Reconstructing missing data sequences of environmental data by means of spatial correlation

G. Albano, M. La Rocca, M.L. Parrella, C. Perna
University of Salerno - Italy

TIES-GRASPA 2017
Bergamo, July 24-26

Outline of the talk

1. Framework setup
   - Motivations and goals
   - Literature review on imputation methods for time series
2. Our proposal
   - Modelling by a *Generalized SDPD* model
   - The iterative imputation procedure
   - Theoretical foundations
3. Performance and validation
   - Application to environmental spatio-temporal data
   - Results from a simulation study

Motivations

- A huge number of epidemiological studies establishes the link between air pollution and health (see, for example, [Biggeri et al. (2004)] and [Raaschou-Nielsen et al. (2013)])
- In 2013 the International Agency for Research on Cancer and the World Health Organization established that ***airborne particulates is a Group 1 carcinogen***.
- As a consequence, in the air quality directive (2008/EC/50), the EU has set two limit values for particulate matter (PM_{10}):
  1. *the PM_{10} daily mean value may not exceed 50 micrograms per cubic metre ($\mu g/m^3$) more than 35 times in a year*,
  2. *the PM_{10} annual mean value may not exceed 40 micrograms per cubic metre ($\mu g/m^3$).*
- **In this context, missing data represent the main problem to monitoring the air quality.**

Goals

- In the general contexts of multivariate time series, characterized by both serial and cross-correlation, ignoring the missing values can lead to bias and error during data mining.
- It may happen that not only isolated values but **also long temporal sequences of values may miss**. In such cases, it is quite impossible to reconstruct the missing sequences basing on the serial dependence structure alone.
- **Cross-correlation and serial correlation are strictly interconnected and this property should be guaranteed by the imputation model**.
- In this paper we propose a new procedure for estimating even long missing sequences in time series, taking into account both the serial correlation and the cross correlation of data, simultaneously.

A brief literature review

- Various techniques have been proposed to impute missing values in environmental data, but they perform well only when the number of missing values is small (see [Junninen et al. (2004)]; [Norazian et al. (2008)]; [Fitri et al. (2010)]).
- A lot of models have been proposed in literature taking into account the spatio-temporal dependence of PM_{10} located in near places, but they do not focus on missing sequences (see [Cameletti et al. (2011)]).
- In the context of PM_{10} concentration, an imputation technique based on linear spatial regression, where no spatial weighting structure is assumed for the imputed data, is proposed by [Pollice & Lasinio (2009)].
- Recently, some new approaches combine two different imputation methods in separate stages, accounting for cross-correlation and serial correlation separately ([Liu et al. (2014)], [Oehmcke et al. (2016)])

R packages for imputation methods

- There are several imputation methods implemented in R software packages, but very few are suitable for multivariate time series.
- The most popular R packages available on the CRAN are `AmeliaII`, `mice`, `VIM`, `missMDA` and `imputeTS` ([Honaker et al. (2011)], [van Buuren & Groothuis-Oudshoorn (2011)], [Kowarik & Templ (2016)], [Josse & Husson (2016)] and[Moritz & Bartz-Beielstein (2017)]).
- Among these, the only ones that give direct support for dependent data are the packages `imputeTS` and `AmeliaII` ([Moritz & Bartz-Beielstein (2017)] and [Honaker et al. (2011)]).
- The `imputeTS` package can handle with univariate time series imputation and include multiple imputation algorithms.
- The `AmeliaII` package is designed to impute cross-sectional data and it also considers the case of longitudinal data.

Our proposal: imputing basing on a G-SDPD model

- We use a *Generalized Spatial Dynamic Panel Data* model for PM_{10} time series to manage a network of near monitoring stations.
- The G-SDPD model belongs to the family of *spatial econometric models* born with [Anselin (1988)] and has been first proposed and analysed by [Dou, Parrella & Yao (2016)]:

$$\mathbf{y}_t = D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t + D(\boldsymbol{\lambda}_1)\mathbf{y}_{t-1} + D(\boldsymbol{\lambda}_2)\mathbf{W}\mathbf{y}_{t-1} + \mathbf{u}_t, \qquad (1)$$

  where $D(\cdot)$ are diagonal matrices from vectors $\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2$ and the error process $\mathbf{u}_t$ may include other (spurious) cross-correlation, but we assume it is not serially correlated.
- Note that,
  1. $D(\boldsymbol{\lambda}_0)\mathbf{W}\mathbf{y}_t$ captures the pure spatial effects;
  2. $D(\boldsymbol{\lambda}_1)\mathbf{y}_{t-1}$ captures the pure dynamic effects;
  3. $D(\boldsymbol{\lambda}_2)\mathbf{W}\mathbf{y}_{t-1}$ captures the spatial-dynamic effects.

Estimation methods

- Following [Dou, Parrella & Yao (2016)], we derive the Yule-Walker equation system

$$(\mathbf{I} - D(\boldsymbol{\lambda}_0)\mathbf{W})\boldsymbol{\Sigma}_1 = (D(\boldsymbol{\lambda}_1) + D(\boldsymbol{\lambda}_2)\mathbf{W})\boldsymbol{\Sigma}_0.$$

- The i-th row of the above multivariate equation system is

$$(\mathbf{e}_i' - \lambda_{0i}\mathbf{w}_i')\boldsymbol{\Sigma}_1 = (\lambda_{1i}\mathbf{e}_i' + \lambda_{2i}\mathbf{w}_i')\boldsymbol{\Sigma}_0, \quad i = 1, \ldots, p.$$

