

Modeling particulate matter data in Milan monitoring network

Stima di un modello per i dati delle polveri fini per la rete di monitoraggio di Milano

Alessandro Fassò, Orietta Nicolis¹

Dipartimento di Ingegneria Gestionale e dell'Informazione

Università di Bergamo

alessandro.fasso@unibg.it, orietta.nicolis@unibg.it

Riassunto: Le polveri fini (PM_{10}) rappresentano una forma di inquinamento dell'aria fra i più pericolosi per la salute. Poiché la sua misurazione è avvenuta in Italia solo recentemente, non si dispone ancora di una vasta rete di monitoraggio. L'obiettivo di questo lavoro è di trovare un modello in grado di prevedere i livelli dei PM_{10} nell'area di Milano utilizzando tutte le informazioni disponibili (sia sui PM_{10} che su altri inquinanti strettamente correlati a questi). A tal fine si utilizza un modello di regressione *state space* multivariato a parametri variabili (TVP) in cui la variabile dipendente è il livello medio giornaliero del PM_{10} al tempo t in una data centralina ed i regressori sono sia i PM_{10} rilevati in centraline spazialmente vicine che i valori degli ossidi di azoto (NO_x). La bontà dei risultati ottenuti out - of - sample rispetto ad un modello di regressione multipla, suggerisce l'adeguatezza dell'approccio state-space e costituisce un punto di partenza per un'ulteriore espansione del modello.

Keywords: Particular matter data; forecasting; TVP state space model; Kalman filter.

1. Introduction

In Italy, fine particulate matters (PM_{10}) are being monitored only recently. Since these monitoring networks are often sparsely distributed over the land, it is of interest to assess spatio temporal uncertainty and correlation between particulate matters and other quantities.

In this work we consider the small scale spatio temporal relation between PM_{10} and nitrogen oxides (NO_x) observed in data from Milan area. Since only few pollution stations are available, detailed spatial modelling is quite difficult but we show that spatial forecast of PM_{10} is still possible using concomitant observations.

We then use a multivariate state space model as a preliminary analysis of spatio temporal variability of PM_{10} . This is intended to be the first step to extend the model into two directions: first modelling PM_{10} in the basin of Po river where the network is composed by sparsely distributed clusters of small scale nets (metropolitan areas) and then calibrating the measurement instruments. The latter problem, also called spatio temporal calibration, has been considered in the rainfall data literature (see, Brown *et al.* (2001)) and has serious consequences in air quality standard assessment.

Finally, these results are useful in environmental engineering for assessing if a more dense network is important and/or some more information may be picked up from the

¹Indirizzo per corrispondenza: Viale Marconi, 5, 24044 Dalmine (BG), Italy

²Il lavoro è stato svolto con fondi MIUR - COFIN 2002.

network at hand: using NO_x data to forecast PM_{10} and/or using some mobile stations for preliminary estimation.

The data considered here consist of daily averages of hourly readings of NO_x and PM_{10} expressed in $\mu g/m^3$ for seven stations, numbered from 1 to 7 in the sequel of the paper, and located in the following places: Milano Juvara, Milano Verziere, Pioltello, Meda, Vimercate, Magenta and Trezzo. The data range from January 1th, 2002 to August 25th, 2003 for a total of 602 days.

The work is organized as follows. Section 2 is a preliminary analysis of the pollution data. In section 3 we introduce the model. The results and the conclusions are in Sections 4 and 5.

2. Preliminary analysis

The idea of using NO_x in order to forecast PM_{10} is based on the observed correlations between these quantities in Milan area. In particular, these correlations are in the range 0.44-0.67 and are generally higher than the corresponding correlations between PM_{10} and other air pollutants, e.g. carbon oxide and other nitrogen oxides (NO and NO_2). Since these correlations are unsatisfactory for direct use in forecasting, we then consider the following repeated multiple regression

$$y_t(s) = \beta(s)' x_t(s) + \varepsilon_t(s) \quad (1)$$

where t is time, $s = 1, \dots, 7$ is the monitor identifier, x_t is a vector that contains all NO_x readings and those PM_{10} readings at $(-s)$ that is at each station but the station s . The model parameters are given by the vector β and ε_t is a independent and identically distributed disturbance. This gives a relevant increase in fitting with $0.83 \leq R^2 \leq 0.97$. The analysis of correct and false alarm statistics is also satisfactory by considering the warning and attention levels of Italian regulations, $50 \mu g/m^3$ and $75 \mu g/m^3$ on daily averages, respectively.

In order to introduce the dynamical approach of next section, we note that the above model is not able to cover with high serial and spatial correlation and missing values handling.

3. The Model

We consider the time varying parameter model (TVP) defined the following state space representation

$$\begin{aligned} y_t &= \beta_t' x_t + \varepsilon_t \\ \beta_t &= \mu + \Phi \beta_{t-1} + \eta_t \end{aligned} \quad (2)$$

where Φ is a diagonal matrix with elements ϕ_j and the error terms are Gaussian white noises with covariance matrices $Var(\varepsilon_t) = R$ and $Var(\eta_t) = Q$ (see, Harvey(1989) and Kim and Nelson (1999), for details on state-space models). This model generalizes the multiple regression model with respect to: spatial correlation of errors (R and Q), serial correlation and missing values handling. We use two different setups for this model: univariate estimation model and multivariate representation.

