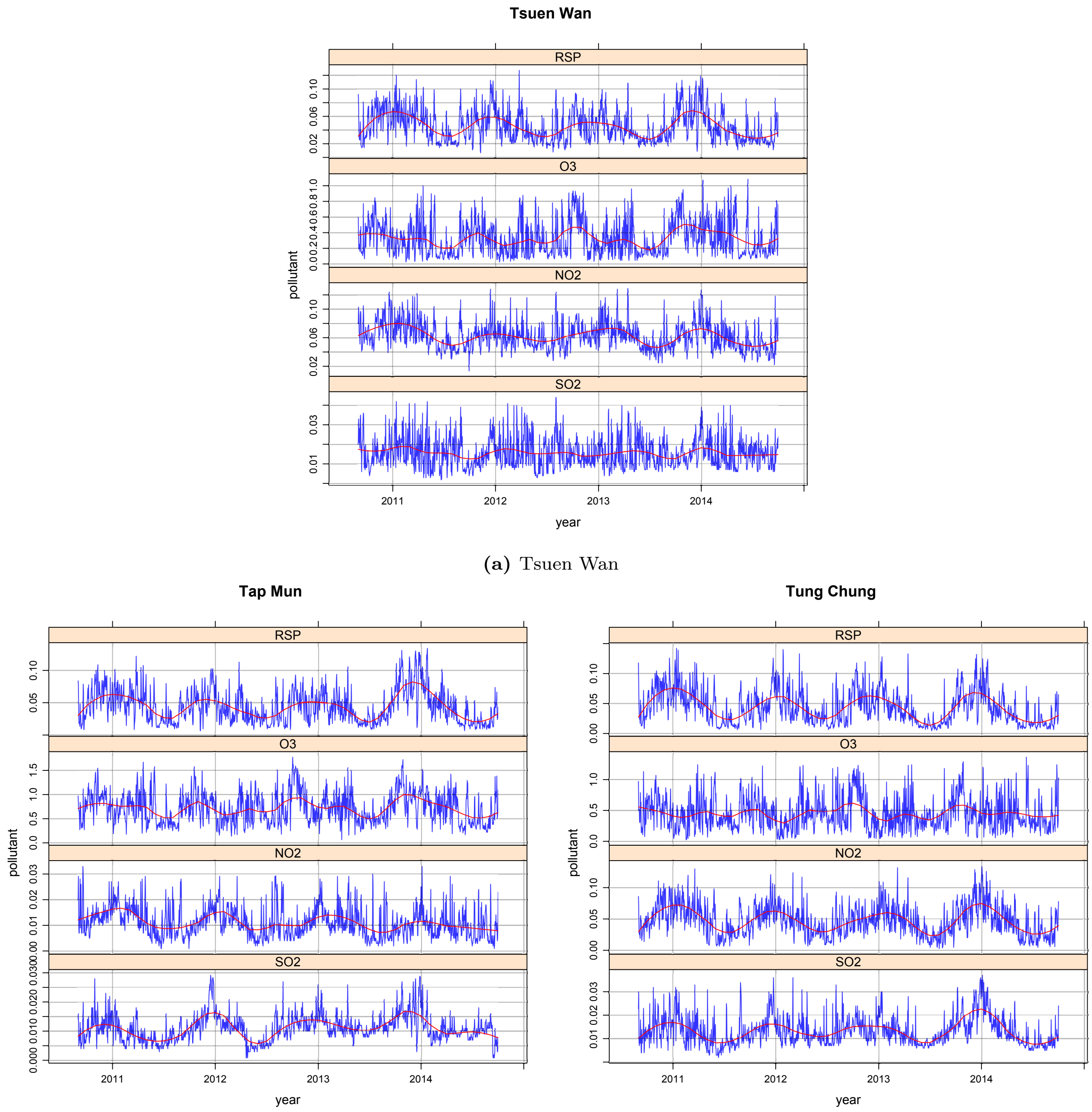
Tunnicliffe Wilson G, Reale M, Haywood J, 2015. *Models for dependent time series.* Boca Raton: CRC Press.

