Air Quality Forecasting

Datena Amedeo, Dell'Atti Martina, Palummo Alessandro

7 January 2021

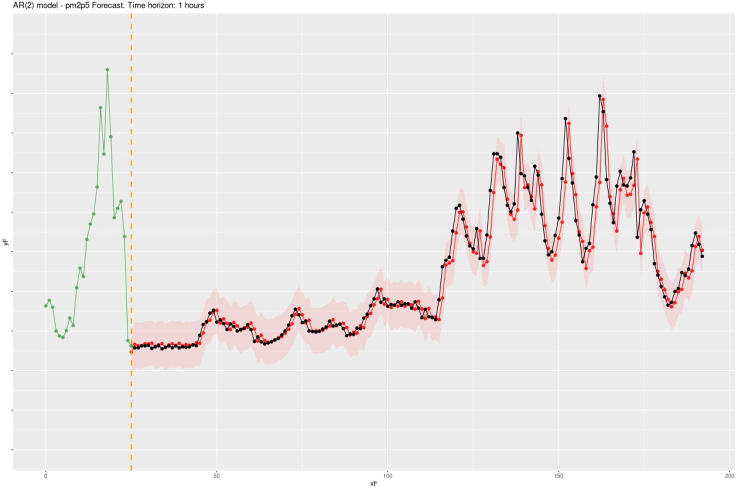
POLITECNICO
MILANO 1863

When we met before

- Our goal is to predict future PM values.
- Our dataset collects values of PM 2.5, PM 1, PM 4 once per hour from different sensors in Milan, called Arianna.
- Other variables of our dataset are
  - latitude and longitude of sensors
  - temperature
  - humidity
  - wind
  - rain
- We start from an **univariate analysis** considering only data from one sensor in the period from 01-09-2020 to 27-10-2020.

AR(2)

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \qquad \epsilon_t | \sigma^2 \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$



$$y_t|y_{t-1},..,y_{t-p},\boldsymbol{X},\beta,\sigma^2 \sim \mathcal{N}(y_t|\boldsymbol{f_t}^T\beta,\sigma^2) \qquad \epsilon_t|\sigma^2 \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma^2)$$

$$\Rightarrow p(\boldsymbol{Y}|\boldsymbol{y_{t-1}},..,\boldsymbol{y_{t-p}},\boldsymbol{X},\beta,\sigma^2) = \prod_{t=p_{max}}^{n+p_{max}} \mathcal{N}(y_t|\boldsymbol{f_t}^T\beta,\sigma^2) = \mathcal{N}_n(\boldsymbol{Y}|\boldsymbol{F}^T\beta,\sigma^2)$$

$$\begin{cases} \boldsymbol{Y}|\boldsymbol{y_{t-1}},..,\boldsymbol{y_{t-p}},\boldsymbol{X},\beta,\sigma^2 \sim \mathcal{N}_n(\boldsymbol{Y}|\boldsymbol{F}^T\beta,\sigma^2) \\ \beta|\sigma^2 \sim \mathcal{N}_k(\boldsymbol{\mu_0},\sigma^2 B_0) \\ \sigma^2 \sim inverse-gamma(\frac{\nu_0}{2},\frac{\nu_0\sigma_0^2}{2}) \end{cases}$$

- $\boldsymbol{Y} = [y_{p\,max},..,y_{p_{max}+n}]$, where $p_{max} = max(p,p_1,p_2,p_3,p_4)$
- $\boldsymbol{X} = [\boldsymbol{x_1},\boldsymbol{x_2},\boldsymbol{x_3},\boldsymbol{x_4}]$ is the vector of covariates: temperature, humidity, rain, wind, where $\boldsymbol{x_i} = [x^i_{t-1},..,x^i_{t-p_i}]$
- $\boldsymbol{f_t^T} = [1,y_{t-1},..,y_{t-p},\boldsymbol{X}]$ is the autoregressive part together with the vector of covariates
- $\boldsymbol{F} = [\boldsymbol{f_{p_{max}}},...,\boldsymbol{f_{p_{max}+n}}]$

