Comparative Analysis of Hydrocarbon Concentrations Air Quality Monitoring Sites

Federico Berton Giachetti, Carlotta Francia, Sebastian Leiva, Filippo Marino, Alessandro Piana, Fiamma Ruscito

Politecnico di Milano

February 15, 2023

Outline

Dataset Presentation

Daily concentrations from February 2018 to September 2022 of $m = 16$ different pollutants in two sites:

- Milano, Via Pascal
- Schivenoglia (Mantova)

After visualization and data exploration we considered a time and scale transformation as well as a reduction in the number of pollutants.

Dataset Presentation

Weekly measures of the log concentration from February 2018 to September 2022 of $m = 10$ different pollutants in two sites.

Pollutants

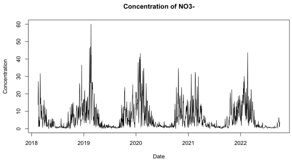
Cl^-	NO_3^-	SO_4^{2-}
NH_4^+	K^+	Mg^{2+}
Levoglucosano	Na^+	Ca^{2-}

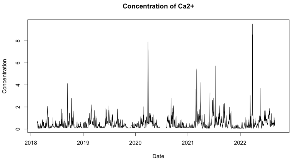
To enhace the performance of our models we added meteorological information of each site for each date:

Meteorological variables

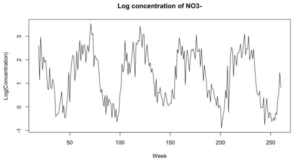
Temperature	Rain	Humidity	Clouds	Wind

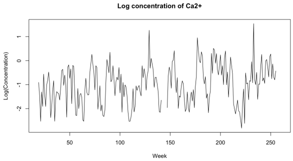
Pollutants before pre-processing





Pollutants after logarithmic transformation and weekly averaging





Main Objective

Main Goal

Comparative analysis of hydrocarbon concentrations observed at two air quality monitoring sites.

Specific goals

Model the multivariate time series through a suitable Bayesian tool and study the correlation structure of pollutants

Methods

1. Preliminary analysis (Bayesian Autoregression)
2. Multivariate Bayesian Structural Time Series
3. Multivariate Autoregressive State-space Models

Preliminary analysis

For an initial result we considered a basic Bayesian linear model with lagged time series as covariates to have an autoregressive model.

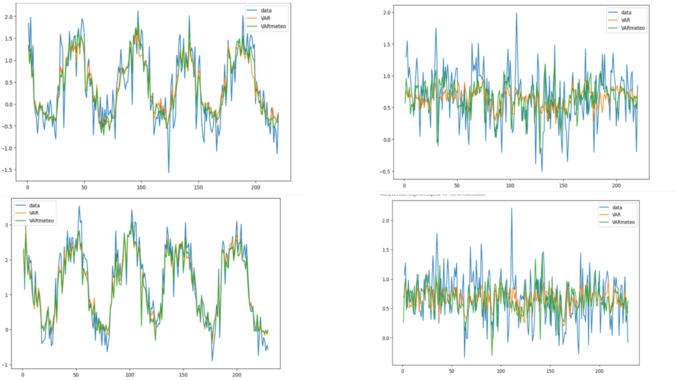
$$
\begin{aligned}
y_t^{(i)} &= \vec{\beta}^{(i)} \vec{Z}_t + \epsilon_t^{(i)}, & \vec{Z}_t &= (y_{t-1}^{(1)}, \dots, y_{t-1}^{(m)})^T \\
\epsilon_t^{(i)} &\overset{\text{iid}}{\sim} N(0, \sigma^2), & \sigma^2 &\sim InvGamma(1, 2) \\
\beta_j^{(i)} \mid \sigma_j^2 &\sim N(0, \sigma_j^2), & \sigma_j^2 &\overset{\text{iid}}{\sim} InvGamma(1, 2)
\end{aligned}
$$

We want to study the relation between pollutants using the resulting posterior estimates of $\beta_j^{(i)}$.

Preliminary analysis

The model was fitted once using only the lagged time series as predictors and once using the meteorological data as well.

Fit of the model



Preliminary analysis

Our analysis revealed a significant influence between the majority of the examined hydrocarbons and the fitted values.