The vector $(\lambda_{0i}, \lambda_{1i}, \lambda_{2i})'$ *is estimated by least squares method*

$$\min_{\lambda_{0i},\lambda_{1i},\lambda_{2i}} \|\widehat{\boldsymbol{\Sigma}}_1^T(\mathbf{e}_i - \lambda_{0i}\mathbf{w}_i) - \widehat{\boldsymbol{\Sigma}}_0(\lambda_{1i}\mathbf{e}_i + \lambda_{2i}\mathbf{w}_i)\|_2^2,$$

$$(\widehat{\lambda}_{0i}, \widehat{\lambda}_{1i}, \widehat{\lambda}_{2i})' = (\widehat{\mathbf{X}}_i'\widehat{\mathbf{X}}_i)^{-1}\widehat{\mathbf{X}}_i'\widehat{\mathbf{Y}}_i, \quad i = 1, 2, \ldots, p,$$

where $\widehat{\mathbf{X}}_i = \left(\widehat{\boldsymbol{\Sigma}}_1'\mathbf{w}_i, \widehat{\boldsymbol{\Sigma}}_0\mathbf{e}_i, \widehat{\boldsymbol{\Sigma}}_0\mathbf{w}_i\right)$ *and* $\widehat{\mathbf{Y}}_i = \widehat{\boldsymbol{\Sigma}}_1'\mathbf{e}_i$.

The new iterative imputation procedure

1. *Start, at iteration* 0, *by initializing*

$$\widetilde{\mathbf{y}}_t^{(0)} = \delta_t * \mathbf{y}_t, \qquad t = 1, \ldots, T,$$

where $*$ *is the Hadamard product and* δ_t *is a vector of zeroes/ones that identifies the missing values in* $\mathbf{y}_t$

2. *Then, the generic iteration* s, *with* $s \geq 1$, *is:*
   a) *estimate* $(\widehat{\lambda}_0^{(s-1)}, \widehat{\lambda}_1^{(s-1)}, \widehat{\lambda}_2^{(s-1)})$ *using* $\{\widetilde{\mathbf{y}}_1^{(s-1)}, \ldots, \widetilde{\mathbf{y}}_T^{(s-1)}\}$;
   b) *compute, for* $t = 1, \ldots, T$,

$$\begin{aligned}
\widehat{\mathbf{y}}_t^{(s)} &= D(\widehat{\lambda}_0^{(s-1)})\mathbf{W}\widetilde{\mathbf{y}}_t^{(s-1)} + D(\widehat{\lambda}_1^{(s-1)})\widetilde{\mathbf{y}}_{t-1}^{(s-1)} + D(\widehat{\lambda}_2^{(s-1)})\mathbf{W}\widetilde{\mathbf{y}}_{t-1}^{(s-1)} \\
\widetilde{\mathbf{y}}_t^{(s)} &= \delta_t * \mathbf{y}_t + (\mathbf{1} - \delta_t) * \widehat{\mathbf{y}}_t^{(s)},
\end{aligned}$$

3. *Stopping rule:* $\|\widetilde{\mathbf{y}}_t^{(s)} - \widetilde{\mathbf{y}}_t^{(s-1)}\|_2^2 \leq \epsilon$, *with* ϵ *small enough.*

Theorethical foundations

The following assumptions are required for the consistency of the imputation procedure:

- A1 The spatial weight matrix $\mathbf{W}$ has zero main diagonal elements; moreover, matrix $S(\lambda_0) = (\mathbf{I} - D(\lambda_0)\mathbf{W})$ is invertible.
- A2 The disturbance ε_t satisfies $Cov(\mathbf{y}_{t-1}, \varepsilon_t) = 0$. Moreover, the process $\mathbf{y}_t$ is strictly stationary and α-mixing.
- A3 The rank of matrix $(\mathbf{\Sigma}_1'\mathbf{w}_i, \mathbf{\Sigma}_0\mathbf{e}_i, \mathbf{\Sigma}_0\mathbf{w}_i)$ is equal to 3, for all i.
- A4 $\rho\left(D(\lambda_0)\mathbf{W}\right) < 1$ and $\rho\left(D(\lambda_1) + D(\lambda_2)\mathbf{W}\right) < 1$.

Theorem

Under the assumptions A1-A4, for $s \to \infty$ and $T \to \infty$ the difference $\|\widetilde{\mathbf{y}}_t^{(s)} - \widetilde{\mathbf{y}}_t^{(s-1)}\|_2^2$ goes to zero and the iterative procedure converges to a unique solution $\widetilde{\mathbf{y}}_t^{(\infty)}, t = 1, \ldots, T$.

Application to environmental spatio-temporal data

- In our real data analysis, we consider daily PM_{10} data (in $\mu g/m^3$) by gravimetric instruments at 24 sites in Piemonte.
- Data were provided from the website of *Agenzia Regionale per la Protezione Ambientale (ARPA)*, Piemonte.

Station	missings	%	Station	missings	%
Alba Tanaro	34	5.17	Cuneo Alpini	32	4.86
Alessandria-D'Annunzio	27	4.10	Druento La Mandria	21	3.19
Arquata Scrivia Minzoni	175	26.60	Novara Verdi	97	14.74
Asti Baussano	3	0.46	Novi Ligure Gobetti	65	9.88
Biella Sturzo	17	2.58	Pinerolo Alpini	43	6.53
Borgaro Torinese Caduti	62	9.42	Torino Consolata	32	4.86
Borgomanero Molli	16	2.43	Torino Grassi	224	34.04
Borgosesia-Tonella	25	3.80	Torino Lingotto	59	8.97
Carmagnola I Maggio	18	2.74	Torino Rebadeugo	53	8.05
Casale Monferrato Castello	18	2.74	Torino Rubino	25	3.80
Cerano Bagno	14	2.13	Tortona Carbone	174	26.44
Cossato Pace	34	5.17	Vercelli CONI	15	2.28

