A multivariate approach to the analysis of air quality in a high environmental risk area

Alessio Pollice* Giovanna Jona Lasinio** Serena Arima**

*Dipartimento di Scienze Statistiche "Carlo Cecchi"
Università degli Studi di Bari

**Dipartimento di Statistica, Probabilità e Statistiche applicate
Università di Roma "La Sapienza"

19th annual meeting of the International Environmetrics Society
Kelowna, June 8-13 2008

Topics

1. Introduction
2. Modelling issues
3. 3 pollutants daily concentrations - Taranto, 2005-2007

Location of the Taranto area

Sponsor and main objectives

Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

Sponsor and main objectives

Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

Sponsor and main objectives

Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

Sponsor and main objectives

Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

Sponsor and main objectives

Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

Sponsor and main objectives

Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

Sponsor and main objectives

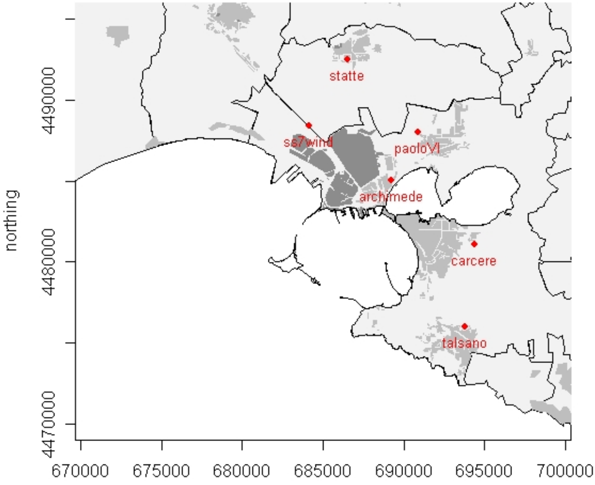
Regional Environmental Protection Agency

- TIES2007, Mikulov: A. Pollice, G. Jona Lasinio *"Spatial analysis of PM10 concentrations with seasonal adjustment"*
- TIES2008, Kelowna: Multi-pollutant spatio-temporal extension, *still a work in progress*

Main objectives:

- summarize the behaviour of pollution diffusion processes over the area of the municipality for a study period
- integrate pollution and meteorological data
- compare alternative approaches to the Bayesian modelling of multivariate spatio-temporal pollution data

The ARPA network - 6 monitoring stations



The database

- Study period: 1 jan 2005 - 31 dec 2007
- Three pollutants:
  - PM10 - particulate matter
  - SO2 - sulphur dioxide
  - NO2 - nitrogen dioxide
- Daily averages of hourly concentration levels
- Normalizing transformations:
  - PM10, NO2 $\quad log(\text{daily average})$
  - SO2 $\quad \sqrt{log(\text{daily average})}$

The database

- Study period: 1 jan 2005 - 31 dec 2007
- Three pollutants:
  - PM10 - particulate matter
  - SO2 - sulphur dioxide
  - NO2 - nitrogen dioxide
- Daily averages of hourly concentration levels
- Normalizing transformations:
  - PM10, NO2 $\quad log(\text{daily average})$
  - SO2 $\quad \sqrt{log(\text{daily average})}$

The database

- Study period: 1 jan 2005 - 31 dec 2007
- Three pollutants:
    - PM10 - particulate matter
    - SO2 - sulphur dioxide
    - NO2 - nitrogen dioxide
- Daily averages of hourly concentration levels
- Normalizing transformations:
    - PM10, NO2 $\quad log(\text{daily average})$
    - SO2 $\quad \sqrt{log(\text{daily average})}$

The database

- Study period: 1 jan 2005 - 31 dec 2007
- Three pollutants:
  - PM10 - particulate matter
  - SO2 - sulphur dioxide
  - NO2 - nitrogen dioxide
- Daily averages of hourly concentration levels
- Normalizing transformations:
  - PM10, NO2 $\quad log(\text{daily average})$
  - SO2 $\quad \sqrt{log(\text{daily average})}$

The database - meteo

- Hourly meteorological data for 3 monitoring stations were made available including: temperature, relative humidity, pressure, rain, solar radiation, wind speed and direction
- A complete daily database was obtained by:
  - Choosing one of the three stations as the main source of data
  - Combining with more reliable pressure and solar radiation measurements recorded by each of the other two monitors
  - Obtaining daily averages by:
    - aritmetic mean (temperature, relative humidity, pressure)
    - geometric mean (wind speed, solar radiation)
    - circular mean (wind direction)
    - mode (wind direction - quadrants)
    - maximum (wind speed)
    - sum (rain)
  - Imputing missing daily values by averaging hourly data recorded 12h before and after the gap. Missing daily rain levels were imputed as averages of those recorded at the other two stations

The database - meteo

- Hourly meteorological data for 3 monitoring stations were made available including: temperature, relative humidity, pressure, rain, solar radiation, wind speed and direction
- A complete daily database was obtained by:
  - Choosing one of the three stations as the main source of data
  - Combining with more reliable pressure and solar radiation measurements recorded by each of the other two monitors
  - Obtaining daily averages by:
    - aritmetic mean (temperature, relative humidity, pressure)
    - geometric mean (wind speed, solar radiation)
    - circular mean (wind direction)
    - mode (wind direction - quadrants)
    - maximum (wind speed)
    - sum (rain)
  - Imputing missing daily values by averaging hourly data recorded 12h before and after the gap. Missing daily rain levels were imputed as averages of those recorded at the other two stations

EDA - pollutants

- AR(1) time dependence with similar coefficients for the 18 series (3 pollutants $\times$ 6 sites) AR(1)
- Spatial dependence doesn't follow a well defined parametric model
- Space-time separability was checked SEP
- Marginal pollutant dependence

	PM10	SO2	NO2
PM10	$.53^2$	.02	.07
SO2	.08	$.39^2$	.04
NO2	.21	.18	$.62^2$

- Missing daily averages (%)

	Archimede	Carcere	PaoloVI	SS7wind	Statte	Talsano
PM10	321 (29)	98 (09)	143 (13)	183 (17)	199 (18)	20 (02)
SO2	183 (17)	109 (10)	176 (16)	206 (19)	93 (08)	25 (02)
NO2	209 (19)	120 (11)	202 (18)	214 (20)	159 (15)	71 (06)