$$p = p_1 = p_2 = p_3 = p_4 = 7$$
$$\Downarrow$$
$$y_t = \beta_0 + \beta_1 y_{t-1} + ... + \beta_7 y_{t-7} + \beta_8 x^1_{t-1} + ... + \beta_{14} x^1_{t-7} + ... + \beta_{29} x^4_{t-1} + ... + \beta_{35} x^4_{t-7} + \epsilon_t$$

$$\begin{cases} \boldsymbol{Y}|\boldsymbol{y_{t-1}},..,\boldsymbol{y_{t-p}},\boldsymbol{X},\boldsymbol{\beta},\sigma^2 \overset{i.i.d.}{\sim} \mathcal{N}_n(\boldsymbol{Y}|\boldsymbol{F}^T\boldsymbol{\beta},\sigma^2) & \\ \beta_j|\lambda \overset{i.i.d.}{\sim} \mathcal{N}(0,\frac{1}{\lambda}) & j = 1,\ldots,k \\ \lambda \sim gamma(a_\lambda, b_\lambda) & \\ \sigma^2 \sim inverse-gamma(a_{\sigma^2}, b_{\sigma^2}) & \end{cases}$$

$$y_t = \beta_1 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_9 x^1_{t-1} + \beta_7 x^2_{t-1} + \beta_{18} x^2_{t-2} + \beta_{19} x^2_{t-3} + \beta_{23} x^4_{t-1} + \beta_{24} x^3_{t-7} + \epsilon_t$$

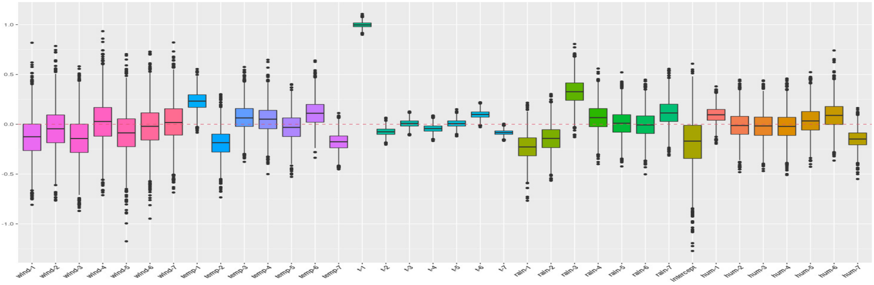


Figure: Boxplots from Ridge regression

Significant covariates (considering 90% credible intervals) from Ridge are:

- lag 1-2 for the autoregressive part $\Rightarrow$ y_{t-1}, y_{t-2}
- lag 1 for humidity $\Rightarrow$ x^1_{t-1}
- lag 1-2-3 for rain $\Rightarrow$ $x^2_{t-1}, x^2_{t-2}, x^2_{t-3}$
- lag 1-2 for temperature $\Rightarrow$ x^3_{t-1}, x^3_{t-2}
- the covariate wind seems to be not significant