Table 1: Parameter estimates and their standard deviations ($\hat{\sigma}$) for gauge $s = 1$.

n. obs=400	$\sigma_\epsilon =$	3.9678	$R^2 = 96.22\%$				
lik = -1183.87	Bias =	0.1512					
			(-s)				
Parameter	Intercept	2	3	4	5	6	7
$\hat{\sigma}_\beta$	0.0000	0.0991	0.0001	0.00339	0.0000	0.0001	0.0000
$\hat{\phi}$	0.6020	0.7608	-0.8839	0.9106	-0.8963	-0.3357	0.8477
$se(\hat{\phi})$	0.0236	0.0305	0.0344	0.0613	0.0352	0.0161	0.0592
$\hat{\mu}$	78.3431	0.1517	0.5313	0.0102	0.2010	0.0419	-0.1051
$se(\hat{\mu})$	1.2631	0.0217	0.0679	0.0067	0.0884	0.0353	0.0290

The first step is performed by the direct generalization of the repeated multiple regression model (1). In this case y_t is the univariate series of PM_{10} at each fixed station $s = 1, \dots, 7$ and x_t includes PM_{10} readings at $(-s)$, all NO_x readings and a constant. In this approach seven models with the same structure are fitted to the seven stations.

The second step is performed by stacking the seven models (2) into the following multivariate representation

$$\begin{aligned} Y_t &= B_t X_t + \varepsilon_t \\ B_t &= \mu + \Phi B_{t-1} + \eta_t \end{aligned} \quad (3)$$

with “spatial covariances” R and Q given by the empirical estimates from step one. In this way, during forecasting of station s , say, we consider $y_t(s)$ as missing and we “reconstruct” $y_t(s) \forall t$ using only $y_t(-s)$ and NO_x . This is possible because the multivariate Kalman filter approach (see, Stroud *et al.* (2001) and Wikle and Cressie (1999)) allows us to learn from forecasting errors of the other stations $(-s)$.

4. Estimation and Data Analysis

We utilized time series of length $n = 400$ for the parameter estimation and in-sample forecasting. Whenever the in-sample performance was better using nitrogen oxides and particulates together, the out-of-sample R^2 showed that a model with only PM_{10} is better. As an example in Table 1, we show the parameter estimates of Equation (2) for gauge n. 1. In particular, all ϕ coefficients, being all in the range $(-1, 1)$, characterize stable dynamics for the state equations. Moreover, the variances of the state equation errors, $\sigma_\beta^2 = \text{diag}(Q)$, are generally very low. Especially the variances for the intercept, gauge n.5 and 7 are zero, indicating β_t -terms which are constant but an initial transient. Analogous results are obtained for the other 6 gauges and all seven R^2 are greater than 0.9.

In order to assess the spatial forecasting performance, we consider the out-of-sample model behaviour for $t = 401, \dots, 602$. To do this, for each fixed station s , we use data $y_{t'}(-s)$, $t' \leq t$, to forecast $y_t(s)$ which is considered as missing by the Kalman filter for all $t \geq 401$.

In Table 2, we compare the out-of-sample performance of the repeated regression model (1), the multivariate TVP model (3) with spatial covariances estimated on the observed residuals in the first column, and the TVP model with assumed spatial uncorrelation in the second column. It can be seen that TVP models are generally better than

Table 2: R^2 values.

		TVP empirical R,Q	TVP diagonal R,Q	Repeated Multiple Regression
1	Mi-Juvara	95.28%	95.59%	88.76%
2	Mi-Verziere	95.44%	94.75%	87.94%
3	Pioltello	95.91%	95.67%	92.47%
4	Meda	89.35%	88.15%	75.99 %
5	Vimercate	93.05%	90.29%	92.60%
6	Magenta	87.02%	87.57%	86.44%
7	Trezzo	85.79%	84.12%	75.40 %
	Mean	91.69%	90.88%	85.66%

the repeated regression model. In particular, the empirical covariance model is always better than the repeated regression with a mean difference about 6%. The second best is the TVP with diagonal covariance which is close to the previous TVP model and always better than the repeated regression except for Vimercate ($s = 5$).

5. Conclusions

TVP models considered improve over regression. The time varying coefficients of our model are very useful for improving spatial forecasting. These results can be used for giving “estimated” particulate concentrations where a mobile station is situated for the short time required in model estimation. After this the model is capable to give forecasts without any monitors in place.

References

- Brown, P.E., Diggle, P.J., Lord, M.E., Young, P.C. (2001), Space-time calibration of radar rainfall data. *Applied Statistics*, 50, 2, 221-241.
- Harvey, A. (1989) *Forecasting, Structural Time Series Models, and the Kalman Filter*, Cambridge University Press.
- Kim, C. J. and Nelson, C. R. (1999) *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*, MIT Press, Cambridge, Massachusets.
- Stroud, J.R. and Muller, P. and Sansò, B. (2001) Dynamic models for Spatio-temporal data, *Journal of the Royal Statistical Society, Series B*, 63, 815-829.
- Wikle, C.K. and Cressie, N. (1999) A dimension-reduced approach to space-time Kalman filtering, *Biometrika*, 86, 815-829.