EDA - pollutants

- AR(1) time dependence with similar coefficients for the 18 series (3 pollutants $\times$ 6 sites) AR(1)
- Spatial dependence doesn't follow a well defined parametric model
- Space-time separability was checked SEP
- Marginal pollutant dependence

	PM10	SO2	NO2
PM10	$.53^2$	.02	.07
SO2	.08	$.39^2$	.04
NO2	.21	.18	$.62^2$

- Missing daily averages (%)

	Archimede	Carcere	PaoloVI	SS7wind	Statte	Talsano
PM10	321 (29)	98 (09)	143 (13)	183 (17)	199 (18)	20 (02)
SO2	183 (17)	109 (10)	176 (16)	206 (19)	93 (08)	25 (02)
NO2	209 (19)	120 (11)	202 (18)	214 (20)	159 (15)	71 (06)

EDA - pollutants

- AR(1) time dependence with similar coefficients for the 18 series (3 pollutants $\times$ 6 sites) AR(1)
- Spatial dependence doesn't follow a well defined parametric model
- Space-time separability was checked SEP
- Marginal pollutant dependence

	PM10	SO2	NO2
PM10	$.53^2$	.02	.07
SO2	.08	$.39^2$	.04
NO2	.21	.18	$.62^2$

- Missing daily averages (%)

	Archimede	Carcere	PaoloVI	SS7wind	Statte	Talsano
PM10	321 (29)	98 (09)	143 (13)	183 (17)	199 (18)	20 (02)
SO2	183 (17)	109 (10)	176 (16)	206 (19)	93 (08)	25 (02)
NO2	209 (19)	120 (11)	202 (18)	214 (20)	159 (15)	71 (06)

EDA - pollutants

- AR(1) time dependence with similar coefficients for the 18 series (3 pollutants × 6 sites) AR(1)
- Spatial dependence doesn't follow a well defined parametric model
- Space-time separability was checked SEP
- Marginal pollutant dependence

	PM10	SO2	NO2
PM10	$.53^2$	.02	.07
SO2	.08	$.39^2$	.04
NO2	.21	.18	$.62^2$

- Missing daily averages (%)

	Archimede	Carcere	PaoloVI	SS7wind	Statte	Talsano
PM10	321 (29)	98 (09)	143 (13)	183 (17)	199 (18)	20 (02)
SO2	183 (17)	109 (10)	176 (16)	206 (19)	93 (08)	25 (02)
NO2	209 (19)	120 (11)	202 (18)	214 (20)	159 (15)	71 (06)

EDA - pollutants

- AR(1) time dependence with similar coefficients for the 18 series (3 pollutants $\times$ 6 sites) AR(1)
- Spatial dependence doesn't follow a well defined parametric model
- Space-time separability was checked SEP
- Marginal pollutant dependence

	PM10	SO2	NO2
PM10	$.53^2$	.02	.07
SO2	.08	$.39^2$	.04
NO2	.21	.18	$.62^2$

- Missing daily averages (%)

	Archimede	Carcere	PaoloVI	SS7wind	Statte	Talsano
PM10	321 (29)	98 (09)	143 (13)	183 (17)	199 (18)	20 (02)
SO2	183 (17)	109 (10)	176 (16)	206 (19)	93 (08)	25 (02)
NO2	209 (19)	120 (11)	202 (18)	214 (20)	159 (15)	71 (06)

EDA - influential explanatory variables

- Conditional OLS estimates were obtained for the 3 pollutants at the 6 sites with weekday and month calendar variables and all the meteo covariates as explanatory variables
- Pollutant concentration levels were overall significantly affected by the effects of:
  - weekday
  - month
  - temperature
  - humidity
  - rain
  - maximum wind speed
  - wind direction quadrant

EDA - influential explanatory variables

- Conditional OLS estimates were obtained for the 3 pollutants at the 6 sites with weekday and month calendar variables and all the meteo covariates as explanatory variables
- Pollutant concentration levels were overall significantly affected by the effects of:
  - weekday
  - month
  - temperature
  - humidity
  - rain
  - maximum wind speed
  - wind direction quadrant

Two Hierarchical Bayesian multivariate space-time models

(I) Le & Zidek (2006)

- Semi-parametric nonstationary anisotropic spatial covariance structure
- Conditional independence of pollutant concentrations over time given covariates
- Only staircase and systematic patterns of missing data
- Software package implementing the model available at `http://enviRo.stat.ubc.ca`

(II) Shaddick & Wakefield (2002)

- Exponential spatial covariance structure
- First order random walk nonstationary temporal structure
- Any pattern of missing data
- Can be implemented in WinBUGS

Two Hierarchical Bayesian multivariate space-time models

(I) Le & Zidek (2006)

- Semi-parametric nonstationary anisotropic spatial covariance structure
- Conditional independence of pollutant concentrations over time given covariates
- Only staircase and systematic patterns of missing data
- Software package implementing the model available at `http://enviRo.stat.ubc.ca`

(II) Shaddick & Wakefield (2002)

- Exponential spatial covariance structure
- First order random walk nonstationary temporal structure
- Any pattern of missing data
- Can be implemented in WinBUGS

(I) Le & Zidek, 2006 - *notation*

- p pollutants
- r regressors
- t time points
- g monitoring stations (gauged sites)
- u prediction points (ungauged sites)
- $s = g + u$ spatial locations

- *spt*-dimensional response vector Y contains normalized daily mean pollutant concentrations
- $(spt \times spr)$-dimensional matrix $Z = I_{sp} \otimes \tilde{Z}$ contains *sp* replicates of common time-varying covariates $\tilde{Z}$ measured at one site

(I) Le & Zidek, 2006 - *notation*

- p pollutants
- r regressors
- t time points
- g monitoring stations (gauged sites)
- u prediction points (ungauged sites)
- $s = g + u$ spatial locations

- spt-dimensional response vector Y contains normalized daily mean pollutant concentrations
- $(spt \times spr)$-dimensional matrix $Z = I_{sp} \otimes \tilde{Z}$ contains sp replicates of common time-varying covariates $\tilde{Z}$ measured at one site

(I) Le & Zidek, 2006 - *notation*

- p pollutants
- r regressors
- t time points
- g monitoring stations (gauged sites)
- u prediction points (ungauged sites)
- $s = g + u$ spatial locations

- spt-dimensional response vector Y contains normalized daily mean pollutant concentrations
- $(spt \times spr)$-dimensional matrix $Z = I_{sp} \otimes \tilde{Z}$ contains sp replicates of common time-varying covariates $\tilde{Z}$ measured at one site

