FACULTY
OF ELECTRICAL
ENGINEERING
CTU IN PRAGUE

BACHELOR'S THESIS

Andrej Vnuk

Statistical models of PM_x pollution

Department of Radio Engineering

Supervisor of the bachelor's thesis: doc. Ing. Stanislav Vítek, Ph.D.

Study programme: Electrical engineering, electronics and communication technology

Prague 2020

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Vnuk** Jméno: **Andrej** Osobní číslo: **409054**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra radioelektroniky**

Studijní program: **Elektrotechnika, elektronika a komunikační technika**

II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Statistické modely šíření PMx částic

Název bakalářské práce anglicky:

Statistical models of PMx pollution

Pokyny pro vypracování:

Cíl práce je návrh statistického modelu, který bude schopen krátkodobé predikce stavu znečištění ovzduší polétavým prachem v daném místě. Při vypracování práce se řiďte následujícími pokyny:
1) seznamte se s problematikou měření PMx v atmosféře,
2) proveďte studii existujících medod a publikovaných řešení,
3) na základě studie navrhněte model pro predikci koncentrace PMx,
4) model ověřte na dostupné datová sadě,
5) zhodnoťte dosažené výsledky.

Seznam doporučené literatury:

[1] APPICE, Annalisa, et al. Data mining techniques in sensor networks: Summarization, interpolation and surveillance. Springer Science & Business Media, 2013.
[2] KUTNER, Michael H., et al. Applied linear statistical models. New York: McGraw-Hill Irwin, 2005.
[3] JAN, Zídek. Dolování znalostí z bezdrátových senzorových sítí. ČVUT, 2019.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

doc. Ing. Stanislav Vítek, Ph.D., katedra radioelektroniky FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) bakalářské práce:

Datum zadání bakalářské práce: **14.02.2020** Termín odevzdání bakalářské práce: **14.08.2020**

Platnost zadání bakalářské práce: **30.09.2021**

doc. Ing. Stanislav Vítek, Ph.D. podpis vedoucí(ho) práce	doc. Ing. Josef Dobeš, CSc. podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Student bere na vědomí, že je povinen vypracovat bakalářskou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací.
Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v bakalářské práci.

Datum převzetí zadání Podpis studenta

I declare that I carried out this bachelor's thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Czech Technical University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

I would like to thank the following people for helping with this thesis:
My supervisor doc. Ing. Stanislav Vítek, Ph.D., for his support, encouragement and patience. David Roesel, for his grammar corrections and inputs. And Šimon Knoška, for his insights about consistent usage of math symbols and other valuable advice.

Abstract: The main focus of this thesis is air pollution, methods of its measuring and statistical models, which can be used to forecast it. The first chapter is dedicated to the identification of the key pollutants and their effects on the environment and human health. Further, we present the common air quality index, which is used to indicate the level of pollution. In the second chapter, we outline different methods of measuring the concentration of individual pollutants via low-cost sensors as well as in professional measuring stations. The next chapter describes statistical methods of analysing and forecasting univariate time series via classical decomposition and the Box-Jenkins method. Additionally, we present machine learning methods (neural networks and decision trees), which can be used for modelling of the time series. These methods are then applied in the fourth chapter to predict the concentration of PM_{10} for a short horizon based on data collected from measuring stations in Prague. Next, we evaluate the accuracy of the predictions and present the best performing approach. The resulting models can be used in the development of the service, that would warn residents of cities about potentially dangerous air conditions.

Keywords: air pollution, statistical model, time series, PM_x

Abstrakt: Hlavnou témou tejto práce je znečistenie ovzdušia, metódy jeho merania a štatistické modely, ktoré môžu byť použité pri jeho predpovedaní. Prvá kapitola je venovaná identifikácii kľúčových znečisťujúcich látok a ich vplyvov na životné prostredie a ľudské zdravie. Ďalej uvádzame všeobecný index kvality ovzdušia, ktorý sa používa na indikovanie miery znečistenia. V druhej kapitole sú načrtnuté rôzne metódy merania koncentrácie jednotlivých znečisťujúcich látok pomocou bežne dostupných senzorov ako aj s použím profesionálnych meracích staníc. Nasledujúca kapitola opisuje štatistické metódy analýzy a predpovede jednorozmerných časových radov pomocou klasickej dekompozície ako aj s využitím Boxovho-Jenkinsovho prístupu. Navyše sa venujeme metódam strojového učenia (neurálne siete a rozhodovacie stromy), ktoré môžu byť použité na modelovania časového radu. Tieto metódy následne používame v štvrtej kapitole na predpovedanie koncentrácie PM_{10} v krátkom horizonte, na základe dát zozbieraných z meracích staníc v Prahe. Ďalej vyhodnocujeme presnosť týchto predpovedí a prezentujeme prístupy, ktoré si pri predpovediach viedli najlepšie. Výsledné modely môžu byť použité pri vývoji služby, ktorá by varovala obyvateľstvo miest pred potenciálne nebezpečnými stavmi ovzdušia.