The PM_{10} data

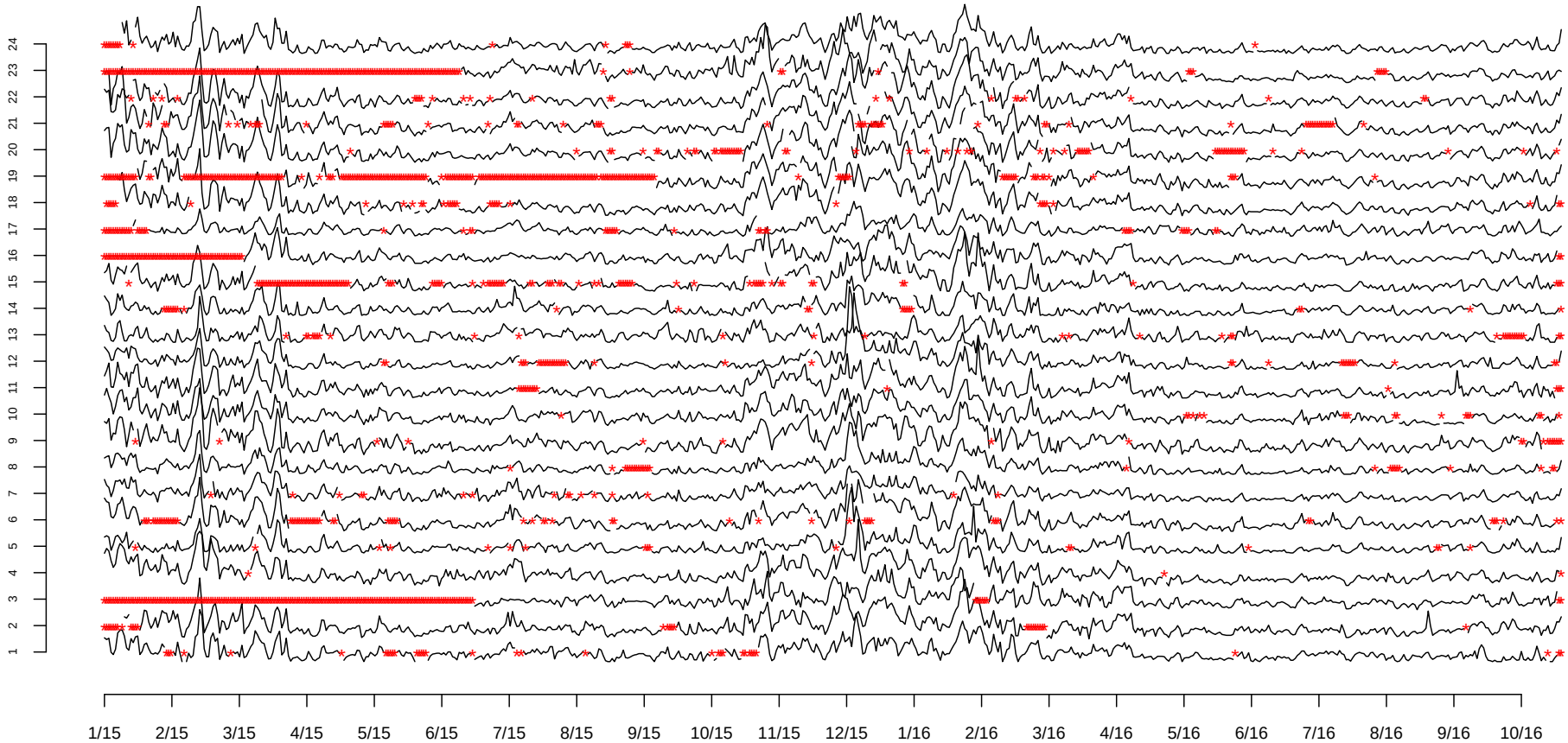

Figure: Plots of the 24 time series for PM_{10} data, observed daily from January 2015 to October 2016 at the Italian stations listed in the previous Table. The red points indicate the missing values.

Settings

- We use two different spatial matrices:
  1. $\mathbf{W}_1$ is a normalized sample correlation matrix of $\mathbf{y}_t$,
  2. $\mathbf{W}_2$ is a spatial matrix based on the geographical distance between the stations.
- Here and in the simulation study, we set the maximum number of iterations to $N = 100$ corresponding to a convergence error value of $\epsilon = O(10^{-10})$.