(I) Le & Zidek, 2006 - *the model*

Level I: data process

$$Y|Z,\beta,\Sigma \sim N_{spt}(Z\beta, I_t \otimes \Sigma)$$

- regression coefficients in β vary over sites
- Σ between sites/pollutants covariance matrix
- Kronecker structure $\Longrightarrow$ $Y|Z$ are independent over time

Level II: conjugate prior distributions

$$\beta|\beta_0,\Sigma,F \sim N_{rst}(\beta_0, F^{-1} \otimes \Sigma)$$

$$\Sigma|\Theta,\delta \sim \mathrm{IW}(\Theta,\delta)$$

- F^{-1} among covariates variance component of β
- GIW can be substituted to IW in case of staircase missing data

(I) Le & Zidek, 2006 - *the model*

Level I: data process

$$Y|Z, \beta, \Sigma \sim N_{spt}(Z\beta, I_t \otimes \Sigma)$$

- regression coefficients in β vary over sites
- Σ between sites/pollutants covariance matrix
- Kronecker structure $\Longrightarrow$ $Y|Z$ are independent over time

Level II: conjugate prior distributions

$$\beta|\beta_0, \Sigma, F \sim N_{rst}(\beta_0, F^{-1} \otimes \Sigma)$$

$$\Sigma|\Theta, \delta \sim \mathrm{IW}(\Theta, \delta)$$

- F^{-1} among covariates variance component of β
- GIW can be substituted to IW in case of staircase missing data

(I) Le & Zidek, 2006 - *estimation & prediction*

- The predictive distribution is a multivariate t-distribution depending on hyperparameters β_0, F, Θ and δ
- Two-step hyperparameter estimation procedure
  - Gauged sites: EM marginal likelihood maximization (empirical Bayes/type-II MLE)
  - Ungauged sites: spatial covariance and cross-covariance matrices are obtained by the Sampson-Guttorp method, introducing nonstationarity and anisotropy of the spatial fields (Sampson & Guttorp, 1992) SG

(I) Le & Zidek, 2006 - *estimation & prediction*

- The predictive distribution is a multivariate t-distribution depending on hyperparameters β_0, F, Θ and δ
- Two-step hyperparameter estimation procedure
  - Gauged sites: EM marginal likelihood maximization (empirical Bayes/type-II MLE)
  - Ungauged sites: spatial covariance and cross-covariance matrices are obtained by the Sampson-Guttorp method, introducing nonstationarity and anisotropy of the spatial fields (Sampson & Guttorp, 1992) SG

(II) Shaddick & Wakefield, 2002 - *the model*

Level I: data process

$$Y|\mu, \tau_p \sim N_{spt}(\mu, I_s \otimes \tau_p \otimes I_t)$$

- $\mu = Z\beta + \theta_{pt} \otimes u_s + u_{pt} \otimes \epsilon_s$
  - $(spt \times r)$-dimensional matrix Z contains (possibly) time-varying and spatially varying covariates
  - θ_{pt} joint effect of pollutant and time
  - ϵ_s error term including the spatial effect
- τ_p diagonal matrix of the pollutants residual variances

(II) Shaddick & Wakefield, 2002 - *the model*

Level I: data process

$$Y|\mu, \tau_p \sim N_{spt}(\mu, I_s \otimes \tau_p \otimes I_t)$$

- $\mu = Z\beta + \theta_{pt} \otimes u_s + u_{pt} \otimes \epsilon_s$
    - $(spt \times r)$-dimensional matrix Z contains (possibly) time-varying and spatially varying covariates
    - θ_{pt} joint effect of pollutant and time
    - ϵ_s error term including the spatial effect
- τ_p diagonal matrix of the pollutants residual variances

(II) Shaddick & Wakefield, 2002 - *the model*

Level II: prior distributions

- $\beta \sim N_r$
- $\theta_{p,t'} | \theta_{p,t'-1}, \tau_\theta \sim N_p(\theta_{p,t'-1}, \tau_\theta), \qquad t' = 2, \ldots, t$
- $\epsilon_s | \sigma_\epsilon, \Sigma \sim N(0_s, \sigma_\epsilon^2 \Sigma), \quad \Sigma_{s',s''} = \exp(-\phi d_{s's''}), \quad s', s'' = 1, \ldots, s$
- $\tau_{p'} \sim \text{Gamma}, \qquad p' = 1, \ldots, p$

Level III: hyperpriors

- $\tau_\theta \sim \text{Gamma}$
- $\sigma_\epsilon^{-1} \sim \text{Gamma}$
- $\phi \sim U[0, 1]$

(II) Shaddick & Wakefield, 2002 - *the model*

Level II: prior distributions

- $\beta \sim N_r$
- $\theta_{p,t'} | \theta_{p,t'-1}, \tau_\theta \sim N_p(\theta_{p,t'-1}, \tau_\theta), \qquad t' = 2, \ldots, t$
- $\epsilon_s | \sigma_\epsilon, \Sigma \sim N(0_s, \sigma_\epsilon^2 \Sigma), \quad \Sigma_{s',s''} = \exp(-\phi d_{s's''}), \quad s', s'' = 1, \ldots, s$
- $\tau_{p'} \sim \text{Gamma}, \qquad p' = 1, \ldots, p$

Level III: hyperpriors

- $\tau_\theta \sim \text{Gamma}$
- $\sigma_\epsilon^{-1} \sim \text{Gamma}$
- $\phi \sim U[0, 1]$

(II) Shaddick & Wakefield, 2002 - *estimation & prediction*

- Analitically intractable joint posterior distribution of model parameters, but posterior samples can be generated by MCMC
- Pollutant concentrations at unmonitored days (NA's) or sites (ungauged prediction points) can be treated as unknown parameters: samples from their posterior distribution can be generated

(II) Shaddick & Wakefield, 2002 - *estimation & prediction*

- Analitically intractable joint posterior distribution of model parameters, but posterior samples can be generated by MCMC
- Pollutant concentrations at unmonitored days (NA's) or sites (ungauged prediction points) can be treated as unknown parameters: samples from their posterior distribution can be generated