Kľúčové slová: znečistenie ovzdušia, štatistický model, časový rad, PM_x

Contents

Introduction

The climate change caused by human activities is currently one of the major threats to our civilisation. Pollution of air by greenhouse gases and other substances is the main cause of Earth's rising temperature which leads to warming of the oceans, melting of the glaciers and ice caps, and changes of weather. Furthermore, polluted air has adverse effects on human health, causes various respiratory and cardiovascular diseases and generally shortens the life expectancy of the population. The thorough observation and measurement of air pollution is a key element in efforts to halt or even reverse these changes.

In this work, we aim to provide basic information about the main air pollutants, their sources and their impact on the environment. Further, we discuss methods used to detect and measure individual components of air pollution and describe statistical models, which can be used to model their behaviour. The main objective is then to use these models to forecast the future evolution of concentration of $\mathrm{PM_x}$ pollutant and evaluate their accuracy. The results of this work may be used in the development of an information system, which will provide warnings about worsening pollution in the observed area.

1. Air pollution

Air pollution represents a major environmental threat to the health of the population as it increases the risk of cardiovascular and respiratory diseases and may lead to premature death [1]. In this chapter, we outline the main pollutants defined by the European Union, their sources and their effects on the environment. Later we describe the air quality index, which is commonly used to indicate the level of air pollution in a given area.

1.1 Main categories of air pollutants

The European Union (EU) identifies the following seven main air pollutants, excluding greenhouse gases [2]:

1. Ammonia (NH_3)
2. Carbon monoxide (CO)
3. Nitrogen oxides (NO_x)
4. Non-methane volatile organic compounds (NMVOC)
5. Ozone (O_3)
6. Sulphur dioxide (SO_2)
7. Particulate matter (PM_x)

These pollutants affect human health in various ways and high concentrations of them contribute to climate change. Locally, accumulation of the pollutants leads to smog situations (mainly in urban areas) and can cause acute respiratory problems. Due to this, multiple organisations measure their levels and World Health Organization (WHO) has set safe limits for the exposure to them [1].

Ammonia

Under normal conditions, ammonia is a colourless gas with a distinctively pungent smell. It is highly corrosive and is soluble in water. The main source of NH_3 is the decomposition of biological waste, mainly from agriculture. It is commonly used as a fertiliser, as the main reactant in the synthesis of nitric acid and as a cleaning reagent. It is highly toxic to the water organisms and causes eutrophication of the aquatic ecosystem, which results in excessive growth of algae. For humans, exposition to ammonia leads to the irritation of the eyes and mucous membranes, pulmonary oedema and in high concentrations to death [3].

Carbon monoxide

CO is a colourless and odourless gas that is released when fossil fuels are burned. In high concentrations it impairs the amount of oxygen transported in the bloodstream to critical organs which may impact people, who already have problems with the oxygenation of blood (e.g. people with certain types of heart disease) [1].

Nitrogen oxides

Nitrogen oxides are group gases and compounds composed of nitrogen and oxygen produced mainly during the combustion of fossil fuels in power plants and combustion engines. Out of all the nitrogen oxides, the nitrogen dioxide (NO_2) has the most adverse effect on human health and when present in higher concentrations it causes inflammation of the airways and increases symptoms of bronchitis and asthma. Nitrogen oxides contribute to the acidification of soil and water which causes damage to vegetation and organisms resulting in decreased biodiversity [4].

Non-methane volatile organic compounds

NMVOCs are a collection of various organic compounds that display similar behaviour in the atmosphere. They are emitted mainly by processes which use industrial solvents or burn oxygenated fuels and certain types are hazardous to human health. These compounds are precursors to PM_x and ground level ozone [5].