Traceplots and posterior density plots of significant parameters

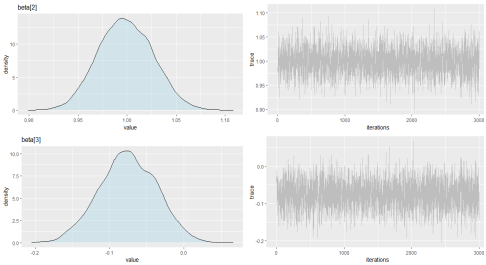


Figure: lag 1-2 autoregressive

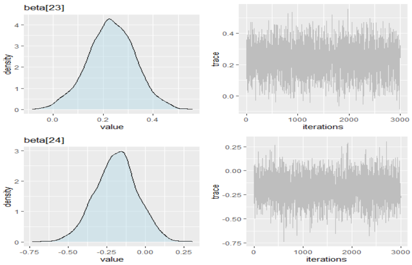


Figure: lag 1-2 temperature

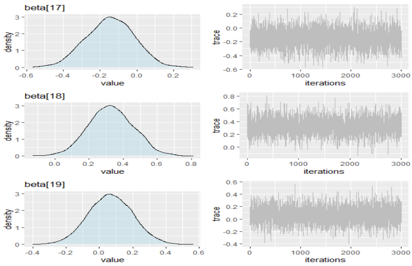


Figure: lag 1-2-3 rain

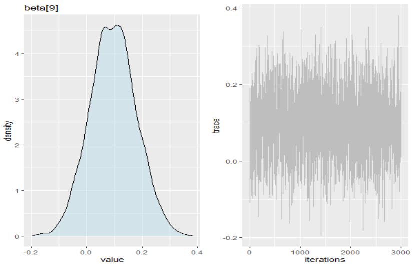


Figure: lag 1 humidity

Hourly SARX: ARX with hourly seasonal effect

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \qquad t = p_{max} : n, \quad p_{max} = max(p, p_1, p_2, p_3)$$

$$\mu_t = \boldsymbol{f_t}^T \boldsymbol{\alpha} + \boldsymbol{x_t}^T \boldsymbol{\beta} + S_t$$

$$S_t = \sum_{i=0}^{T-1} \gamma_i \delta_t^i \qquad T = 24$$

$$\delta_t^i = \begin{cases} 1 & \text{if at time t is hour i} \\ 0 & \text{otherwise} \end{cases}$$

- $\boldsymbol{f_t}^T = (1, y_{t-1}, ..., y_{t-p})$ is the autoregressive term
- $\boldsymbol{x_t}^T = (X_{t-1}^1, ..., X_{t-p_1}^1, ..., X_{t-1}^3, ..., X_{t-p_3}^3)$ is the regressive term

$$\boldsymbol{\mu} = \boldsymbol{F}^T\boldsymbol{\alpha} + \boldsymbol{X}^T\boldsymbol{\beta} + \Delta\boldsymbol{\gamma}$$

Δ is a $24 \times n$ matrix having for each row t, 1 in position j if $\delta_t^j = 1$, 0 otherwise

$$\begin{cases} \boldsymbol{Y}|\boldsymbol{F},\boldsymbol{\alpha},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{\Delta},\boldsymbol{\gamma},\sigma^2 \sim \mathcal{N}_n(\boldsymbol{Y}|\boldsymbol{\mu},\sigma^2) & \\ \sigma^2 \sim inverse-gamma(a_\sigma^2, b_\sigma^2) & \\ \alpha_i \overset{i.i.d.}{\sim} \mathcal{N}(a_0,\sigma_\alpha^2) & i = 0,..,p \\ \beta_j \overset{i.i.d.}{\sim} \mathcal{N}(b_0,\sigma_\beta^2) & j = 1,..,k, \\ & k = p_1 + p_2 + p_3 \\ \gamma_h|\sigma_\gamma \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\gamma) & h = 0,..,23 \\ \sigma_\gamma \sim inverse-gamma(a_\gamma, b_\gamma) & \end{cases}$$

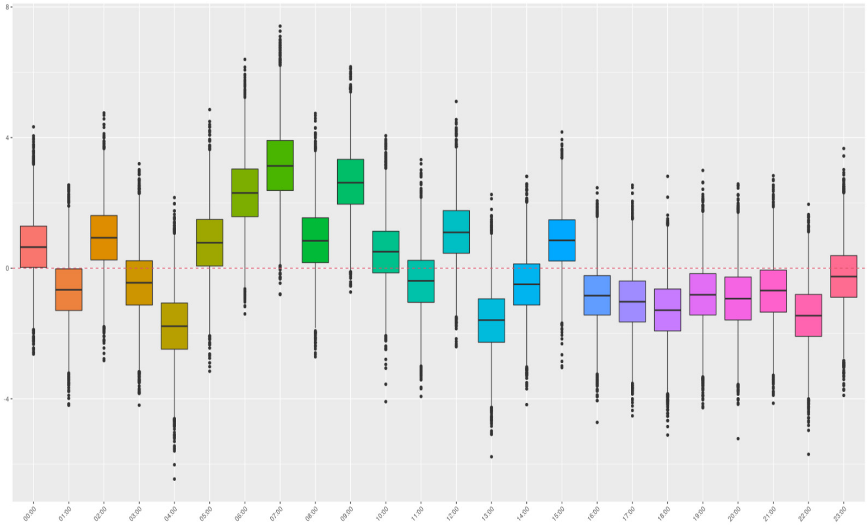


Figure: Boxplots of gamma

Daily SARX: ARX with daily seasonal effect

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\mu_t = \boldsymbol{f_t}^T \boldsymbol{\alpha} + \boldsymbol{x_t}^T \boldsymbol{\beta} + S_t$$

