Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 2028–2043
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Data analysis using regression models with missing observations and long-memory: an application study

Pilar Iglesias[a], Hector Jorquera[b], Wilfredo Palma[a,*]

[a]*Department of Statistics, P. Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, Chile*
[b]*Department of Chemical & Bioprocess Engineering, P. Universidad Católica de Chile, Santiago, Chile*

Received 12 November 2003; received in revised form 21 March 2005; accepted 21 March 2005
Available online 29 April 2005

Abstract

The objective of this work is to propose a statistical methodology to handle regression data exhibiting long memory errors and missing values. This type of data appears very often in many areas, including hydrology and environmental sciences, among others. A generalized linear model is proposed to deal with this problem and an estimation strategy is developed that combines both classical and Bayesian approaches. The estimation methodology proposed is illustrated with an application to air pollution data which shows the impact of the long memory in the statistical inference and of the missing values on the computations. From a Bayesian standpoint, genuine priors are considered for the parameters of the model which are justified within the context of the air pollution model derivation.

Keywords: ARFIMA model; Bayesian estimation; Kalman filter; Long memory processes; Parameter estimation; Regression model

1. Introduction

Long-memory time series models have received considerable attention in the last decade, see for example Doukhan et al. (2003) and references therein. Regression models with long memory errors have been discussed by several authors, including Yajima (1988, 1991),

* Corresponding author. Tel.: +562 686 4506; fax: +562 686 6229.
E-mail address: wilfredo@mat.puc.cl (W. Palma).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.03.007

Koul and Mukherjee (1993), Dahlhaus (1995), Hall et al. (1995) and Robinson and Hidalgo (1997), among others. For instance, Yajima (1991) establishes asymptotic properties of least squares estimates and best linear estimates. From a Bayesian perspective, the problem of estimation of long-memory models has been approached by Carlin et al. (1985), Koop et al. (1997) and Pai and Ravishanker (1998), among others. Computational techniques based on Monte Carlo Markov Chains (MCMC), see for example Robert and Casella (2004), are frequently used in these works, but these procedures are usually computationally demanding. Philippe and Rousseau (2002) discuss an asymptotic Bayesian analysis of Gaussian long memory processes. A Bayesian nonparametric approach to the problem is studied by Petris (1997), including processes with a linear regression term for the mean. All these methods apply strictly to complete regression data sets. However, often in practice the available data are incomplete with missing values in the response or in the explanatory variables. This situation can be found in many scientific areas, ranging from physics to economics. In this paper, we provide a strategy to make inference about the regression coefficients for ARFIMA regression models with missing values by combining classical and Bayesian methods.

The remaining of the paper is organized as follows. In the next section we describe the model considered. Section 3 discusses estimation techniques combining classical and Bayesian approaches. We also describe the assumptions of the procedures used at each stage, particularly the handling of the high number of missing data which is typical for air quality data records. In Section 4, the methodology proposed in Section 3 is applied to the analysis of the air pollution in Santiago, Chile. Concluding remarks are presented in Section 5.

2. Model

In this section we propose a strategy to estimate the parameters of a long memory regression model and to make inference about them. We begin our discussion by describing the linear model used, introducing a family of long-memory processes and finally proposing estimation methodologies.

2.1. Linear models

We consider the following linear model for the data,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y} = \left(y_{t_1}, \ldots, y_{t_n}\right)'$ is a vector formed by the dependent variable observed at times t_j, $j = 1, \ldots, n$, $\mathbf{X}$ is a matrix of size $n \times r$ whose columns contain the independent variables observed at times t_j, $j = 1, \ldots, n$, $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_r\right)'$ is the vector of coefficients and $\boldsymbol{\varepsilon} = \left(\varepsilon_{t_1}, \ldots, \varepsilon_{t_n}\right)'$ is the vector of regression errors.

It is assumed that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Gamma})$, where $\boldsymbol{\Gamma}$ is a positive definite symmetric matrix. Under these conditions, the inverse of $\boldsymbol{\Gamma}$ exists and it can be decomposed as

$$\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{\Gamma}^{-\frac{1}{2}}. \tag{2}$$

Therefore, by using the following transformation of regression (1) we have:

$$\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{y} = \boldsymbol{\Gamma}^{-\frac{1}{2}}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\Gamma}^{-1/2}\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$. The customary assumption is that the errors are uncorrelated leading to the standard ordinary least squares (OLS) setup where the variance–covariance matrix is the identity matrix. However, a key aspect of this work is to consider the possibility that the data under study display a more complex error covariance structure. To model these errors, we consider a class of long-memory time series, the so-called ARFIMA processes described in Section 2.2.

There are many techniques to verify whether the assumption of non-correlated errors is appropriate or not, including the Durbin–Watson and the Ljung–Box portmanteau tests. These procedures are based on the analysis of the autocorrelation function (ACF) of the residuals of the regression model. In particular, the ACF of the residuals can be used to detect the presence of long-memory behavior, see for example Beran (1994).

Our analysis starts assuming that the errors are uncorrelated and hence an OLS approach is taken. Afterwards, the residuals from the OLS fitting are studied and the appropriateness of a more general model is considered.

In the next subsection we discuss briefly the family of time series models used in this study.

2.2. *Long-memory processes*

Long-memory time series are characterized by autocorrelation decaying to zero at an hyperbolic rate. In contrast, the ACF of a short-memory process decays at an exponential rate. For instance, this is the case of an ARMA model. A class of models allowing long-memory behavior is the autoregressive fractionally integrated moving average (ARFIMA) processes, described by

$$\Phi(B)(1-B)^d \varepsilon_t = \Theta(B)\eta_t, \tag{4}$$

where $d < 1/2$, $\{\eta_t\}$ is a white noise sequence, B is the backshift operator $B\varepsilon_t = \varepsilon_{t-1}$, $\Phi(B) = 1 + \phi_1 B + \cdots + \phi_p B^p$ is the AR operator, $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ is the MA operator and $(1-B)^d = \sum_{k=0}^{\infty} \pi_k B^k$ is the fractional difference operator, with $\pi_k = \Gamma(k-d)/\Gamma(k+1)\Gamma(-d)$, where $\Gamma(\cdot)$ denotes the Gamma function. If $d \in (0, 1/2)$, the resulting is a *long-memory* process with non-summable correlations, that is, $\sum_{k=0}^{\infty} |\gamma(k)| = \infty$, where $\gamma(k)$ is the autocorrelation of order k. If $d < 0$, the result is an *intermediate-memory* process with zero spectral density at frequency zero and summable correlations, $\sum_{k=0}^{\infty} |\gamma(k)| < \infty$, cf., Brockwell and Davis (1991). Note that when $d = 0$, (4) corresponds to an ARMA model.

