ENVIRONMETRICS
Environmetrics 2008; **19**: 582–601
Published online 18 December 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/env.899

Bayesian analysis of dynamic factor models: an application to air pollution and mortality in São Paulo, Brazil

T. Sáfadi[1*,†] and D. Peña[2]

[1]*Departamento de Ciências Exatas, Universidade Federal de Lavras, 37200-000, Lavras, Minas Gerais, Brazil*
[2]*Departamento de Estadística, Faculdad de Ciencias Sociales, Universidad Carlos III de Madrid, Getafe, Spain*

SUMMARY

The Bayesian estimation of a dynamic factor model where the factors follow a multivariate autoregressive model is presented. We derive the posterior distributions for the parameters and the factors and use Monte Carlo methods to compute them. The model is applied to study the association between air pollution and mortality in the city of São Paulo, Brazil. Statistical analysis was performed through a Bayesian analysis of a dynamic factor model. The series considered were minimal temperature, relative humidity, air pollutant of PM_{10} and CO, mortality circulatory disease and mortality respiratory disease. We found a strong association between air pollutant (PM_{10}), Humidity and mortality respiratory disease for the city of São Paulo.

KEY WORDS: air pollution; data augmentation; factor model; Gibbs sampler; VAR model

1. INTRODUCTION

A main problem in building a model for a vector of time series is that the number of parameters grows with the square of the dimension of the vector. Therefore models for dimension reduction are needed to model a large number of time series. A useful tool for dimension reduction in time series are dynamic factor models (Geweke and Singleton, 1981; Peña and Box, 1987; Stock and Watson, 1988; Molenaar *et al*., 1992; Forni *et al*., 2000; Peña and Poncela, 2004, 2006), among others. Dimension reduction can be improved by incorporating useful prior information and thus the Bayesian analysis of dynamic factor models seems a promising line of research. The standard factor model has a very appropriate structure for MCMC methods, and this has been illustrated by Lee and Shi (2000), who derived a joint estimation for the factor scores and structural parameters, and by Lopes and West (2004), who used reversible jump MCMC methods to estimate the number of factors of made model assessment. From the dynamic point of view (West *et al*., 1999; Aguilar and West, 2000) are important references, and recently Lopes and Carvalho (2007) have proposed a spatial dynamic factor models. In this paper, we extend some of these results by developing a full Bayesian approach in estimating jointly

*Correspondence to: T. Sáfadi, Departamento de Ciências Exatas, Universidade Federal de Lavras, 37200-000, Lavras, Minas Gerais, Brasil.
†E-mail: safadi@ufla.br

Received 31 January 2007
Accepted 23 October 2007

the parameter vector θ and the factor scores in the dynamic factor model when the factors follow a VAR(p) model.

We apply the dynamic factor model to investigate the relationship between time series of mortality (circulatory and respiratory disease) and time series of air pollutants (PM_{10}, CO). We consider for the analysis minimal temperature, relative humidity, PM_{10}, CO, mortality respiratory disease and mortality circulatory disease in the city of São Paulo, Brazil, from 1994 to 1997. The relationship between air pollutant levels and mortality has been previously studied in the city of Madrid, Spain, during the period 1986–1992 by Odriozot *et al*. (1998). They considered multivariate integrated moving-average (ARIMA) models to adjust season, temperature, relative humidity and influenza. Pope and Dockery (1992) studied the association between daily changes in respiratory health and respirable particulate pollution (PM_{10}) in Utah Valley during the winter of 1990–1991. During the study period, 24-h PM_{10} concentrations ranged from 7 to 251 $\mu g/m^3$. Participants included symptomatic and asymptomatic samples of fifth- and sixth-grade students. Large associations between the incidence of respiratory symptoms, especially cough and PM_{10} pollution were also observed for both samples. Immediate and delayed PM_{10} effects were observed. Respiratory symptoms were more closely associated with 5-day moving-average PM_{10} levels than with concurrent-day levels. These associations were also observed at PM_{10} levels below the 24-h standard of 150 $\mu g/m^3$. This study indicates that both symptomatic and asymptomatic children may suffer acute health effects of respirable particulate pollution, with symptomatic children suffering the most.

In this paper, we have two main contributions. First, we develop a full Bayesian approach in estimating jointly the parameter vector θ and the factor scores in the dynamic factor model when the factors follow a VAR(p) model. Second, we show how this model can be applied to investigate the relationship between air pollution and mortality and found a strong relationship between these variables in the city of São Paulo, in Brazil.

The rest of the paper is organized as follows. In Section 2, we present the dynamic factor model and in Section 3 we developed its Bayesian analysis. Section 4 analyses the pollution and mortality in São Paulo, Brazil , considering univariate and multivariate autoregressive model and a dynamic factor model. Finally, Section 5 presents some conclusions. The derivations for the posteriors distributions required for the MCMC algorithm are given in the Appendix A.

2. DYNAMIC FACTOR MODELS

We consider in this paper a factor model given by the following two equations:

$$\begin{aligned} y_t &= \beta + Cf_t + e_t, \\ f_t &= \sum_{i=1}^{p} \rho_i f_{t-i} + w_t \end{aligned} \tag{1}$$

where y_t is a $q \times 1$ vector of observed time series, β is the $q \times 1$ mean vector and C is a $q \times k$ matrix of unknown constants, the factor loading matrix. The specific components, e_t, are independent normal q-vectors with $e_t \sim N(0, \Gamma)$, where Γ is a diagonal matrix, $\Gamma = \text{diag}(\psi_1, \psi_2, \ldots, \psi_q)$. The factors f_t are represented by a $k \times 1$ vector which follows a multivariate autoregressive model, where the AR matrices, ρ_i are diagonal matrices with $\rho_i = \text{diag}(\rho_{i1}, \ldots, \rho_{ik})$, $i = 1, \ldots, p$ and $\{\rho_{1j}, \rho_{2j}, \ldots, \rho_{pj}\}$ satisfy the

DOI: 10.1002/env

stationary condition, $j = 1, \ldots, k$ and w_t are independent normal k-vectors with $w_t \sim N(0, I_k)$, where I_k is the identity matrix, and e_t and w_s are independent for all t and s.

From this model we have that $f_t \sim N(0, U)$ and $U = \sum_{i=1}^{p} \rho_i' U \rho_i + I_k$, for all t and $y_t \sim N(\beta, \Sigma)$, where Σ satisfies $\Sigma = CUC' + \Gamma$.

In practical problems, specially with large values of q, the number of factors k will often be small relative to q, so that much of the variance structure is explained by the common factors. As is well known, the k-factor model must be further constrained to define a unique model free from identification problems. A solution adopted here is to constrain the matrix C so that it is a block lower triangular matrix, assumed to be of full rank. That is,

$$C = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ c_{21} & 1 & 0 & \ldots & 0 \\ c_{31} & c_{32} & 1 & \ldots & 0 \\ : & : & : & \ldots & : \\ c_{k1} & c_{k2} & c_{k3} & \ldots & 1 \\ : & : & : & \ldots & : \\ c_{q1} & c_{q2} & c_{q3} & \ldots & c_{qk} \end{pmatrix}$$

This form is used (e.g. Geweke and Zhou, 1996; Aguilar and West, 2000; Lee and Shi, 2000), and provides both identification and useful interpretation of the factor model. From a Bayesian point of view this is equivalent to assigning fixed values to these parameters with probability one, and in the analysis they are not estimated.

The number of factor can be selected by analysing the eigenvalues and eigenvectors of the autocovariance matrices, as shown by Peña and Box (1987). Also, the number of factors can be obtained by a model selection criteria which approximates the posterior probability of the models such as BIC. If the number of factors is supposed unknown, it could be treated as a parameter. An alternative that will be explored elsewhere is to fully account for the uncertainty in the number of common factors by using for instance reversible jump MCMC, as suggested by Lopes and West (2004). However, our objective here is to develop the joint estimation of the parameter vector and the factor scores in this model and this is the objective of the next section.