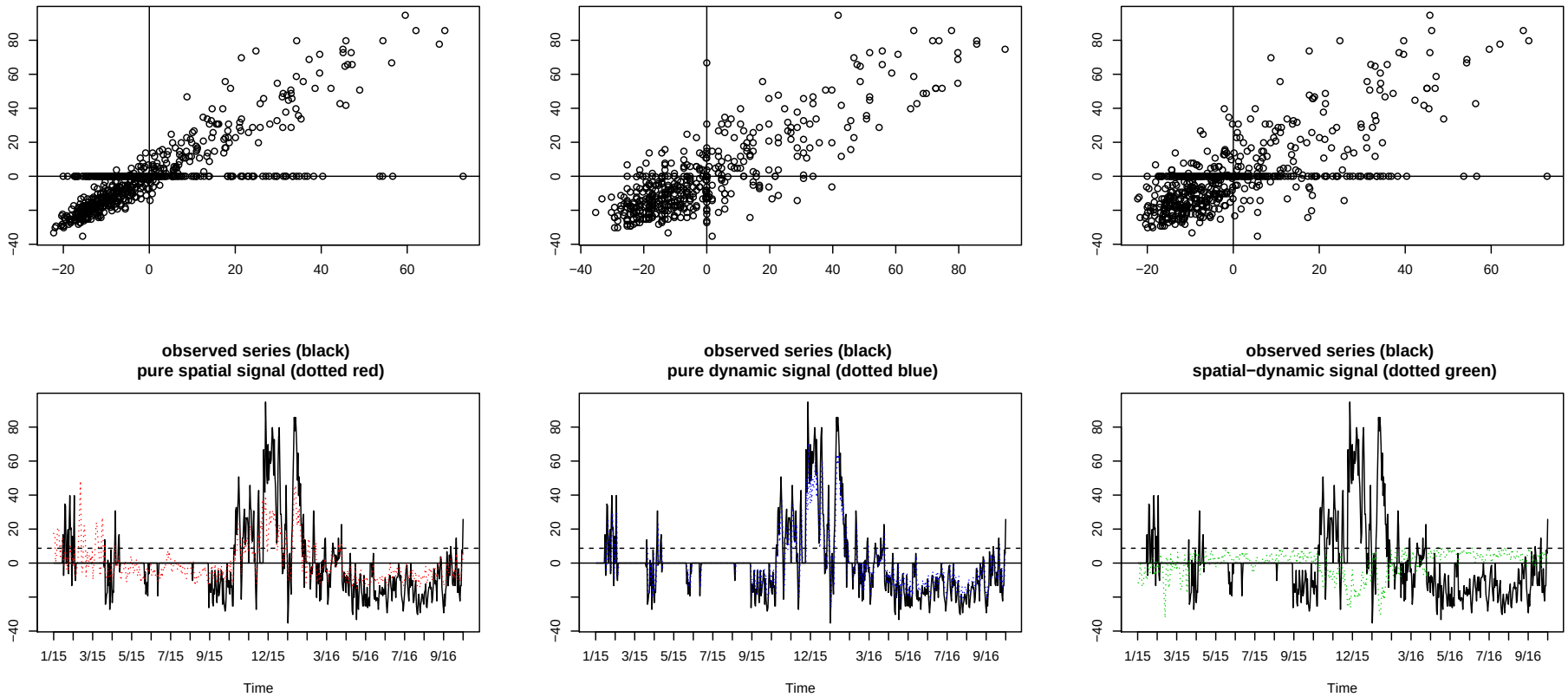

Figure: Results for Torino-Grassi station, **at the first step of the iterative procedure**

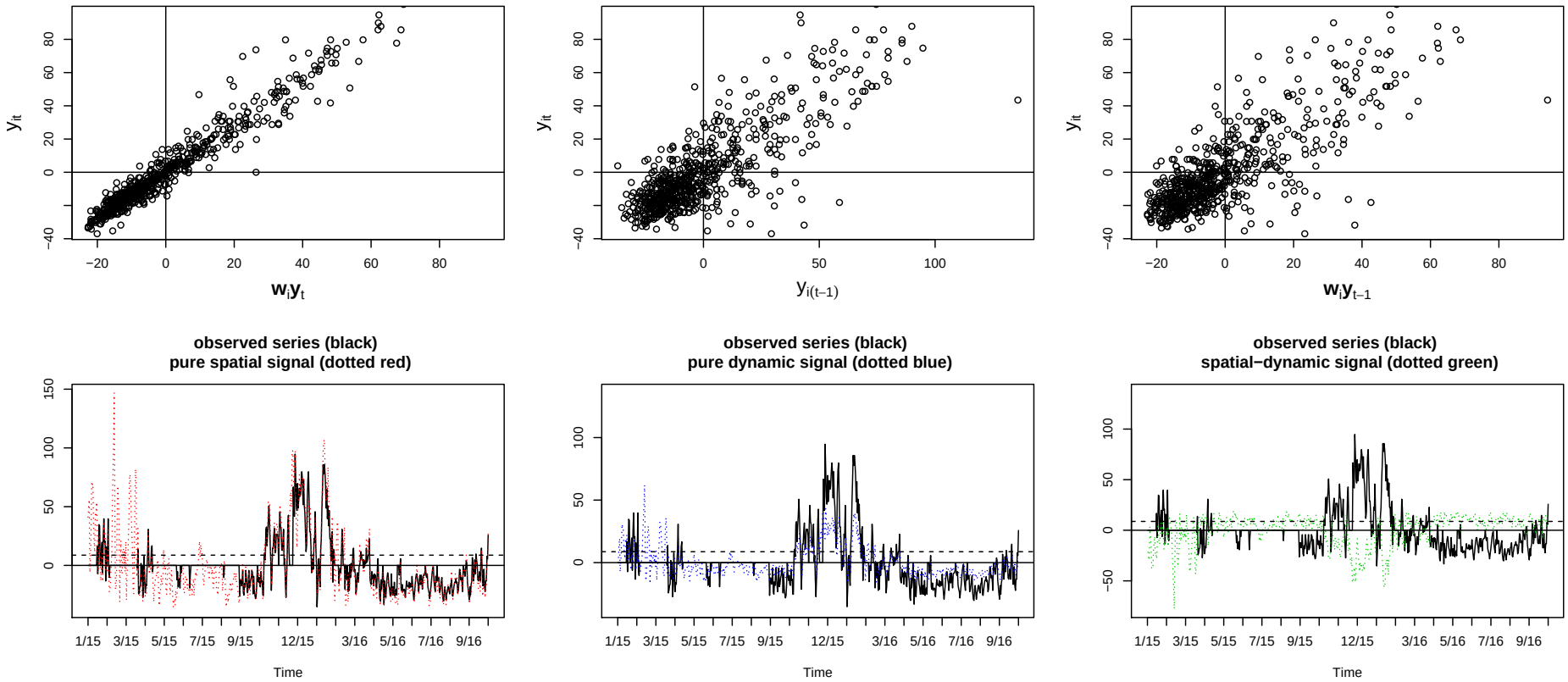

Figure: Results for Torino-Grassi station, **at the final step of the iterative procedure**

