POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Clustering PM_{10} and other cute stuff

Project Report of Bayesian Statistics - Mathematical Engineering

Giulia Mezzadri, Ettore Modina, Oswaldo Jesus Morales Lopez, Federico Angelo Mor, Abylaikhan Orynbassar, Federica Rena

Professor:
Alessandra Guglielmi

Tutors:
Michela Frigeri
Alessandro Carminati

Academic year:
2023-2024

Abstract[a] In this project we undertake a comprehensive clustering analysis of PM10 levels in the Lombardy region (Italy), employing four different Bayesian models to account for the complex nature of our data, which comprise spatio-temporal measurements of PM10, together with many other environmental variables, collected from various monitoring stations displaced across the entire region. The main objective was to leverage on covariates, station locations and time trends to cluster weekly PM10 data over a one-year period. Our analysis revealed distinct clusters for each time step, with a noteworthy influence of morphological terrain characteristics (e.g. altitude, wind speed) and anthropological factors (e.g. agricultural activities, vehicles and road transports, etc.).
The analysis was executed concurrently across a set of four models, to study the different interactions and combinations of spatio-temporal aspects and covariates information. Despite some variations among the models, that however highlighted peculiar patterns and characteristics which each model independently dwelt on, a unanimous consensus emerged regarding the overall division between the stations. This study contributes valuable insights into the delicate interaction of spatial, temporal, and covariate variables in shaping PM10 levels, providing a robust foundation for understanding the clustering dynamics in the Lombardy region.

[a] See `https://github.com/federicomor/progetto-bayesian` for all the project codes and `https://federicomor.github.io/assets/figures/visualize.html` for the visualization page.

1. Introduction

[Com17] Particulate matter with a diameter of 10 micrometers or less, known as PM10, comprises small airborne particles sourced from various origins, posing potential health risks upon inhalation due to their ability to deeply penetrate the respiratory system. The meticulous monitoring of PM10 levels is imperative for comprehensive air quality assessment and the safeguarding of public health.
This paper embarks on a project with the overarching aim of identifying both natural and anthropogenic factors contributing to elevated PM10 levels. Employing a clustering analysis, our objective is to delineate distinct regions within Lombardy, unveiling discernible patterns influencing particulate levels.
Drawing upon data from the Agrimonia project, which encompasses diverse measurements, our focus centers on weekly averages across a one-year timeframe. Our analytical approach involves the utilization of various models, including DRPM and SPPM, alongside additional models for covariate selection.

In subsequent sections, we delve into the dataset cleaning process, present individual analyses for each model, and expound on our interpretation of results. Visualization plays a pivotal role in our exploration, with a particular emphasis on manual (visual) interpretation to extract nuanced insights.
It is essential to acknowledge the inherent limitations of our approach. Notably, our lack of technical expertise in the phenomenon necessitated a wholly data-driven analysis, underscoring the importance of contextualizing our findings within this parameter.

2. The Dataset

The Agrimonia [FRFM+23] dataset spans from January 1, 2016, to December 31, 2021, recording observations from a network of 141 stations. The dataset encompasses measurements from five distinct covariate groups: air quality (AQ), weather and climate (WE), pollutants' emissions (EM), livestock (LI), and land and soil characteristics (LA). In total, there are 38 covariates, with our focal variable falling under AQ: AQpm10.
To facilitate our analysis, the data underwent a weekly averaging process. Additionally, a logarithmic transformation was applied to the "AQPM10'" variable to achieve a normal distribution, enhancing the suitability of the data for subsequent statistical analyses. This comprehensive dataset forms the foundation for our investigation into the factors influencing PM10 levels in the Lombardy region.

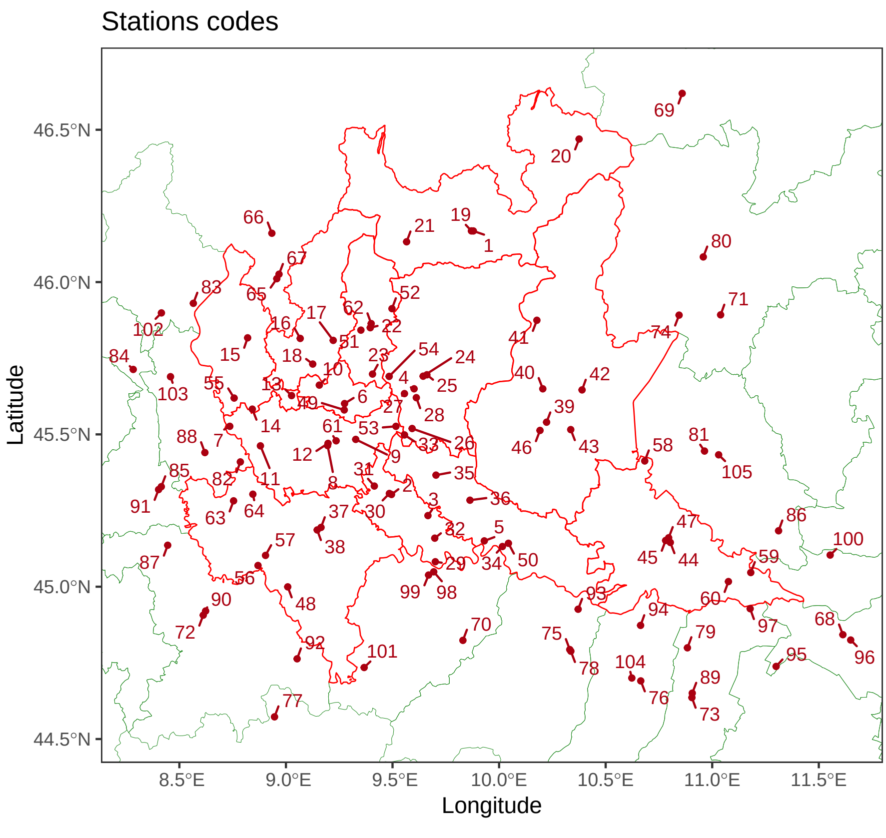

Figure 1: Map of the 105 selected stations after the data preprocessing. For now im adding a lot of images but if they dont fit then remove it. Also here im trying a more compact caption, let me know which one do you like more.

2.1. ecc

3. Data inspection and preprocessing

3.1. Missing data

The analysis and treatment of missing data proved to be necessary to reduce the dataset and at the same time allowed us to gain a better understanding of which information was at our disposal.
The first selection was performed visually through a heatmap colouring the columns differently based on the number of missing values to see a general pattern (see Figure 2). This allowed us to perform a first elimination of those covariates presenting too high a number of NAs, specifically those referring to air quality. Then we divided the dataset by station and plotted the missing data on each, realising some of them didn't collect data about pm10 and so we ended up removing 36 stations and only keeping 105, still a representative set. We also noticed some temporal patterns on NAs, for example on some covariates the missing data were concentrated on the last year or so.
At that point we focused on the years with less missing values to select the one to perform clustering on: the best choice according to this criterion would have been the year 2020, but we excluded it since because of the covid-19 pandemic it could be considered a particular isolate case, so in the end we selected the year 2018. We came back at the station-wise analysis and performed the final selection, looking also at graphs considering combinations of variables.
We ended up with 39 covariates of which 4 still presented missing values on three stations. Being just a few problematic cases regarding stations outside the lombardy area of interest, we didn't deem necessary to remove them completely but filled in the missing values with an average of the values of the three closest stations on the map.