3. BAYESIAN ESTIMATION OF THE DYNAMIC FACTOR MODEL

As (Lee and Shi, 2000), we developed a procedure based on data augmentation. The essential idea is to determine posterior distributions for all unknown parameters conditional on the latent factor and then the conditional distribution of the latent factor given the observable and the other parameters. That is, the observable data are 'augmented' by samples from the conditional distribution for the factor given the data and the parameters of the model. Specifically, the joint posterior distribution for the unknown parameters and the unobserved factors can be sampled by using a Markov Chain Monte Carlo procedure on the full set of conditional distributions. So, the conditional distribution for the factors $F = (f_{p+1}, \ldots, f_n)$ and the parameters $\theta = (\beta, C, \Gamma, \rho)$ are $\theta|F, Y$ and $F|\theta, Y$, where $Y = (y_{p+1}, \ldots, y_n)$, conditioning on the

DOI: 10.1002/env

p first observations can be approximated by

$$\begin{aligned}
&\mathcal{L}(Y, F|\beta, C, \Gamma, \rho, Y_1, \ldots, Y_p, f_1, \ldots, f_p) \\
&\quad \propto \prod_{t=p+1}^{n} |\Gamma|^{-1/2} \exp\left\{-\frac{1}{2}(y_t - \beta - Cf_t)'\Gamma^{-1}(y_t - \beta - Cf_t)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(f_t - \sum_{i=1}^{p} \rho_i f_{t-i})'(f_t - \sum_{i=1}^{p} \rho_i f_{t-i})\right\} \qquad (2)
\end{aligned}$$

We assume independent prior distributions given as $P(C)P(\rho)P(\beta) \propto$ constant, $\gamma_i = \psi_i^{-1} \sim \Gamma(\alpha_0, \beta_0)$, so that the distribution of ψ_i is an Inverse Gamma, for each components of $\Gamma = \text{diag}\{\psi_1, \psi_2, \ldots, \psi_q\}$.

To implement the Gibbs sampler, we need to derive the full conditional posterior distribution of each parameter given all the others parameters. We present here the full conditional distribution for $\theta = (\beta, C, \Gamma, \rho)$ and F. The derivations for the posteriors are in the Appendix A.

(a) Posterior for $\beta|C, F, \Gamma, \rho, Y$

$$\beta|C, F, \Gamma, \rho, Y \sim N\left(\frac{1}{n-p}\left(\sum_{t=p+1}^{n}(y_t - Cf_t), \frac{\Gamma}{n-p}\right)\right.$$

(b) Posterior for $C|Y, F, \rho, \Gamma, \beta$

$$C_i^*|Y, F, \rho, \Gamma, \beta \sim N\left(\left[\left(\sum_{t=p+1}^{n} f_t f_t'\right)^{-1}\left(\sum_{t=p+1}^{n} f_t(y_{it} - \beta_i)\right)\right], \left(\gamma_i^{-1}\left(\sum_{t=p+1}^{n} f_t f_t'\right)^{-1}\right)\right)$$

where $C_i^{*'}$ is the ith row of C, $i = 1, \ldots, q$.

(c) Posterior for $\Gamma|C, \rho, \beta, F, Y$

We have, for $\Gamma = \text{diag}(\gamma_1^{-1}, \ldots, \gamma_q^{-1})$

$\gamma_i|C, \rho, \beta, F, Y \sim \Gamma(\alpha_i, \beta_i)$

where for each $i = 1, \ldots, q$, $\alpha_i = (n - p + 2\alpha_0)/2$ and $\beta_i = [2\beta_0 + \sum_{t=p+1}^{n}(y_{it} - \beta_i - C_i^{*'} f_t)'(y_{it} - \beta_i - C_i^{*'} f_t)]/2$.

(d) Posterior for $\rho|\Gamma, \beta, C, F, Y$

Let $B_t = [\text{diag}(f_{t-1}), \text{diag}(f_{t-2}), \ldots, \text{diag}(f_{t-p})]$ a $(k \times kp)$ matrix and $\rho_v = (\rho_{11}, \ldots, \rho_{1p}, \rho_{21}, \ldots, \rho_{kp})'$ a $(kp \times 1)$ vector. Note that $\sum_{i=1}^{p} \rho_i f_{t-i} = B_t \rho_v$, then

$$\rho_v|\Gamma, \beta, C, F, Y \sim N\left(\left[\left(\sum_{t=p+1}^{n} B_t' B_t\right)^{-1}\left(\sum_{t=p+1}^{n} B_t' f_t\right)\right], \left(\sum_{t=p+1}^{n} B_t' B_t\right)^{-1}\right)$$

DOI: 10.1002/env

This distribution is the multivariate normal truncated so that $(\rho_{1i}, \rho_{2i}, \ldots, \rho_{pi})$ satisfy the stationary condition for $i = 1, \ldots, k$. This distribution complete the Gibbs sampler for $\theta = (\beta, C, \Gamma, \rho)$. The full posterior for the factors is given by:

(e) Posterior for $F|\theta, Y$

We have, for $F = \{f_t, t = 1, \ldots, n\}$

$$f_t|\beta, \Gamma, C, Y \sim N\left(\left(U^{-1} + C'\Gamma^{-1}C\right)^{-1}\left(C'\Gamma^{-1}(y_t - \beta)\right), \left(U^{-1} + C'\Gamma^{-1}C\right)^{-1}\right)$$

$t = 1, \ldots, p$, and for each $t = p + 1, \ldots, n$,

$$f_t|f_{t-1}, \ldots, f_{t-p}, \rho, \beta, \Gamma, C, Y \sim N\left(H^{-1}\left(C'\Gamma^{-1}(y_t - \beta) + \sum_{i=1}^{p} \rho_i f_{t-i}\right), H^{-1}\right)$$

where $H = (C'\Gamma^{-1}C + I_k)$.

From results given in (a–d), the derivation of the posterior distribution $P(\theta|F, Y)$ is completed. This distribution will be extended to handle the general situation with fixed known elements in C as follows: Let $\delta_{ij} = 0$ if c_{ij} is a fixed parameter of C and $\delta_{ij} = 1$ if c_{ij} is an unknown parameter for, $i = 1, \ldots q, j = 1, \ldots, k$ and $r_i = \delta_{i1} + \delta_{i2} + \cdots + \delta_{ik}$. Moreover, let D_i' be the row vector that contains the unknown parameters in C_i and F_i be the $r_i \times n$ submatrix of $F(k \times n)$ such that for $j = 1, \ldots, k$, all the rows corresponding to $\delta_{ij} = 0$, $j = 1, \ldots, k$ are deleted, and $\tilde{y}_{it} = y_{it} - \beta_i - \sum_{j=1}^{k} c_{ij} f_{jt}(1 - \delta_{ij})$. Then, the components of the posterior for γ_i and C_i' will be changed by D_i', F_i and $\tilde{y}_{it}$. In this case, there are a total of $(qk - k(k + 1)/2 + pk + q)$ free unknown structural parameters in this model.

4. POLLUTION AND MORTALITY IN SÃO PAULO, BRAZIL

São Paulo has a state air pollution controlling agency (CETESB) with 11 monitoring stations that provide daily records of sulfur dioxide (SO_2) (24-h mean), carbon monoxide (CO) (greatest 8-h moving average), inhalable particulate matter (PM) less than 10 μm in diameter, PM_{10}, (24-h mean) and ozone (O_3)(24-h peak) concentrations. The measurements are based on different time intervals mostly because the health standards required by Brazilian legislation were defined using those time windows.

However, not all stations provide measurements of all the pollutants. Because the trend and the variability of the pollutant concentrations are similar for all stations, that is, when pollution levels increase in the central area, there is a proportional increase in the suburbs, the values records obtained were averaged and considered indicative of the citywide status. Particles in the air come form a range of sources and range widely in size and chemical composition from place to place and time-to-time. Natural sources include pollen and sea spray; industrial sources include combustion processes, quarrying and aggregate handling and transport sources include diesel vehicle exhaust emission and dust from tyre and brake wear. Other sources such as smoking produces by far the greatest concentration of particles ingested in those who smoke.

DOI: 10.1002/env

PM_{10} is defined as PM with a mass median aerodynamic diameter less than 10 μm. In other words, these are the (smaller) particles that make it through some type of pre-separator (removes large particle) and are collected on a sampling medium filter. PM_{10} is therefore PM which is very small, remains suspended in the air for periods and is easily inhaled into the deep lung. Among the regulated pollutants, PM_{10} is the only chemical non-specific agent. All these characteristics of PM_{10} (and more) have made identifying health effects associated with environmental levels of PM_{10} a significant issue. Increase death (mortality) and disease (morbidity) have been linked to periods of time of high outdoor PM_{10} concentrations. It is difficult to estimate the relative contribution to the particles in the air for any one place, but in 1995, transport was the biggest source of primary particles, accounting for some 25%, power generation emitting 15%, mining and quarrying emitting 12% and domestic and commercial heating 11%. The effects are likely to vary depending on what the particles in the air actually are, and because it is difficult to do studies on large populations, which can distinguish between different mixtures, the findings tend to be rather generalized. It is now accepted widely that populations living in areas with higher airborne particle concentrations show a range of differences in health to otherwise similar people. The effects include higher death rates, respiratory and circulatory effects and cancer.