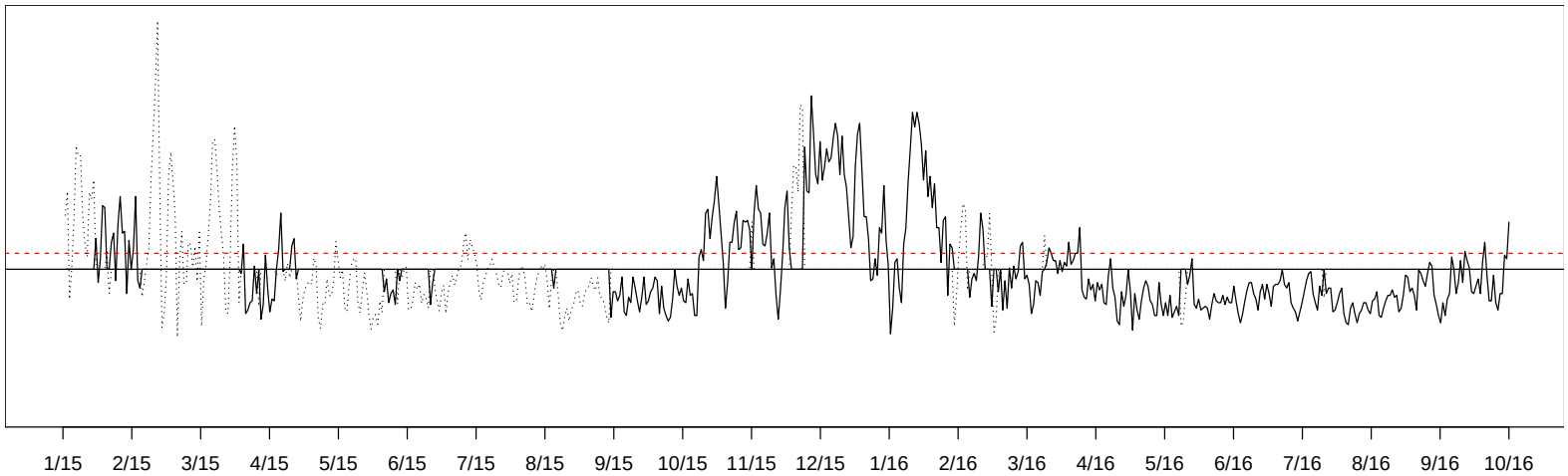

Figure: Final output of our iterative procedure for the station Torino-Grassi.

Number of days exceeding the threshold $50\mu g/m^3$

	Station	Number of days exceeding the threshold $50\mu g/m^3$					
		original data		imputed with W_1		imputed with W_2	
1	Alba Tanaro	34.95		34.95		34.95	
2	Alessandria-DAnnunzio	62.13	(+27)	67.67	(+33)	67.67	(+33)
3	Arquata Scrivia Minzoni	30.51		46.6	(+12)	46.6	(+12)
4	Asti Baussano	65.46	(+30)	65.46	(+30)	65.46	(+30)
5	Biella Sturzo	15.53		15.53		15.53	
6	Borgaro Torinese Caduti	51.59	(+17)	54.92	(+20)	54.92	(+20)
7	Borgomanero Molli	19.41		19.41		19.41	
8	Borgosesia-Tonella	19.41		19.41		19.41	
9	Carmagnola I Maggio	76	(+41)	77.1	(+42)	77.1	(+42)
10	Casale Monferrato Castello	47.71	(+13)	47.71	(+13)	47.71	(+13)
11	Cerano Bagno	51.59	(+17)	52.14	(+17)	52.14	(+17)
12	Cossato Pace	26.07		26.07		26.07	
13	Cuneo Alpini	13.31		13.31		13.31	
14	Druento La Mandria	18.31		18.31		18.31	
15	Novara Verdi	37.17	(+2)	41.6	(+7)	41.6	(+7)
16	Novi Ligure Gobetti	33.28		41.05	(+6)	40.49	(+5)
17	Pinerolo Alpini	8.88		10.54		11.09	
18	Torino Consolata	68.23	(+33)	71	(+36)	71	(+36)
19	Torino Grassi	62.13	(+27)	92.64	(+58)	92.08	(+57)
20	Torino Lingotto	59.91	(+25)	64.35	(+29)	64.35	(+29)
21	Torino Rebadeugo	72.11	(+37)	82.65	(+48)	82.65	(+48)
22	Torino Rubino	59.91	(+25)	62.13	(+27)	62.13	(+27)
23	Tortona Carbone	39.94	(+5)	61.02	(+26)	60.46	(+25)
24	Vercelli CONI	39.38	(+4)	41.6	(+7)	41.6	(+7)

Comparison with `AmeliaII`'s imputation procedure

- The imputation model in Amelia assumes that the complete data are multivariate normal. It draws imputations of the missing values using a bootstrapping approach, the EMB (expectation-maximization with bootstrapping) algorithm.
- To deal with time series, `AmeliaII` builds a general model of patterns within variables across time by creating a sequence of polynomials of the time index, up to the k-th order ($k \leq 3$).
- Moreover, to improve multivariate time-series imputation, `AmeliaII` can also include lags and leads of certain variables into the imputation model.
- In our analysis, for the Amelia procedure, we consider $m = 5$ imputations (finally averaged), a polynomial path of order $k = 3$ and we add the lagged values and leads in the imputation model.