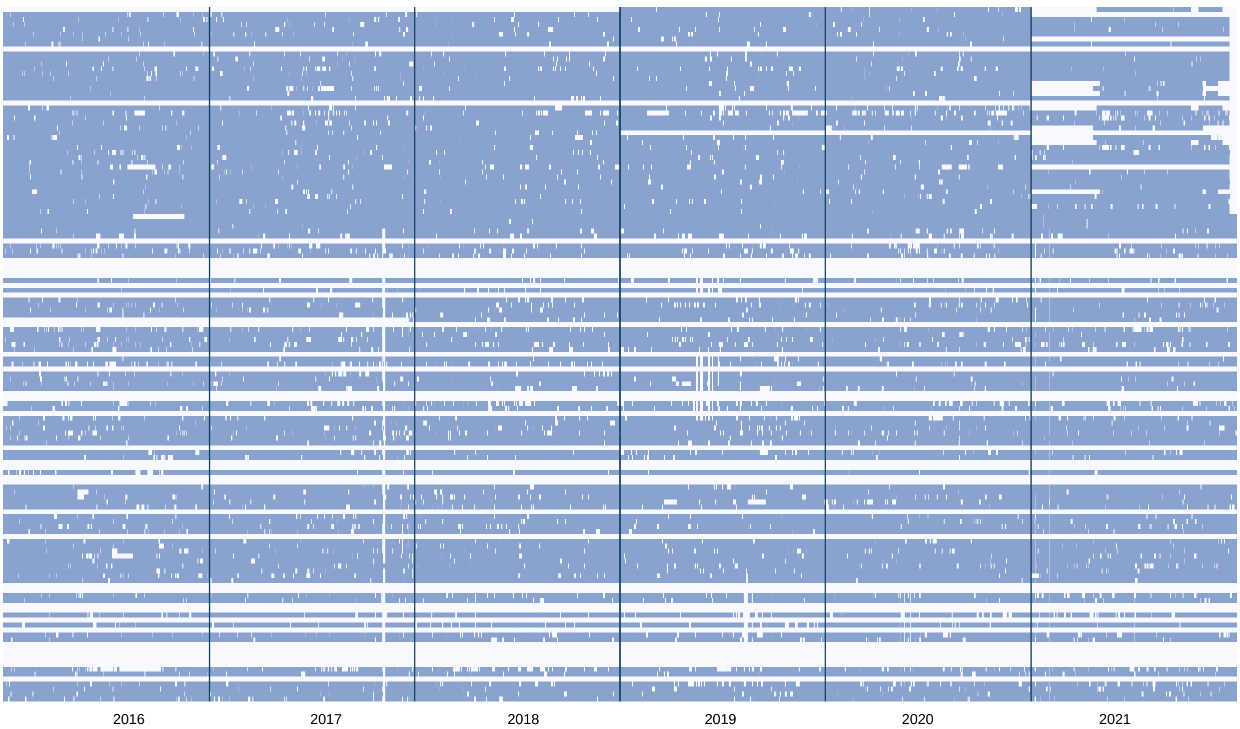

Figure 2: Heatmap of the missing data (in white) of the PM_{10} values in the available dataset. On the rows there are all the original 141 stations, while on the columns the 2192 days composing the six years.

4. Models

4.1. sPPM model

4.2. DRPM model

The second model we focused on, outside of the `PPMSuite` package, is the Dependent Modeling of Temporal Sequences of Random Partitins (DRPM), developed in [PQD22]. The main objective of the authors was to define a spatio-temporal model capable of performing "smooth" clusterings, i.e. a framework which would favour a

gentle evolution in time of the units allocations, rather than abrupt (and therefore less interpretable) changes in them. This result was clearly reached also in our analysis, as we will describe more precisely in section 5.2, where we experienced a more regular trend in the clusters definition for the DRPM model with respect to the other ones.
The model used in our study, fully detailed in (1), starts with assuming that there is a first order dependence between clusters, meaning that the conditional distribution of ρ_t given $\rho_{t-1}, \dots, \rho_1$ will just depend on ρ_{t-1}. Then the idea is using a temporal dependence parameter $\alpha \in 0, 1$ to controls the level of flexibility in the cluster allocation variables: the higher is α, the higher is the tendency of units to remain in their current cluster, so clusters ρ_{t+1} will be similar to ρ_t. Conversely, when α approaches 0, we would get more independent clusters. In this way the clusters allocations variables $\boldsymbol{c}_t$ will follow a temporal Random Partition Model (the entry tRPM in the model formulation) driven by the sequence of α_ts and the Dirichlet dispersion parameter M.

$$\begin{aligned}
Y_{it}|Y_{it-1}, \boldsymbol{\mu}_t^\star, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{\eta}, \boldsymbol{c}_t &\overset{\text{ind}}{\sim} \mathcal{N}(\mu_{c_{it}t}^\star + \eta_{1i}Y_{it-1}, \sigma_{c_{it}t}^{2\star}(1-\eta_{1i}^2) \quad i = 1, \dots, n \quad \text{and} \quad t = 2, \dots, T \\
Y_{i1} &\overset{\text{ind}}{\sim} \mathcal{N}(\mu_{c_{i1}1}^\star, \sigma_{c_{i1}1}^{2\star}) \\
\xi_i = \text{Logit}(\tfrac{1}{2}(\eta_{1i}+1) &\overset{\text{ind}}{\sim} \text{Laplace}(a, b) \\
(\mu_{jt}^\star, \sigma_{jt}^\star) &\overset{\text{ind}}{\sim} \mathcal{N}(\theta_t, \tau_t^2) \times \mathcal{U}(0, A_\sigma) \\
\theta_t|\theta_{t-1} &\overset{\text{ind}}{\sim} \mathcal{N}((1-\phi_1)\phi_0 + \phi_1\theta_{t-1}, \lambda^2(1-\phi_1^2)) \\
(\theta_1, \tau_t) &\sim \mathcal{N}(\phi_0, \lambda^2) \times \mathcal{U}(0, A_\tau) \\
(\phi_0, \phi_1, \lambda) &\sim \mathcal{N}(m_0, s_0^2) \times \mathcal{U}(-1, 1) \times \mathcal{U}(0, A_\lambda) \\
\{\boldsymbol{c_t}, \dots, \boldsymbol{c_T}\} &\sim \text{tRPM}(\boldsymbol{\alpha}, M) \quad \text{with} \quad \alpha_t \overset{\text{iid}}{\sim} \text{Beta}(a_\alpha, b_\alpha)
\end{aligned} \tag{1}$$