-	Cl^-	NO_3^-	SO_4^{2-}	NA^+	NH_4^+	K^+	Mg^{2+}	Ca^{2-}	Levgl
Cl^-	0.3582	-0.0126	-0.0764	0.009	0.068	0.1519	0.1718	-0.023	0.2063
NO_3^-	0.0224	0.5917	-0.1555	-0.0219	0.0833	-0.1362	-0.3497	0.0149	0.2217
SO_4^{2-}	-0.0284	0.0498	0.1671	-0.01	0.0443	-0.1151	-0.1553	0.02	0.0754
NA^+	0.1167	-0.0888	-0.0253	0.12	0.0101	0.0784	0.29	0.049	0.0338
NH_4^+	0.0217	0.2356	-0.0711	-0.0322	0.1509	-0.1036	-0.2082	-0.031	0.2401
K^+	0.0091	-0.0533	-0.0907	-0.0239	0.0049	0.3507	0.1501	-0.0405	0.2776
Mg^{2+}	0.0307	-0.1814	-0.1452	-0.1183	0.0916	0.0591	0.809	0.0256	0.0737
Ca^{2-}	-0.0332	-0.1036	-0.0113	-0.0248	-0.0016	0.0286	0.1769	0.3852	0.0245
Levgl	0.0895	0.0044	-0.1768	0.0512	0.032	0.1045	0.0338	-0.1462	0.8127

Table 2: Schivenoglia β

This model, however, is simple and does not capture precisely the correlation structure of the pollutants.

MBSTS - Model description

The model decomposes the time series as a sum of five components as follows:

$$\vec{y_t} = \vec{\mu}_t + \vec{\tau}_t + \vec{\omega}_t + \vec{\xi}_t + \vec{\epsilon}_t$$

where

1. $\vec{y_t}$ is the vector of pollutants at time t
2. $\vec{\mu}_t$ is the trend component
3. $\vec{\tau}_t$ is the seasonal component
4. $\vec{\omega}_t$ is the cycle component
5. $\vec{\xi}_t$ is the regression component
6. $\vec{\epsilon}_t$ is the error term

MBSTS - Regression component

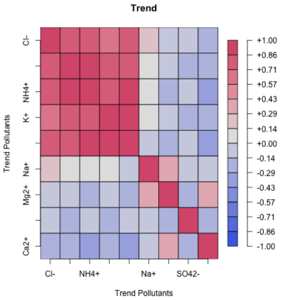
The model for the regression component is given by a spike and slap prior, allowing for automatic variance selection during training.

$$\vec{\xi} = (\xi_t^{(1)}, \dots, \xi_t^{(m)})^T \qquad \xi_t^{(i)} = \vec{\beta}_i^T x_t^{(i)}$$

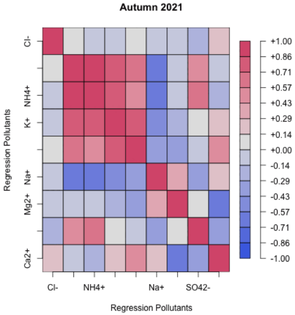
$$\gamma_i \sim \pi_i^{\gamma_i}(1 - \pi_i)^{1-\gamma_i} \qquad \beta \mid \gamma \sim N_K(b, (\kappa X^T X / n)^{-1})$$

MBSTS - Results

After training we separated the components by year and season to analyze their correlation structure.
For Schivenoglia we can see a clear similar behavior of the first 5 pollutants.



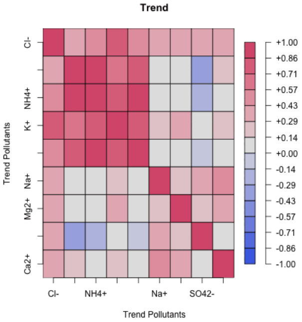
(a) Trend for all time



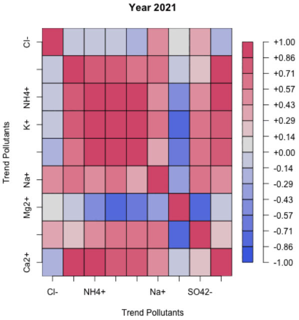
(b) Trend for autumn 2021

MBSTS - Results

For Pascal we can see a similar structure, however less strong as Schivenoglia. This could be explained by the urban condition versus the rural condition of both sites.



(a) Trend for all time



(b) Trend for 2021

MARSS - Model description

State Transition Equation:

$$\mathbf{x}_t = B_t \mathbf{x}_{t-1} + u_t + C_t c_t + G_t w_t$$

$$\mathbf{W}_t \sim \mathrm{MVN}(0, Q_t)$$

Measurement Equation:

$$\mathbf{y}_t = \mathbf{Z}_t \mathbf{x}_t + \mathbf{a}_t + \mathbf{D}_t \mathbf{d}_t + \mathbf{H}_t \mathbf{v}_t$$

$$\mathbf{V}_t \sim \mathrm{MVN}(0, \mathbf{R}_t)$$

Setting of initial values:

$$X_0 \sim \mathsf{MVN}(x_0, V_0)$$

MARSS - Bayesian Framework

State Transition Equation:

$$\mathbf{x}_t = \mathbf{x}_{t-1}$$

Prior for initial state:

$$x_0 \sim \mathcal{N}(0, 4)$$

Likelihood:

$$y[, t] \sim \text{Multi\_normal}(Z \cdot x[, t], R)$$

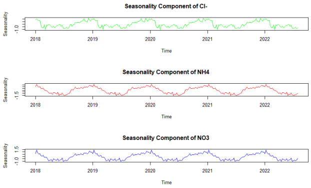
Log-likelihood:

$$\text{log\_lik}[t] = \text{multi\_normal\_lpdf}(y[, t] \mid Z \cdot x[, t],\ R);$$