Validating the imputation procedures

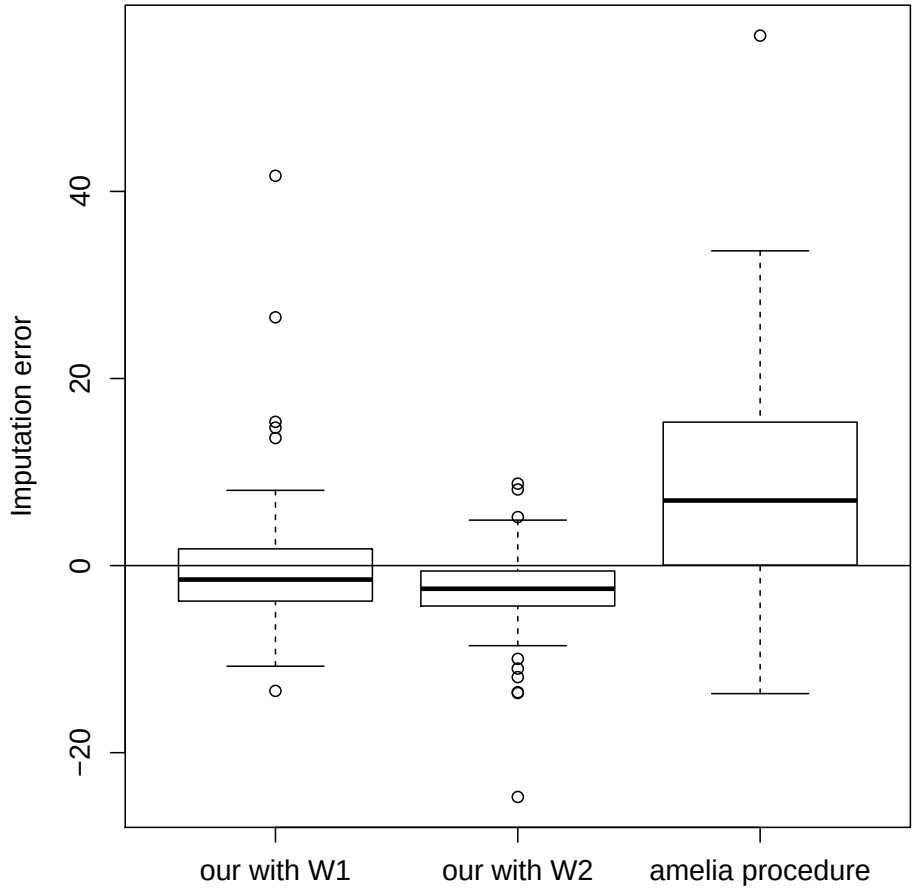

Figure: From PM_{10} data, 50 observed values where removed and considered as missing (chosen randomly in the dataset, among which 30 as a single missing sequence). The boxplots show the distribution of the imputation error for the 50 simulated missing values.

A simulation study: settings (1)

- We simulate $N = 400$ replications of the model with $p = 30$ locations and $T = (50, 100, 500, 1000)$ observations.
- The spatial matrix **W** has been randomly generated as a full rank symmetric matrix and has been row-normalized.
- The parameters λ of the G-SDPD model have been randomly generated in the interval $[-0.6, 0.6]$.
- The error component ε_t has been generated from a multivariate normal distribution, with mean vector zero and diagonal variance-covariance matrix, with heteroscedastic variances $(\sigma_1^2, \ldots, \sigma_p^2)$, where the standard deviations have been generated randomly from a Uniform distribution $U(0.5; 1.5)$.

Performance with different percentages of missings

- Over the total number of observations of the multivariate time series, $T \times p$, the 1%, 5%,..., 70% are randomly chosen and considered as missing.
- For example, there are 750 missing values when $T = 50$, $p = 30$ and the percentage is 50%.
- Among these missing values, we always simulate a sequence which is long the 1%, 5%,..., 70% of the time series length T, respectively.
- For example, when the time series length is $T = 100$ and the percentage is 50%, the missing sequence includes 50 sequential values.

A simulation study: settings (2)

- For each monte carlo replication, indexed by $r = 1, \ldots, N$, the imputation error have been calculated as

$$e_{it}^{(r)} = y_{it} - \tilde{y}_{it}^{(r)}, \qquad \forall i, t \in M,$$

- We derive the average estimation error and the average squared error,

$$ASE_{it} = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(y_{it} - \tilde{y}_{it}^{(r)}\right)^2}, \qquad \forall i, t \in M$$

- We report the mean (and standard deviations in brackets) of the standardized $ASE_{it}/\sigma_{\varepsilon_i}$, calculated over $i, t \in M$,