Model comparison: advantages

- Model I (LZ)
  - Semiparametric nonstationary anisotropic spatial covariance structure
  - Explicit analytic expression of the predictive distribution (no MCMC!)
  - Implementation in R
- Model II (SW)
  - Inclusion of a spatial trend as a function of the coordinates
  - Accounts for time variability
  - Allows any missing data pattern

Model comparison: advantages

- Model I (LZ)
  - Semiparametric nonstationary anisotropic spatial covariance structure
  - Explicit analytic expression of the predictive distribution (no MCMC!)
  - Implementation in R
- Model II (SW)
  - Inclusion of a spatial trend as a function of the coordinates
  - Accounts for time variability
  - Allows any missing data pattern

Model comparison: disadvantages

- Model I (LZ)
  - Need for sparse missing data imputation
  - Time conditional independence assumption: need to filter the time variability
- Model II (SW)
  - Big MCMC issues (sensitivity to prior specification and slow convergence: *chains are still running in Rome!!*)

Model comparison: disadvantages

- Model I (LZ)
  - Need for sparse missing data imputation
  - Time conditional independence assumption: need to filter the time variability
- Model II (SW)
  - Big MCMC issues (sensitivity to prior specification and slow convergence: *chains are still running in Rome!!*)

Model I (LZ) *Missing data & temporal correlation*

- An iterative procedure based on function `krige.bayes` in the `R` library `geoR` (Diggle & Ribeiro, 2002) is used to reconstruct the daily spatial fields of each pollutant (Pollice & Jona Lasinio, 2008) PJL
  - Missing data predictions are obtained within a daily spatial leave-one-out scheme
  - Priors are set by posterior estimates obtained on the previous day (sort of order 1 type dependence, with spatial covariance parameter estimates depending stochastically on those of the day before)
  - Predictions are recursively repeated until convergence
- Residuals of AR(1) models fitted to each pollutant concentration data are obtained
- The space-time model is fitted to such imputed residuals

Model I (LZ) *Missing data & temporal correlation*

- An iterative procedure based on function `krige.bayes` in the R library `geoR` (Diggle & Ribeiro, 2002) is used to reconstruct the daily spatial fields of each pollutant (Pollice & Jona Lasinio, 2008) PJL
  - Missing data predictions are obtained within a daily spatial leave-one-out scheme
  - Priors are set by posterior estimates obtained on the previous day (sort of order 1 type dependence, with spatial covariance parameter estimates depending stochastically on those of the day before)
  - Predictions are recursively repeated until convergence
- Residuals of AR(1) models fitted to each pollutant concentration data are obtained
- The space-time model is fitted to such imputed residuals

Model I (LZ) *Missing data & temporal correlation*

- An iterative procedure based on function `krige.bayes` in the R library `geoR` (Diggle & Ribeiro, 2002) is used to reconstruct the daily spatial fields of each pollutant (Pollice & Jona Lasinio, 2008) PJL
  - Missing data predictions are obtained within a daily spatial leave-one-out scheme
  - Priors are set by posterior estimates obtained on the previous day (sort of order 1 type dependence, with spatial covariance parameter estimates depending stochastically on those of the day before)
  - Predictions are recursively repeated until convergence
- Residuals of AR(1) models fitted to each pollutant concentration data are obtained
- The space-time model is fitted to such imputed residuals

Model I (LZ) *Missing data & temporal correlation*

- An iterative procedure based on function `krige.bayes` in the `R` library `geoR` (Diggle & Ribeiro, 2002) is used to reconstruct the daily spatial fields of each pollutant (Pollice & Jona Lasinio, 2008) PJL
  - Missing data predictions are obtained within a daily spatial leave-one-out scheme
  - Priors are set by posterior estimates obtained on the previous day (sort of order 1 type dependence, with spatial covariance parameter estimates depending stochastically on those of the day before)
  - Predictions are recursively repeated until convergence
- Residuals of AR(1) models fitted to each pollutant concentration data are obtained
- The space-time model is fitted to such imputed residuals

Model I (LZ) *Prediction implementation and assessment*

- The predictive distribution is obtained for a 400 points square interpolation grid, giving:
  - Daily expectations
  - Mean, variances and quantiles of 1000 daily simulations
- Estimates of AR(1) coefficients for the three pollutants are used to put back the temporal component
- Normalizing transformations are used to back-transform to the original scale
- Observed values are compared to predictions at the nearest grid-points GRIDP
  - Model validation statistics (Carrol & Cressie, 1996) CC
  - Credibility intervals of predictions are obtained by the quantiles of the simulations

Model I (LZ) *Prediction implementation and assessment*

- The predictive distribution is obtained for a 400 points square interpolation grid, giving:
  - Daily expectations
  - Mean, variances and quantiles of 1000 daily simulations
- Estimates of AR(1) coefficients for the three pollutants are used to put back the temporal component
- Normalizing transformations are used to back-transform to the original scale
- Observed values are compared to predictions at the nearest grid-points GRIDP
  - Model validation statistics (Carrol & Cressie, 1996) CC
  - Credibility intervals of predictions are obtained by the quantiles of the simulations

Model I (LZ) *Prediction implementation and assessment*

- The predictive distribution is obtained for a 400 points square interpolation grid, giving:
  - Daily expectations
  - Mean, variances and quantiles of 1000 daily simulations
- Estimates of AR(1) coefficients for the three pollutants are used to put back the temporal component
- Normalizing transformations are used to back-transform to the original scale
- Observed values are compared to predictions at the nearest grid-points GRIDP
  - Model validation statistics (Carrol & Cressie, 1996) CC
  - Credibility intervals of predictions are obtained by the quantiles of the simulations

Model I (LZ) *Prediction implementation and assessment*

- The predictive distribution is obtained for a 400 points square interpolation grid, giving:
  - Daily expectations
  - Mean, variances and quantiles of 1000 daily simulations
- Estimates of AR(1) coefficients for the three pollutants are used to put back the temporal component
- Normalizing transformations are used to back-transform to the original scale
- Observed values are compared to predictions at the nearest grid-points GRIDP
  - Model validation statistics (Carrol & Cressie, 1996) CC
  - Credibility intervals of predictions are obtained by the quantiles of the simulations

Model I (LZ) *Assessing predictions - overall*

- Model validation statistics CC

	CR_1	CR_2	CR_3
PM10	$< e - 04$	0.56	0.22
SO2	$< e - 04$	0.59	0.20
NO2	$< e - 04$	0.63	0.30
best	0	1	small

- Credibility intervals coverage (%)