Wolstenholme R, Walden A, 2015. An efficient approach to graphical modeling of time series. *Signal Processing, IEEE Transactions on* **63**(12): 3266–3276.

SUPPORTING INFORMATION

Additional information, including the dataset analyzed and the program scripts used, may be found in the online supplements of this article at the publisher's website.

FIGURES



(a) Tsuen Wan

(b) Tap Mun

(c) Tung Chung

Figure 1. Daily averages of SO_2, NO_2, O_3 and RSP with LOESS smoother

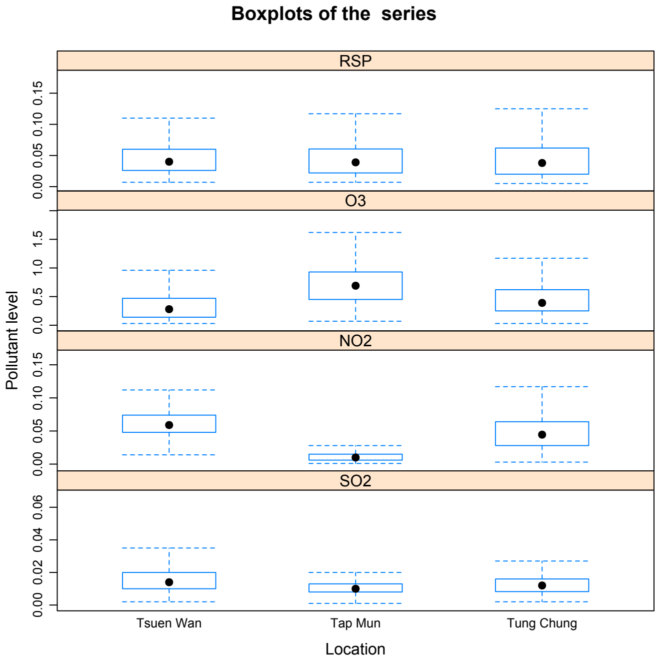


Figure 2. Boxplot of the 4 pollutants at the 3 stations

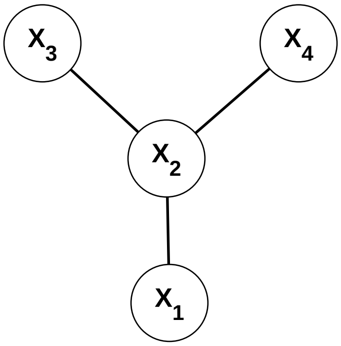
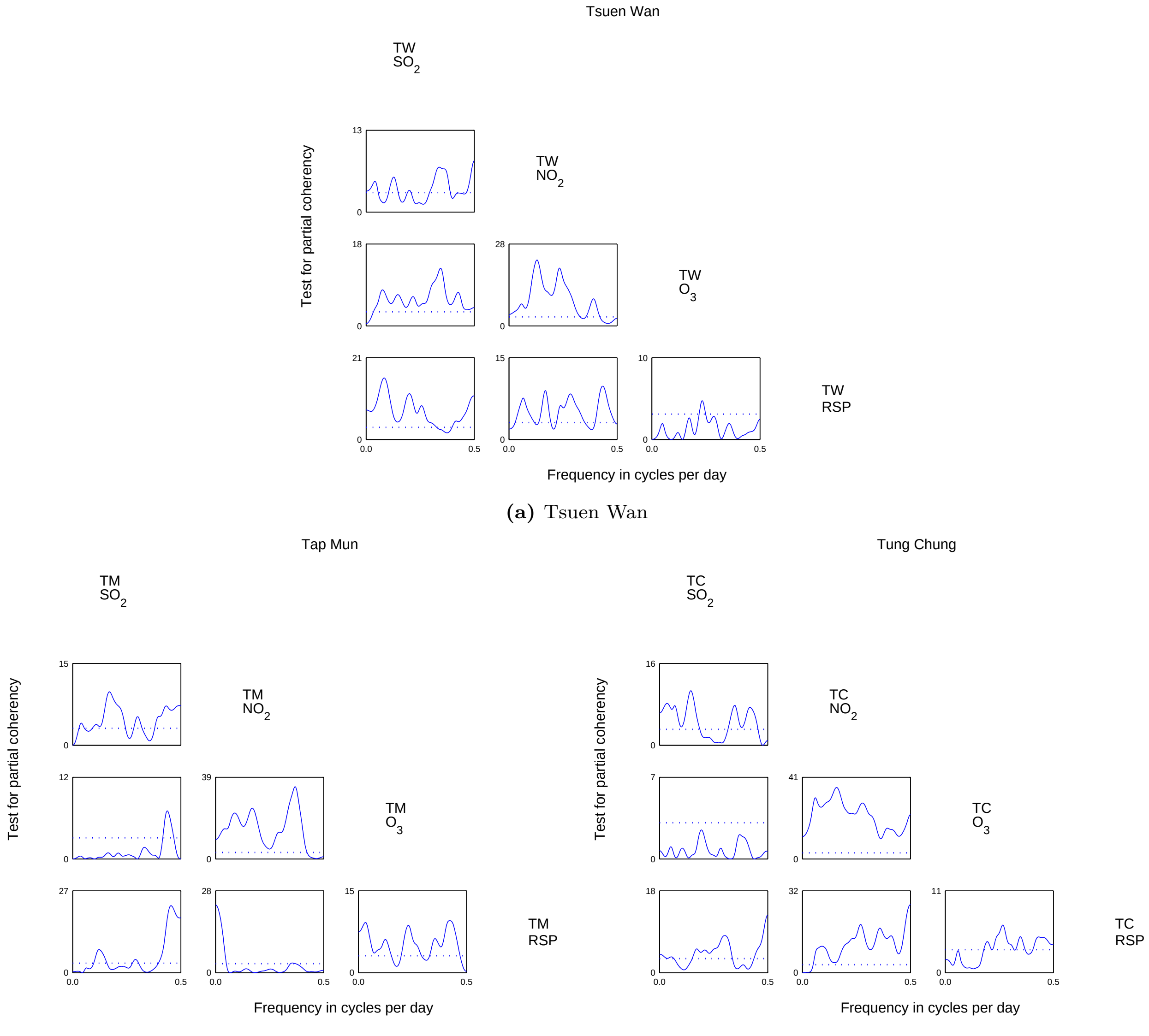


Figure 3. The graphical model of the 4 time series in Equation (18) in Section 3.5. This demonstrates that X_2 is partially correlated to X_1, X_3 and X_4, but X_1, X_3 and X_4 themselves are not partially correlated.