The autocovariance function of an ARFIMA(p, d, q) process can be found in Sowell (1992). For a fractional noise process, i.e. ARFIMA$(0, d, 0)$, the autocovariance function can be written as

$$\gamma(k) = \sigma^2 \frac{\Gamma(1-2d)\Gamma(k+d)}{\Gamma(1-d)\Gamma(k-d+1)\Gamma(d)}. \tag{5}$$

This expression is used in Section 4 to construct the variance–covariance matrix of the regression errors starting from the fitted ARFIMA model.

Assuming this autocovariance structure of the regression errors we can write their variance–covariance matrix as follows:

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{\boldsymbol{\theta}}(0) & \gamma_{\boldsymbol{\theta}}(t_1 - t_2) & \cdots & \gamma_{\boldsymbol{\theta}}(t_1 - t_n) \\ \gamma_{\boldsymbol{\theta}}(t_2 - t_1) & \gamma_{\boldsymbol{\theta}}(0) & \cdots & \gamma_{\boldsymbol{\theta}}(t_2 - t_n) \\ \vdots & \vdots & & \vdots \\ \gamma_{\boldsymbol{\theta}}(t_n - t_1) & \gamma_{\boldsymbol{\theta}}(t_n - t_2) & \cdots & \gamma_{\boldsymbol{\theta}}(0) \end{pmatrix},$$

where $\gamma_{\boldsymbol{\theta}}(\cdot)$ denotes the autocovariance function determined by the ARFIMA model with parameters $\boldsymbol{\theta} = (d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$.

3. Estimation methodology

3.1. Estimating the ACF

In the presence of missing data the usual estimate for the autocorrelation function of the process $\{\varepsilon_t\}$ (the residuals in this context)

$$\hat{\rho}(k) = \frac{\sum_{t=1}^{n-k} (\varepsilon_t - \hat{\mu})(\varepsilon_{t+k} - \hat{\mu})}{\sum_{t=1}^{n} (\varepsilon_t - \hat{\mu})^2}, \tag{6}$$

where $\hat{\mu} = \frac{\sum_{t=1}^{n} \varepsilon_t}{n}$, cannot be used since some of the ε_t are not available. Instead, following Yajima and Nishino (1999), we use the following approach. Consider the process $z_t = \varepsilon_t a_t$, where $a_t = 1$ if ε_t is observed and $a_t = 0$ if ε_t is missing. Then, an estimate of the ACF of the process $\{\varepsilon_t\}$ is

$$\tilde{\rho}(k) = \frac{\sum_{t=1}^{n-k} a_t a_{t+k} (z_t - \tilde{\mu})(z_{t+k} - \tilde{\mu}) \sum_{t=1}^{n} a_t^2}{\sum_{t=1}^{n} (z_t - \tilde{\mu})^2 \sum_{t=1}^{n-k} a_t a_{t+k}}, \tag{7}$$

where $\tilde{\mu} = \frac{\sum_{t=1}^{n} z_t}{\sum_{t=1}^{n} a_t}$. We use this estimate to analyze the correlation structure of the residual of the regression model.

3.2. Estimating $\boldsymbol{\theta}$

In order to fit the linear model, we first estimate the ARFIMA parameters $\boldsymbol{\theta} = (d, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ and the variance σ^2 directly from the residuals of the OLS model. Since there is a large number of missing values in the data, we use a Kalman filter approach as proposed by Palma and Chan (1997) and Palma and Del Pino (1999). This procedure calculates maximum likelihood estimates by means of a space state representation of the ARFIMA model and then computing the Gaussian log-likelihood function by means of Kalman filter recursive equations, see also Chapter 12 of Brockwell and Davis (1991).

After finding an estimate for $\boldsymbol{\theta}$, the error variance–covariance matrix may be calculated by applying formula (5) for an ARFIMA(0, d, 0) model—or formula (8) of Sowell (1992) for the general ARFIMA(p, d, q) model—to the sample $\boldsymbol{y}=\left(y_{t_1}, \ldots, y_{t_k}\right)'$ to obtain $\widehat{\mathrm{Var}}(\boldsymbol{y})=\widehat{\boldsymbol{\Gamma}}=\widehat{\gamma}_{\boldsymbol{\theta}}\left(|t_i - t_j|\right)$, replacing $\boldsymbol{\theta}$ by its estimated value $\boldsymbol{\theta}^*$ and $\boldsymbol{\sigma}$ by $\boldsymbol{\sigma}^*$. Once the symmetric matrix $\boldsymbol{\Gamma}$ is obtained, its square root $\boldsymbol{\Gamma}^{\frac{1}{2}}$ is computed by means of the Cholesky decomposition algorithm.

3.3. Estimating $\boldsymbol{\beta}$

To estimate the coefficients of the regression model, we may use a classical or Bayesian approach. If a classical approach is adopted, then the likelihood estimation from a generalized linear model must be obtained. The generalized least square (GLS) estimate is in this case

$$\hat{\boldsymbol{\beta}} = \left[\mathbf{X}\boldsymbol{\Gamma}\left(\boldsymbol{\theta}^*, \boldsymbol{\sigma}^*\right)^{-1}\mathbf{X}'\right]^{-1}\mathbf{X}'\boldsymbol{\Gamma}\left(\boldsymbol{\theta}^*, \boldsymbol{\sigma}^*\right)^{-1}\boldsymbol{y}. \tag{8}$$

here $\boldsymbol{\Gamma}\left(\boldsymbol{\theta}^*, \boldsymbol{\sigma}^*\right)$ is the error covariance matrix and $\boldsymbol{\theta}^*$ and $\boldsymbol{\sigma}^*$ are chosen by the method described in Section 3.2.