Carbon monoxide, CO, is a colourless, odourless gas produced from the incomplete burning of virtually any combustible product. In European urban areas, CO is produced almost entirely (90%) from road traffic emissions. It survives in the atmosphere for a period of approximately 1 month but is eventually oxidized to carbon dioxide (CO_2). It may accumulate indoors as a result of tobacco smoking, poorly ventilated appliances and attached garages. Carbon monoxide enters the blood from the lungs and combines with haemoglobin, blocking the blood's ability to carry oxygen to body cells. Symptoms of carbon monoxide exposure may mimic influenza and include fatigue, headache, dizziness, nausea and vomiting, mental confusion and rapid heart rate. Depending on the level of exposure, carbon monoxide can be immediately fatal. This can lead to a significant reduction in the supply of oxygen to the heart, particularly in people suffering from heart disease. Long-term, low-level exposures to carbon monoxide by pregnant woman have the potential to injure the developing foetus.

Air pollution and child mortality in São Paulo, from 1994 to 1996 was studied by Conceição *et al*. (2001). Statistical analysis was performed through a generalized additive model considering a Poisson response distribution and a log link. Explanatory variables were time, temperature, humidity and pollutant concentrations. Safadi and Morettin (2001) considered the open loop threshold autoregressive models to study the daily number of deaths caused by heart problems and the minimal temperature in São Paulo, from 1994 to 1997. They noted that for temperatures between 12.9 and 15.23°C there was an increase of 300 deaths.

In this work, we consider weekly data for minimal temperature (Temp), relative humidity (Humid), PM_{10}, CO, mortality respiratory disease (Resp) and mortality circulatory disease (Card) in the city of São Paulo during the period from January, 1994 to December, 1997, with a total of 208 points. Unfortunately, O_3 and SO_2 were not included in the present analysis because there was so many missing points.

Figure 1 displays time series graphics of the weekly Temp, Humid, PM_{10}, CO, Resp and Card and their autocorrelation function, ACF, are in Figure 2. In some of them it is easy to see a period of 52 weeks (a year), but in others this periodicity is not clear. However, if we build the periodogram for each series, see Figure 3, the seasonal period of 52 weeks is clearly seen in all of them.

DOI: 10.1002/env

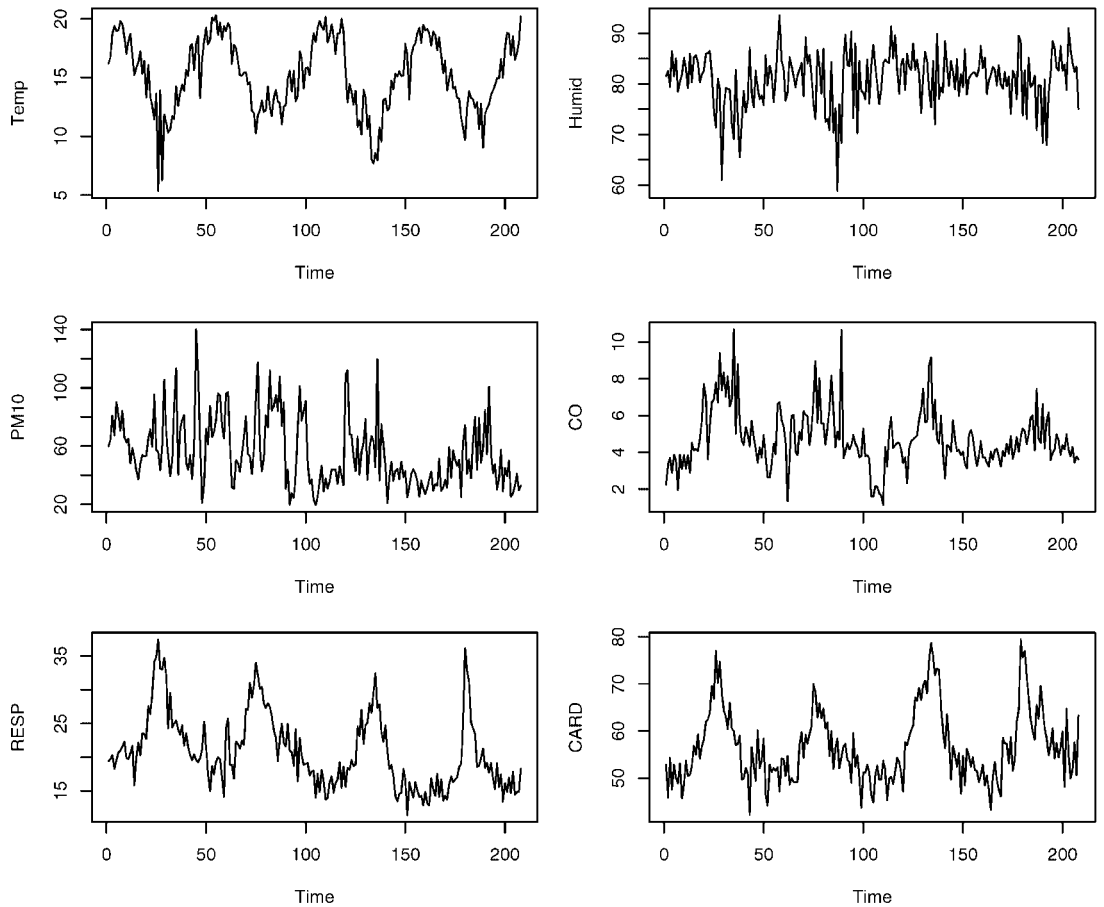

Figure 1. Weekly data for Temp, Humid, PM_{10}, CO, Resp and Card (1994–1997)

As the seasonal component seems to be stable we made seasonal adjustment by fitting a term of the form $\mu + \sum_{j=1}^{26}(A_j \cos(2\pi jt/52) + B_j \sin(2\pi jt/52)), t = 1, \ldots, 208$ and computing the seasonally adjusted series by

$$z_t = y_t - \sum_{j=1}^{26} \left(\widehat{A_j} \cos(2\pi jt/52) + \widehat{B_j} \sin(2\pi jt/52)\right)$$

Table 1 presents the significant parameters μ, A_j and B_j, $j = 1, \ldots, 28$, for each of the series. The seasonally adjusted series are showed in Figure 4, and their ACF and PACF plots are given in Figures 5 and 6. In order to decide if these series are stationary a Dickey-Fuller test was applied to these seasonally adjusted series. The results are given in Table 2, where we see that the hypothesis of a unit root is rejected by this test for all the series.

