POLITECNICO DI MILANO

Comparative Analysis of Hydrocarbon Concentrations in Milano and Schivenoglia (MN) Air Quality Monitoring Sites

Bayesian Statistics Project

Tutor: Prof. Matteo Gianella

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

Academic Year 2023/2024

Contents

Abstract

This study conducts a comparative analysis of hydrocarbon concentrations observed at two air quality monitoring sites, Milano and Schivenoglia (Mantova), spanning the period from 2018 to 2022. The dataset encompasses multiple time series of various hydrocarbons of PM10, providing valuable insights into the air quality dynamics of these regions. Milano's data cover the period from January 2013 to September 2022, while Schivenoglia's data range from February 2018 to October 2022. To ensure consistency in our analysis, we focus on the overlapping time frame from 2018 to 2022.

Abstract

Questo studio conduce un'analisi comparativa delle concentrazioni di idrocarburi osservate in due siti di monitoraggio della qualità dell'aria, Milano e Schivenoglia (Mantova), per un periodo compreso tra il 2018 e il 2022. Il set di dati comprende più serie temporali di vari idrocarburi del PM10, fornendo preziose indicazioni sulle dinamiche della qualità dell'aria in queste regioni. I dati di Milano coprono il periodo da gennaio 2013 a settembre 2022, mentre quelli di Schivenoglia vanno da febbraio 2018 a ottobre 2022. Per garantire la coerenza della nostra analisi, ci concentriamo sulla sovrapposizione temporale tra il 2018 e il 2022.

1 Introduction

Monitoring air quality is crucial for assessing environmental health and implementing effective pollution control measures. Hydrocarbons, emitted from various sources such as vehicular traffic, industrial activities and natural processes are significant contributors to air pollution. Understanding the temporal trends and spatial variations of hydrocarbon concentrations is essential for devising targeted mitigation strategies. In this study, we compare the hydrocarbon levels recorded at two distinct monitoring sites, Milano and Schivenoglia (Mantova), over the period 2018-2022. Our objective is to develop suitable temporal models for these time series and make a comparison across two locations to identify pollutants with similar temporal behavior and correlation among them.

2 Dataset presentation

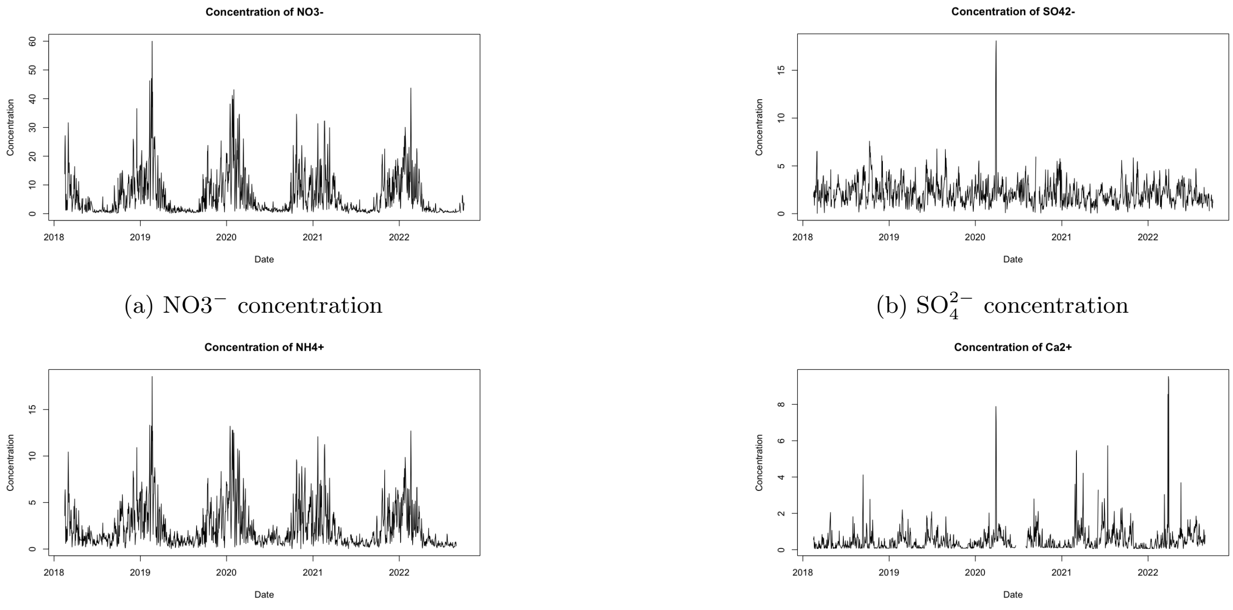
Air quality measurement, particularly regarding PM10, refers to the monitoring of particulate matter (PM) with a diameter of less than 10 micrometers. From the various components included in PM10, hydrocarbon pollutants due to combustion are a particularly important source of contamination. Monitoring PM10 levels and understanding its components are essential for assessing air quality, protecting public health, and implementing effective pollution control measures. The dataset represents a comprehensive collection of the concentration of 16 different pollutants. In addition, weather variables such as temperature, humidity, rain and wind have been considered [6].

2.1 Data Analysis and Preprocessing

The data set encompasses a comprehensive collection of data pertaining to air quality monitoring and we focused on the concentration of 9 distinct pollutants. The selection process involved eliminating 7 features deemed less relevant due to consistently low values or constant trends over almost the entire temporal span. The specific components in the dataset are the following:

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

Figure 1 presents four time series corresponding to various pollutants for the selected dates. It is evident that some of them exhibit clear seasonal behaviors (left), while others display less predictable fluctuations (right). Moreover, the original dataset lacks registrations for every day, often resulting in several consecutive days with missing values, as observed in the Ca^{2-} concentration in Figure 1 between 2020 and 2021.