If a Bayesian approach is adopted at this stage, then we need to elicit a prior distribution for $\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2\right)$. Let us assume that

$$\boldsymbol{y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 \sim N\left(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Gamma}\left(\boldsymbol{\theta}^*, \boldsymbol{\sigma}\right)\right). \tag{9}$$

The following prior distributions are considered

$$\boldsymbol{\beta} \sim N\left(\mathbf{b}_0, \mathbf{B}_0\right), \boldsymbol{\sigma}^2 \sim IG\left(v_{10}, v_{20}\right), \tag{10}$$

where $IG(a, b)$ denotes the inverted gamma distribution with density proportional to

$$\left(\boldsymbol{\sigma}^2\right)^{-(a+1)}\mathrm{e}^{-b/\boldsymbol{\sigma}^2}, \tag{11}$$

and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r)$ and $\boldsymbol{\sigma}^2$ are assumed to be mutually independent. Even in this case (Normal model), it is not possible to obtain analytical expressions for the posterior distributions. Thus, in order to implement a Bayesian analysis we may use a MCMC-type approach. A well known MCMC technique is the Gibbs sampler. The standard implementation of this method requires sampling from the conditional posterior distribution for each parameter. In the present case, the conditional posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ can be easily obtained from the normal theory (see for example, O'Hagan, 1994). Hence, with $\boldsymbol{y} = \left(y_{t_1}, \ldots, y_{t_n}\right)'$ denoting the observed responses, the following conditional distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$

are obtained:

$$\begin{aligned}
&\boldsymbol{\beta} \mid \mathbf{X}, \boldsymbol{y}, \boldsymbol{\sigma}^2 \sim N(\mathbf{b}, \mathbf{B}) \\
&\boldsymbol{\sigma}^2 \mid \mathbf{X}, \boldsymbol{y}, \boldsymbol{\beta} \sim IG(v_1, v_2) \\
&\mathbf{B} = \left[\mathbf{B}_0^{-1} + \boldsymbol{\sigma}^{-2}\left(\mathbf{X}'\mathbf{M}^{-1}\left(\boldsymbol{\theta}^*\right)\mathbf{X}\right)^{-1}\right]^{-1}, \\
&\mathbf{b} = \mathbf{B}\left[\mathbf{B}_0\mathbf{b}_0 + \left(\mathbf{X}'\mathbf{M}^{-1}\left(\boldsymbol{\theta}^*\right)\mathbf{X}\right)^{-1}\hat{B}\right], \\
&v_1 = v_{10} + \frac{n}{2}, \quad v_2 = \frac{S(\boldsymbol{y}, \boldsymbol{\beta})}{2} + v_{20},
\end{aligned} \tag{12}$$

where $\hat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{M}^{-1}\left(\boldsymbol{\theta}^*\right)\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}^{-1}\left(\boldsymbol{\theta}^*\right)\boldsymbol{y}$ and $S(\boldsymbol{y}, \boldsymbol{\beta}) = (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{M}^{-1}\left(\boldsymbol{\theta}^*\right)(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$ with $\mathbf{M}\left(\boldsymbol{\theta}^*\right) = \boldsymbol{\sigma}^{-2}\boldsymbol{\Gamma}\left(\boldsymbol{\theta}^*, \boldsymbol{\sigma}^2\right)$. The standard Gibbs Sampling or the grouped Gibbs Sampling are used to obtain the posterior required, see for example Chen et al. (2000). In particular, we have implemented this approach in a Fortran code.

4. Application to air pollution data analysis

4.1. Introduction

In this section we apply the methodology described above to a severe air pollution problem affecting many cities around the world. In fact, urban air pollution is still on the rise at many cities worldwide, or has experienced only small improvements (Fenger, 1999; Mage et al., 1996; Elsom, 1996). One of the main goals is to assess how different sources and processes contribute to ambient particulate matter (PM) levels, because this is the airborne pollutant that is associated with the most severe public health impacts, see for example Vedal (1997) and Cifuentes et al. (2000). Given the complexity of the different sources and processes involved, deterministic, physically based models for estimating ambient PM levels are still under development—Carmichael et al. (1999) and Jacobson (1999)—and have been tested only in few selected cities at which intensive measurement campaigns have been performed, cf. Arnold et al. (2003).

In this work we propose a linear model to analyze PM pollution levels in Santiago, Chile, on the basis of a physical model derived by Jorquera (2002). This model is based on a simpler approach but it produces the required output: To relate ambient PM levels to the emissions levels within the city. Then, the resulting model is estimated and analyzed by means of the techniques discussed in Section 3.

4.2. The data

The dataset under study comes from the *Ambient Air Quality Monitoring Network* (MACAM in Spanish) in Santiago, Chile. The following ambient variables were obtained from the MACAM network (http://www.sesma.cl): fine particulate matter PM2.5 $\left(\mu g/m^3\right)$, carbon monoxide CO (ppm), sulfur dioxide SO_2 (ppb), nitrogen dioxide NO_2 (ppb), nitric oxide NO (ppb), total hydrocarbons THC (ppm), ozone O_3 (ppb), precipitation PRE (mm),