DOI: 10.1002/env

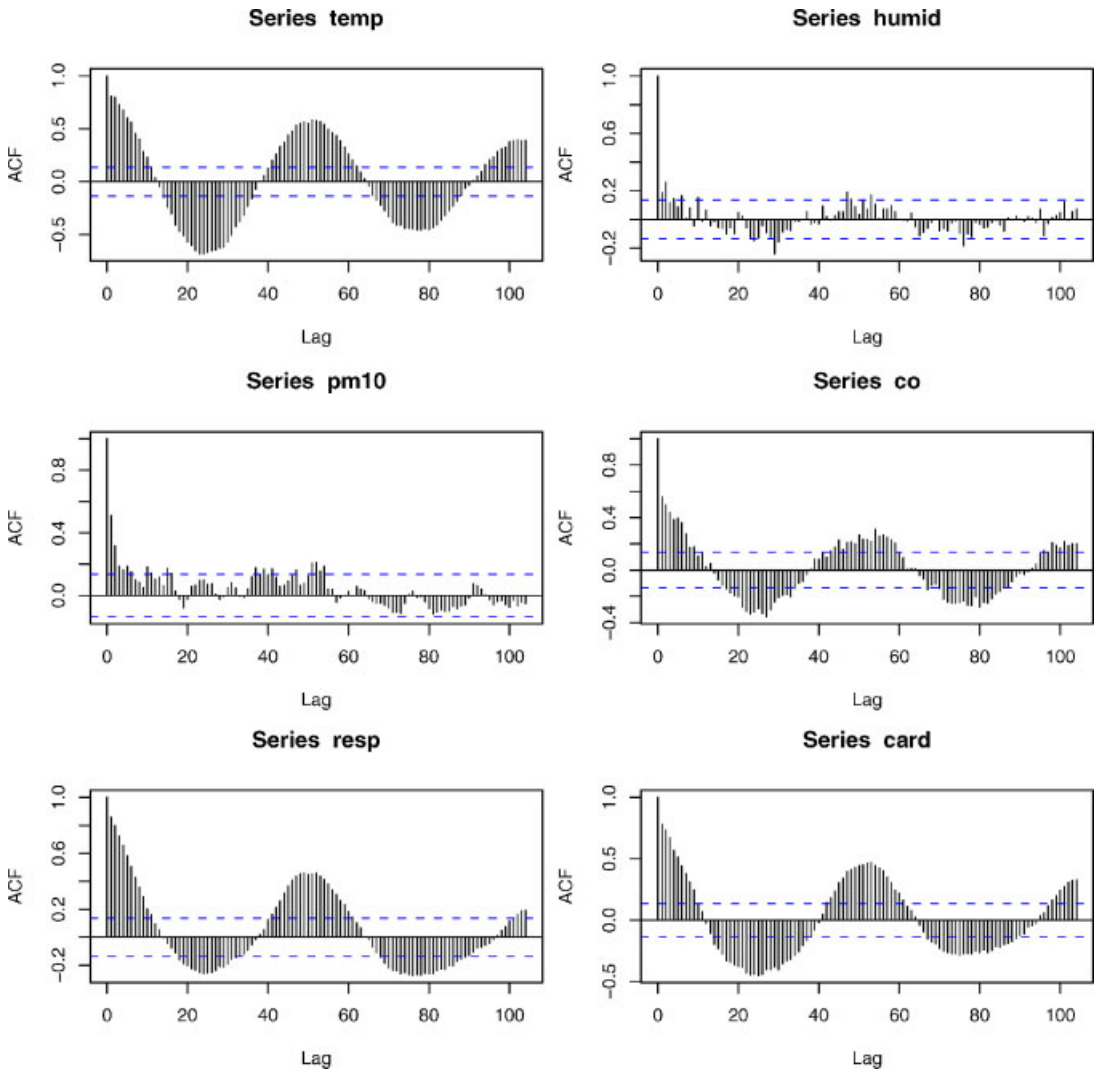

Figure 2. The autocorrelation function for Temp, Humid, PM_{10}, CO, Resp and Card

Univariate autoregressive models were fitted to the seasonally adjusted series, and Table 3 shows the coefficients for the AR(2) models selected for each series. We also fitted multivariate autoregressive models and Table 4 presents the BIC values for VAR (vector autoregressive) models of orders $p = 1, 2, 3, 4$. The minimum value of the BIC criterion corresponds to a VAR(1) model, which is the one selected, and includes 63 parameters: 21 for the residual covariance matrix, 6 for the mean and 36 for the AR(1) parameter matrix.

The Bayesian analysis of the dynamic factor model is started considering a 1-factor model, where the factor follows a AR(2) model and the loading matrix is complete, that is, without constrain. The priors are as before and for the covariance matrix we suppose $\alpha_0 = 4$ and $\beta_0 = 3$. For the analysis,

DOI: 10.1002/env

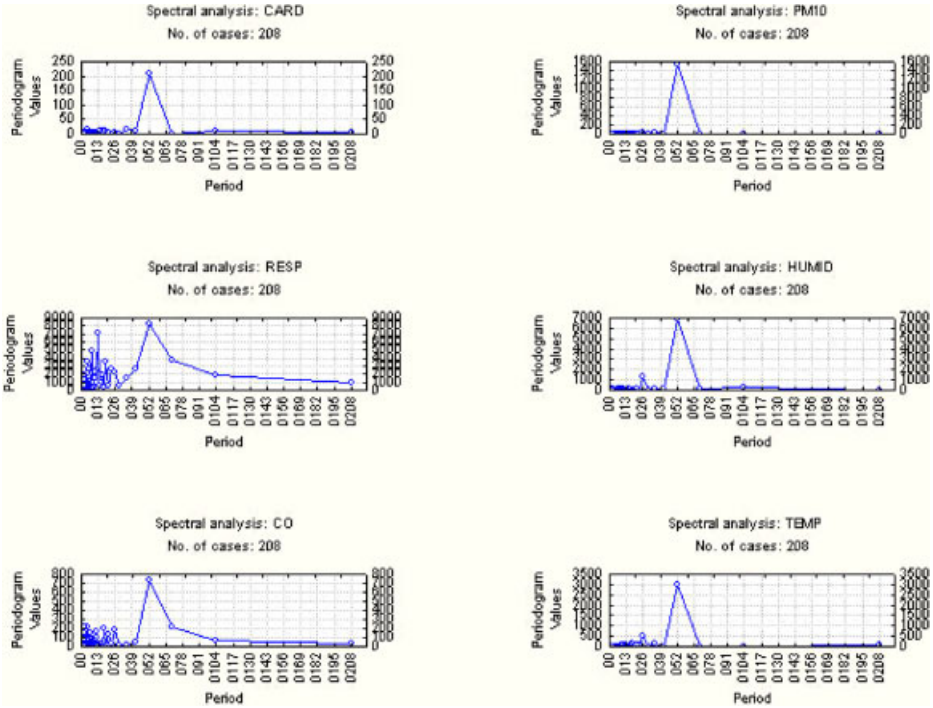

Figure 3. Periodogram for Temp, Humid, PM_{10}, CO, Resp and Card records

Table 1. Coefficients μ, A and B and standard error in parentheses

	Temp	Humid	PM_{10}	CO	Resp	Card
μ	15.216	80.874	56.186	4.751	20.843	56.713
	(0.105)	(0.351)	(1.522)	(0.091)	(0.27)	(0.323)
$\cos(2\pi t/52)$	3.477	1.273	−6.862	−1.265	−5.237	−7.90.5
	(0.149)	(0.500)	(2.166)	(0.129)	(0.379)	(0.457)
$\sin(2\pi t/52)$	1.668	2.160		−0.551	−0.812	−2.299
	(0.149)	(0.497)		(0.128)	(0.377)	(0.457)
$\cos(4\pi t/52)$	−0.458				2.218	3.511
	(0.149)				(0.377)	(0.456)
$\sin(4\pi t/52)$		−1.349	4.511			
		(0.497)	(2.154)			
$\cos(6\pi t/52)$					−1.139	
					(0.377)	
$\cos(8\pi t/52)$			−8.060			
			(2.152)			
$\cos(18\pi t/52)$			5.047			
			(2.153)			
$\sin(20\pi t/52)$		1.361	−4.64			
		(0.487)	(2.151)			
$\sin(30\pi t/52)$		−1.029				
		(0.498)				

DOI: 10.1002/env

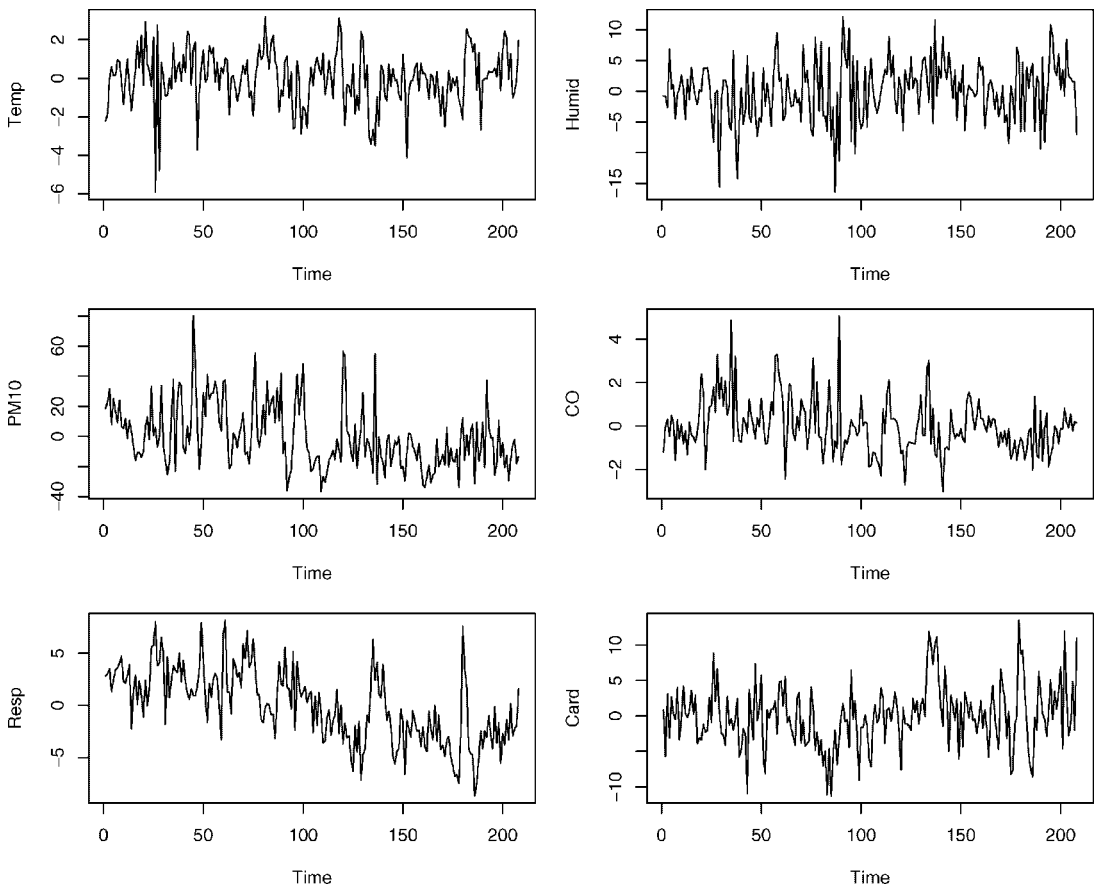

Figure 4. Seasonally adjusted series (Temp, Humid, PM_{10}, CO, Resp, Card)

we considered two parallel chains and verify the convergence by the Gelman-Rubin criterion (Gelman and Rubin (1992)) with 15 000 iterations each, skip the first 50% and for the remaining observations, took one in each 15. The posterior mean of the factor loading matrix C and its normalization form, C^*, are given in Table 5.