(a) NO3^- concentration

(b) SO_4^{2-} concentration

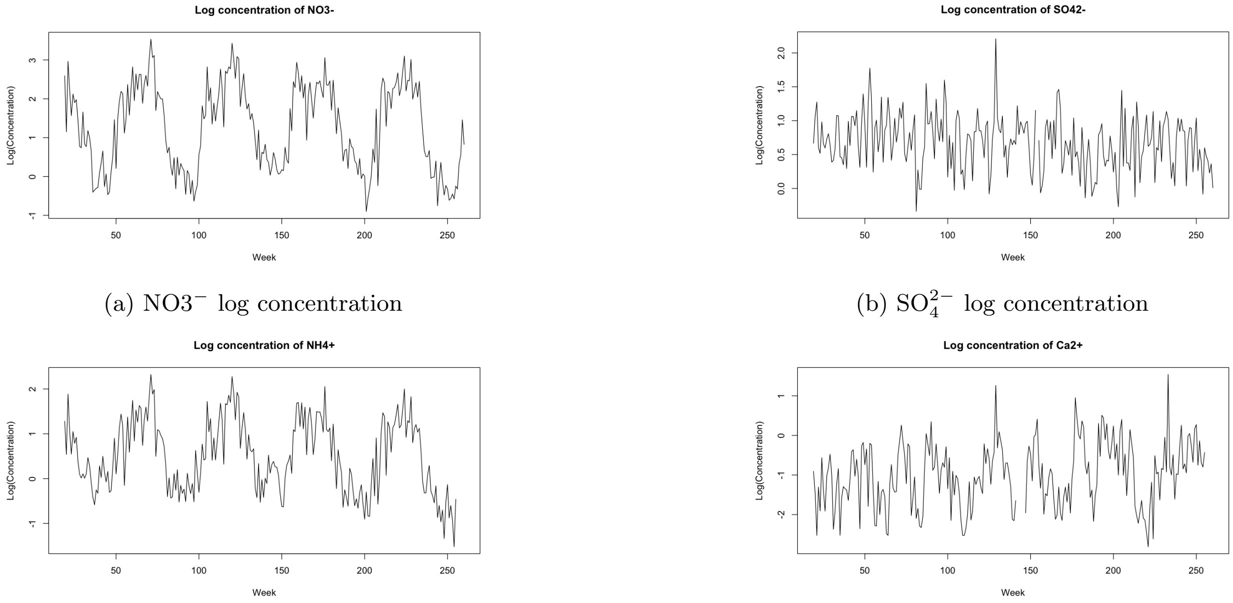
(c) NH_4^+ concentration

(d) Ca^{2-} concentration

Figure 1: Historical daily pollutant concentration (Schivenoglia)

To ensure a consistent analysis, we focus on the shared time frame between 15^{th} February 2018 and 5^{th} September 2022. We preprocess the data to handle missing values, outliers, and any inconsistencies that may affect the accuracy of our analysis. Missing observations are treated as "holes" in the hydrocarbon data. Left censoring, resulting from detection limit truncation, poses another challenge, which is addressed by substituting constant values with their threshold value.

To address these data issues a key preprocessing step involved aggregating the daily hydrocarbon concentration values into weekly averages. This consolidation reduced the granularity of the time series data, facilitating smoother modeling and analysis. Consequently, the original time series, comprising 1665 daily observations, was condensed to 238 weekly observations. This aggregation not only streamlined the dataset but also minimized the computational complexity of subsequent modeling techniques. The logarithmic transformation was then applied to the weekly hydrocarbon concentration values. This transformation offers several advantages, including maintaining positivity constraints inherent in pollutant concentrations and achieving symmetry in data distribution. Moreover, log-transformed data are often more suitable to Gaussian-based modeling frameworks, facilitating parameter estimation and interpretation.



(a) $NO3^-$ log concentration

(b) SO_4^{2-} log concentration

(c) NH_4^+ log concentration

(d) Ca^{2-} log concentration

Figure 2: Historical logarithm of pollutant concentration (Schivenoglia)

In our analysis, we employed three distinct approaches, each leveraging different modeling techniques to address the complexities of the hydrocarbon time series data and achieve our research objectives. In our analysis, we utilized two distinct packages: MBSTS (Multivariate Bayesian Structural Time Series) [5] and MARSS [2] (Multivariate Auto-Regressive State-Space Model), in addition to drawing insights from the research conducted by Korobilis-Pettenuzzo (2019) [3].

3 A preliminary analysis

Koriobilis-Pettenuzzo's paper proposes a simulation-free estimation algorithm for vector autoregressions (VARs) that allows fast approximate calculation of marginal parameter posterior distributions. Taking inspiration from the model in the paper, we ran an autoregressive model that takes into account every pollutant in order to extract the β coefficients. Obviously in the bayesian context we are estimating the marginal probability of the β, so we reported their mean value. This approach has been used to initially understand how the pollutants influence each other.

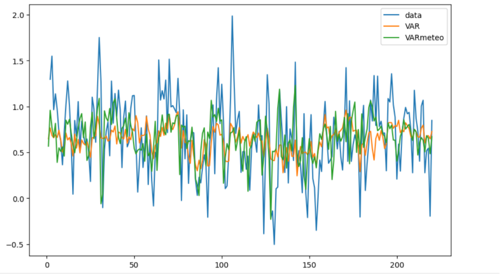
$$\begin{aligned} y_t^{(i)} &= \vec{\beta}^{(i)} \vec{Z}_t + \epsilon_t^{(i)}, & \vec{Z}_t &= (y_{t-1}^{(1)}, \ldots, y_{t-1}^{(m)})^T \\ \epsilon_t^{(i)} &\overset{\text{iid}}{\sim} N(0, \sigma^2), & \sigma^2 &\sim InvGamma(1,2) \\ \beta_j^{(i)} \mid \sigma_j^2 &\sim N(0, \sigma_j^2), & \sigma_j^2 &\overset{\text{iid}}{\sim}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

The hyperparameters of the gamma have been chosen to be non-informative for all the sigma's and m are the number of pollutants.

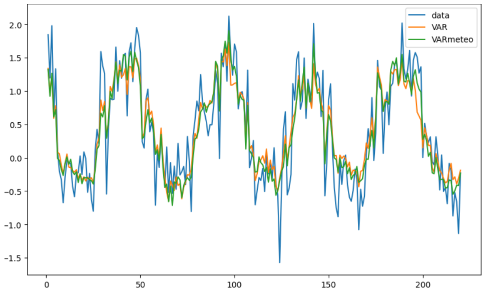
3.1 Application to BVAR estimation

[4] [1] We initiated the implementation of a simplistic models for both sites incorporating a single time-lagged component (AR(1)) to derive the coefficient β, thereby initiating a preliminary examination of correlation with a specific emphasis on the estimated β value.