Then they model the target variable Y_{it} with a Normal distribution driven by cluster specific parameters, which are $\boldsymbol{\mu}_t^\star$ and $\boldsymbol{\sigma}_t^{2\star}$. For that mean they actually have an interesting modeling: there is an autoregressive part both in the observations and the parameter levels. Indeed, the Y_{it} depend on Y_{it-1} through the parameter η_{i1}, while the $\mu_{jt}^\star$ depend on another parameter θ_t which incorporates the autoregressive structure.
This deepening level allowed us to test different combinations of models and to select the best one which would suit our data. Through their developed package `drpm` on `R`, we fitted 8 different models based on the binary choice available for those three key parameters: the α could be set constant or varying in time, while the η_{1i} and ϕ_1 to be present (and therefore introducing the autoregressive design) or not. Then we selected the best model according to the LPML and WAIC metrics and performed the final fit.

model name				LPML	WAIC
model	η:`No`	ϕ:`Yes`	α:`Yes`	**1077.64**	**-2366.48**
model	η:`No`	ϕ:`No`	α:`Yes`	950.17	-2117.36
model	η:`Yes`	ϕ:`No`	α:`No`	724.34	-1474.02
model	η:`No`	ϕ:`Yes`	α:`No`	693.04	-1458.70
model	η:`Yes`	ϕ:`No`	α:`Yes`	605.32	-1287.13
model	η:`No`	ϕ:`No`	α:`No`	504.41	-1129.83
model	η:`Yes`	ϕ:`Yes`	α:`No`	445.16	-913.62
model	η:`Yes`	ϕ:`Yes`	α:`Yes`	403.05	-1264.03

Table 1: Metrics values computed for the DRPM model selection. Higher LPML and lower WAIC values indicate a better fit.

Then on the final model ecc

4.3. Gaussian PPMx model

4.4. Curve PPMx model

4.5. Linear Model

We thought it could be useful to implement also a simpler baseline model to use for comparison, variable selection and to better understand the data, that could allow to try a vast range of methods faster than more complex models.

Being simpler and allowing to include covariates, we chose the linear approach, actually implementing a linear model for each station. The models considered the numerical covariates linearly but tried to allow more variability on the time considering also its sine, cosine and square.

The first idea was to try a clustering on the linear model, maybe grouping together the stations according to their betas but it was soon discarded as we deemed it redundant, since we already had different models more fitted for clustering, and to focus on variable selection instead. The model was firstly implemented through jags to try and use a selection method since with so many covariates it would have been extremely long and computationally heavy to try methods based on partitions of covariates or spike and slab. The method returned a matrix with confidence quantiles on the columns and the covariates on the rows, with value 1 if the covariate was kept at that confidence, 0 otherwise. Our idea was to choose a quantile, keep only the corresponding column as a vector, and sum element-wise for all stations, expecting an higher value in correspondence of a useful covariate to select, and a significantly lower for covariates discarded by most stations. Unfortunately this first attempt did not lead us to any solid conclusion since, even changing the threshold and hyperparameters, the final vector had very similar values on all covariates, usually between 50 and 60, suggesting all had been selected only for about half the stations, and none was significantly more important than the others.

We then tried bayesian lasso and horseshoe methods, using the corresponding R packages and the same approach as before, since both methods returned a binary vector indicating whether or not to keep a variable. Horseshoe analysis was inconclusive, discarding all covariates, while lasso showed a great weight on the total precipitations, and a more moderate but still interesting on the livestock, lvi and hvi (related to total green area per unit horizontal ground surface area for low and high vegetation type) variables.

5. Models comparison

5.1. Cluster trends

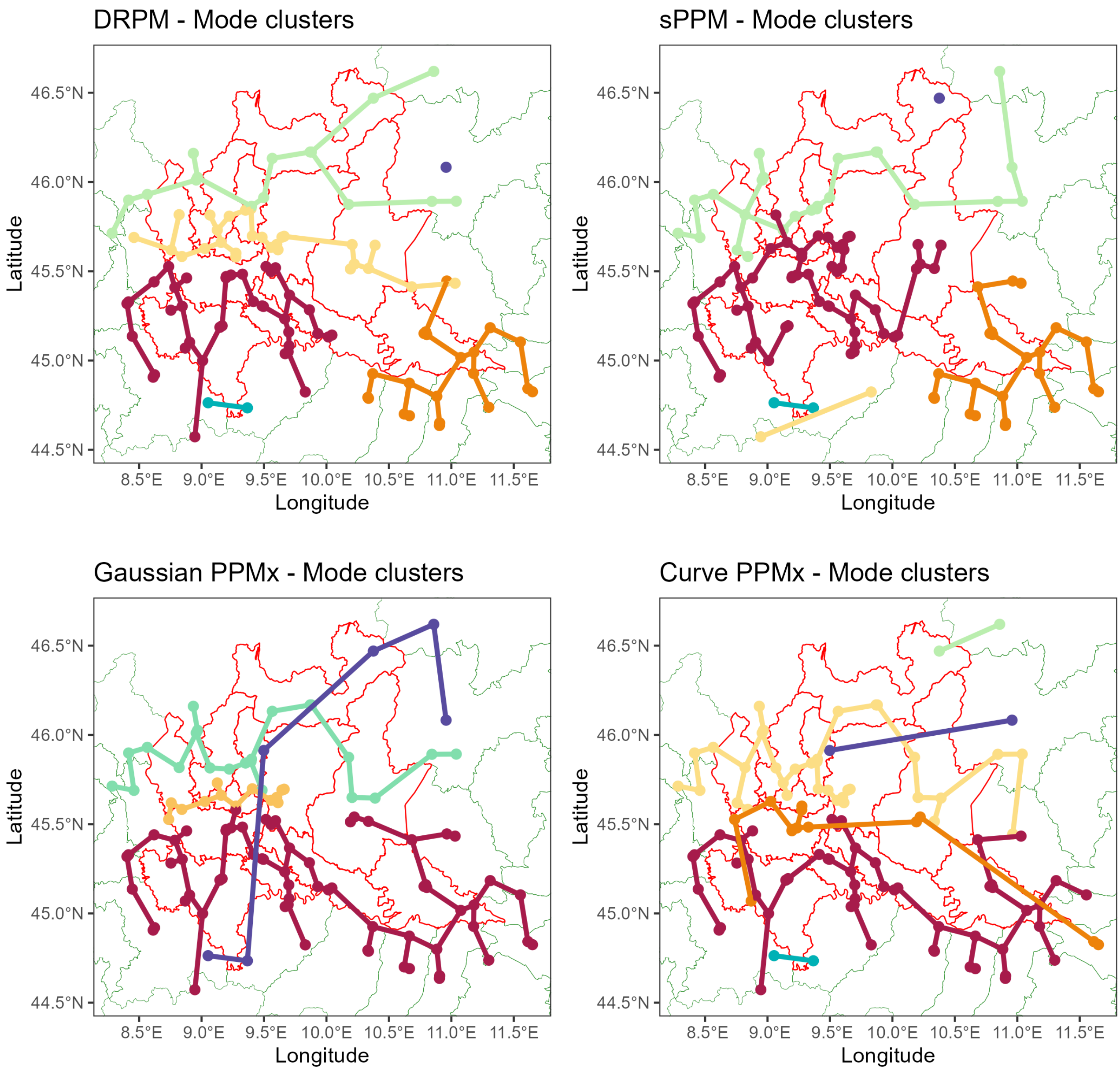

Figure 3: Maps of the most frequent clusters, throughout the 53 weeks of 2018, for all the models. See `https://federicomor.github.io/assets/figures/visualize.html` for a more detailed analysis of the plots (e.g. all the week by week clusterings).