This matrix indicates that the load associated with the pollutant PM_{10} is much larger than the others indicating that this first factor is mostly associated to this variable and also, but with less weight to the variable Resp. The correlation between the factor and the variable PM_{10} was very high and the plot of the two series indicates a strong relationship. Thus, we conclude that the first factor is mostly dominated by the variable PM_{10} and mortality respiratory disease, variable Resp. To simplify the computations as this factor is well identified we decide to continue the analysis eliminating the effect of this factor. We do this by deleting the variable PM_{10} from the analysis and by eliminating from the series Resp the effect due to PM_{10}. Thus, we fit the regression $\text{Resp} = 0.0093 - 0.0525\text{PM}_{10} + \text{Res}$, and in the rest of this analysis we consider the five series Temp, Humid, CO, Card and Res, where Res is the residual of the regression between Resp and PM_{10}.

DOI: 10.1002/env

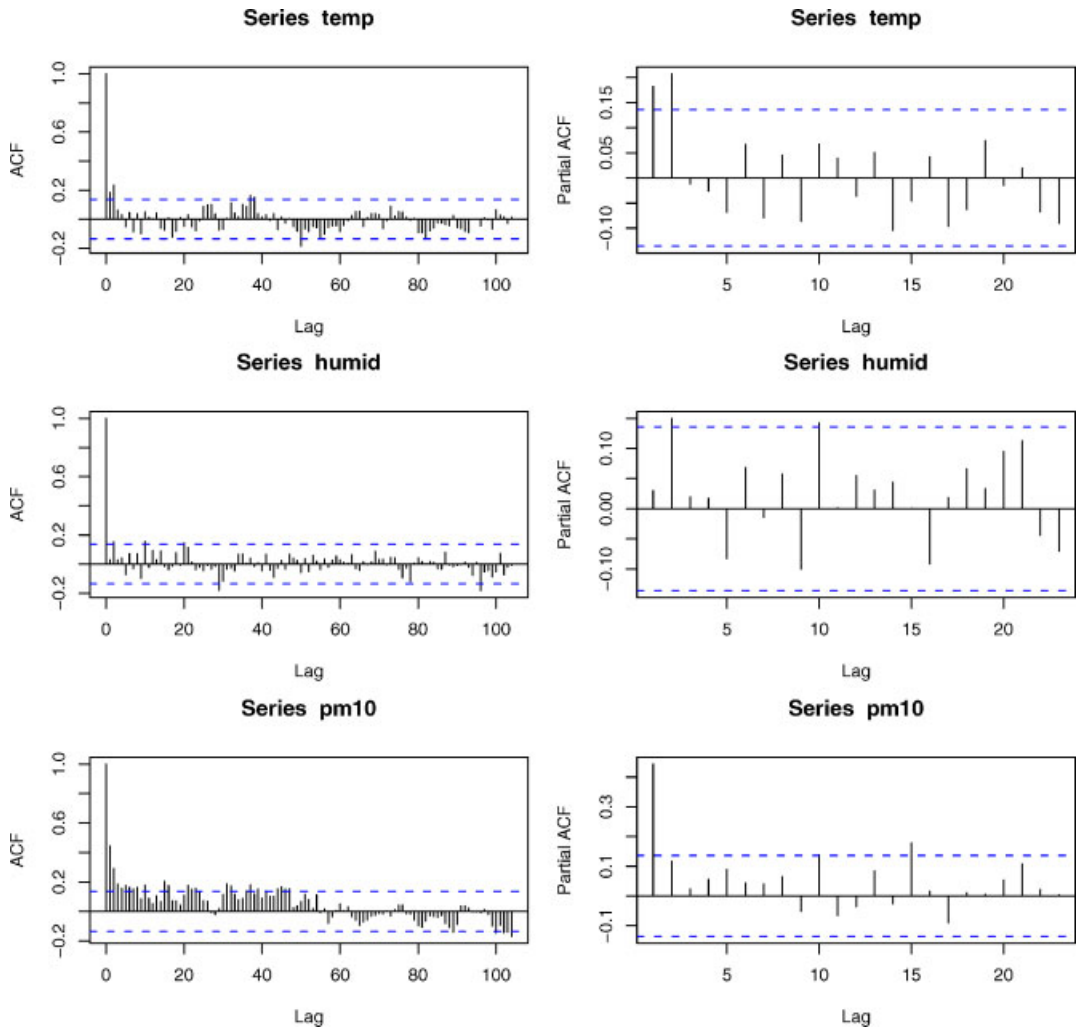

Figure 5. Autocorrelation and partial autocorrelation function for the seasonally adjusted series (Temp, Humid, PM_{10})

For this five series we fitted factor models with k factors following an autoregressive AR(p) model, for $k = 1, 2$ and $p = 1, 2$. Table 6 shows the BIC values for each case and we conclude that a 2-factor model following a AR(1) model seems appropriate for the data. Table 7 shows the posterior mean for the factor-loading matrix, C, for this model. The first factor has positive association with Temp and Humid and negative with the three mortality variables. Thus, it represent an indicator of health clime conditions and takes into account that mostly low temperature, and also although in less degree low humidity, are associated to higher mortality, especially for circulatory disease in the city of São Paulo. The second factor separates the two types of mortality considered and associates high humidity to smaller mortality for respiratory disease. Note that the effect of the variable CO is small on the two factors which are associated mostly with the weather and mortality variables.

DOI: 10.1002/env

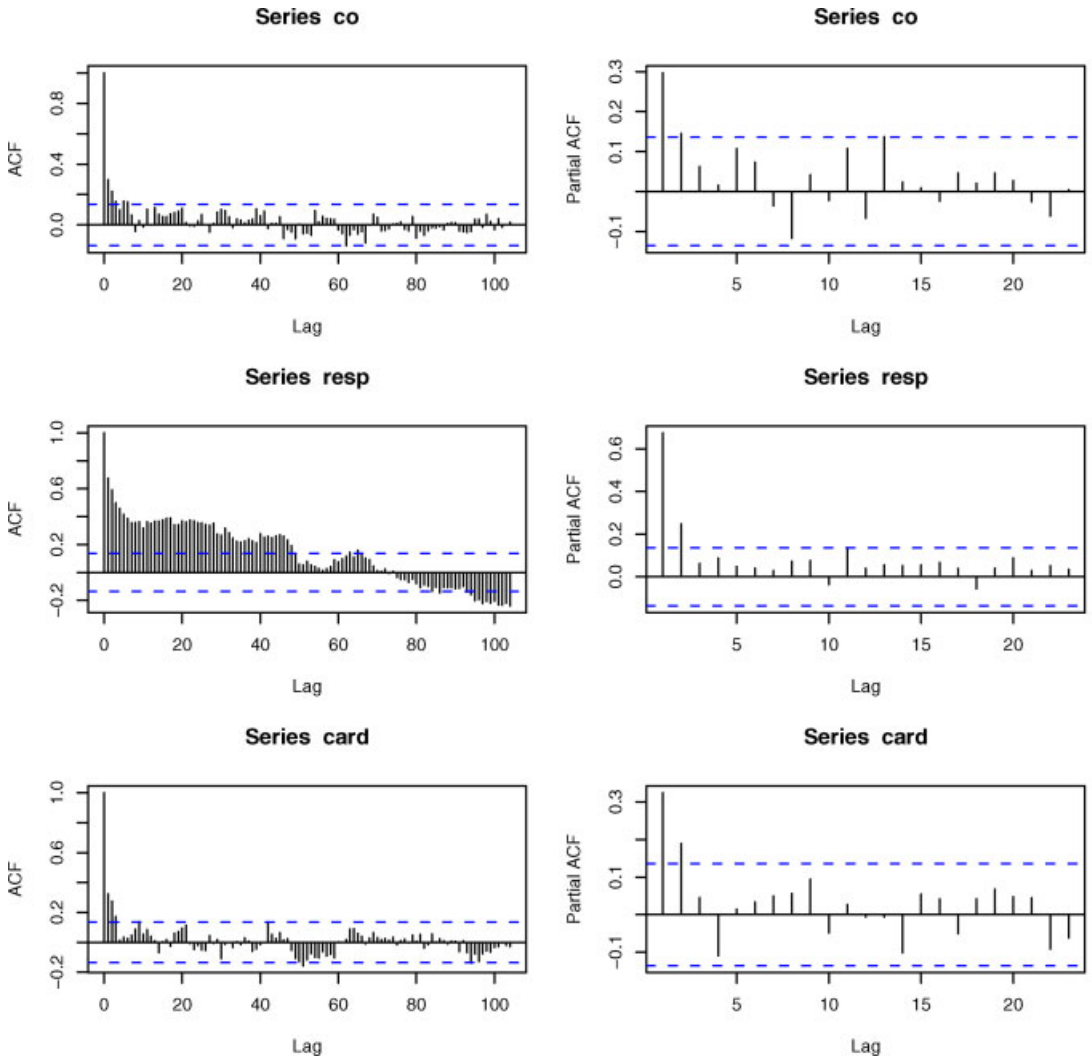

Figure 6. Autocorrelation and partial autocorrelation function for the seasonally adjusted series (CO, Resp, Card)