$$\beta_i \sim \mathcal{N}(0, 1)$$



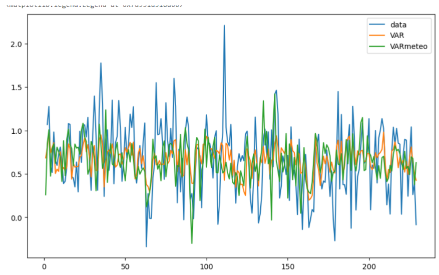
(a) Pascal SO_4^{2-} concentration



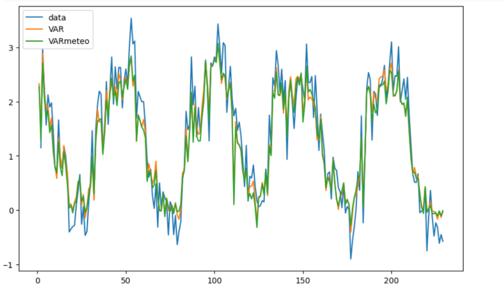
(b) Pascal NO_3^- concentration

-	Cl^-	NO_3^-	SO_4^{2-}	NA^+	NH_4^+	K^+	Mg^{2+}	Ca^{2-}	Levgl
Cl^-	0.5061	0.0181	-0.0197	-0.0038	0.0255	0.1678	0.0171	-0.0327	0.1299
NO_3^-	0.1254	0.5827	-0.1672	-0.0284	0.0513	-0.0276	-0.1555	-0.0109	0.1608
SO_4^{2-}	0.1157	-0.0024	0.216	-0.0375	-0.0697	-0.0255	-0.1767	0.0121	-0.0081
NA^+	0.0241	-0.0637	0.0747	0.3943	0.0405	0.0914	-0.0545	0.0324	0.0259
NH_4^+	0.1192	-0.1918	-0.0394	-0.0514	0.145	-0.0037	-0.1427	-0.019	0.1522
K^+	0.1713	-0.0539	-0.0642	-0.0229	-0.009	0.2449	0.0611	-0.0392	0.2184
Mg^{2+}	0.0191	-0.0773	-0.147	0.073	-0.0014	0.0902	0.3526	0.0251	0.0598
Ca^{2-}	0.0369	0.0053	0.0018	0.1127	-0.018	0.0424	0.0566	0.2560	0.0083
Levgl	0.0717	0.0576	-0.1976	-0.051	0.0618	0.1945	0.0143	-0.2041	0.7609

Table 1: Pascal β



(a) Schivenoglia SO_4^{2-} concentration



(b) Schivenoglia NO_3^- concentration

-	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

Based on the inspiration drawn from the Korobilis-Petenuzzo paper, we developed a linear model incorporating a designed matrix consisting of pollutants with a lag of 1 time unit. This approach allowed us to investigate the extent to which various pollutants influence the fitted hydrocarbons. Furthermore, we augmented the model by incorporating meteorological variables such as temperature, humidity, and wind, enabling us to assess their impact on pollutant concentrations.

Our analysis revealed a significant influence between the majority of the examined hydrocarbons and the fitted values. Additionally, we observed that the inclusion of exogenous variables contributed to better modeling the behavior of the time series data as a whole, espetially the ones that doesn't seem to have periodic patterns, such as SO42+.

4 MBSTS

The multivariate Bayesian structural time series (MBSTS) model (Qiu et al., 2018) deals with inference and prediction for multiple correlated time series, where one also has the choice of using a different candidate pool of contemporaneous predictors for each target series. The MBSTS model has wide applications and is ideal for feature selection, time series forecasting, nowcasting, inferring causal impact, and others. Particularly, the package mbsts developed by (Ning and Qiu, 2023) in R is a way to apply the theoretical framework of the model into our data set.

The MBSTS model is constructed as the sum of several components which represent the different elements of each time series. It is possible to select only a subset of the components to fit the model as well as to adjust their corresponding hyperparameters. The time series can be decomposed as,

$$\vec{y_t} = (y^{(1)}, \ldots y^{(m)})^T \tag{1}$$

$$\vec{y_t} = \vec{\mu_t} + \vec{\tau_t} + \vec{\omega_t} + \vec{\xi_t} + \vec{\epsilon_t}, \quad t = 1, 2, \ldots, n \tag{2}$$

where $\vec{y_t}, \vec{\mu_t}, \vec{\tau_t}, \vec{\omega_t}, \vec{\xi_t}, \vec{\epsilon_t}$ are m-dimensional vectors, representing target time series, a linear trend component, a seasonal component, a cyclical component, a regression component, and an observation error term. A brief explanation of each component is presented in the following sections.

4.1 Model description

4.1.1 Trend

A trend is the long-term growth of time series, and it can be further decomposed into two components: level and slope. Level represents the actual mean value of the trend and slope represents the tendency to grow or decline from the trend. The trend component $\vec{\mu}_t$ is generated as:

$$\begin{aligned} \vec{\mu}_{t+1} &= \vec{\mu}_t + \vec{\delta}_t + \vec{u}_t, & \vec{u}_t &\overset{\text{iid}}{\sim} \mathcal{N}_m(0, \Sigma_\mu) \\ \vec{\delta}_{t+1} &= \vec{D} + \vec{\rho}(\vec{\delta}_t - \vec{D}) + \vec{v}_t, & \vec{v}_t &\overset{\text{iid}}{\sim} \mathcal{N}_m(0, \Sigma_\delta) \end{aligned}$$

where $\vec{\delta}_t$ is a m-dimensional vector that represents the short-term slope, while $\vec{D}_t$ represents the long-term slope. The parameter $\vec{\rho}$ is a $m \times m$ diagonal matrix that incorporates the learning rates at which local trend is updated. Both Σ_μ and Σ_δ are $m \times m$ covariance matrices assumed to be diagonal, with entries distributed as Inverse-Gamma.

4.1.2 Seasonality

Seasonality is a characteristic of a time series in which the data has regular and predictable changes that recur every period. The seasonal component $\vec{\tau}_t = [\tau_t^1, ..., \tau_t^m]^T$ is generated as follows:

$$\vec{\tau}^{(i)}_{t+1} = -\sum_{k=0}^{S_i-2} \vec{\tau}^{(i)}_{t-k} + \vec{\omega}^i_t \tag{3}$$

$$\vec{\omega}_t = (\omega^1_t, , \omega^m_t)^T \overset{\text{iid}}{\sim} \mathcal{N}_m(0, \Sigma_\tau)$$

where S_i represents the number of seasons for y^i and $\tilde{\tau}_t$ s a m-dimensional vector denoting their joint contribution to the observed target time series $\tilde{y}_t$. The matrix $\Sigma_{tau}$is assumed to be diagonal, with entries distributed as Inverse-Gamma.