5.2. ARI results

ecc [HA85] ecc ecc

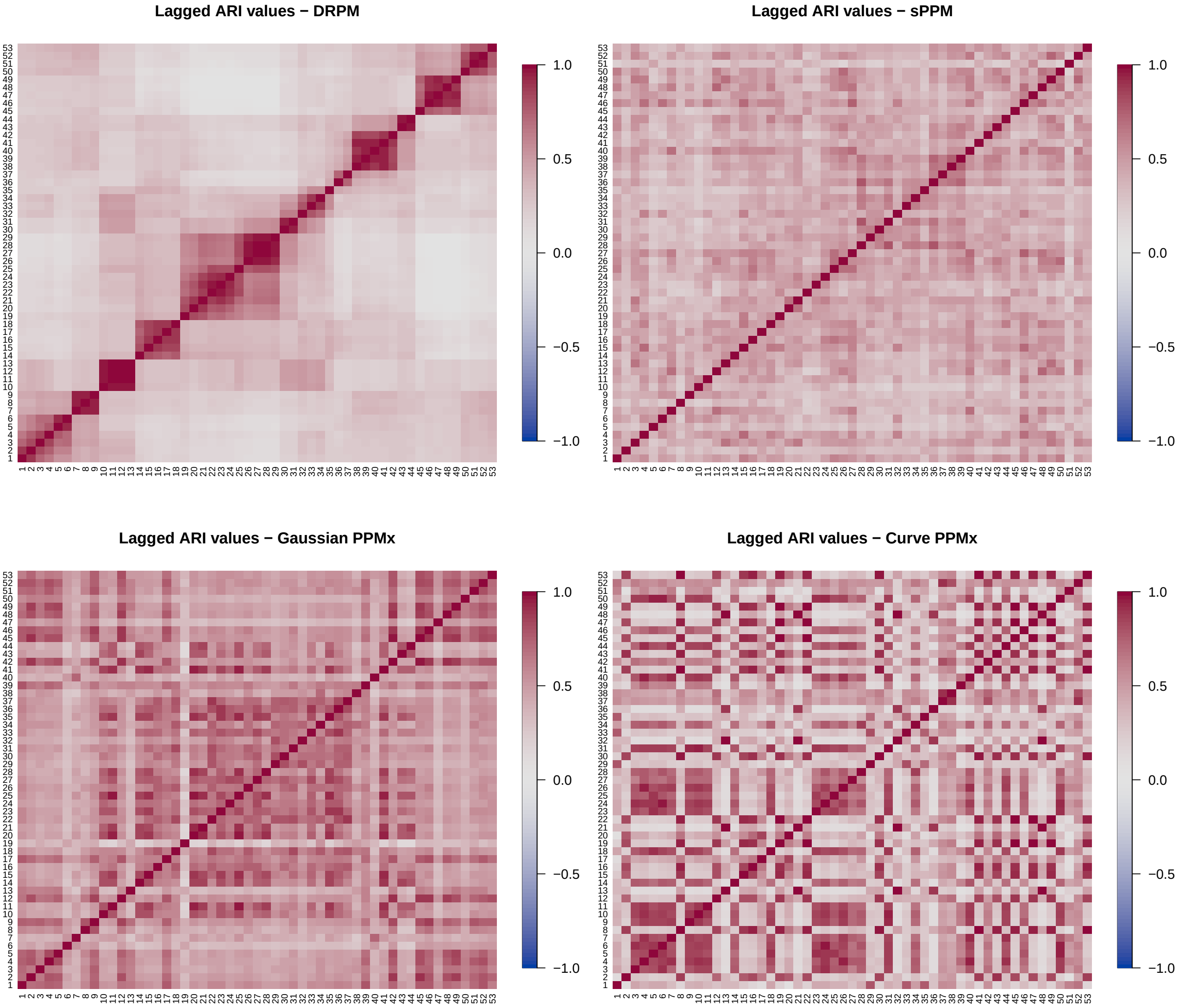

Figure 4: Lagged ARI values of the four models. The ARI values are bounded above by one and have zero expected value.

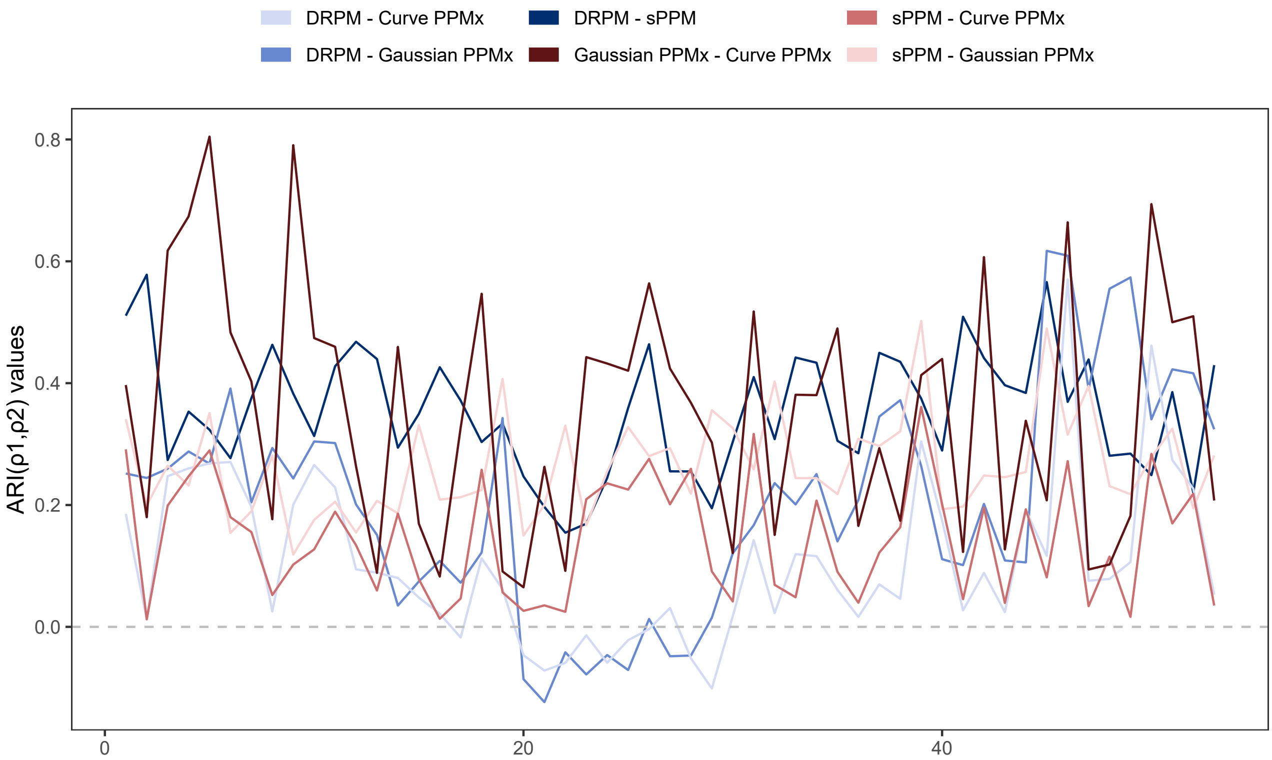

Figure 5: Plot of the ARI values for each pairwise comparison of the models, for all the weeks. Higher values mean an higher concordance in the clustering definitions, while lower values indicate a lower one, i.e. disagreement in the clustering assignments. The ARI values are bounded above by one and have zero expected value.