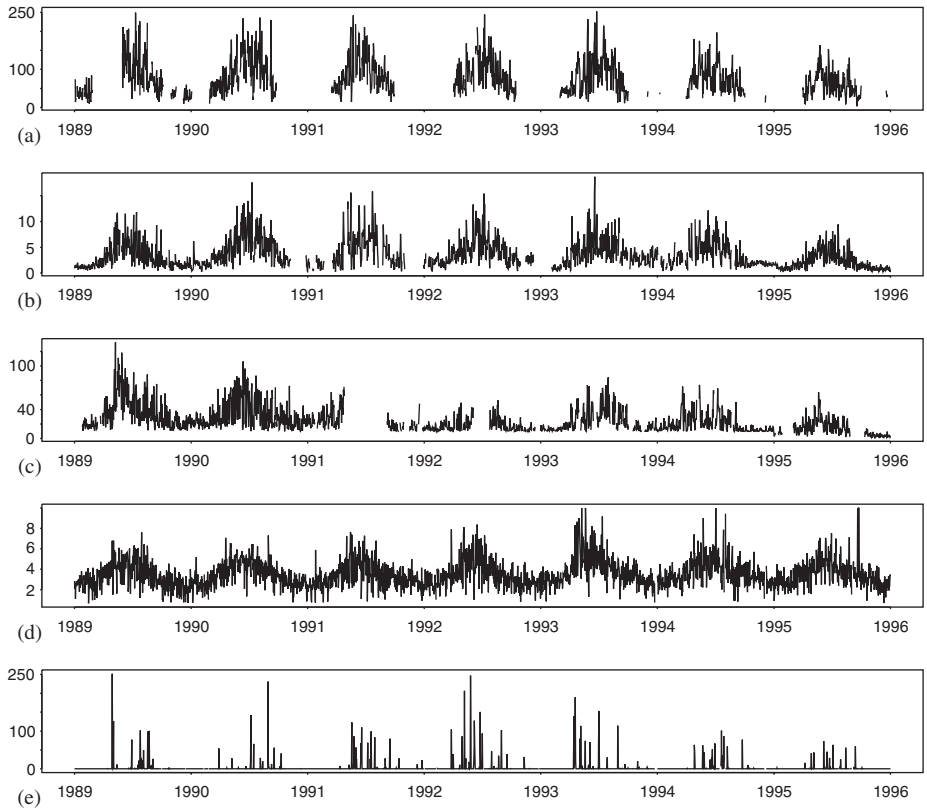

Fig. 1. Air pollution data: (a) PM2.5, (b) CO, (c) SO2, (d) 1/WS and (e) PRE/WS.

wind speed WS (m/seg), where ppm and ppb indicates parts per million and parts per billion in volume, respectively. The PM data are taken on a daily average basis, whereas the rest of the variables are measured each hour, so some preprocessing was performed to render all variables on the same temporal resolution for the purpose of analysis. One problem here is that we have a high number of missing data in the MACAM records for the period analyzed, from January 1, 1989 to December 31, 1995, so only 531 daily cases could be obtained after the data preprocessing step. The data actually used in the final model is displayed in Fig. 1.

4.3. Air pollution data modeling

In what follows, we describe the assumptions used in the modeling process. First, we begin by writing simultaneous mass balances for all pollutants being considered. Then, after assuming perfect mixing of pollutants over the city (the so-called box model approach) it is possible to construct a regression model that relates pollutants concentrations and meteorological variables. The following assumptions are made in this derivation: (i) CO

Table 1
Estimated parameters of the OLS model

	$\widehat{Intercept}$	$\widehat{CO}$	$\widehat{SO2}$	$\widehat{WS}^{-1}$	$\widehat{PRE \times WS}^{-1}$
Coefficients	8.2219	7.8415	0.7514	19.4270	−0.9621
p-value	0.1227	0.0000	0.0000	0.0054	0.1102

emissions are dominated by mobile sources and thus are dispersed in much the same way as the PM emissions coming from mobile sources (tailpipe exhaust), (ii) SO_2 emissions are dominated by stationary sources and thus are dispersed like the PM emissions coming from those sources, and (iii) Other PM emissions coming from construction activities, industrial handling and processing of dusty materials, street dust resuspension, etc. are called fugitive emissions and they are uncorrelated with CO and SO_2 emissions, that is, they have temporal and spatial patterns independent of those in (i) and (ii).

Provided that those assumptions are met in Santiago, the following model is obtained (Jorquera, 2002):

$$PM_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 CO_t + \boldsymbol{\beta}_2 SO2_t + \boldsymbol{\beta}_3 (WS_t)^{-1} + \boldsymbol{\beta}_4 PRE_t (WS_t)^{-1} + \boldsymbol{\varepsilon}_t,$$

where $t = 1, \ldots, 531$. Here the left-hand side stands for PM2.5 measured as daily averages, so in the right-hand side we have an intercept, and terms proportional to the daily averages of CO, SO_2, 1/(wind speed) and daily precipitation/(wind speed), respectively. Furthermore, the error components $\boldsymbol{\varepsilon}_t$ is assumed first to be uncorrelated, but this assumption is later removed, in order to consider the correlation structure of the data.

4.3.1. Classical approach

In order to fit a regression model to the air pollution data, we start with an OLS approach, assuming that the observations are not correlated. This model was fitted using the statistical package Splus. The estimated parameters are given in Table 1.

As expected, the coefficient associated to precipitation is negative, since raindrops washout particles down to the ground, and the associated winds improve ventilation in the city cleaning up the air. In addition, the coefficients associated to CO and SO_2 are positive, since those chemicals are released along with the particulate matter emissions from the mobile and stationary sources, respectively. However, observe that both the intercept and the precipitation variable are not significant at the 5% level.

In order to assess the level of collinearity of the covariables in the regression, we calculated the *condition number* of the matrix $\mathbf{X}$ as suggested by Belsley et al. (1980):

$$K(\mathbf{X}) = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and the smallest singular values of $\mathbf{X}$. Observe that as discussed by Rawlings (1988, p. 274), the collinearity analysis must be conducted on the standardized design matrix which corresponds to the matrix $\mathbf{X}$ with every column divided by the square root of the sum of squares of the elements of that column. Thus, the sum of