4.1.3 Cycle

The cyclical effect refers to regular or periodic fluctuations around the trend, revealing a succession of phases of expansion and contraction. In contrast to seasonality that is always of fixed and known periods, a cyclic pattern exists when data exhibits ups and downs that are not of fixed periods. The formulation of the cycle component is as follows:

$$\vec{\omega}_{t+1} = \widehat{\vec{\varrho} cos(\lambda)} \vec{\omega}_t + \widehat{\vec{\varrho} sin(\lambda)} \vec{\omega}^*_t + \vec{\kappa}_t \tag{4}$$

$$\vec{\omega}^*_{t+1} = -\widehat{\vec{\varrho} cos(\lambda)} \vec{\omega}_t + \widehat{\vec{\varrho} sin(\lambda)} \vec{\omega}^*_t + \vec{\kappa}^*_t \tag{5}$$

$$\vec{\kappa}_t \overset{\text{iid}}{\sim} \mathcal{N}_m(0, \Sigma_\omega) \qquad \vec{\kappa}^*_t \overset{\text{iid}}{\sim} \mathcal{N}_m(0, \Sigma_\omega)$$

Here, $\vec{\rho}$ is an $m \times m$-dimensional diagonal matrix as the damping factors for each target series such that $0 < \rho_{ii} < 1$. The term $\sin(\lambda)$ (resp. $\cos(\lambda)$) is a $m \times m$-dimensional diagonal matrix whose diagonal entries equal to $\sin(\lambda_{ii})$ (resp. $\cos(\lambda_{ii})$), where $\lambda_{ii} = 2\pi/q_i$ is the frequency with q_i being a period such that $0 < \lambda_{ii} < \pi$. In the cycle component, the covariance matrix for the error terms $\vec{\kappa}_t$ and $\vec{\kappa}*_t$ are assumed to be $m \times m$-dimensional diagonal matrices with entries distributed as Inverse-Gamma.

4.1.4 Regression component

For the regression component, each pollutant potentially has a unique set of regressors, thus we denote as $\vec{\beta}_i = (\beta_{i1}, \ldots, \beta_{i\kappa_i})^T$ the coefficients for the κ_i covariates $\vec{x}_t^{(i)} = (x_{i1}, \ldots, x_{i\kappa_i})^T$ to predict the i-th pollutant. The regression component is defined as,

$$\begin{aligned}\vec{\xi} &= (\xi_t^{(1)}, \ldots, \xi_t^{(m)})^T \\ \xi_t^{(i)} &= \vec{\beta}_i^T x_t^{(i)}\end{aligned}$$

The regression component considers a spike and slab prior, meaning that variable selection is done automatically during training. The latent variables $\gamma_i \in \{0, 1\}$ denote the inclusion or not of a predictor to the model. Specifically, the spike prior is written as:

$$\gamma_i \sim \pi_i^{\gamma_i}(1 - \pi_i)^{1-\gamma_i}$$

while the slab prior, conditional to γ, is

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

Finally, the error term has the following prior distribution:

$$\begin{aligned}&\vec{\epsilon}_t \sim N_m(0, \Sigma_\epsilon) \\ &\Sigma_\epsilon | \gamma \sim IW\left(v_0, (v_0 - m - 1)(1 - R^2)\Sigma_y\right)\end{aligned}$$

4.2 Simulation

The seasonal component is formulated as a sum of coefficients denoted by $\vec{\tau}_{t-k}^{(i)}$, with the specific number of coefficients represented by S_i. In our analysis, the pollutants exhibit a recurring seasonal pattern that repeats each year, resulting in $S_i = 53$ weeks. Notably, the computational demands associated with model training proved to be substantial, with the time required for fitting the seasonal coefficients exceeding 12 hours. Consequently, we made the decision to exclude this component from our model, opting instead to rely on the regression, trend, and cycle components to effectively capture and address this observed behavior.

In order to have a vector autoregressive structure we selectd as covariates for each time series the time shifted pollutants and the meteorological variables. This can be formulated as

$$y_t^{(i)} = \beta_1 y_{t-1}^{(1)} + \cdots + \beta_i y_{t-1}^{(i)} + \cdots + \beta_m y_{t-1}^{(m)} + \vec{\beta}' \vec{X}_t$$

where $\vec{X}_t$ is the vector of meteorological variables at time t.

After different simulations we realized during variable selection the coefficients
Hyperparameters for the cycle, trend and regression were adjusted through experimental iterations to assure that convergence is achieved and that at least one regressor is selected for each pollutant. The latter condition was satisfied through the tuning of the parameter prior inclusion probability, considered by us as 1 for each coefficient corresponding to a lag regressor predicting itself. The tuned parameters, considered the same for each pollutant, are:

μ	1
ρ	0
Seasonality	NULL
v_ρ	0.9
λ	(2*pi)/4

We trained our model using 2000 Markov Chains (Burn-in: 500) and assuming 12 degrees of freedom for in the InvWish as the prior for the covariance residuals

4.3 Results

By looking at the plots of the pollutants we were able to recognize five of them that present a clear common pattern. These are: $Cl^-, NO_3^-, NH_4^+, K^+,$ Levoglucosano. For this reason, we ordered the data set putting these time series on the first positions, as to see the resulting matrices in an easier way.

4.3.1 Schivenoglia

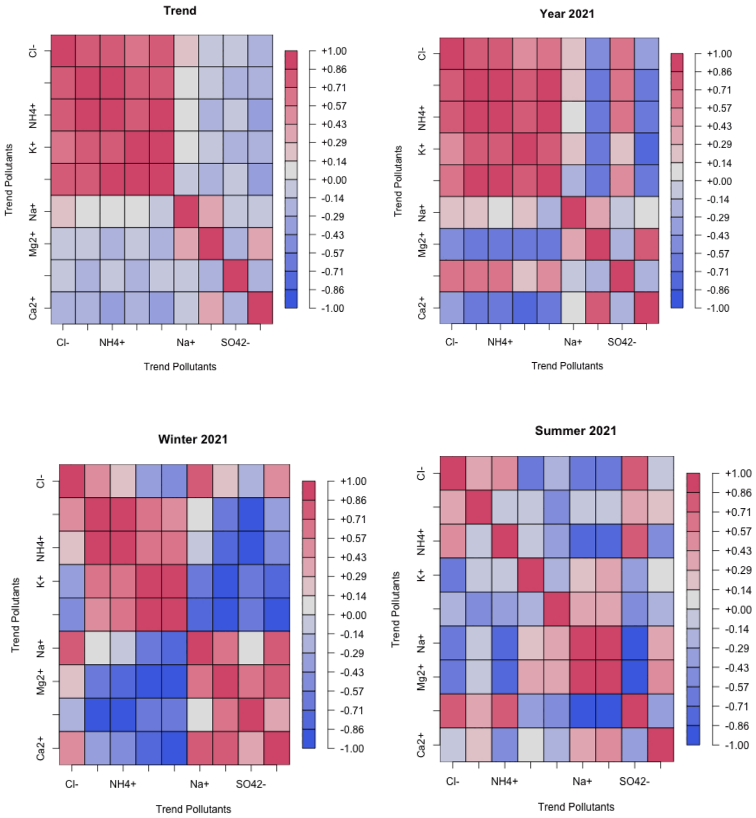
Regarding the correlation matrices of the trend components we can observe the following:

1. The first five pollutants seem to be positively correlated from one another in the whole period and it doesn't change when analyzed by year.
2. While analyzing the individual seasons the general correlation pattern is preserved in a lesser extent. This is due to the fact that each season is composed by only 13 weeks, meaning that small alterations could change the correlation structure.
3. Summer is the only season for which a pattern is not identifiable, this could be explained by the fact that during this period the concentration of pollutants is low overall, meaning that any change in the environment reates a significant effect.

For what concerns the regression component we found out that the same five pollutant, with the exception of Cl^-, are again positively correlated between them.

In the cycle component the correlation structure is not preserved because this component captures the remaining variability of the time series that fluctuates in not fixed periods.

These conclusions can be drawn by looking at the matrices below:



Trend

Year 2021

Winter 2021

Summer 2021

Figure 5: Correlation matrices of the trend component

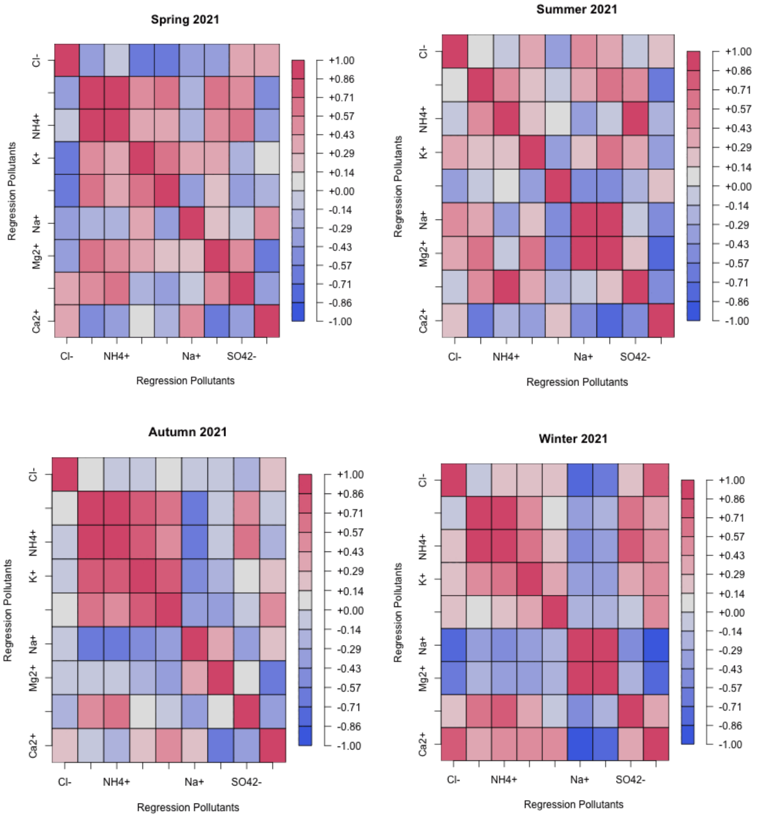
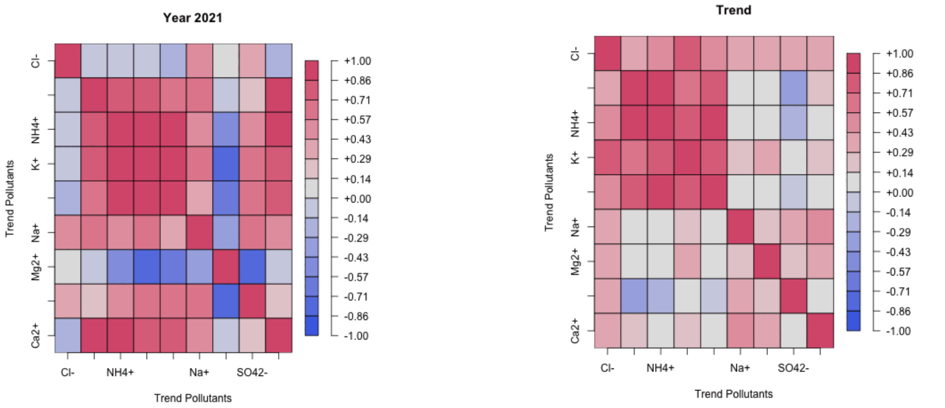


Figure 6: Regression component for each season of 2021

4.3.2 Pascal

For what concerns Pascal (Milan), we haven't found the same correlation pattern between pollutants that we found in Schivenoglia. For trend and regression, NO_3^-, NH_4^+, Levoglucosano have an high positive correlation along years, except for 2019. Conversely, like for Schivenoglia, the correlation for the cycle component has not a recurrent structure.



4.4 Conclusion

Analyzing the two sites, we have found two clusters for trend and regression correlation matrices. We can see that the correlation is higher in winter, when the concentrations of pollutants are elevated, while in summer the correlation is very distant from the identified pattern. This can be explained since the concentrations is very low, and they may depends on unpredictable factors. The two identified cluster, in Schivenoglia and in Milan (via Pascal), are different since we are comparing a rural place with a city, but we can see that three pollutants are recurrent: NO_3^-, NH_4^+ and Levoglucosano. We can say nothing about cycle component since it captures the remaining variability of the time series that fluctuates in not fixed periods, resulting erratic. We performed this analysis thanks to mBSTS package, that turned out to be useful to decompose our time series using a bayesian approach, but it resulted very expensive by a computational point of view, so we had to exclude seasonality. To solve this problem, we considered meteorological data, that showed a seasonal trend.

5 MARSS

MARSS is a package in R that provides Multivariate Autoregressive State-space Models for Analyzing Time-series Data and in particular it performs maximum-likelihood parameter estimation for constrained and unconstrained linear multivariate autoregressive state-space models.