6. Analysis of the results

7. Conclusions

8. Further developements

Several avenues for a further development remain, presenting opportunities to enhance the depth and precision of our analysis:

- *Utilize Previous Year Data for Model Priors:* consider incorporating data from the preceding year to establish priors for the models. While our current dataset facilitated model convergence, integrating historical data could offer additional insights and refine the robustness of our findings.
- *Distinguish Between Weekends and Weekdays:* exploring the impact of human activities on PM10 levels by stratifying the analysis between weekends and weekdays. This differentiation may uncover patterns associated with specific human-related factors, contributing to a more nuanced understanding of particulate matter dynamics.
- *Ensemble Modeling:* exploring the potential benefits of ensemble modeling by combining the outputs of different models. This approach can enhance the overall accuracy and reliability of our clustering analysis. By leveraging the strengths of individual models, we can obtain a more comprehensive and robust estimation of the identified clusters.

These proposed extensions aim to further refine our methodology, enrich the interpretability of results, and provide a more comprehensive understanding of the intricate factors influencing PM10 levels in the Lombardy region.

References

[Com17] European Commission. Infringement actions for excessive levels of pm10 in italy, 2017. `https://ec.europa.eu/commission/presscorner/detail/ET/IP_17_1046`.

[CTAa] CTAN. BiBTeX documentation.

[CTAb] CTAN. pgf – create PostScript and PDF graphics in TEX.

[FRFM+23] A. Fassò, J. Rodeschini, A. Fusta Moro, Q. Shaboviq, P. Maranzano, M. Cameletti, F. Finazzi, N. Golini, R. Ignaccolo, and P. Otto. Agrimonia: Open access dataset correlating livestock and air quality in the lombardy region, italy (3.0.0), 2023.

[HA85] Lawrence J. Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[Knu74] Donald E. Knuth. Computer programming as an art. *Commun. ACM*, pages 667–673, 1974.

[Knu92] Donald E. Knuth. Two notes on notation. *Amer. Math. Monthly*, 99:403–422, 1992.

[Kot15] Stefan Kottwitz. *LaTeX Cookbook*. Packt Publishing Ltd, 2015.

[Lam94] Leslie Lamport. *LaTeX: A Document Preparation System*. Pearson Education India, 1994.

[OPHS95] Tobias Oetiker, Hubert Partl, Irene Hyna, and Elisabeth Schlegl. The not so short introduction to latex2ε. *Electronic document available at http://www. tex. ac. uk/tex-archive/info/lshort*, 1995.

[PQD22] Garrit L. Page, Fernando A. Quintana, and David B. Dahl. Dependent modeling of temporal sequences of random partitions. *Journal of Computational and Graphical Statistics*, 31(2):614–627, 2022.

A. Appendix A

A pivotal component of a robust statistical analysis lies in the effective interpretation of results. To address this crucial aspect, we meticulously constructed a library of auxiliary functions, empowering us to visually scrutinize various facets of our research.
Given the inherent temporal and spatial dimensions of our dataset, we opted for a dynamic approach, creating videos instead of static images to seamlessly navigate the temporal component.
For the visualisation of spatial variables, we devised two principal tools: a grid map and an expanding circles plot.
1. **Grid Map:** - This tool harnesses a distinct dataset featuring measurements across the entire region, organized on a grid of evenly spaced points. It offers a panoramic overview of key variables, such as Altitude and Weather measurements, providing a comprehensive understanding of spatial patterns. 2. **Expanding Circles Plot:** - Focused on station-level measurements, this tool illustrates the magnitude of variables by employing radius and color intensity of circles centered around each station. This approach grants us insights into localized patterns, enhancing our comprehension of variable distribution across the region.
To enhance the clarity of our cluster representations, we devised a function that establishes connections between stations within the same cluster. These connections are formed by solving a minimum spanning tree, a strategic approach chosen to yield a more organized and visually coherent representation of the clusters.

B. MCMC diagnostics

We now insert the plots that we used to check the convergence of the MCMC values generated by the models fit functions.
We used trace plots to ensure that the chosen values for the burn-in were high enough in order to remove any unstable behaviour in the iterates. However, even after really high burn-in periods (e.g. 60000 in DRPM), occasionally there were still some oscillations in the iterates (but by looking at the y axis we see that there is no significant variation afterall). We think that this small issue is due to the complexity of the models, since they span on lots of subjects (the 105 stations) and several time instants (the 53 weeks), together with implementing a deep hierarchical structure.

We also looked at ACF (Auto-Correlation Function) plots to tune the thinning parameter by seeing the trend of the auto correlation on subsequent iterates.
Finally, we propose the plots on two relevant parameters of each model and on weeks from 1 to 50 with jumps of length 7 for summary purposes.