		Nominal			
		50	80	90	95
Empirical	PM10	79.6	96.0	98.6	99.5
	SO2	73.4	97.9	99.9	100
	NO2	72.3	95.3	98.9	99.8

Model I (LZ) *Assessing predictions - overall*

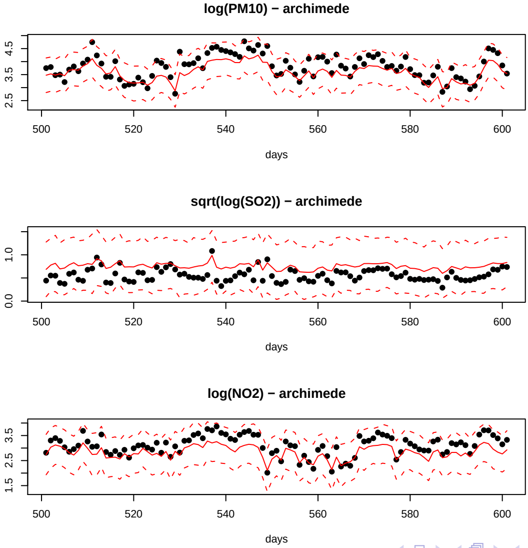
- Model validation statistics CC

	CR_1	CR_2	CR_3
PM10	$< e-04$	0.56	0.22
SO2	$< e-04$	0.59	0.20
NO2	$< e-04$	0.63	0.30
best	0	1	small

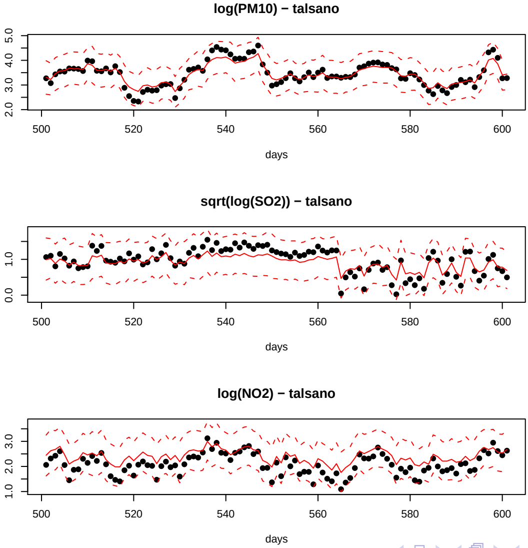
- Credibility intervals coverage (%)

		Nominal			
		50	80	90	95
Empirical	PM10	79.6	96.0	98.6	99.5
	SO2	73.4	97.9	99.9	100
	NO2	72.3	95.3	98.9	99.8

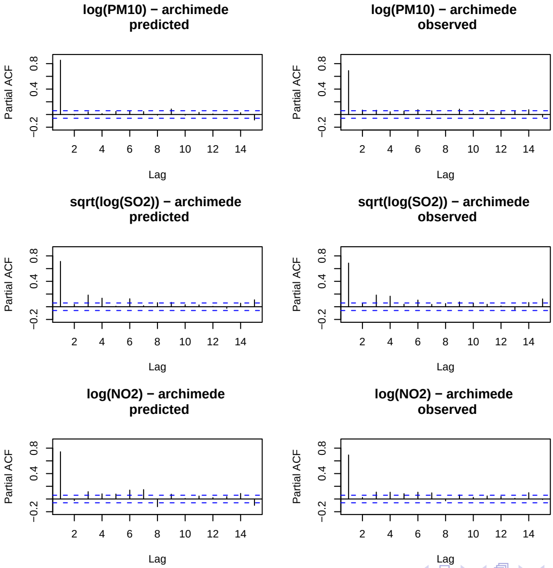
Model I (LZ) *Assessing predictions - time patterns* BT1



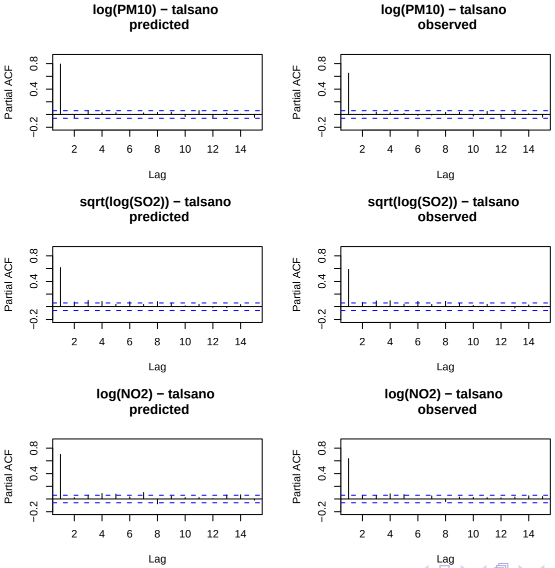
Model I (LZ) *Assessing predictions - time patterns* BT6



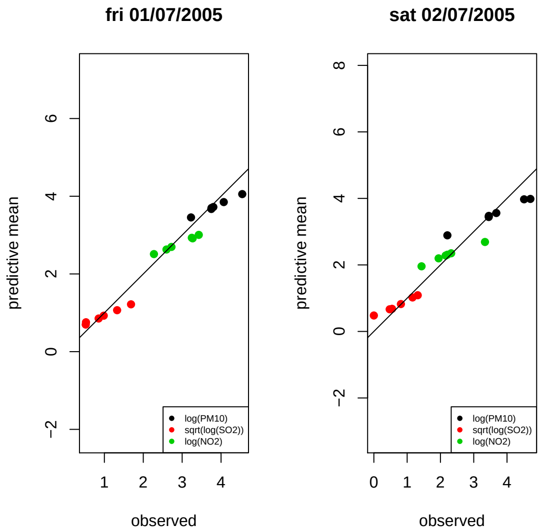
Model I (LZ) *Assessing predictions - time patterns*