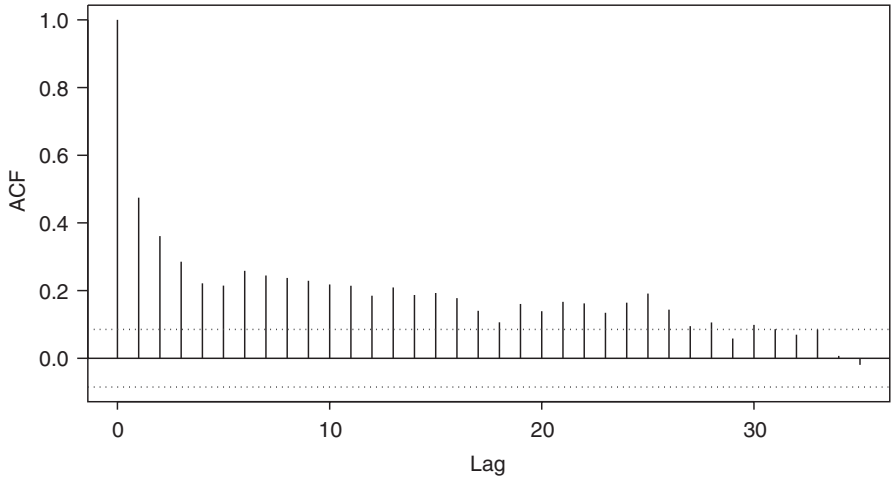

Fig. 2. Sample autocorrelation function of ordinary least squared residuals.

squares of each column is equal to one. This standardization prevent the singular values from being dominated by the *scale* of one or more independent variables.

In our case, the condition number is $K(\mathbf{X}) = 25.9$. Following, Belsley et al. (1980) this number indicates a low level of linear dependence among the covariables which will not affect the estimation or inference.

Even though this model could be useful to measure the contribution of the CO and SO_2 emissions and the precipitation levels to the pollution, it is not statistically sound since its residuals are highly correlated, see Fig. 2. This strong correlation may greatly affect the inference about the parameters of the model. To further investigate the correlation structure of the residuals we employ two techniques to assess the presence of long-memory behavior. First, Fig. 3 shows a log-var plot, see Beran (1994). The variance of the mean of a long-memory series $\{y_1, \ldots, y_k\}$, $\bar{y}_k = (y_1 + \cdots + y_k)/k$ is proportional to

$$\text{Var}\,(\bar{y}_k) \sim ak^{2d-1},$$

where a is a constant and d is the long-memory parameter. Now by taking logarithm in both sides we obtain

$$\log\,[\text{Var}\,(\bar{y}_k)] \sim \log(a) + (2d-1)\log(k).$$

Consequently, by regressing $\log\,[\text{Var}\,(\bar{y}_k)]$ on $\log(k)$ for several choices of sample size k, we obtain a log-var plot. Observe that if there is no long-memory behavior, the slope of the line should be -1 (dotted line in Fig. 3). On the other hand, the presence of long-range dependence is shown by departures from the log-var line resulting from the data (heavy line in Fig. 3).

In addition, we have plotted the periodogram of the residuals from the OLS regression, see Fig. 4. As pointed out by Brockwell and Davis (1991, Chapter 13), the spectral density

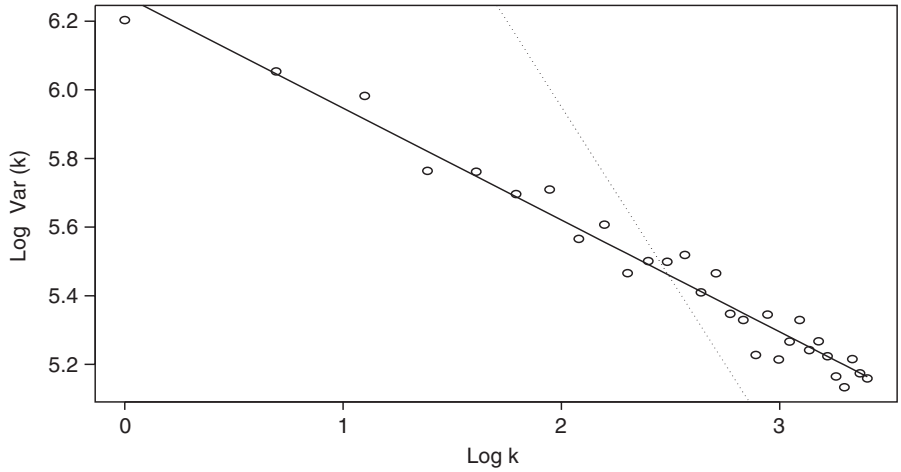

Fig. 3. Log-var plot of ordinary least squared residuals.

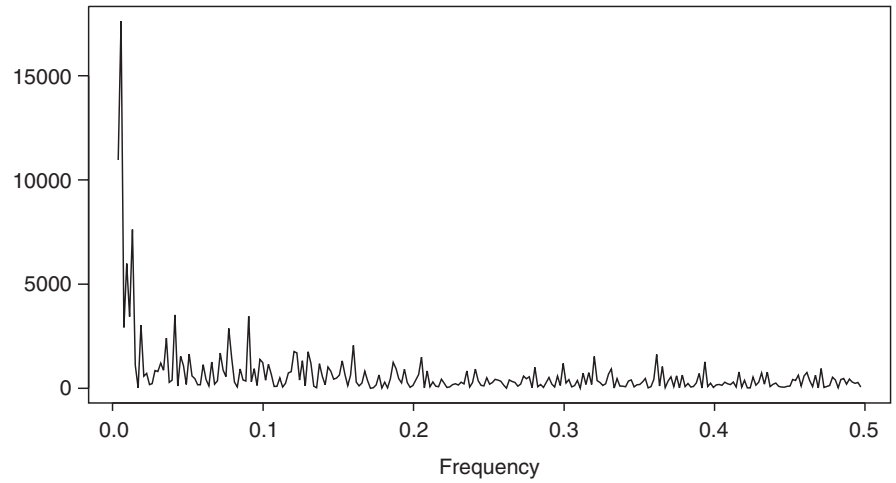

Fig. 4. Periodogram of ordinary least squared residuals.

of a long-range dependent process displays a pole at zero frequency. Note that the spectrum in Fig. 4 displays a sharp peak at zero frequency.