Mean values and standard deviations of $ASE_{it}/\sigma_{\varepsilon_i}$

	2 % of missings							
	$T = 50$		$T = 100$		$T = 500$		$T = 1000$	
our with $\mathbf{W}$	1.084	(0.071)	1.051	(0.099)	1.028	(0.086)	1.026	(0.083)
our with $\widehat{\mathbf{W}}$	1.053	(0.119)	1.043	(0.131)	1.064	(0.111)	1.073	(0.105)
Amelia	1.502	(0.268)	1.469	(0.284)	1.445	(0.271)	1.474	(0.274)
	5 % of missings							
	$T = 50$		$T = 100$		$T = 500$		$T = 1000$	
our with $\mathbf{W}$	1.112	(0.137)	1.059	(0.065)	1.034	(0.079)	1.032	(0.085)
our with $\widehat{\mathbf{W}}$	1.062	(0.172)	1.054	(0.12)	1.083	(0.11)	1.083	(0.108)
Amelia	1.454	(0.274)	1.472	(0.283)	1.492	(0.28)	1.47	(0.27)
	10 % of missings							
	$T = 50$		$T = 100$		$T = 500$		$T = 1000$	
our with $\mathbf{W}$	1.146	(0.119)	1.088	(0.101)	1.049	(0.086)	1.043	(0.087)
our with $\widehat{\mathbf{W}}$	1.094	(0.153)	1.075	(0.132)	1.088	(0.115)	1.089	(0.113)
Amelia	1.531	(0.272)	1.481	(0.282)	1.471	(0.271)	1.479	(0.275)
	30 % of missings							
	$T = 50$		$T = 100$		$T = 500$		$T = 1000$	
our with $\mathbf{W}$	1.232	(0.133)	1.152	(0.112)	1.095	(0.109)	1.09	(0.108)
our with $\widehat{\mathbf{W}}$	1.168	(0.156)	1.137	(0.145)	1.124	(0.133)	1.126	(0.129)
Amelia	1.477	(0.27)	1.476	(0.26)	1.484	(0.27)	1.485	(0.265)
	50 % of missings							
	$T = 50$		$T = 100$		$T = 500$		$T = 1000$	
our with $\mathbf{W}$	1.358	(0.197)	1.225	(0.14)	1.141	(0.13)	1.134	(0.13)
our with $\widehat{\mathbf{W}}$	1.356	(0.271)	1.246	(0.222)	1.165	(0.152)	1.162	(0.147)
Amelia	1.498	(0.257)	1.497	(0.261)	1.494	(0.266)	1.498	(0.269)

Computational times 2% of missings

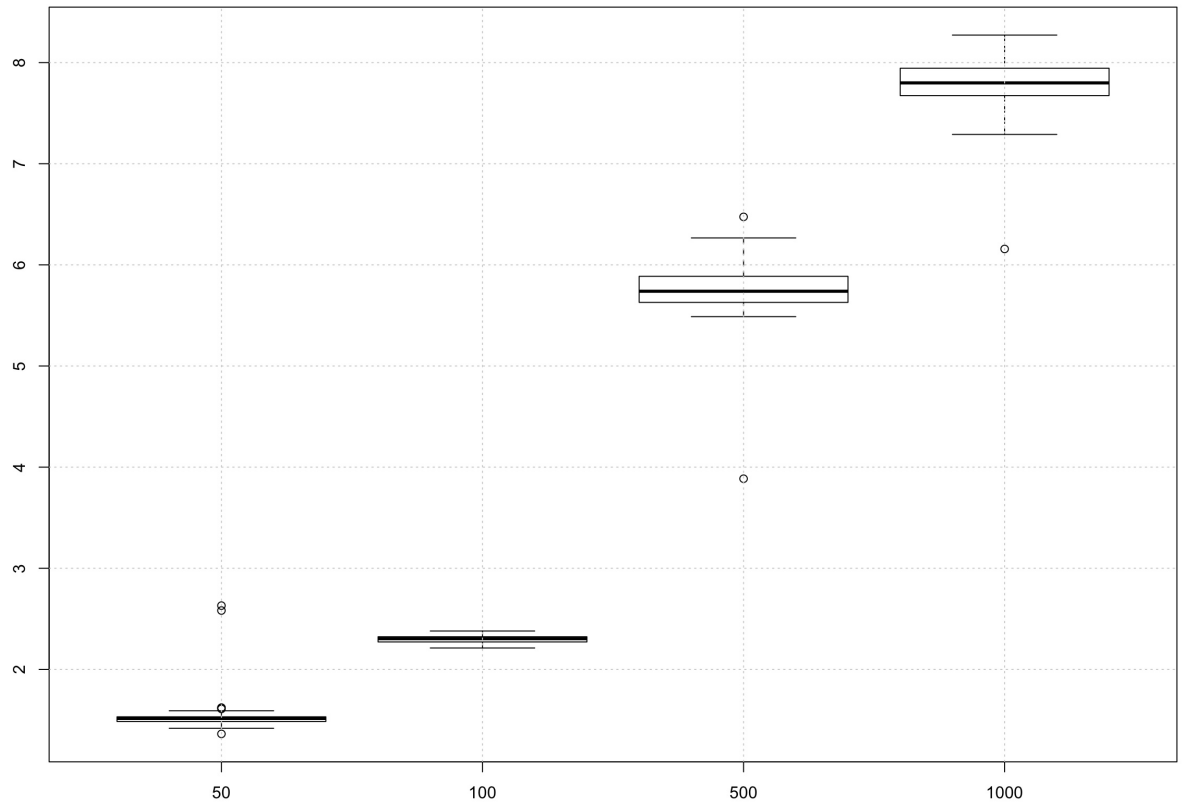

Figure: Ratio of the computational times for Amelia and our procedures, without bootstrapping.