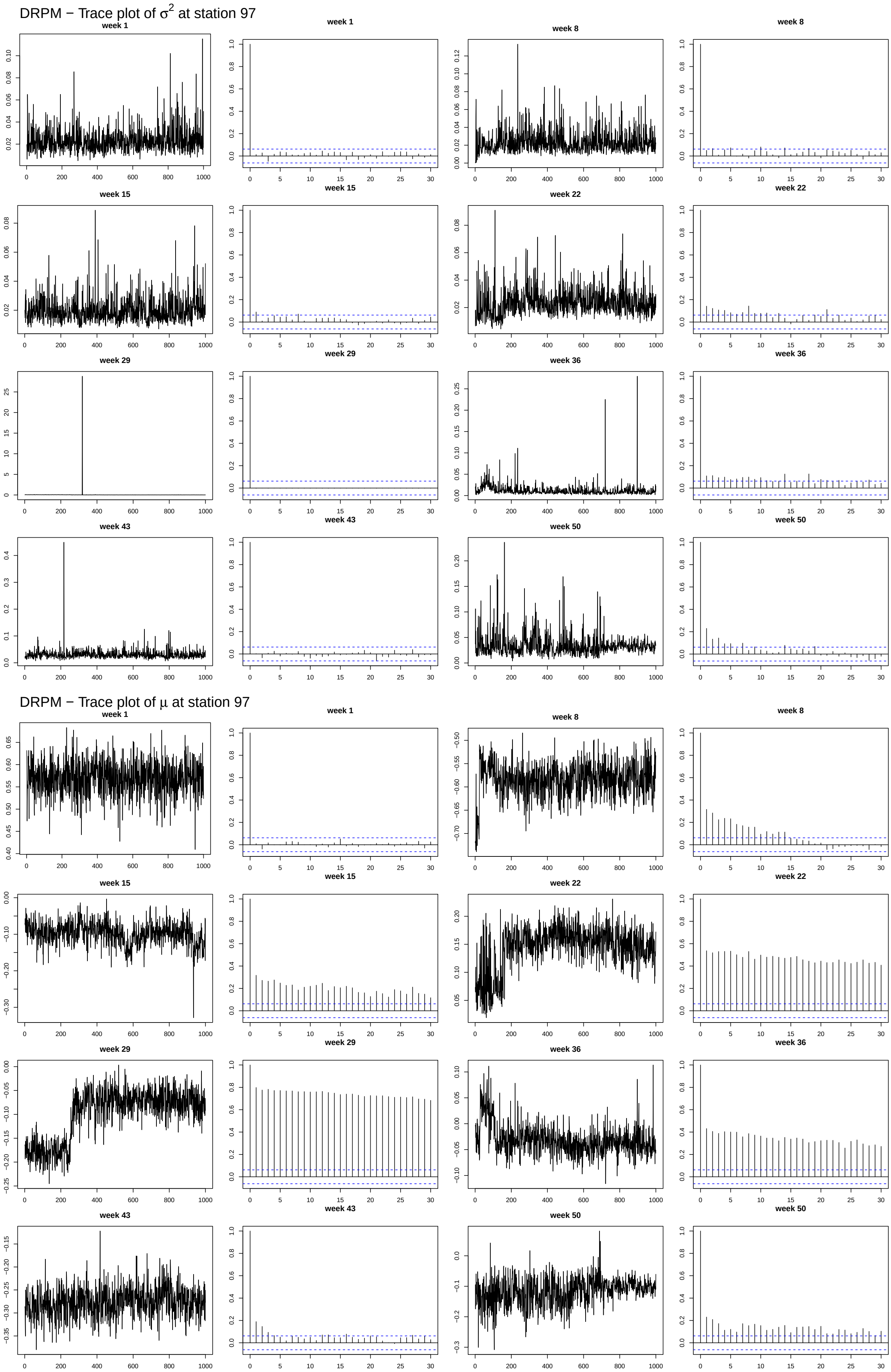

Figure 6: Trace and ACF plots of parameters σ^2 and μ of the DRPM model.

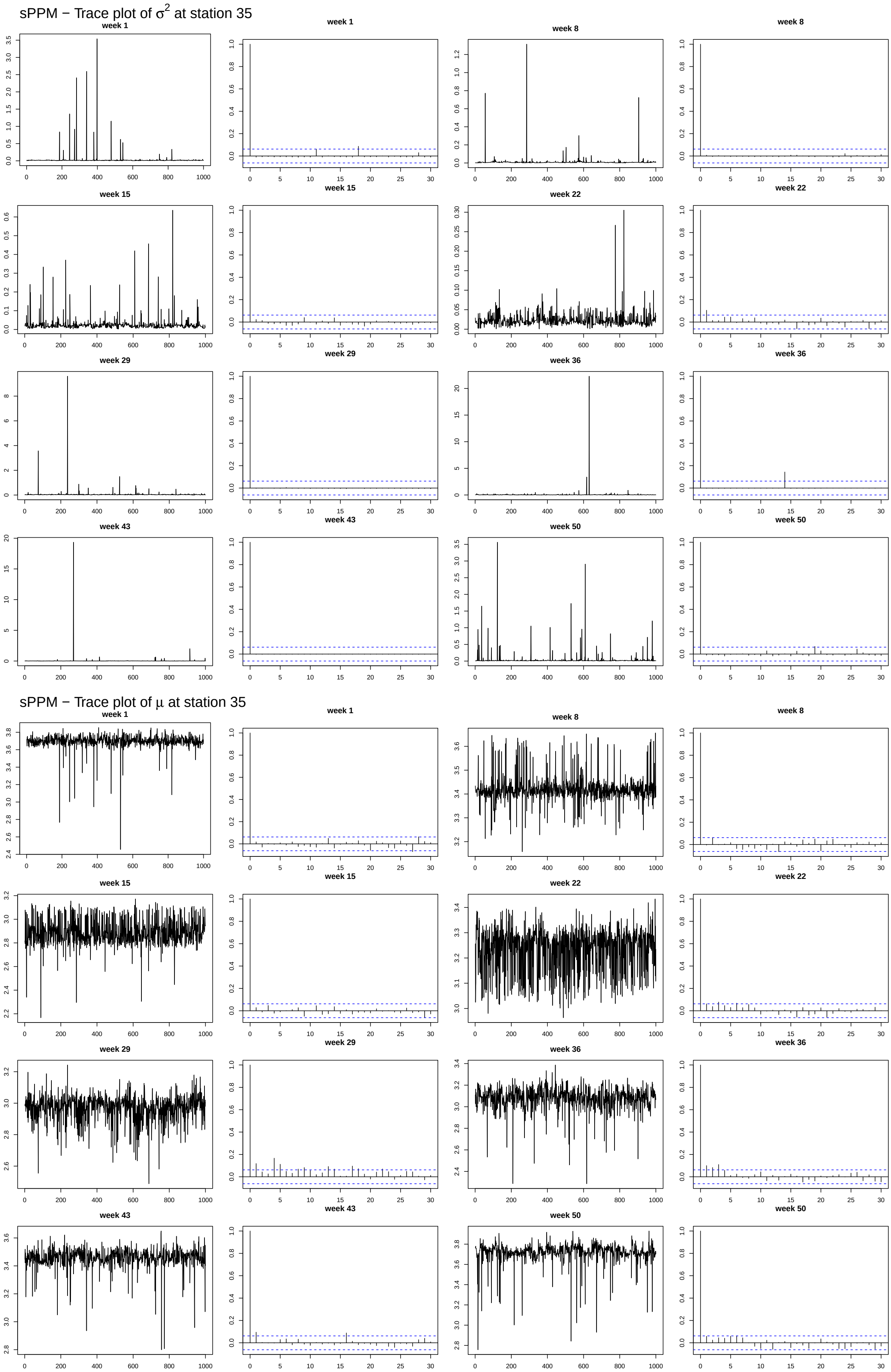

Figure 7: Trace and ACF plots of parameters σ^2 and μ of the sPPM model.