Ozone

Ozone at the ground level is a secondary pollutant, which is not emitted directly but is formed a via photochemical reaction of various precursor pollutants such as nitrogen oxides and NMVOCs from traffic and industry. Sunlight is required for this reaction to occur, therefore highest levels of ozone are created during sunny weather. Prolonged exposure to ozone may cause breathing problems, trigger asthma and lead to lung disease. Aside from these, ozone is one of the most important greenhouse gases which impact climate change [1].

Sulphur dioxide

Sulphur dioxide is a colourless gas with a sharp odour which is produced during the combustion of fossil fuels and smelting of mineral ores that contain sulphur. Higher levels of SO_2 irritate the eyes, cause inflammation of the respiratory tract and may worsen asthma and chronic bronchitis. When it reacts with water vapour in air it forms sulphuric acid, which is the main component of acid rain.

Particulate matter

Particulate matter is a mixture of solid and liquid particles of various size suspended in the air, which serves as a common proxy indicator for air pollution. According to [1], they impact more people than any other pollutant and have the most serious effects on human health. They originate from both organic and inorganic substances and usually contain sulfates, nitrates, ammonia, sodium chloride, black carbon, mineral dust and water. The main emitters of PM_x are combustion engines, solid-fuel burning for energy production, as well as other industrial activities. We distinguish between two major categories of particulate matter – PM_{10}, which have a diameter of less than 10 µm and $PM_{2.5}$, which have a diameter of less than 2.5 µm. When PM_{10} is inhaled it penetrates and lodges deep inside the lungs where it causes irritation and inflammation and leads to lung diseases and cancer. Even more dangerous are $PM_{2.5}$ as they are capable of passing through the lung barrier thus enters the bloodstream, where they contribute to the development of various cardiovascular diseases. Small particulate pollution has health impacts even at very low concentrations – no threshold has been identified below which no damage to health is observed.

1.2 Common air quality index

The EU uses the common air quality index (CAQI) to evaluate the level of pollution in observed area and to compare measurements from different countries between each other. This system evaluates the average values of five key pollutants and grades their level from very low to very high according to the calculation grid. The overall CAQI is then the highest grade received across all pollutants. There are two different calculation grids, one for city background and one for roadside, which differs slightly in limit values for individual categories as well as in observed pollutants. The CAQI calculation grid for city background measurements is shown in table 1.1.

Pollution	**PM$_{10}$**		**PM$_{2,5}$**		**NO$_2$**	**SO$_2$**	**CO**	**O$_3$**
	1h	24h	1h	24h	1h	1h	8h	1h
Very low	0 25	0 15	0 15	0 10	0 50	0 50	0 5000	0 60
Low	25 50	15 30	15 30	10 20	50 100	50 100	5000 7500	60 120
Medium	50 90	30 50	30 55	20 30	100 200	100 350	7500 10000	120 180
High	90 180	50 100	55 110	30 60	200 400	350 500	10000 20000	180 240
Very high	>180	>100	>110	>60	>400	>500	>20000	>240

Table 1.1: The CAQI calculation grid for urban areas. All values are in µm and are measured as an average value over the stated period [6].

2. Sensors used in air pollution detection

In this chapter, we discuss different methods of measuring the concentration of pollutants in the air. Firstly, we describe various kinds of microsensors, which are usually used in amateur measurements or as supplementary data sources for more precise methods. Later we briefly outline methods used in measuring stations, which provide data for weather services and other government agencies. In the last section, we mention satellites used for large scope measurement of air pollution.

2.1 Air quality microsensors

In recent years we have seen increased demand for low-cost air quality microsensors, which can be used for rough measurements either in non-professional settings or for educational purposes. Due to their small dimensions and low weight, these sensors are also considered for usage when construction of a larger measuring station is not viable or for portable measuring devices [7]. Their utilisation for regulatory purposes is currently not considered by the EU, mainly due to strict requirements for data quality. However, they can be used in combination with traditional methods to increase the density of data collection and thus enrich spatial models of air pollution [8].

Based on the method by which they measure pollutant concentrations, we can split low-cost microsensors into the following categories [9]:

1. **Electrochemical sensors** for measuring nitrogen oxides, SO_2, O_3 and CO.