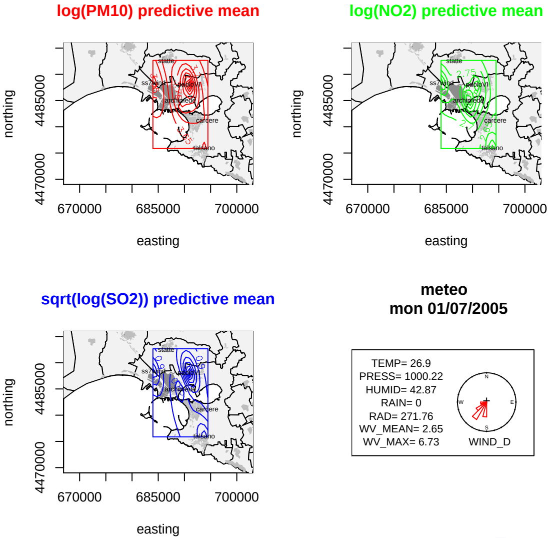
Model I (LZ) *Assessing predictions - time patterns*



Model I (LZ) *Assessing predictions - spatial patterns*

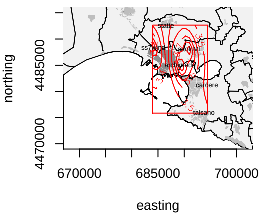


Model I (LZ) *Spatial patterns* STD1 CI1 map1 BT

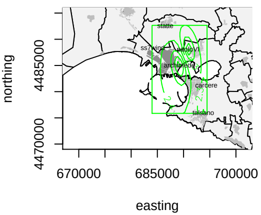


Model I (LZ) *Spatial patterns* STD2 CI2 map2 BT

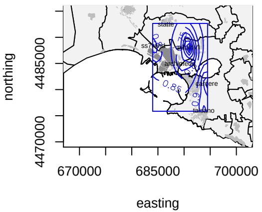
log(PM10) predictive mean



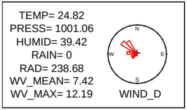
log(NO2) predictive mean



sqrt(log(SO2)) predictive mean



meteo
mon 02/07/2005



Conclusions

- We introduce several tools to analyze air quality data dense in time and sparse in space
- Among these tools we propose an original data imputation procedure and we organize several EDA procedures to elicit model elements
- Two estimation models were considered and one (LZ) was deeply explored
  - LZ model has the advantage of fast computation and good estimation quality
  - However, due to the many steps required to reach predictions, the evaluation of their uncertainty is not very reliable

Conclusions

- We introduce several tools to analyze air quality data dense in time and sparse in space
- Among these tools we propose an original data imputation procedure and we organize several EDA procedures to elicit model elements
- Two estimation models were considered and one (LZ) was deeply explored
  - LZ model has the advantage of fast computation and good estimation quality
  - However, due to the many steps required to reach predictions, the evaluation of their uncertainty is not very reliable

Conclusions

- We introduce several tools to analyze air quality data dense in time and sparse in space
- Among these tools we propose an original data imputation procedure and we organize several EDA procedures to elicit model elements
- Two estimation models were considered and one (LZ) was deeply explored
  - LZ model has the advantage of fast computation and good estimation quality
  - However, due to the many steps required to reach predictions, the evaluation of their uncertainty is not very reliable

Further developments

- Verify if Model II (SW) can provide a more accurate uncertainty evaluation (overcoming the “multiple steps” problem)
- Model II (SW) structure, few questions to be answered:
  - suitability of the random walk structure for the joint effect of pollutant and time (θ_{pt})
  - relevance of the role of categorical covariates in improving estimates
  - effect of categorical covariates in slowing down the convergence of MCMC chains
- Model II (LZ): sensitivity analysis to choose the “best” number of grid points and also the “best” degree of smoothing
- Suggest the best protocol according to final users needs

Further developments

- Verify if Model II (SW) can provide a more accurate uncertainty evaluation (overcoming the "multiple steps" problem)
- Model II (SW) structure, few questions to be answered:
  - suitability of the random walk structure for the joint effect of pollutant and time (θ_{pt})
  - relevance of the role of categorical covariates in improving estimates
  - effect of categorical covariates in slowing down the convergence of MCMC chains
- Model II (LZ): sensitivity analysis to choose the "best" number of grid points and also the "best" degree of smoothing
- Suggest the best protocol according to final users needs

Further developments

- Verify if Model II (SW) can provide a more accurate uncertainty evaluation (overcoming the "multiple steps" problem)
- Model II (SW) structure, few questions to be answered:
  - suitability of the random walk structure for the joint effect of pollutant and time (θ_{pt})
  - relevance of the role of categorical covariates in improving estimates
  - effect of categorical covariates in slowing down the convergence of MCMC chains
- Model II (LZ): sensitivity analysis to choose the "best" number of grid points and also the "best" degree of smoothing
- Suggest the best protocol according to final users needs

Further developments

- Verify if Model II (SW) can provide a more accurate uncertainty evaluation (overcoming the "multiple steps" problem)
- Model II (SW) structure, few questions to be answered:
  - suitability of the random walk structure for the joint effect of pollutant and time (θ_{pt})
  - relevance of the role of categorical covariates in improving estimates
  - effect of categorical covariates in slowing down the convergence of MCMC chains
- Model II (LZ): sensitivity analysis to choose the "best" number of grid points and also the "best" degree of smoothing
- Suggest the best protocol according to final users needs

Essential references

- Carroll, S.S., Cressie, N. (1996) *A comparison of geostatistical methodologies used to estimate snow water equivalent.* Wat. Resour. Bull., **32**, 267-278.
- Diggle, P.J., Ribeiro Jr, P.J. (2002) *Bayesian inference in Gaussian model-based geostatistics.* Geographical and Environmental Modelling, **6**, 129-146.
- Le, N.D., Zidek, J.V. (2006) *Statistical Analysis of Environmental Space-Time Processes.* Springer.
- Pollice A., Jona Lasinio G. (2008) *Two approaches to imputation and adjustment of air quality data from a composite monitoring network.* GRASPA Working Paper, **30**, `www.graspa.org`.
- Sahu, S.K., Mardia K.V. (2005) *A Bayesian kriged Kalman model for short-term forecasting of air pollution levels.* Appl. Statist., **54**, 223-244.
- Sampson P., Guttorp P. (1992) *Nonparametric estimation of non stationary spatial structure.* JASA, **87**, 108-119.

thank you for your attention

Sampson & Guttorp, 1992

- Iterative two-step approach based on multidimensional scaling to obtain virtual locations for which the isotropy assumption is appropriate and on thin-plate splines to estimate the smooth mapping between original geographic locations and the new ones
- an isotropic variogram model is fitted using the observed correlation and distances of the new locations
- the smooth mapping function, together with the isotropic variogram model estimates the spatial dispersion between the stations and the ungauged sites

The method implies the separability of between sites and between pollutants covariances

Pollice & Jona Lasinio, 2008

The usual LME model is chosen as the daily spatial interpolation model (Diggle and Ribeiro, 2007).