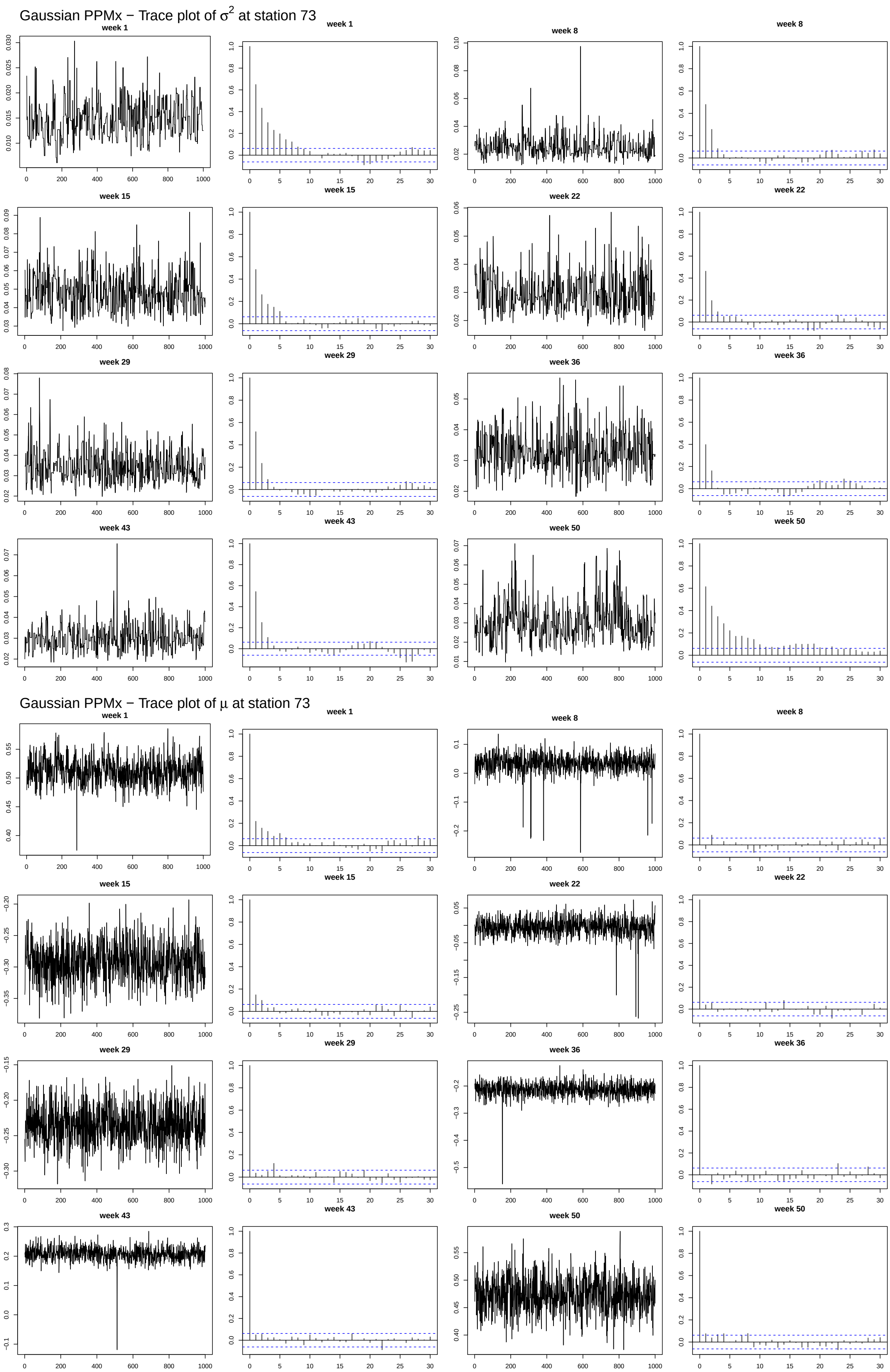

Figure 8: Trace and ACF plots of parameters σ^2 and μ of the Gaussian PPMx model.

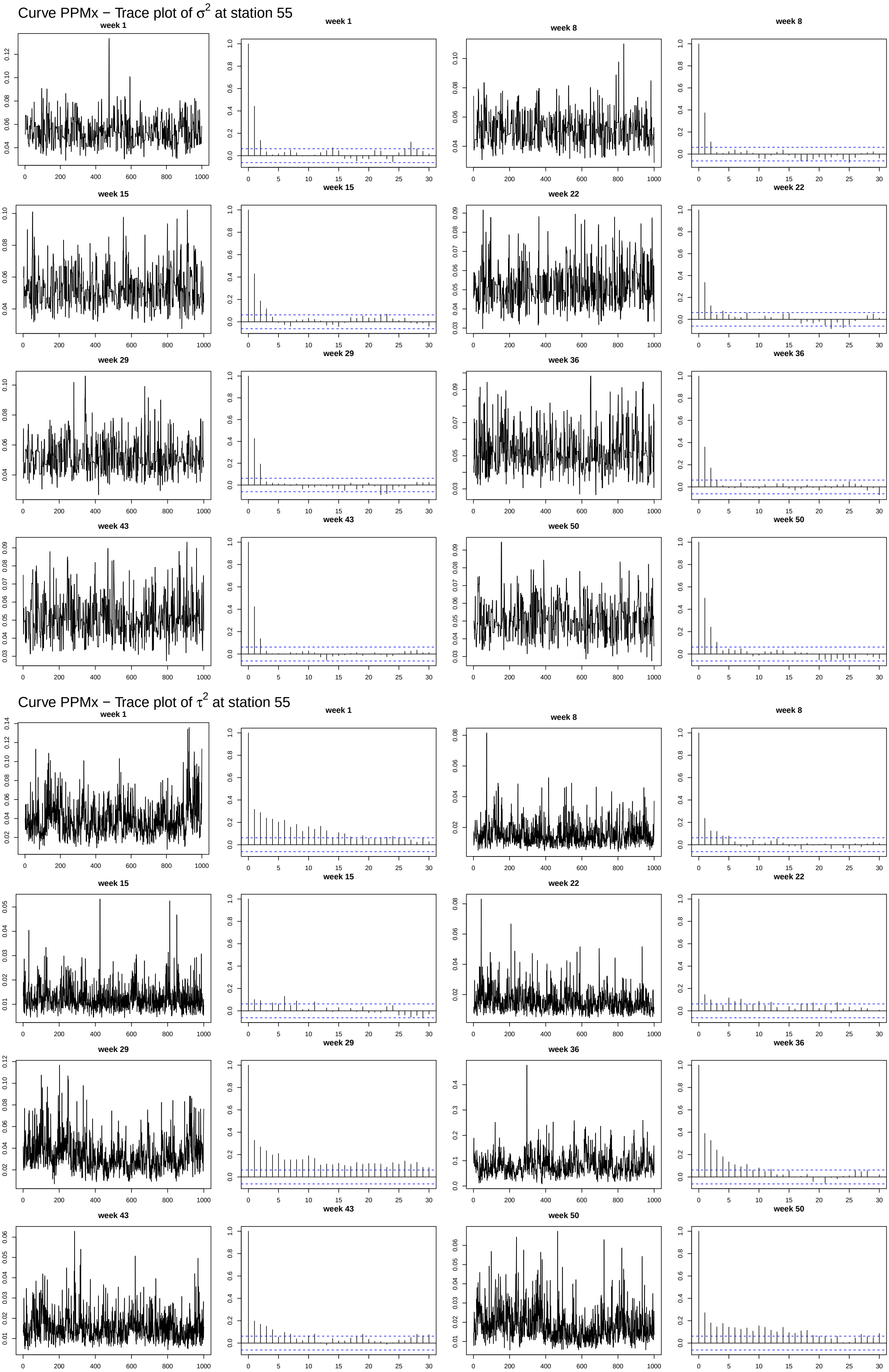

Figure 9: Trace and ACF plots of parameters σ^2 and τ^2 of the Curve PPMx model.