5.1 MODEL

The default MARSS model form is ”marxss”, which is Multivariate Auto-Regressive eXogenous inputs State-Space model:

State Transition Equation:

$$\mathbf{x}_t = B_t\mathbf{x}_{t-1} + u_t + C_t c_t + G_t w_t \tag{6}$$

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

Measurement Equation:

$$\mathbf{y}_t = \mathbf{Z}_t\mathbf{x}_t + \mathbf{a}_t + \mathbf{D}_t\mathbf{d}_t + \mathbf{H}_t\mathbf{v}_t \tag{7}$$

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

Equation (6) describes the evolution of the state variable $\mathbf{x}_t$ (m x 1 vector) over time, these variables are latent states as they represent unobserved (hidden) variables that evolve over time. In the context of pollutant concentrations, hidden states could represent, for example, the presence of an unmeasured pollutant source, the level of pollutant dispersion in the air, or other dynamics not directly observable.
B is a m x m matrix representing the autoregressive coefficients in a multivariate autoregressive state-space model. u is a m x 1 vector r expressing the state noise (the mean level/trend). C is m x c matrix that describes the relationship between the latent states and a c x 1 vector of covariates that I want to include in my model. G is a m x m matrix of the loading weights for the state noise. w is a m x 1 vector that represents the state noise.In conclusion $\mathbf{x}_t$ depends on the previous state $\mathbf{x}_{t-1}$, an input u_t, a deterministic component

C_tc_t, and a random noise term G_tw_t, where w_t is normally distributed with mean 0 and covariance matrix Q_t.

Equation (7) represents the observed measurements $\mathbf{y}_t$ (n x 1 vector), which depends on the current state $\mathbf{x}_t$, a deterministic component $\mathbf{a}_t$, another component related again to possible extra variables that I might want to include in my model $\mathbf{D}_t\mathbf{d}_t$, and a random noise term $\mathbf{H}_t\mathbf{v}_t$, where $\mathbf{v}_t$ is normally distributed with mean 0 and covariance matrix $\mathbf{R}_t$. Components: Z is n x n matrix design matrix representing the relationship between observed data and latent states.
Setting of initial values:

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

5.2 Bayesian MARSS Model

We implement MARSS model in Stan for a Bayesian approach instead of the frequentist approach of MARSS original package.
We started by implementing a simple model without any extra metereogical variables.

5.2.1 The model

State Transition Equation (simple AR(1)):

$$\mathbf{x}_t = \mathbf{x}_{t-1} \tag{8}$$

Where $\mathbf{x}$ will be a M x TT matrix (TT: time instants , M:number of latent states).

Prior for initial states:

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

We choose 4 for the variance of x0 to respect the range of data, since they take values between -4 and 4, using a non-informative distribution.

Likelihood:

$$y[,t] \sim \mathrm{Multi\_normal}(Z \cdot x[,t], R) \tag{9}$$

y is a N x TT matrix (N : number of pollutants) and Z is a matrix N x M that has 1 if the pollutant belongs to the group.

Log-likelihood:

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

Our main objective is to make inference on the variance-covariance matrix R that represents the measurement noise covariance matrix.
First of all we decided a prior for R and we opted for an Inverse-Wishart and we chose hyperparameters that reflect minimal prior information.

For this reason we selected as prior for R:

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

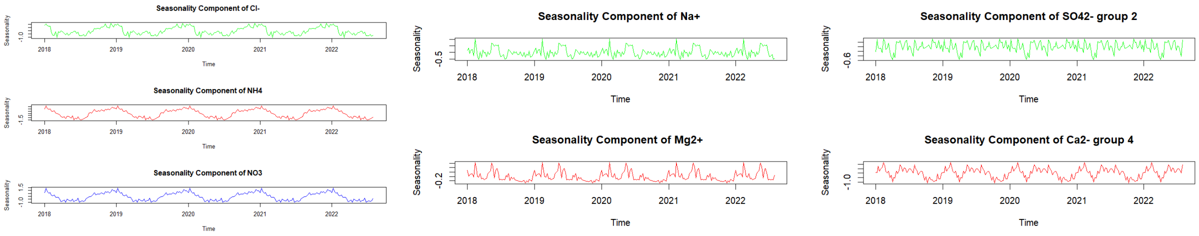
Where N is the dimensionality of R (number of pollutants).

5.3 Analysis of the dataset Schivenoglia for the model

When dealing with 9 pollutants with different behaviour, using multiple hidden states could be advantageous for modelling the complexity of the system. Each hidden state can be interpreted as a representing subset of pollutants with similar dynamics. We tried to understand these dynamics doing some qualitative analysis on the seasonality of each pollutants, in order to understand in how many groups they could be divided.

5.3.1 Analysis of seasonality

In order to understand the possible common dynamics of our pollutants we studied the seasonality of each pollutant by decomposing our time series in seasonality, trend and by focusing on the seasonality. We noticed that there could be spotted four different behaviours from a seasonal point of view and for this reason we divided our pollutants in four groups.



(a) Group 1 Seasonality

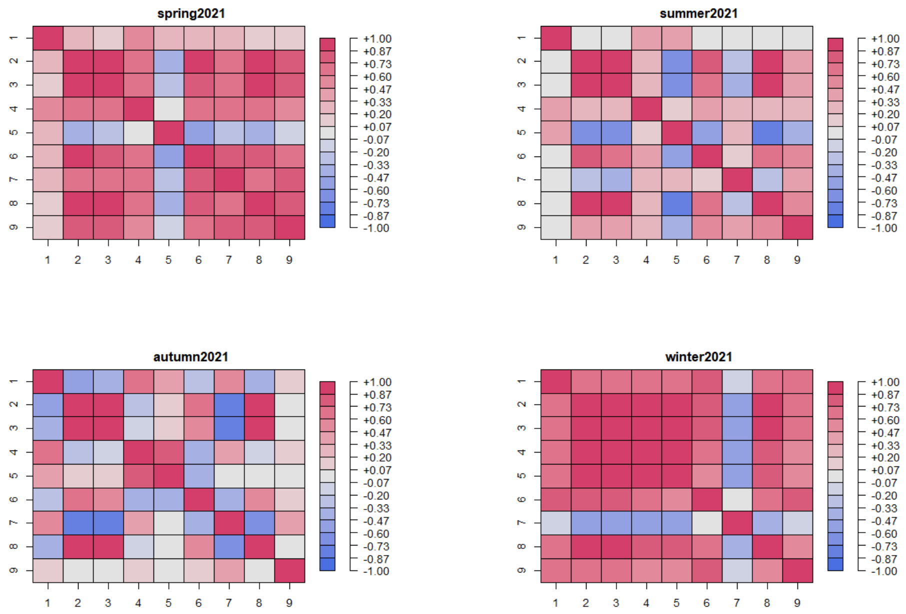
(b) Group 3 Seasonality

(c) Group 2 and 4 Seasonality

From this analysis we can conclude that we need four different latent states. We can now apply our Bayesian model to this dataset keeping in mind the above conclusions.