Table 2. Dickey–Fuller test for the seasonally adjusted series

Series	ADF	Critical value
Temp	−5.959	−3.464 (1%)
Humid	−6.161	−2.876 (5%)
PM_{10}	−4.548	
CO	−4.548	
Resp	−3.256	
Card	−5.310	

DOI: 10.1002/env

Table 3. AR(2) coefficients for the seasonally adjusted series, standard error in parentheses

Series	AR(1)	AR(2)
Temp	0.137	0.211
	(0.068)	(0.068)
Humid	0.025	0.152
	(0.069)	(0.069)
PM_{10}	0.3919	0.1168
	(0.087)	(0.090)
CO	0.255	0.145
	(0.069)	(0.069)
Resp	0.505	0.249
	(0.068)	(0.067)
Card	0.270	0.191
	(0.069)	(0.069)

Table 4. BIC values for VAR(p) model

p	1	2	3	4
BIC	15.079	15.673	16.260	16.963

Table 5. Posterior mean for the parameters of the factor loading matrix

	C	C^*
Temperature	0.0788	0.0022
Humid	0.8090	0.0228
PM_{10}	−5.3579	−0.1510
CO	−0.2352	−0.0066
Resp	−2.4173	−0.0681
Card	−0.4575	−0.0129

Table 6. BIC values for the k-factor model following a AR(p) model

k	1	1	2	2
p	1	2	1	2
BIC	9.024	9.013	7.127	7.343

Table 7. Posterior mean of the factor loading matrix, C, for a 2-factor model

	C	
Temperature	1	0
Humid	0.3065	1
CO	−0.1965	−0.1473
Card	−3.3776	0.8670
Res	−1.3887	−2.0888

DOI: 10.1002/env

Table 8. Posterior mean of the covariance matrix, Γ, autoregressive parameters, standard deviation and Gelman and Rubin factor, R

Parameter	Mean	Std.	R
ψ_{11}	1.8423	0.3798	1.018
ψ_{22}	23.8141	2.4089	1.004
ψ_{33}	1.5107	0.1608	1.002
ψ_{44}	6.2570	5.1457	1.014
ψ_{55}	2.2884	2.1953	1.052
ρ_1	0.3995	0.0709	1.018
ρ_2	0.5650	0.0796	1.002

Table 8 gives the posterior mean for the covariance matrix, Γ, autoregressive coefficients, ρ_{ij}, standard deviation and Gelman and Rubin factor. We note a bigger variance for Humidity and a correlation significant of a week for the factors. Figures 7 and 8 show the posterior histogram for the autoregressive parameters, ρ, and for the matrix Γ.

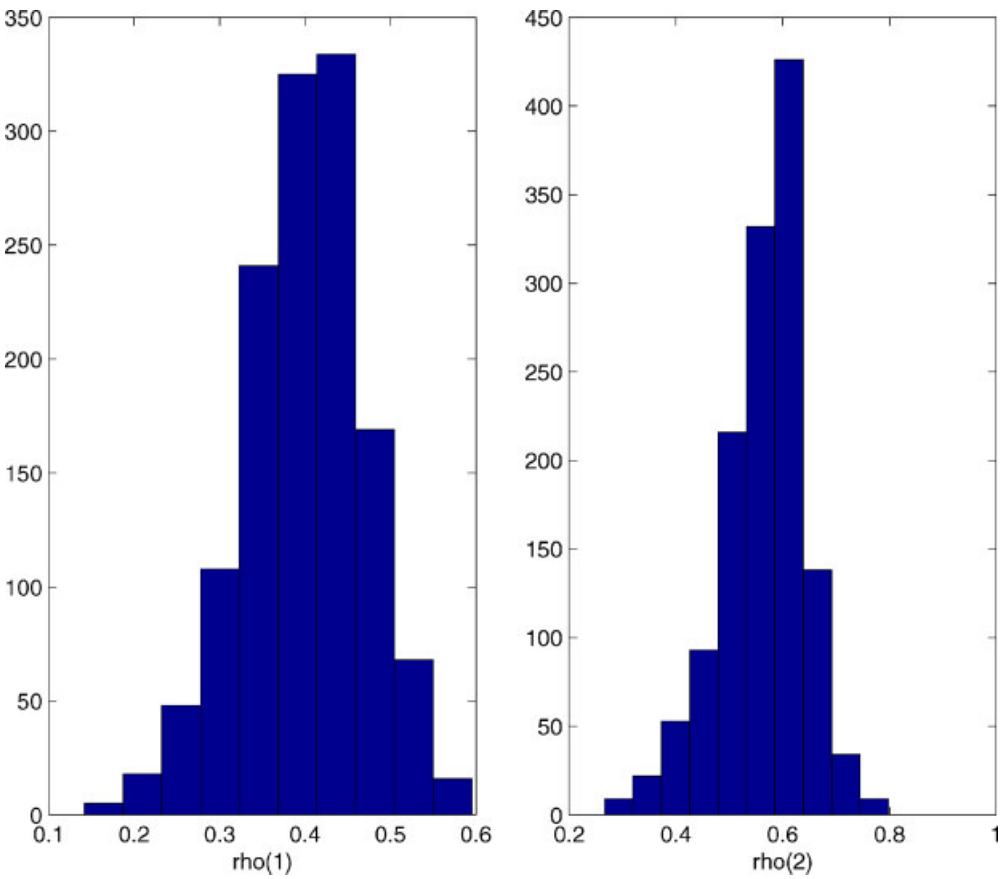

Figure 7. Marginal posteriors of the autoregressive parameters- ρ

DOI: 10.1002/env

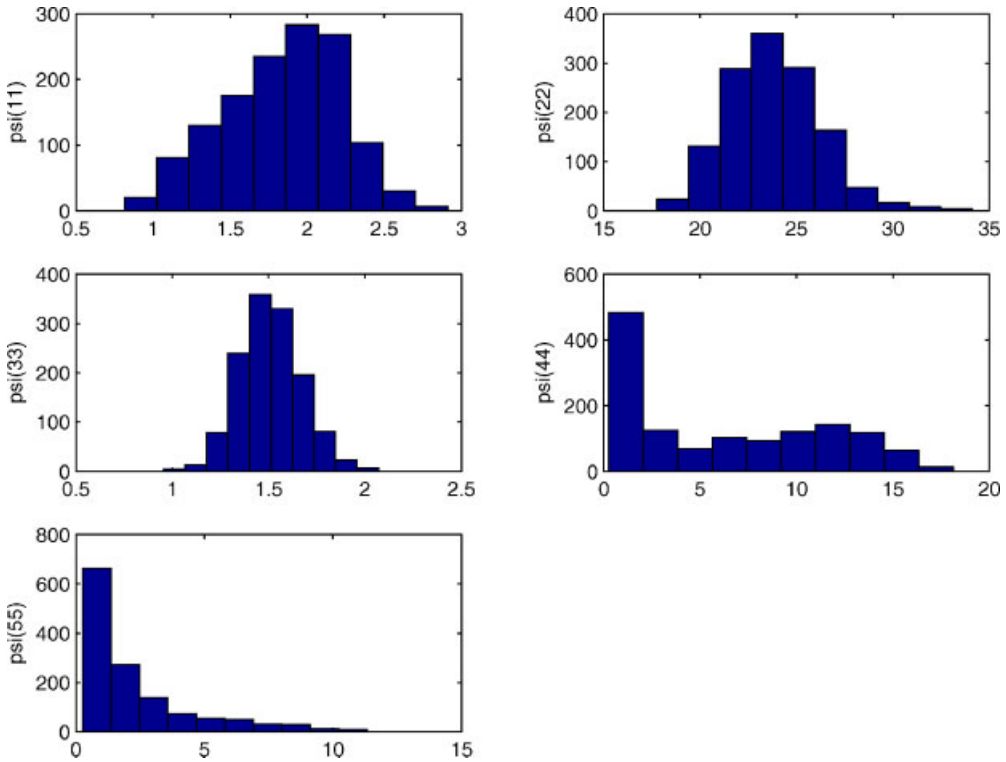

Figure 8. Marginal posteriors of the Gamma matrix

We conclude that the dynamic factor model has been able to fit the time series of weather variables, air pollutants and mortality (respiratory and circulatory) with much less parameters than the VAR(1) model and provide a more clear interpretation of the relationships among the variables.