In order to account for the strong autocorrelation of the data, an ARFIMA regression model is fitted following the strategy described in Section 2. Thus, we first estimate the parameter $\boldsymbol{\theta}$ directly from the residuals of the OLS model, using a Kalman filter approach for handling the large number of missing values in the data. The resulting model—selected using Akaike's information criterion—is an ARFIMA$(0, d, 0)$ with $\hat{d} = 0.38$ and $t_d = 8.5$. The error variance–covariance matrix was calculated by applying formula (5) to the sample

Table 2
Estimated parameters of GLS model using ARFIMA error structure

	$\widehat{Intercept}$	$\widehat{CO}$	$\widehat{SO2}$	$\widehat{WS}^{-1}$	$\widehat{PRE \times WS}^{-1}$
Coefficients	16.63	6.49	0.67	11.32	−1.29
p-value	0.0006	0.0000	0.0000	0.0000	0.0000

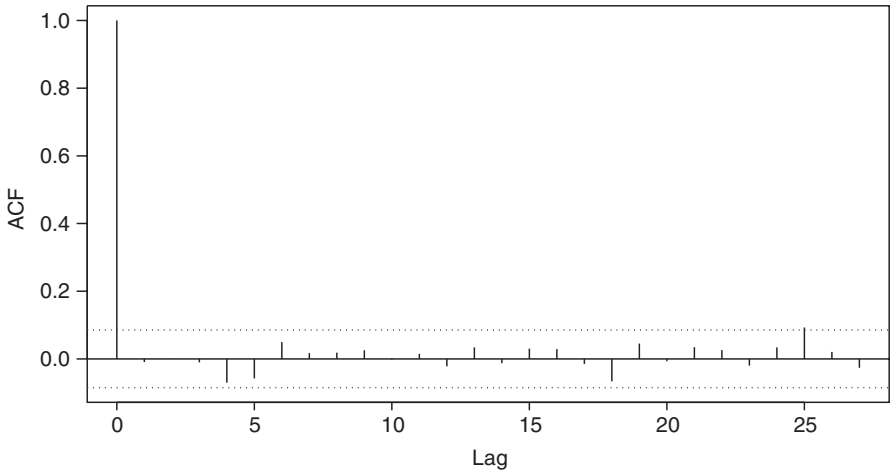

Fig. 5. Sample autocorrelation function of ARFIMA residuals.

$\boldsymbol{y} = \left(y_{t_1}, \ldots, y_{t_k}\right)'$ to obtain $\widehat{\mathrm{Var}}(\boldsymbol{y}) = \widehat{\boldsymbol{\Gamma}} = \gamma_{\hat{d}}\left(|t_i - t_j|\right)$, for $i, j = 1, \ldots, 531$, where $\hat{d}$ is the estimated value of d. Once the symmetric matrix $\widehat{\boldsymbol{\Gamma}}$ of dimension 531×531 is obtained, its square root is computed by means of the Cholesky decomposition algorithm. This calculation was carried out by using two programs, in S-plus and Fortran, respectively. Both approaches gave similar results. The estimates from the generalized linear model are shown in Table 2.

Comparing Tables 1 and 2, observe that all the estimated coefficients are similar, except the corresponding to the intercept and WS^{-1}. However, a major difference among these models appears when analyzing their residuals. Observe from Fig. 5 that the residuals from the GLS model are no longer correlated, indicating that the GLS fitting seems to be adequate.

4.3.2. Bayesian approach

With the objective of obtaining Bayesian solutions for the estimation problem, we elicited the prior distribution guided by the work presented in Jorquera (2002), where the model was derived from physical principles of PM mass conservation. In this modeling framework, $\boldsymbol{\beta}_1$ is proportional to the ratio of CO to PM tailpipe emissions from mobile sources; $\boldsymbol{\beta}_2$ is proportional to the ratio of SO_2 to PM emissions from stationary sources; the term containing $\boldsymbol{\beta}_3$ lumps together the fugitive PM emissions and the physicochemical generation

of PM in the troposphere, so these three coefficient must be essentially positive. In addition, information about the relative magnitudes of the emission of CO, SO_2 and PM in Santiago (CONAMA, 2003), plus the chemical composition of PM characteristic for Santiago (Artaxo et al., 1999) and literature values for deposition velocities (Seinfeld and Pandis, 1998) led us to the following prior specification:

$$\boldsymbol{\beta}_1 \sim N(8; 4), \quad \boldsymbol{\beta}_2 \sim N(0.4; 0.2) \quad \text{and} \quad \boldsymbol{\beta}_3 \sim N(5; 4). \tag{13}$$

These prior distributions are chosen to guarantee that the three coefficients are positive with probability around 98%. For $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, we have used typical estimates of uncertainties in emissions magnitude (20–30%) for the different pollutants; whereas for $\boldsymbol{\beta}_3$ the uncertainties in the magnitude of fugitive PM emissions are typically much higher.

Parameter $\boldsymbol{\beta}_4$ takes into account the scavenging of particles effected by raindrops, hence $\boldsymbol{\beta}_4$ is a negative (sink) term. In addition, experimental values for the scavenging process have been reported in the literature (Seinfeld and Pandis, 1998). The intercept $\boldsymbol{\beta}_0$ stands for other sources of PM, like the natural contribution to PM levels effected by natural sources such as wind erosion and natural PM background, for instance. Thus, the intercept has to be a positive parameter. From the estimated background contribution to PM at Santiago (CONAMA, 2003) and from standard data from literature we arrive at the estimates

$$\boldsymbol{\beta}_0 \sim N(20; 25) \quad \text{and} \quad \boldsymbol{\beta}_4 \sim N(-1; 0.25). \tag{14}$$

Recall that all of these terms come from independent processes: parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ represent PM emissions from mobile, stationary and fugitive sources in the city, that are mutually independent. Parameter $\boldsymbol{\beta}_0$ is a background contribution to ambient PM levels in a city. Thus, we can expect that the information about the magnitude of the anthropogenic PM emissions in a city is independent of the background (upwind) PM levels. In addition, $\boldsymbol{\beta}_0$ should reflect the high seasonality imposed by the meteorology (Jorquera, 2002), hence the high value assigned to its variance.