5.3.2 Correlation matrix in time

In order to state if the correlation between pollutants changes over time, we divided for each year our Schivenoglia dataset based on the different seasons : Winter, Spring, Summer and Autumn on which we applied our Bayesian model. Obtaining the posterior estimation of the matrix R for each iteration we computed the mean obtaining a single R. From this we extracted the correlation matrix. Below are reported the correlation matrices for each season in 2021.



We can clearly conclude that the correlation changes with the seasons, for example winter and autumn are very different between each other.

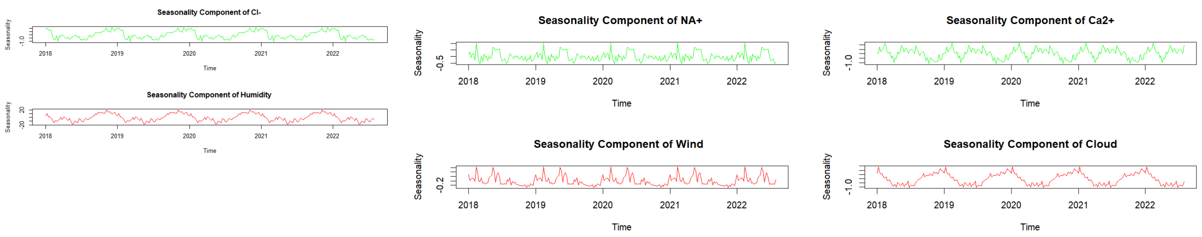
5.3.3 Adding meteorological data

We now want to add some extra information using some variables regarding meteorological events: humidity, clouds and wind. In order to understand where to add them we made a seasonal analysis on this new variables comparing their behaviour with the already existing pollutants.

Our model will change mainly in the Likelihood that becomes:

$$y[,t] \sim \text{Multi\_normal}(Z \cdot x[,t] + D \cdot d[,t]\mathbf{d}_t, R) \tag{10}$$

Where D will be a matrix N x c (c : number of meteorological variables) and d is a c x TT matrix with the values of these meteorological data.



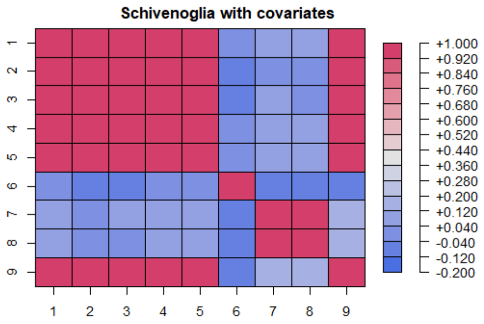
(a) Humidity

(b) Wind

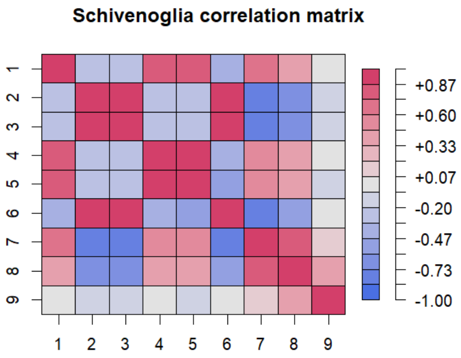
(c) Cloud

From this plots we are able to qualitatively state that some of these new variables can be considered part of the different groups we have found.

Here instead is reported the correlation matrix obtained using all Schivenoglia dataset with weekly means.



(a) Correlation matrix with covariates



(b) Correlation matrix without covariates

5.3.4 Comparison between models with and without meteorological variables

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

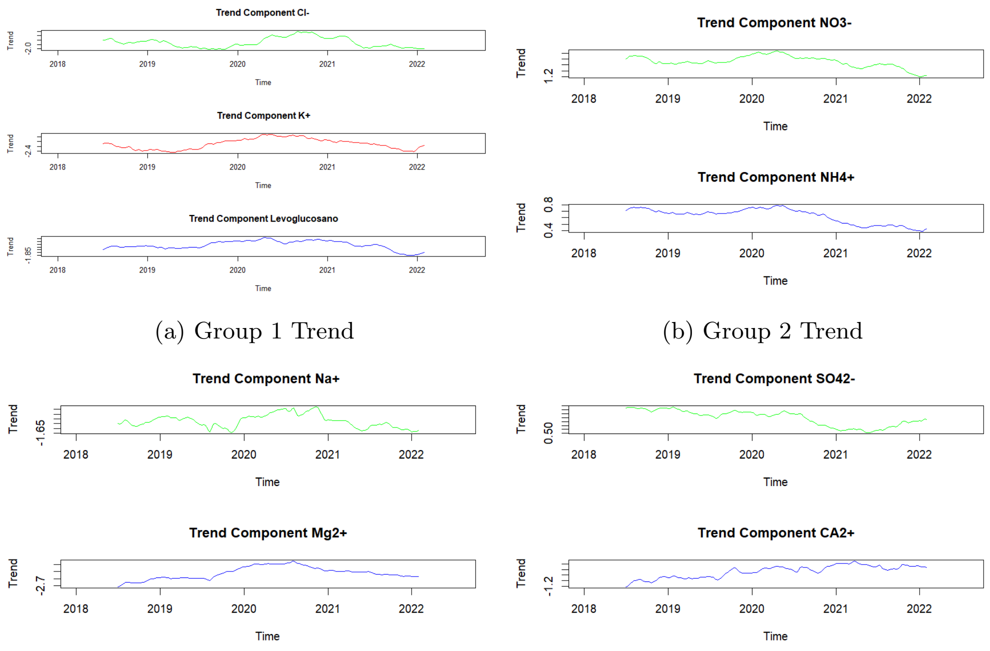
.

5.3.5 Conclusions from these matrices

It is clear that the pollutants that we have selected to belong in the same group have actually a different trend based on the values of the correlation matrices: if the trend is different the correlation is negative, otherwise is positive. For example, if we observe the matrix obtained with weekly means, we can see that NO3 and NH4 are negatively correlated with respect to the other variables belonging to our first group. For this reason we would like to see if by changing the groups and dividing our pollutants in groups not only based on seasonality but also on trend we can obtain a better result.

5.3.6 Analysis of the trend for Schivenoglia dataset

From these graphs we are able to identify five different groups based on both seasonality and trend. So in this case we will have five different latent variables.