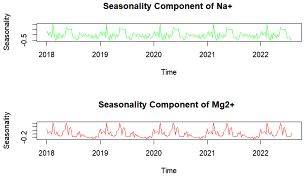
Prior:

$$R \sim \text{Inv-Wishart}(N + 1, \text{diag\_matrix}(N));$$

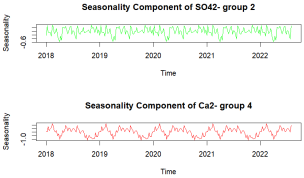
- Analysis of seasonality



(a) Group 1 Seasonality



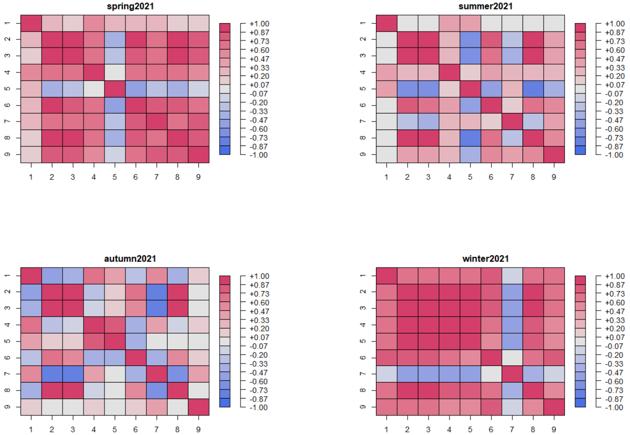
(b) Group 3 Seasonality



(c) Group 2 and 4 Seasonality

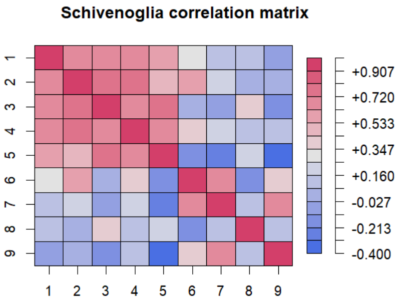
Correlation matrix in time

We can observe that the correlation structure (considering only seasonality) generally changes for each season, while maintaining at the same time some groups with similar behavior.



Total correlation Schivenoglia

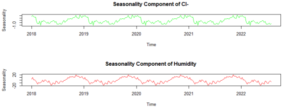
Now when considering seasonality and trend we observe that the correlation is more visible between the pollutants that were assumed to have similar behavior.



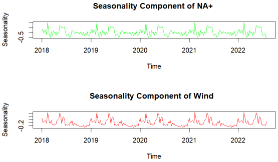
Adding meteorological data

New Likelihood:

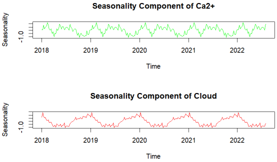
$$y[,t] \sim \text{Multi\_normal}(Z \cdot x[,t] + D \cdot d[,t]\mathbf{d}_t, R)$$



(a) Humidity

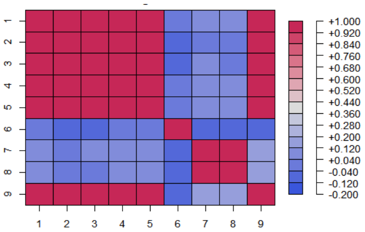


(b) Wind

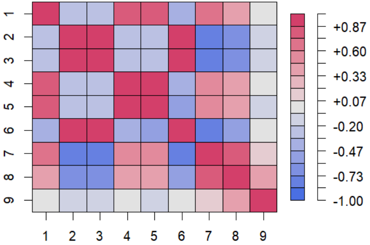


(c) Cloud

Adding meteorological data



(a) Correlation matrix with covariates



(b) Correlation matrix without covariates

We can discard the model with meteorological data as we have evidence to state that these meteorological variables influence too highly the correlation matrix.

Conclusion

1. Each model we have implemented shows different correlation patterns between the various time series.
2. It can be seen that the first five pollutants are well correlated in almost every model. In fact, these are the ones with a fairly clear seasonal pattern and they are often produced by the same combustion processes.
3. Regarding the site on Pascal Street, we can observe clear differences between the matrices generated by the two models.

References

Korobilis D., Pettenuzzo D.
Adaptive hierarchical priors for high-dimensional vector autoregressions
2019

Jinwen Qiu.,Ning Ning.
Package ‘mbsts’
2023

Elizabeth Eli Holmes,Eric J. Ward, Mark D. Scheuerell, Kellie Wills
Package ‘MARSS’
2023