As for parameter $\boldsymbol{\beta}_4$, it is a sink brought in by the rain, so it is a value per unit of precipitation, i.e. is an intensity parameter determined only by the physical attributes of PM: size and solubility in water, so we expect it to be independent of the other model parameters (all particles are washed out by rain regardless of their origin).

Finally, we assume that $\boldsymbol{\beta}$ is independent of $\boldsymbol{\sigma}^2$ and for this parameter we specify a proper non-informative type prior with $\boldsymbol{\sigma}^2 \sim IG(0.0001; 0.0001)$.

The independence assumption is based on the expectation that the residuals reflect local, transient emissions effects in the PM measurements (like traffic jams, urban or forest fires, etc.), that are not accounted for in the average behavior assumed by the model through the $\boldsymbol{\beta}$ parameter.

We implemented a Fortran program to generate chains using Gibbs Sampling. Chains of length 1,000,000 were generated, discarding the 100,000 first observations to eliminate autocorrelation. Thus, chains of length 900,000 were obtained. The generated chains converge adequately. We have found that the *Bayesian using Gibbs sampling* (BUGS) program was unable to process the information required due to size of the matrix. The results are presented in Tables 3 and 4.

From Table 3, observe that classical and Bayesian estimations of the coefficients for the carbon monoxide (CO), the sulfur dioxide (SO_2) and the precipitation/wind speed

Table 3
Estimates from the Normal model without error structure

Coefficients	Posterior mean	Posterior median	Classical estimation
$\widehat{Intercept}$	19	19.1	8.24
$\widehat{CO}$	7.47	7.48	7.85
$\widehat{SO2}$	0.80	0.80	0.79
$\widehat{WS}^{-1}$	4.36	4.31	18.67
$\widehat{PRE \times WS}^{-1}$	−0.788	−0.95	−0.98

Table 4
Posterior summaries using ARFIMA error structure

Coefficients	Posterior mean	Posterior median	Classical estimation
$\widehat{Intercept}$	20.1	20.11	16.63
$\widehat{CO}$	6.45	6.45	6.49
$\widehat{SO2}$	0.57	0.66	0.67
$\widehat{WS}^{-1}$	5.81	5.79	11.32
$\widehat{PRE \times WS}^{-1}$	−1.15	−0.38	−0.96

(PRE/WS) are similar. On the other hand, the estimates for the wind speed (WS) and the intercept are different. When the strong correlation of the observations is accounted for, see Table 4, the classical and Bayesian estimates are similar except for the wind speed variable (WS). A comparison of the results from Tables 3 and 4 indicates that the estimates for the intercept and the wind speed effect (WS) from the classical method seems to change more dramatically. In addition, from Fig. 6, note that the posterior densities of all parameters are more concentrated than the priors, excepting the case for $\boldsymbol{\beta}_3$. We have to bear in mind though, that for a complex terrain domain as Santiago, a single wind speed measure in an open area might not fully represent the wind field across the city as required in the physical derivation.

5. Conclusions

This paper has two main purposes. First, to propose a strategy for the statistical analysis of regression models of environmental variables. One problem with this type of data is the presence of missing values and long-range persistence in the residual series. Therefore, special Kalman filter techniques must be used for parameter estimation. The strategy proposed has the following features: (a) The use of residuals obtained from the OLS estimation of the regression coefficient to estimate the parameters of the error model; (b) the use of a

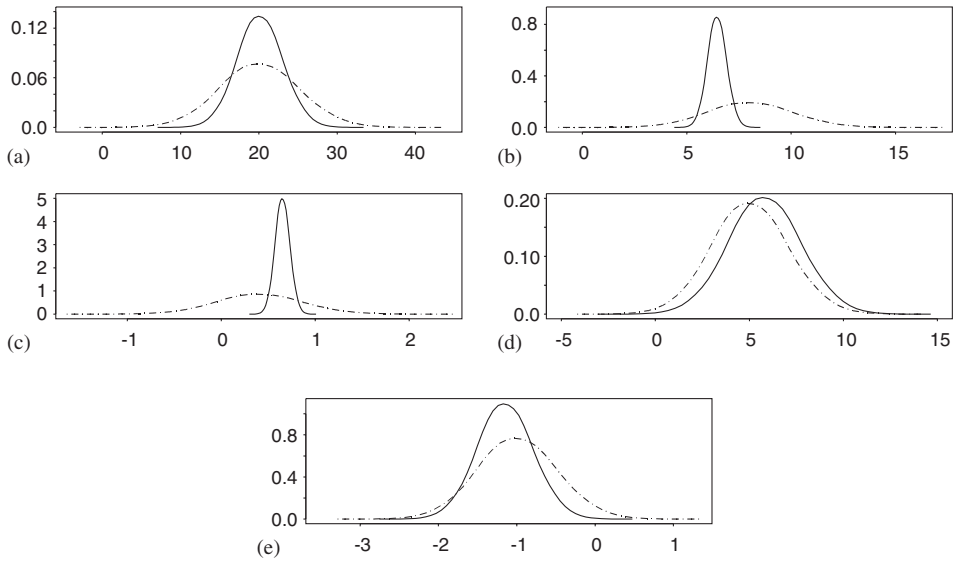

Fig. 6. Prior and posteriori densities for the ARFIMA linear model. Prior: dashed line, Posterior: solid line. (a) β_0, (b) β_1, (c) β_2, (d) β_3 and (e) β_4.

Kalman filter approach to estimate an ARFIMA model with missing data; (c) the use of a Gaussian regression model with covariance matrix following an ARFIMA specification; and (d) the use of classical and Bayesian methods for estimating the regression coefficients on the model adopted in (c). Although this strategy does not consider the simultaneous inference about all the parameters of the model, an advantage is its simplicity for dealing with a complex inference task.

Second, an application of the above methodology was developed for generating a feasible solution to the inference problem on a physical model that describes the behavior of ambient air pollution levels in Santiago. With the classical approach, we have shown that the inference (estimates and *p*-values) can be distorted if the long-memory structure of the errors is not considered. From a conditional Bayesian approach, we have incorporated prior information on the parameters elicited using the model assumptions described in Section 4.3 and the magnitudes of PM source emissions for Santiago.