5. CONCLUSIONS

We have developed a full Bayesian analysis for the dynamic factor model when the factors follow, a AR(p) model. Although in this paper, we have considered that all the factors follow the same AR(p) model the extension to different orders is straightforward. The Gibbs Sampler algorithm was implemented to estimate the parameters. The methodology was applied to analysed the time series of Weather, Pollution and Mortality in the City of São Paulo, Brasil and has allowed to provide a simpler and more parsimonious representation of the data.

ACKNOWLEDGEMENTS

T. Sáfadi gratefully acknowledges financial support from CAPES, Brazil and D. Peña acknowledges financial support from grant SEJ2004-03303, MEC, Spain.

DOI: 10.1002/env

REFERENCES

Aguilar O, West M. 2000. Bayesian dynamic factor model and portfolio allocation. *Journal of Business and Economic Statistics* **18**(3): 338–357.

Conceição GMS, Miraglia SGEK, Kishi HS, Saldiva PHN, Singer JM. 2001. Air pollution and child mortality: a time study in São Paulo, Brazil. *Environmental Health Perspectives* **190**: 347–350.

Forni M, Hallin M, Reichlin L. 2000. The generalized dynamic factor model: identification and estimation. *The Review of Economic and Statistics* **82**: 540–554.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* **7**: 457–511.

Geweke JF, Singleton KJ. 1981. Maximum likelihood confirmatory analysis of economic time series. *International Economic Review* **22**: 37–54.

Geweke JF, Zhou G. 1996. Measuring the pricing error of the arbitrage pricing theory. *The Review of Finantial Studies* **9**: 557–587.

Lee SY, Shi JQ. 2000. Joint Bayesian analysis of factor scores and structural parameters in the factor analysis model. *Annals of the Institue of Statistical Mathematics* **52**(4): 722–736.

Lopes HF, Carvalho CM. 2006. Factor stochastic volatility with time-varying loadings and Markov switching regimes. *Journal of Statistical Planing and Inference* **137**(10): 3082–3091.

Lopes HF, West M. 2004. Model uncertainty in factor analysis. *Statistica Sinica* **14**: 41–67.

Molenaar PCM, De Gooijer JG, Schimitz B. 1992. Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika* **57**: 333–349.

Odriozot JCA, Jimenez JD, Rubio JCM, Pérez IJM, Ortiz MSP, Rodriguez PR. 1998. Air pollution and mortality in Madrid, Spain: a time series analysis. *International Archives of Occupational and Environmental Health* **71**(8): 543–549, Springer-Verlag Heidelberg.

Peña D, Box GEP. 1987. Identifying a simplifying structure in time series. *Journal of the American Statistical Association* **82**(399): 836–843.

Peña D, Poncela P. 2004. Forecasting with nonstationary dynamic factor models. *Journal of Econometrics* **119**: 291–321.

Peña D, Poncela P. 2005. Nonstationary dynamic factor analysis. *Journal of Statistical Planing and Inference* **136**(4): 1237–1257.

Pope CA III, Dockery DW. 1992. Acute health effects of PM10 pollution on symptomatic and asymptomatic children. *The American Review of Respiratory Disease* **145**(5): 1123–1128.

Safadi T, Morettin PA. 2001. Análise bayesiana do modelo "open loop threshold autoregressive". *Revista Brasileira de Estatística* **62**(217): 91–105, Rio de Janeiro.

Stock JH, Watson MW. 1988. Testing for common trends. *Journal of the American Statistical Association* **83**: 1097–1107.

West M, Prado R, Krystal A. 1999. Evaluation and comparison of EEG traces: latent structure in non-stationary time-series. *Journal of the American Statistical Association* **94**: 1083–1095.

APPENDIX A: DERIVATION OF THE POSTERIORS

The conditional posterior distribution for $\theta = (\beta, C, \Gamma, \rho)$ and F is given by

$$P(\theta, F|Z) \quad \alpha \quad L(Y, F|\beta, C, \Gamma, \rho, Y_1, \ldots, Y_p, f_1, \ldots, f_p)P(\theta, F)$$

where, the likelihood $L(Y, F|\beta, C, \Gamma, \rho, Y_1, \ldots, Y_p, f_1, \ldots, f_p)$ is given by Equation (2) and the prior distributions given as: $P(C)P(\rho)P(\beta) \propto$ constant, $\gamma_i = \psi_i^{-1} \sim \Gamma(\alpha_0, \beta_0)$, so that the distribution of ψ_i is an Inverse Gamma, for each components of $\Gamma = \text{diag}\{\psi_1, \psi_2, \ldots, \psi_q\}$. This derivation is done in two steps. First, we derive the complete posterior conditional for each parameter from $\theta = (\beta, C, \Gamma, \rho)$ given all the others and F, then the conditional posterior distribution for F given $\theta = (\beta, C, \Gamma, \rho)$ is derived.

(I) Posterior for $\theta = (\beta, C, \Gamma, \rho)$.

(a) Posterior for mean vector $\beta|\Gamma, C, \rho, F, Y$

$$P(\beta|\Gamma, C, \rho, F, Y) \propto \prod_{t=p+1}^{n} P(y_t|f_t, \rho, C, \Gamma, \beta)P(\beta)$$

DOI: 10.1002/env

$$\propto |\Gamma|^{-(n-p)/2} \exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}(y_t-\beta-Cf_t)'\Gamma^{-1}(y_t-\beta-Cf_t)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(-\beta'\Gamma^{-1}\sum_{t=p+1}^{n}(y_t-Cf_t)-\sum_{t=p+1}^{n}(y_t-Cf_t)'\Gamma^{-1}\beta+\beta'\left((n-p)\Gamma^{-1}\right)\beta\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\beta-\left((n-p)\Gamma^{-1}\right)^{-1}\left(\Gamma^{-1}\sum_{t=p+1}^{n}(y_t-Cf_t)\right)\right]'\right.$$

$$\left.\times\left((n-p)\Gamma^{-1}\right)\left[\beta-\left((n-p)\Gamma^{-1}\right)^{-1}\left(\Gamma^{-1}\sum_{t=p+1}^{n}(y_t-Cf_t)\right)\right]\right\}$$

That is, the β posterior distribution is a Multivariate Normal given by:

$$\beta|C,F,\rho,\Gamma,Y\sim N\left(\frac{1}{n-p}\left(\sum_{t=p+1}^{n}(y_t-Cf_t),\frac{\Gamma}{n-p}\right)\right)$$

(b) Posterior for the factor loading matrix $C|Y,F,\rho,\Gamma,\beta$
If $C_i^{*'}$ is the ith row of C, we have

$$P(C|Y,F,\rho,\Gamma,\beta)\propto\prod_{t=p+1}^{n}|\Gamma|^{-1/2}\exp\left\{-\frac{1}{2}(y_t-\beta-Cf_t)'\Gamma^{-1}(y_t-\beta-Cf_t)\right\}$$

$$\propto |\Gamma|^{-(n-p)/2}\exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}(y_t-\beta-Cf_t)'\Gamma^{-1}(y_t-\beta-Cf_t)\right\}$$

$$\propto\prod_{i=1}^{q}\gamma_i^{(n-p)/2}\exp\left\{-\frac{1}{2}\gamma_i\left[\sum_{t=p+1}^{n}\left(y_{it}-\beta_i-C_i^{*'}f_t\right)'\left(y_{it}-\beta_i-C_i^{*'}f_t\right)\right]\right\}$$

$$\propto\prod_{i=1}^{q}\gamma_i^{(n-p)/2}\exp\left\{-\frac{1}{2}\gamma_i\left[-\sum_{t=p+1}^{n}(y_{it}-\beta_i)f_t'C_i^{*}-C_i^{*'}\sum_{t=p+1}^{n}f_t(y_{it}-\beta_i)'+C_i^{*'}\sum_{t=p+1}^{n}f_tf_t'C_i^{*}\right]\right\}$$

$$\propto\prod_{i=1}^{q}\gamma_i^{n-p/2}\exp\left\{-\frac{1}{2}\gamma_i\left[C_i^{*}-\left(\sum_{t=p+1}^{n}f_tf_t'\right)^{-1}\left(\sum_{t=1}^{n}f_t(y_{it}-\beta_i)\right)\right]'\right.$$

$$\left.\times\left(\sum_{t=p+1}^{n}f_tf_t'\right)\left[C_i^{*}-\left(\sum_{t=p+1}^{n}f_tf_t'\right)^{-1}\left(\sum_{t=p+1}^{n}f_t\,(y_{it}-\beta_i)\right)\right]\right.$$