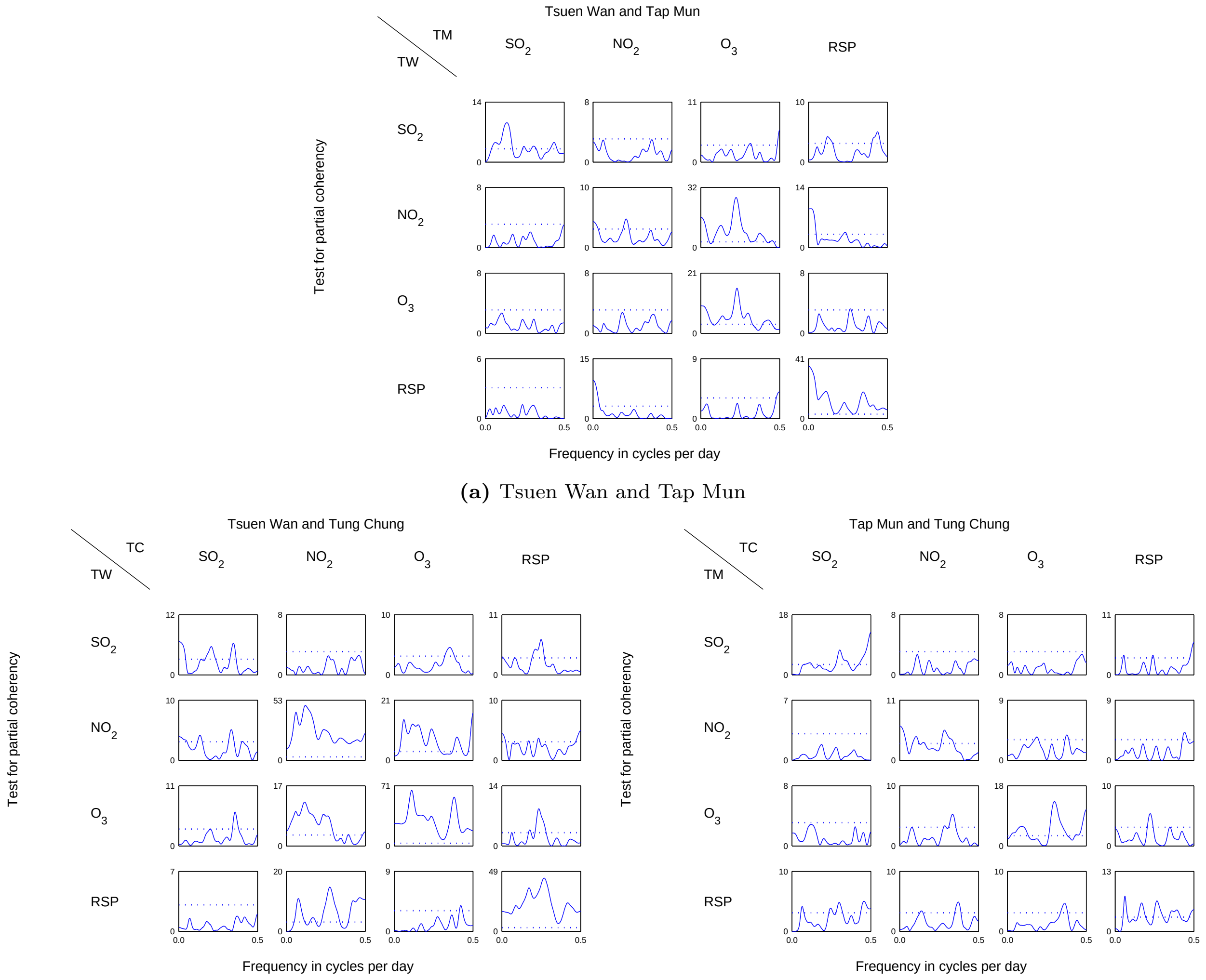


(a) Tsuen Wan

(b) Tap Mun

(c) Tung Chung

Figure 4. Test for partial coherency of pollutants at the same station under the full model. (The dotted line represents a 95% quantile of the $F(2, 88)$ distribution, with a value of 3.10)



(a) Tsuen Wan and Tap Mun

(b) Tsuen Wan and Tung Chung

(c) Tap Mun and Tung Chung

Figure 5. Test for partial coherency of pollutants from different stations under the full model (The dotted line represents a 95% quantile of the $F(2, 88)$ distribution, with a value of 3.10)