From a theoretical standpoint, the treatment here is far from complete because we have not addressed the simultaneous inference of all parameters involved in the model. However, this approach is new and suggests several topics of further study. Thus, we anticipate that tools such as those associated with Bayesian inference of long memory (Koop et al., 1997) can be integrated to obtain a fully Bayesian treatment of the problem. Finally, the ideas presented here can be extended to other regression situations including Gaussian mixture ARFIMA regression models such as the Student-*t* model and to other areas of application.

Acknowledgements

The authors would like to express their gratitude to the Associate Editor and two anonymous referees for helpful and constructive comments, which led to a substantially improved version of this paper. This research was partially supported by Fondecyt grants 1030558 and 1040934. We also thank the computational and data analysis support from graduate students Natalia Bahamonde and Cristian Meza.

References

Arnold, J.R., Dennis, R.L., Tonnesen, G.S., 2003. Diagnostic evaluation of numerical air quality models with specialized ambient observations: testing the Community Multiscale Air Quality modeling system (CMAQ) at selected SOS 95 ground sites. Atmos. Environ. 37, 1185–1198.

Artaxo, P., Oyola, P., Martínez, R., 1999. Aerosol composition and source apportionment in Santiago de Chile. Nuclear Instrum. Methods Phys. Res. B 150, 409–416.

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Beran, J., 1994. Statistics for Long-Memory. Chapman-Hall, New York.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, second ed. Springer, New York.

Carlin, J., Dempster, A.R., Jonas, A.B., 1985. On models and methods for Bayesian time series analysis. J. Econometrics 30, 67–90.

Carmichael, G.R., Sandu, A., Song, C.H., He, S., Phandis, M.J., Daescu, D., Damian-Ioardache, V., Potra, F.A., 1999. Computational challengues of modeling interactions between aerosols and gas phase processes in large scale air pollution models. In: Zlatev, Z. (Ed.), Large-Scale Computations in Air Pollution Modelling. Kluwer Academic, The Netherlands, pp. 99–136.

Chen, M., Shao, Q., Ibrahim, J., 2000. Monte Carlo Methods in Bayesian Computation. Springer, New York.

Cifuentes, L.A., Vega, J., Kopfer, K., 2000. Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago, Chile. J. Air & Waste Management Assoc. 50, 1287–1298.

CONAMA, 2003. National Commission for the Environment, Chile. Background documents (in Spanish) available for download at www.conama.cl/rm.

Dahlhaus, R., 1995. Efficient location and regression estimation for long range dependent regression models. Ann. Statist. 23, 1029–1047.

Doukhan, P., Oppenheim, G., Taqqu, M.S., 2003. Theory and Applications of Long-Range Dependence. Birkhäuser, Boston.

Elsom, D., 1996. Smog Alert. Managing Urban Air Quality. Earthscan, London.

Fenger, J., 1999. Urban air quality. Atmos. Environ. 33, 4877–4900.

Hall, P., Lahiri, S.N., Polzehl, J., 1995. On bandwidth choice in nonparametric regression with both short and long range dependency errors. Ann. Statist. 23, 1921–1936.

Jacobson, M.Z., 1999. Fundamentals of Atmospheric Modeling. Cambridge University Press, New York.

Jorquera, H., 2002. Air quality at Santiago, Chile: A box modeling approach II. PM2.5, coarse and PM10 particulate matter fractions. Atmos. Environ. 36, 331–344.

Koop, G., Ley, E., Osiewalski, J., Steel, M.F.J., 1997. Bayesian analysis of long memory and persistence using ARFIMA models. J. Econometrics 78, 149–169.

Koul, H.L., Mukherjee, K., 1993. Asymptotics of R-, MD- and LAD estimators in linear regression with long range dependent errors. Probab. Theory Related Fields 95, 538–553.

Mage, D., Ozolins, G., Peterson, P., Webster, A., Orthofer, R., Vandeweerd, V., Gwynne, M., 1996. Urban air pollution in megacities of the world. Atmos. Environ. 30, 681–686.

O'Hagan, P., 1994. Bayesian Inference, In: Kendall (Ed.), Advanced Theory of Statistics. vol. 2A, Wiley, New York.

Pai, J.S., Ravishanker, N., 1998. Bayesian analysis of autorregresive fractionally integrated moving-average processes. J. Time Ser. Anal. 19, 99–112.

Palma, W., Chan, N.H., 1997. Estimation and forecasting of long-memory processes with missing values. J. Forecasting 16, 395–410.

Palma, W., Del Pino, G., 1999. Statistical analysis of incomplete long-range dependent data. Biometrika 86, 165–172.

Petris, G., 1997. Bayesian analysis of long memory time series. Ph.D. Dissertation, Institute of Statistics and Reason Sciences, Duke University.

Philippe, A., Rousseau, J., 2002. Non informative priors in the case of Gaussian long-memory processes. Bernoulli 8, 451–473.

Rawlings, J., 1988. Applied Regression Analysis: A Research Tool. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.

Robert, C.P., Casella, G., 2004. Monte Carlo Statistical Methods, second ed. Springer, New York.

Robinson, P.M., Hidalgo, F.J., 1997. Time series regression with long-range dependence. Ann. Statist. 25, 77–104.

Seinfeld, J.H., Pandis, S.N., 1998. Atmospheric Chemistry and Physics, second ed. Wiley, New York

Sowell, F., 1992. Maximum likelihood estimation of stationary univariate fractionally integrated models. J. Econometrics 53, 165–188.

Vedal, S., 1997. Ambient particles and health: lines that divide. J. Air & Waste Management Assoc. 47, 551–581.

Yajima, Y., 1988. On estimation of a regression model with long-memory stationary errors. Ann. Statist. 16, 791–807.

Yajima, Y., 1991. Asymptotic properties of the LSE in a regression model with long-memory stationary errors. Ann. Statist. 19, 158–177.

Yajima, Y., Nishino, H., 1999. Estimation of the autocorrelation function of a stationary time series with missing observations. Sankhyā Ser. A 61, 189–207.