Level I - daily data process: Y is a p-dim GRF representing one pollutant normalized daily mean concentrations

$$Y|\beta, \phi, \tau, \sigma^2 \sim N_p\left(\beta, V_y\left(\frac{\tau^2}{\sigma^2}, \phi\right)\right)$$

Level II - prior specification:

- diffused priors for β and σ^2
- discrete priors on a specified reference grid for covariance structure parameters $\tau^2_{rel} = \tau^2/\sigma^2$ and ϕ

The predictive distribution has to be computed by numerical approximation: values of covariance structure parameters τ^2 and ϕ simulated from their marginal discrete posterior distribution are plugged in the t-type predictive distribution obtained for the fully conjugate case.

Function `krige.bayes` in R library `geoR` is used.

Pollice & Jona Lasinio, 2008

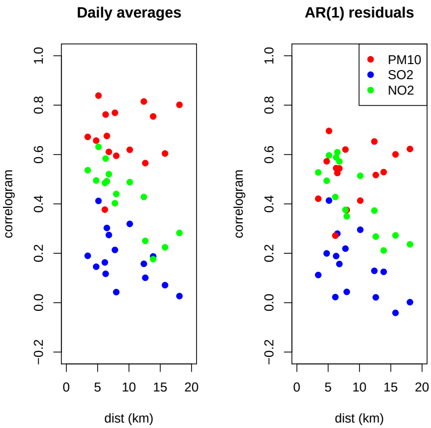
Two daily spatial kinds of models specified as Bayesian LME's are used for missing data imputation: *prediction models* and *estimation models*
Let y be the vector of daily observations and J the set of indices denoting the monitoring stations to be treated.

- Step 0: A discrete uniform prior is chosen for τ^2_{rel} on the interval (0,1) with 0.1 increments, while ϕ is allowed to vary in a discrete sequence between 1 and 7 km with 0.5km incremental value and a reciprocal prior. For day 1 fit the estimation model to vector y where data corresponding to the stations to be treated are omitted. For days 2 to 365 fit the estimation model to vector y of the previous day, where data corresponding to the treated stations (z) are substituted. Obtain daily posterior estimates of ϕ and τ^2_{rel}.

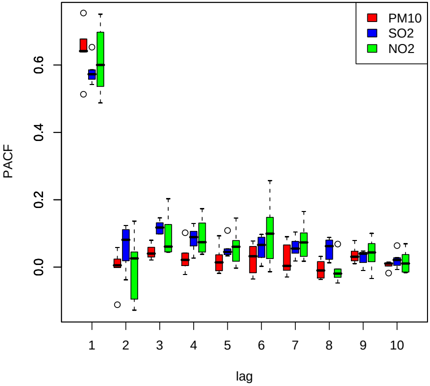
Pollice & Jona Lasinio, 2008

- Step 1: For $i \in J$ let $y_{(i)}$ be obtained by omitting station i in the vector of daily observations y. Iteratively predict each y_i from $y_{(i)}$ using posterior estimates of ϕ and τ^2_{rel} obtained in the previous step in the prior specification of the prediction models. Store predicted values in vector z and substitute them to corresponding values in y.
- Step 2: Store the current z values in z_{old} and repeat step 1 to obtain a new z.
- Step 3: If $|z_{old} - z| < \varepsilon$ ($\varepsilon = 0.0001$) or the iterations number is ≥ 100 stop, otherwise repeat step 2 until convergence.

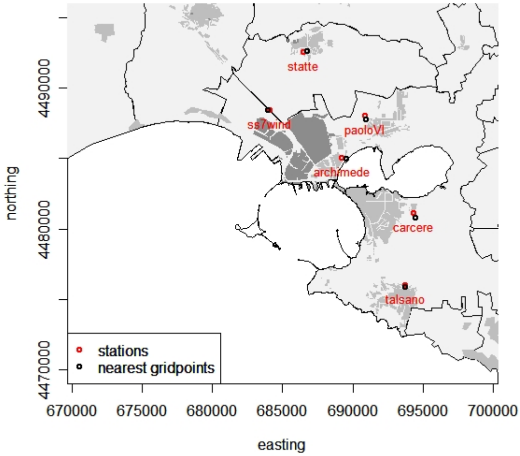
Spatial correlation leakage



PACF's



Assessing predictions



Overall model assessment (Carrol & Cressie, 1996)

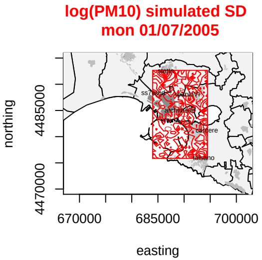
- $$CR_1 = S^{-1} \sum_s \frac{T^{-1} \sum_t \left(Y(s,t) - \hat{Y}(s,t) \right)}{T^{-1} \left(\sum_t \hat{\sigma}^2(s,t) \right)^{1/2}}$$

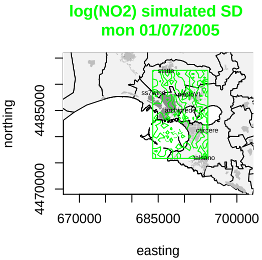
- $$CR_2 = S^{-1} \sum_s \left(\frac{T^{-1} \sum_t \left(Y(s,t) - \hat{Y}(s,t) \right)^2}{T^{-1} \sum_t \hat{\sigma}^2(s,t)} \right)^{1/2}$$

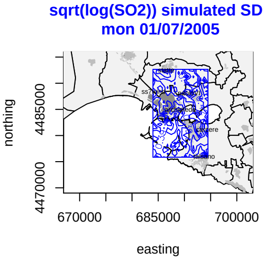
- $$CR_3 = S^{-1} \sum_s \left(T^{-1} \sum_t \left(Y(s,t) - \hat{Y}(s,t) \right)^2 \right)^{1/2}$$

when forecasts are accurate, CR_1 and CR_2 should be close to 0 and 1 respectively; CR_3 provides a "goodness of prediction" and it is expected to be small when predicted values are close to the true values (Sahu & Mardia, 2005)