$$S_t = \sum_{i=0}^{T-1} \gamma_i \delta_t^i \qquad T = 7$$

$$\delta_t^i = \begin{cases} 1 & \text{if at time t is day i} \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \boldsymbol{Y} | \boldsymbol{F}, \boldsymbol{\alpha}, \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\Delta}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_n(\boldsymbol{y} | \boldsymbol{\mu}, \sigma^2)$$

priors as in the hourly case

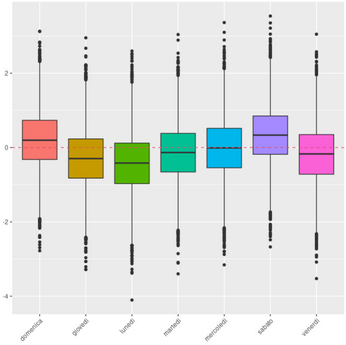


Figure: boxplots of gamma

Model comparison

We compared models on the base of both WAIC and BIC indexes

$$WAIC = \sum_{i=1}^{n} log \left[\frac{1}{M} \sum_{j=1}^{M} f_i(y_i|\theta^{(m)}) \right] - p_{WAIC}$$

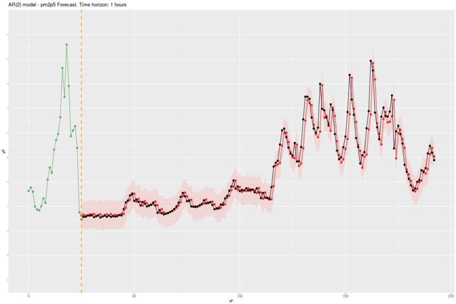
$BIC = 2 \log f(\boldsymbol{x}|\tilde{\theta}) - r \log n \quad \tilde{\theta}:$ MCMC estimate of the posterior mean of θ, r = dim(θ)

model	WAIC	BIC
AR(2)	8948.05	-18405.63
ARX(7)	8830.286	-18175.14
regularized ARX(7)	20150.2	-23111.88
SARX(2,2,1,3)	8820.29	-18122.96

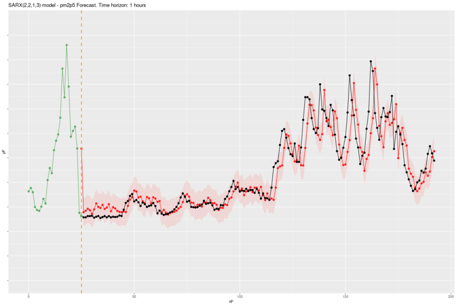
$\Rightarrow$ According to both WAIC and BIC, SARX(2,2,1,3) is the best model

AR vs SARX prediction

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$



$$y_t = \boldsymbol{f}_t^T \boldsymbol{\alpha} + \boldsymbol{x}_t^T \boldsymbol{\beta} + \sum_{i=0}^{23} \gamma_i \delta_t^i + \epsilon_t$$



Why so bad results?

- All the models seen so far are static, in the sense that parameters are kept fixed once the model is fitted.
- Static models are not able to adapt to **big changes in the time series dynamics** (peaks).
- Dynamic parameter models allow to better track the time series as new information is collected, in a sort of **feedback loop** style.

  Next models to investigate:

  - TVAR (Time Varing AR)
  - Exponential smoothing
  - Generic DLM models
  - Multivatiate analysis

Bibliography

- Time Series Modeling, Computation,and Inference. West Mike, Prado Raquel
- Bayesian Forecasting and Dynamic Models. Mike West, Jeff Harrison