DOI: 10.1002/env

Then, for each $i = 1, \ldots, q$,

$$C_i^* | Y, F, \rho, \Gamma, \beta \sim N\left(\left[\left(\sum_{t=p+1}^{n} f_t f_t'\right)^{-1}\left(\sum_{t=p+1}^{n} f_t(y_{it} - \beta_i)\right)\right], \left(\gamma_i^{-1}\left(\sum_{t=p+1}^{n} f_t f_t'\right)^{-1}\right)\right)$$

(c) Posterior for the covariance matrix $\Gamma | C, \rho, \beta, Y, F$

$$P(\Gamma | C, \rho, \beta, Y, F) \propto \prod_{t=p+1}^{n} P(y_t | f_t, C, \Gamma, \beta) P(\Gamma)$$

$$\propto |\Gamma|^{-(n-p)/2} \exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}(y_t - \beta - C f_t)' \Gamma^{-1}(y_t - \beta - C f_t)\right\} P(\Gamma)$$

$$\propto \prod_{i=1}^{q} \gamma_i^{\frac{n-p}{2}} \exp\left\{-\frac{1}{2}\gamma_i \sum_{t=p+1}^{n}\left[\left(y_{it} - \beta_i - C_i^{*'} f_t\right)'\left(y_{it} - \beta_i - C_i^{*'} f_t\right)\right\} \gamma_i^{\alpha_0 - 1} \exp\{-\beta_0 \gamma_i\}$$

$$\propto \prod_{i=1}^{q} \gamma_i^{\frac{n-p+2\alpha_0}{2} - 1} \exp\left\{-\frac{1}{2}\gamma_i\left(2\beta_0 + \sum_{t=p+1}^{n}\left[(y_{it} - \beta_i - C_i^{*'} f_t)'(y_{it} - \beta_i - C_i^{*'} f_t)\right]\right)\right\}$$

So, we have for each $i = 1, \ldots, q$, that γ_i is given by:

$$\gamma_i | C, \rho, \beta, F, Y \sim \Gamma\left(\frac{(n - p + 2\alpha_0)}{2}, \frac{\left[2\beta_0 + \sum_{t=p+1}^{n}(y_{it} - \beta_i - C_i^{*'} f_t)'(y_{it} - \beta_i - C_i^{*'} f_t)\right]}{2}\right).$$

(d) Posterior for the elements of $\rho | \beta, \Gamma, C, F, Y$

Let $B_t = [\text{diag}(f_{t-1}), \text{diag}(f_{t-2}), \ldots, \text{diag}(f_{t-p})]$ a $(k \times kp)$ matrix where

$$\text{diag}(f_{t-i}) = \begin{pmatrix} f_{1t-i} & 0 & 0 & \ldots & 0 \\ 0 & f_{2t-i} & 0 & \ldots & 0 \\ : & : & : & \ldots & : \\ 0 & 0 & 0 & \ldots & f_{kt-i} \end{pmatrix}$$

and $\rho_v = (\rho_{11}, \ldots, \rho_{1p}, \rho_{21}, \ldots, \rho_{kp})'$ a $(kp \times 1)$ vector. Note that $\sum_{i=1}^{p} \rho_i f_{t-i} = B\rho_v$.

$$P(\rho | \beta, \Gamma, C, F, Y) \propto \prod_{t=p+1}^{n} P(f_t | \rho, f_1) P(\rho)$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}\left(f_t - \sum_{i=1}^{p}\rho_i f_{t-i}\right)'\left(f_t - \sum_{i=1}^{p}\rho_i f_{t-i}\right)\right\}$$

DOI: 10.1002/env

Then,

$$P(\rho_v|\beta,\Gamma,C,F,Y) \propto \exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}\left[f_t'B_t\rho_v - \rho_v'B_t'f_t + \rho_v'B_t'B_t\rho_v\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left[\rho_v - \left(\sum_{t=p+1}^{n}B_t'B_t\right)^{-1}\left(\sum_{t=p+1}^{n}B_t'f_t\right)\right]'\left(\sum_{t=p+1}^{n}B'B\right)\right.$$

$$\left.\times\left[\rho_v - \left(\sum_{t=p+1}^{n}B_t'B_t\right)^{-1}\left(\sum_{t=p+1}^{n}B_t'f_t\right)\right]\right\}$$

That is,

$$\rho_v|\Gamma,C,F,Y \sim N\left(\left[\left(\sum_{t=p+1}^{n}B_t'B_t\right)^{-1}\left(\sum_{t=p+1}^{n}B_t'f_t\right)\right], \left(\sum_{t=p+1}^{n}B_t'B_t\right)^{-1}\right)$$

(II) Given $\theta = (\beta, C, \Gamma, \rho)$, we derive the $F|\theta, Y$

(e) Posterior for the factors $F = \{f_t, t = p+1, \ldots, n\}|\theta, Y$

We have, for $f_t, t = 1, \ldots, p$

$$f_t|\beta,\Gamma,C,Y \sim N(U^{-1} + C'\Gamma^{-1}C)^{-1}(C'\Gamma^{-1}(y_t-\beta)), (U^{-1}+C'\Gamma^{-1}C)^{-1})$$

and for each $t = p+1, \ldots, n$,

$$P(F|C,\beta,\rho,\Gamma,Y) \propto \prod_{t=p+1}^{n}P(y_t|f_t,\beta,\Gamma,C,\rho)\prod_{t=p+1}^{n}P(f_t|\rho,f_{t-1},\ldots,f_{t-p})$$

$$\propto |\Gamma|^{-(n-p)/2}\exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}(y_t-\beta-Cf_t)'\Gamma^{-1}(y_t-\beta-Cf_t)\right\}$$

$$\times\exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}\left(f_t-\sum_{i=1}^{p}\rho_i f_{t-i}\right)'\left(f_t-\sum_{i=1}^{p}\rho_i f_{t-i}\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\sum_{t=p+1}^{n}\left[-(y_t-\beta)'\Gamma^{-1}Cf_t - f_t'C'\Gamma^{-1}(y_t-\beta) + f_t'C'\Gamma^{-1}Cf_t\right.\right.$$

$$\left.\left.+ f_t'f_t - f_t'\sum_{i=1}^{p}\rho_i f_{t-i} - \sum_{i=1}^{p}f_{t-i}'\rho_i'f_t\right]\right\}$$

DOI: 10.1002/env

$$\propto \exp\left\{\sum_{t=p+1}^{n}\left[f_t'(C'\Gamma^{-1}C+I_k)f_t - f_t'\left(C'\Gamma^{-1}(y_t-\beta)+\sum_{i=1}^{p}\rho_i f_{t-i}\right)\right.\right.$$

$$\left.\left.-\left((y_t-\beta)'\Gamma^{-1}C+\sum_{i=1}^{p}f_{t-i}'\rho_i'\right)f_t\right]\right\}$$

$$\propto \prod_{t=p+1}^{n}\exp\left\{-\frac{1}{2}\left[\left(f_t-\left(C'\Gamma^{-1}C+I_k\right)^{-1}\left(C'\Gamma^{-1}(y_t-\beta)+\sum_{i=1}^{p}\rho_i f_{t-i}\right)\right)\right]'(C'\Gamma^{-1}C+I_k)\right.$$

$$\left.\times\left[\left(f_t-(C'\Gamma^{-1}C+I_k)^{-1}\left(C'\Gamma^{-1}(y_t-\beta)+\sum_{i=1}^{p}\rho_i f_{t-i}\right)\right)\right]\right\}$$

Then, for each $t = p+1, \ldots, n$, we have

$$f_t|f_{t-1},\ldots,f_{t-p},\rho,\beta,\Gamma,C,Y \propto N\left(H^{-1}\left(C'\Gamma^{-1}(y_t-\beta)+\sum_{i=1}^{p}\rho_i f_{t-i}\right),H^{-1}\right)$$

where $H = (C'\Gamma^{-1}C + I_k)$.

DOI: 10.1002/env