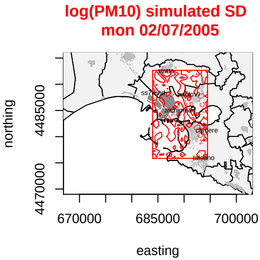
Simulated prediction variability maps

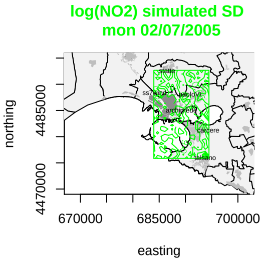


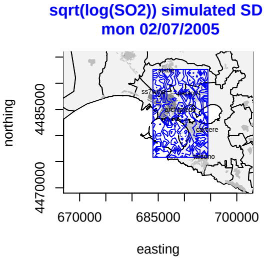




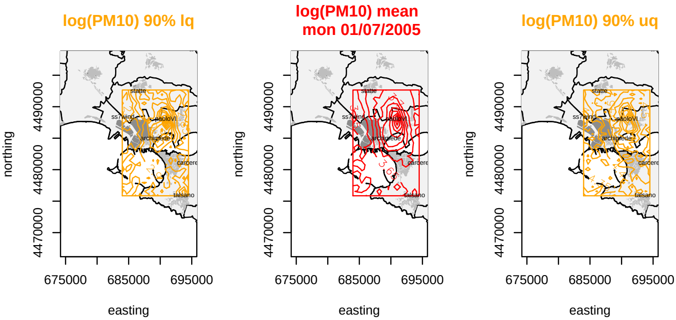
Simulated prediction variability maps



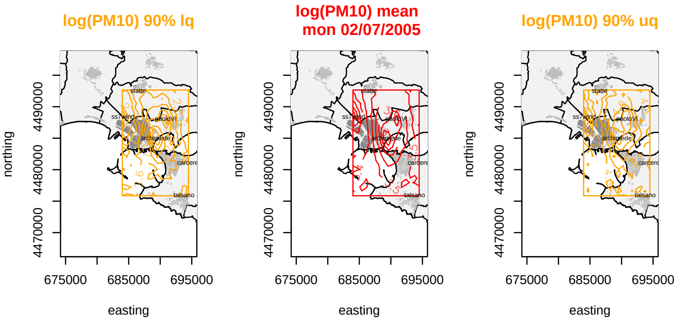




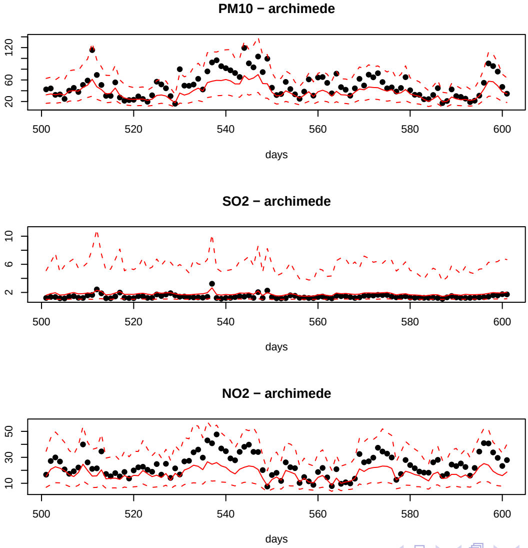
Simulated spatial CI of log PM10 average concentrations



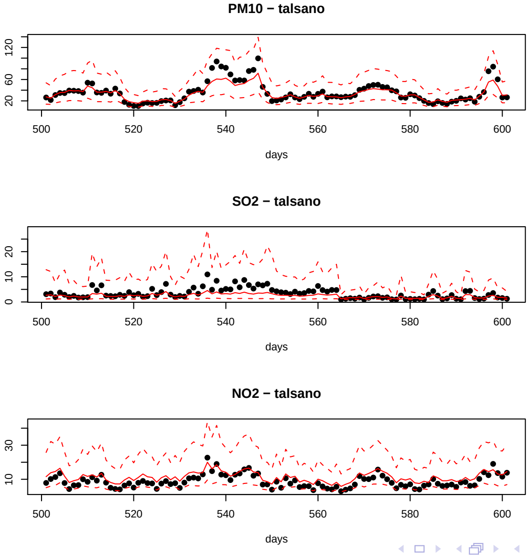
Simulated spatial CI of log PM10 average concentrations



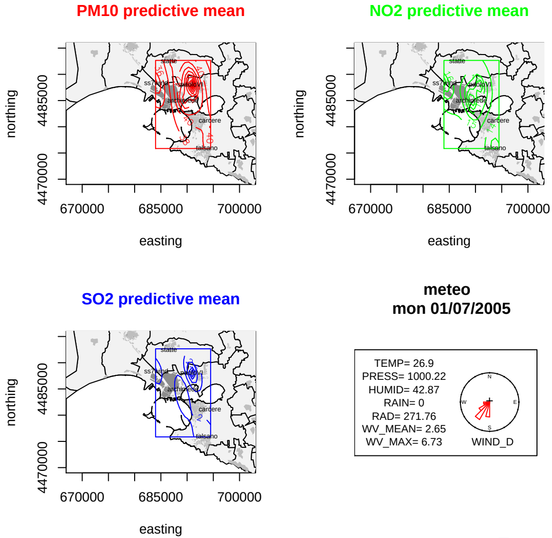
Observed and predicted pollutants concentrations



Observed and predicted pollutants concentrations

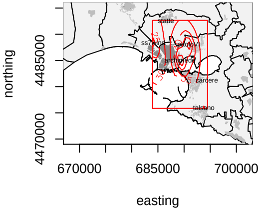


Predicted pollutants concentrations

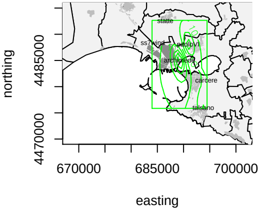


Predicted pollutants concentrations

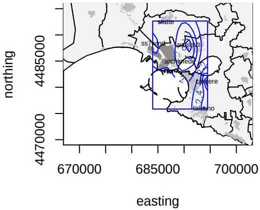
PM10 predictive mean



NO2 predictive mean

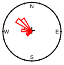


SO2 predictive mean



meteo
mon 02/07/2005

TEMP= 24.82
PRESS= 1001.06
HUMID= 39.42
RAIN= 0
RAD= 238.68
WV_MEAN= 7.42
WV_MAX= 12.19



WIND_D