Computational times: 50% of missings

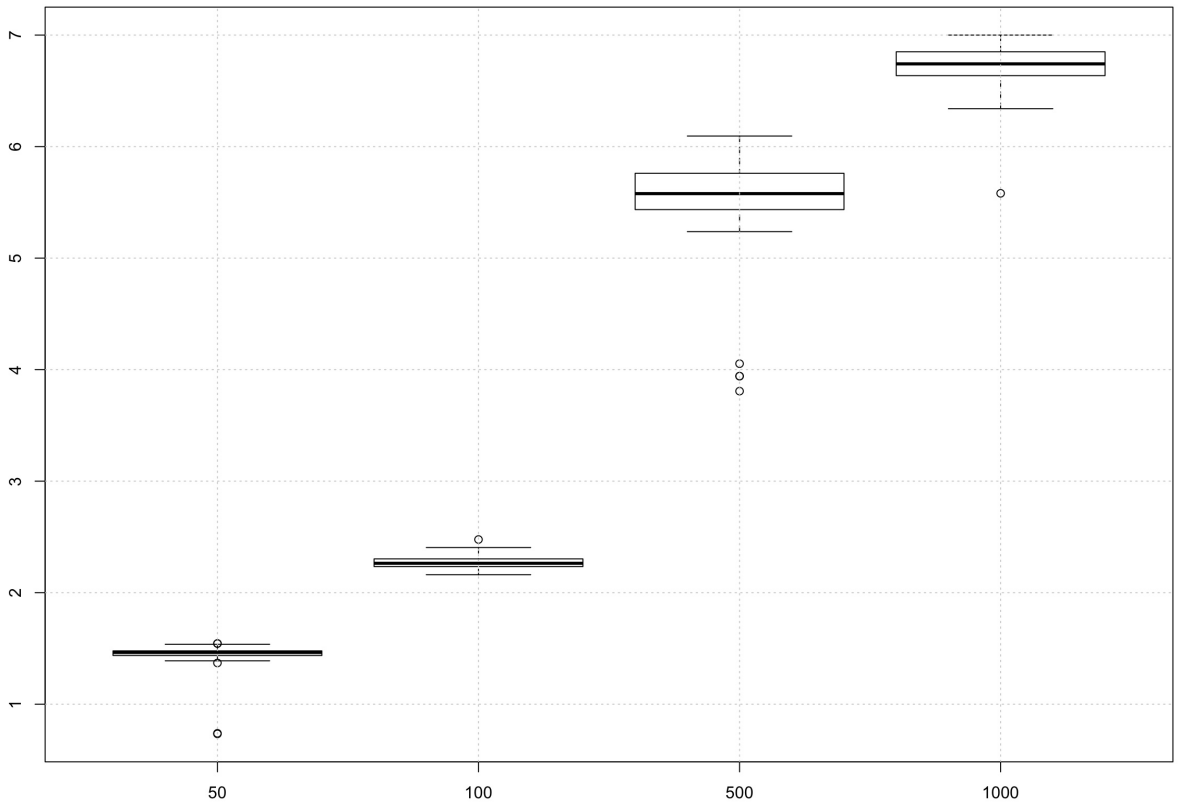

Figure: Ratio of the computational times for Amelia and our procedures, without bootstrapping.

Concluding remarks

- We propose an iterative imputation technique based on the G-SDPD model first introduced in [Dou, Parrella & Yao (2016)], where estimation can be easily implemented since the involved **estimators are obtained in closed form**.
- The procedure is **computationally feasible and scales up to high dimension** of the multivariate time series. Morever, it does not depend on any tuning parameter or any user choice.
- The simulation experiment, in which suitable missing sequences and values were randomly generated shows that our procedure produces estimates very close to the true values.
- Finally, the proposed imputation procedure shows a **good performance even when half of the data is missing**.


Aga, E., Samoli, E., Touloumi, G., Anderson, H. R., Cadum, E. and Forsberg, B. e. a. (2003). Short-term effects of ambient particles on mortality in the elderly: results from 28 cities in the APHEA2 project. The European Respiratory Journal Supplement 40 28s33s.


Anselin, L., ed. (1988). Spatial econometrics: methods and models. Kluwer Aca- demic, The Netherlands.


Biggeri, A., Baccini, M., Accetta, G. and Lagazio, C. (2002). Estimates of short-term effects of air pollutants in Italy. Epidemiologia e Prevenzione 26 203205.


Cameletti, M., Ignaccolo, R. and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. environmetrics 22 985996.


Dou, B., Parrella, M. L. and Yao, Q. (2016). Generalized Yule-Walker Estima- tion for Spatio-Temporal Models with Unknown Diagonal Coefficients. Journal of Econometrics 194 369-382.


Fitri, M. D. N. F., Ramli, N. A., Yahaya, A. S., Sansuddin, N., Ghazali, N. A. and Al Madhoun, W. (2010). Monsoonal differences and probability distribution of PM10 concentration. Environmental Monitoring Assessment 163 655-667.


Honaker, J., King, G. and Blackwell, M. (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software 45 1-47.


Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. Journal of Statistical Software 70 1-31.


Junninen, H., Niska, H., Tuppurrainen, K., Ruuskanen, J. and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38 2895-2907.


Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. Journal of Statistical Software 74 1-16.


Lee, L. F. and Yu, J. (2010). Some recent developments in spatial panel data models. Regional Science and urban Economics 40 255-271.


Liu, S. and Molenaar, P. C. (2014). iVAR: a program for imputing missing data in multivariate time series using vector autoregressive models. Behavior Research Methods 46 1138-1148.


Moritz, S. and Bartz-Beielstein, T. (2017). imputeTS: Time Series Missing Value Imputation in R. To appear on The R Journal.


Norazian, M. N., Shukri, Y. A., Azam, R. N. and Mustafa Al Bakri, A. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. ScienceAsia 34 341-345.


Oehmcke, S., Zielinski, O. and O., K. (2016). kNN Ensembles with Penalized DTW for Multivariate Time Series Imputation. International Joint Conference on Neural Networks (IJCNN), IEEE.


Pollice, A. and Lasinio, G. J. (2009). Two approaches to imputation and ad- justment of air quality data from a composite monitoring network. Journal of Data Science 7 43-59.


Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M. and Weinmayr, G. e. a. (2013). Air pollution and lung cancer incidence in 17 European cohorts: prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). The Lancet Oncology 14 813-822.


van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate Impu- tation by Chained Equations in R. Journal of Statistical Software 45 1-67.