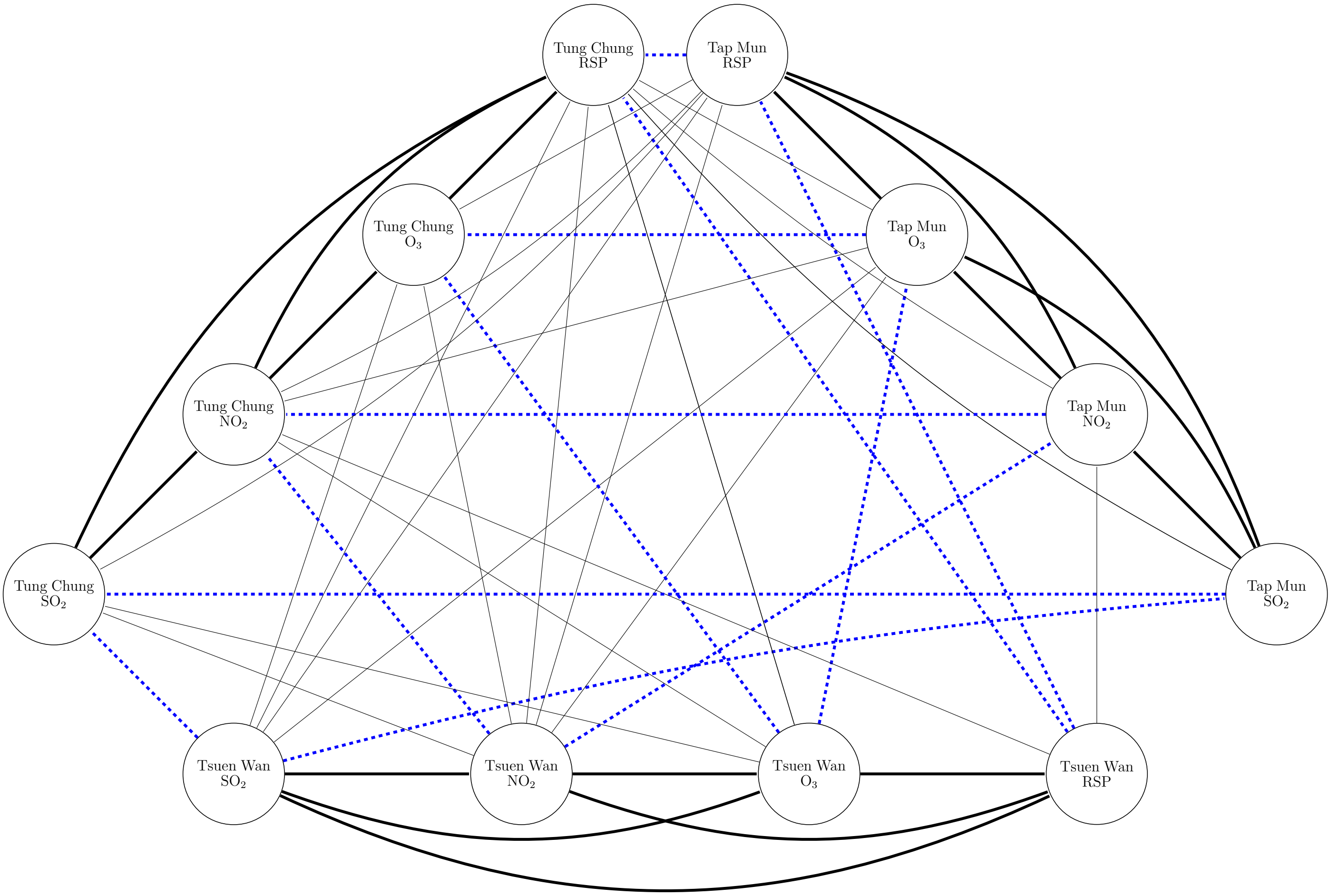


Figure 6. Partial correlation graph for the full model (Black line represents the edge between two different pollutants at two different stations, black bold line represents the edge between two pollutants at the same station and dotted line represent the edge between two stations for the same pollutant)

TABLES

Table 1. Significance of interaction of different pollutants among stations determined by partial coherency

	Test significant	Test not significant
Same station	17	1
Different station	21	15

Table 2. Significance of the interaction between two pollutants

Combination of pollutants	Test significant	Test not significant
NO2/SO2	1	5
SO2/O3	3	3
SO2/RSP	4	2
NO2/O3	5	1
NO2/RSP	6	0
O3/RSP	3	3