======================

POLIMI TEMPLATE EXAMPLE

C. Introduction

This document is intended to be both an example of the Polimi LaTeX template for Master Theses in article format, as well as a short introduction to its use. It is not intended to be a general introduction to LaTeX itself, and the reader is assumed to be familiar with the basics of creating and compiling LaTeX documents (see [OPHS95, Kot15]).
The cover page of the thesis in article format must contain all the relevant information: title of the thesis, name of the Study Programme, name(s) of the author(s), student ID number, name of the supervisor, name(s) of the co-supervisor(s) (if any), academic year.
Be sure to select a title that is meaningful. It should contain important keywords to be identified by indexer. Keep the title as concise as possible and comprehensible even to people who are not experts in your field. The title has to be chosen at the end of your work so that it accurately captures the main subject of the manuscript.
It is convenient to break the article format of your thesis (in article format) into sections and subsections. If necessary, subsubsections, paragraphs and subparagraphs can be used. A new section is created by the command

```
\section{Title of the section}
```

The numbering can be turned off by using `\section*{}`. A new subsection is created by the command

```
\subsection{Title of the subsection}
```

and, similarly, the numbering can be turned off by adding an asterisk as follows

```
\subsection*{}
```

It is recommended to give a label to each section by using the command

```
\label{sec:section_name}%
```

where the argument is just a text string that you'll use to reference that part as follows: *Section C contains* INTRODUCTION . . . .

D. Equations

This section gives some examples of writing mathematical equations in your thesis.
Maxwell's equations read:

$$\begin{cases} \nabla \cdot \boldsymbol{D} = \rho, & (2a) \\ \nabla \times \boldsymbol{E} + \dfrac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}, & (2b) \\ \nabla \cdot \boldsymbol{B} = 0, & (2c) \\ \nabla \times \boldsymbol{H} - \dfrac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}. & (2d) \end{cases}$$

Equation (2) is automatically labeled by `cleveref`, as well as Equation (2a) and Equation (2c). Thanks to the `cleveref` package, there is no need to use `\eqref`. Equations have to be numbered only if they are referenced in the text.
Equations (3), (4), (5), and (6) show again Maxwell's equations without brace:

$$\nabla \cdot \boldsymbol{D} = \rho, \tag{3}$$

$$\nabla \times \boldsymbol{E} + \frac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}, \tag{4}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{5}$$

$$\nabla \times \boldsymbol{H} - \frac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}. \tag{6}$$

Equation (7) is the same as before, but with just one label:

$$\begin{cases} \nabla \cdot \boldsymbol{D} = \rho, \\ \nabla \times \boldsymbol{E} + \dfrac{\partial \boldsymbol{B}}{\partial t} = \boldsymbol{0}, \\ \nabla \cdot \boldsymbol{B} = 0, \\ \nabla \times \boldsymbol{H} - \dfrac{\partial \boldsymbol{D}}{\partial t} = \boldsymbol{J}. \end{cases} \tag{7}$$

E. Figures, Tables and Algorithms

Figures, Tables and Algorithms have to contain a Caption that describes their content, and have to be properly referred in the text.

E.1. Figures

For including pictures in your text you can use `TikZ` for high-quality hand-made figures [CTAb], or just include them with the command

`\includegraphics[options]{filename.xxx}`

Here xxx is the correct format, e.g. `.png`, `.jpg`, `.eps`, ...

Figure 10: Caption of the Figure.

Thanks to the `\subfloat` command, a single figure, such as Figure 10, can contain multiple sub-figures with their own caption and label, e.g. Figure 11a and Figure 11b.

(a) One PoliMi logo.

(b) Another one PoliMi logo.

Figure 11: Caption of the Figure.

E.2. Tables

Within the environments `table` and `tabular` you can create very fancy tables as the one shown in Table 2.

Example of Table (optional)

	column1	**column2**	**column3**
row1	1	2	3
row2	α	β	γ
row3	alpha	beta	gamma

Table 2: Caption of the Table.

You can also consider to highlight selected columns or rows in order to make tables more readable. Moreover, with the use of `table*` and the option `bp` it is possible to align them at the bottom of the page. One example is presented in Table 3.

E.3. Algorithms

Pseudo-algorithms can be written in LaTeX with the `algorithm` and `algorithmic` packages. An example is shown in Algorithm 1.

Algorithm 1 Name of the Algorithm

```
Initial instructions
for for − condition do
    Some instructions
    if if − condition then
        Some other instructions
    end if
end for
while while − condition do
    Some further instructions
end while
Final instructions
```

F. Some further useful suggestions

Theorems have to be formatted as follows:

Theorem F.1. *Write here your theorem.*

Proof. If useful you can report here the proof.

Propositions have to be formatted as follows:

Proposition F.1. *Write here your proposition.*

How to insert itemized lists:

- first item;

	column1	**column2**	**column3**	**column4**	**column5**	**column6**
row1	1	2	3	4	5	6
row2	a	b	c	d	e	f
row3	α	β	γ	δ	ϕ	ω
row4	alpha	beta	gamma	delta	phi	omega

Table 3: Highlighting the columns

- second item.

How to write numbered lists:

1. first item;
2. second item.

G. Use of copyrighted material

Each student is responsible for obtaining copyright permissions, if necessary, to include published material in the thesis. This applies typically to third-party material published by someone else.

H. Plagiarism

You have to be sure to respect the rules on Copyright and avoid an involuntary plagiarism. It is allowed to take other persons' ideas only if the author and his original work are clearly mentioned. As stated in the Code of Ethics and Conduct, Politecnico di Milano *promotes the integrity of research, condemns manipulation and the infringement of intellectual property*, and gives opportunity to all those who carry out research activities to have an adequate training on ethical conduct and integrity while doing research. To be sure to respect the copyright rules, read the guides on Copyright legislation and citation styles available at:

`https://www.biblio.polimi.it/en/tools/courses-and-tutorials`

You can also attend the courses which are periodically organized on "Bibliographic citations and bibliography management".

I. Conclusions

A final section containing the main conclusions of your research/study and possible future developments of your work have to be inserted in the section "Conclusions".

J. Bibliography and citations

Your thesis must contain a suitable Bibliography which lists all the sources consulted on developing the work. The list of references is placed at the end of the manuscript after the chapter containing the conclusions. It is suggested to use the BibTeX package and save the bibliographic references in the file `bibliography.bib`. This is indeed a database containing all the information about the references. To cite in your manuscript, use the `\cite{}` command as follows:
Here is how you cite bibliography entries: [Knu74], or multiple ones at once: [Knu92, Lam94].
The bibliography and list of references are generated automatically by running BibTeX [CTAa].