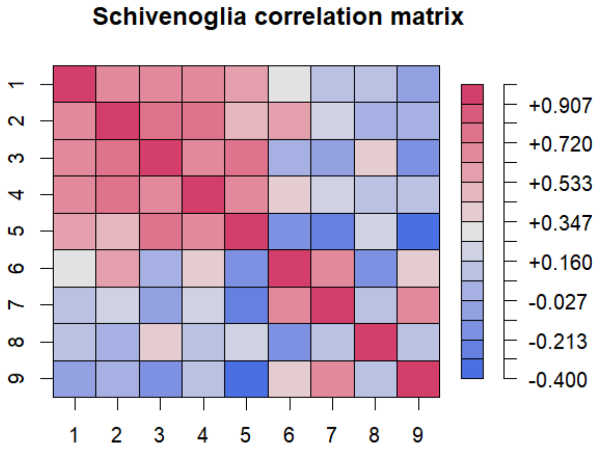
(a) Group 1 Trend

(b) Group 2 Trend

(c) Group 4 Trend

(d) Group 3-5 Trend

5.3.7 Correlation matrix (trend and seasonality groups)



5.3.8 Comparison of the two models

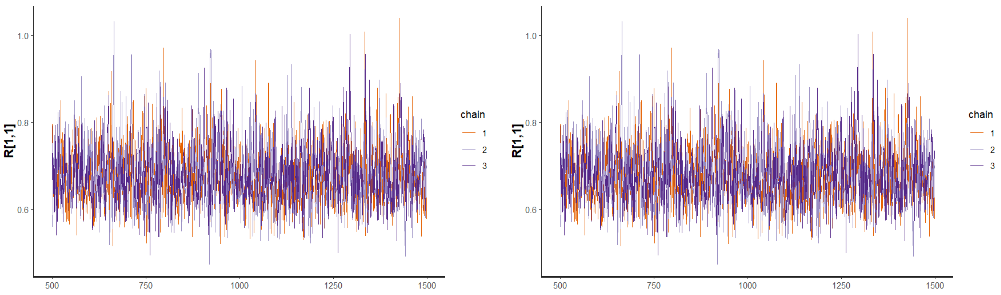
In order to compare the two models obtained we will use the Bayes factor using the log-likelihood.

log_likelihood_difference = log_likelihood_model1 - log_likelihood_model2

Our Bayes factor is bigger than 100 so we have evidence to state that the model considering also the trend is surely better than the one considering only seasonal component.

5.3.9 Convergence Analysis

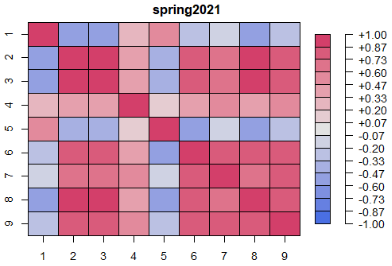
Another important check that needs to be done in order to state if the results that we have obtained are reliable is compute the trace plots.

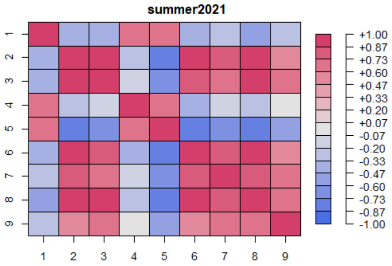


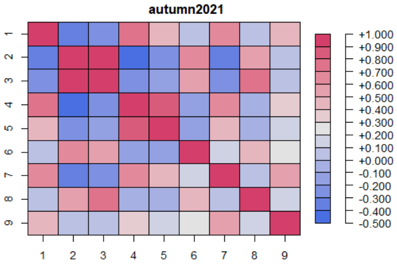
(a) Traceplot of R[1,1] considering seasonality (b) Traceplot of R[1,1] considering trend

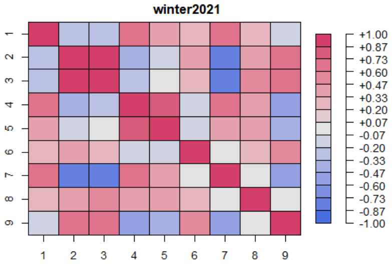
We also obtain for each model an Rhat that is below 1.1 for each member of R.

5.4 Comparison with the dataset Pascal









Comparing Pascal with Schivenoglia for the year 2021 we can see some differences in the single matrices in winter and summer, while they have similar correlation matrices for spring and autumn. This difference is probably due to the fact that in winter in Milan No3- and NH4+

have similar emission sources as traffic emissions that characterize less Schivenoglia emission sources, while K+ and Levuglucosano are more related to biomass burning. In summer the two groups highlighted by our correlation matrix have probably different dominant sources.

6 Final Conclusions

Each model we have implemented shows different correlation patterns between the various time series. In this section, the most sensible thing to make is a comparison between the correlation matrices generated by the trend of the fitted models over the entire period, considering the same site. The most comparable are the Schivenoglia's matrix in section 5.3.7 (concluding model) compared with that in section 4.3.1 of the trend, where we can see a fair similarity.
Therefore, 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. Regarding the site on Pascal Street, we can observe clear differences between the matrices generated by the two models, but it should be noted that the trend grouping was not considered in the MARSS model. Therefore, in this case, it doesn't make much sense to compare the two models.
Furthermore, we can add that both models are sensitive to the differences between the two sites, dictated by the different urban contexts, although they also maintain some similarities.

References

[1] Monica Billio, Roberto Casarin, and Luca Rossini. "Bayesian nonparametric sparse VAR models". In: *Journal of Applied Econometrics* 31.4 (2016), pp. 637–649.

[2] Elizabeth E. Holmes, Eric J. Ward, and Kellie Wills. "MARSS: Multivariate Autoregressive State-space Models for Analyzing Time-series Data". In: *Journal of Statistical Software* 50.1 (2012), pp. 1–19.

[3] Dimitris Korobilis. "VAR Forecasting Using Bayesian Variable Selection". In: *Journal of Applied Econometrics* 28.2 (2013), pp. 204–230.

[4] Nikolas Kuschnig and Lukas Vashold. "BVAR: Bayesian Vector Autoregressions with Hierarchical Prior Selection in R". In: *Journal of Statistical Software* 93.7 (2020), pp. 1–37.

[5] Ning Ning and Jinwen Qiu. "The mbsts package: Multivariate Bayesian Structural Time Series Models in R". In: *Journal of Statistical Software* 91.4 (2019), pp. 1–33.

[6] open-meteo.com. *Historical Weather API*. URL: `https://open-meteo.com/`.