2. **Metal oxide sensors** for measuring NO_2, O_3 and CO.

3. **Photo ionization detectors** used to measure NMVOCs.

4. **Optical particle counters** for PM_x.

5. **Optical sensors** for measuring CO and CO_2.

Electrochemical sensors

These sensors are based on a chemical reaction between gases in the air and the electrodes in a liquid electrolyte inside a sensor. Three electrodes (working, reference and counter) separated by filters are placed in a cell filled with electrolyte. The working electrode is the site for either reduction or oxidation of the chosen gas which generates an electric charge on the surface of the electrode. This charge is balanced by a reaction at the counter electrode, thereby forming a redox pair of chemical reactions and causing current output directly proportional to the concentration of the target gas. The reference electrode is used to maintain potential on the working electrode constant [10].

Metal oxide semiconductor sensors

When a heated metal oxide is exposed to the atmosphere, it changes its resistance based on the concentration of the gases present in the sample. This fact is leveraged by MOS sensors, which are used for measuring the amount of gas in the air by using materials sensitive to the observed pollutant. The sensor is usually heated by an internal heating element to a few hundred degrees centigrade and then it is exposed to a measured sample. At high temperatures, oxygen atoms bond onto the sensor, extracting electrons in the process from the semiconductor's surface. The oxygen then either directly reacts with the ambient gases or these gases are also bonded to the surface, which results in the change of the sensor's resistance [11].

Photo ionization detectors

In a photo ionisation detector, high energy UV photons bombard molecules of the NMVOCs which results in the ejection of electrons and the formation of positively charged ions. These ions then produce electric current, which is proportional to the concentration of the NMVOCs in the measured sample. The main disadvantage of this method is its inability to distinguish between different kinds of NMVOCs – the detector ionises all components that have an lower ionisation energy than the energy of the UV light used [9].

Optical particle counters

These sensors usually work by illuminating particles passing through them by high-intensity light (produced by a laser or an LED) and measuring the resulting scattering. The concentration of the measured particles is proportional to the scattered light intensity. Optical counters can detect particles in an approximate size range of 0.4 µm to 10 µm and usually have upper detection limit of 500-1000 µg/m^3 [12].

Optical sensors

Optical sensors are used to measure the concentration of greenhouse gases via non-dispersive absorption of infrared light. Each greenhouse gas is capable of absorbing specific frequencies of infrared light – the amount of light absorbed is directly proportional to the concentration of said gas. Low-cost sensors which use this method are usually very sensitive to air temperature and humidity therefore frequent calibration of the device is required [12].

2.2 Measuring stations

For the more precise measurements which are required for regulatory purposes, countries usually use a network of automatic measuring stations. These stations are often equipped with larger analysing devices that use advanced methods for detecting air pollutants. The main disadvantage of these is in their size – they can not be built everywhere, therefore the spatial mapping of pollution produced by them is not very detailed. There are 143 measuring stations in the Czech Republic operated by the Czech Hydrometeorological Institute located (CHMI) mainly in large cities. They use the following instruments to measure the concentration of air pollutants [7]:

- **Chemiluminescence NO/NO_2/NO_x Analyzer** to measure concentration of nitrogen oxides
- **UV Absorption O_3 Analyzer** to measure concentration of ozone
- **UV Fluorescence SO_2 Analyzer** to measure of concentration sulfur dioxide
- **Automatic & Real-time Particulate Monitor**, which uses absorption of the β radiation to count PM_x.

2.3 Satellite measurements

The last methods of measuring the air pollution we mention are the satellite measurements operated by NASA and ESA. These agencies have several instruments in the Earth's orbit which observe the distribution of pollution in the troposphere. The NASA Terra project consists of five instruments, two of which are measuring air pollution. The Moderate Resolution Imaging Spectroradiometer (MODIS) measures among other things the properties of aerosols that enter the atmosphere from man-made sources like pollution and biomass burning and natural sources like dust storms, volcanic eruptions, and forest fires [1]. The Measurement of Pollution in the Troposphere (MOPITT) instrument observes the distribution, transport, sources, and sinks of carbon monoxide in the troposphere via gas correlation spectroscopy [2].

The EU's Copernicus Atmospheric Monitoring Services uses several ESA satellites to monitor the environment in Europe. Three of these satellites, Sentinel 4, 5 and 5 precursor, collect information about the air pollution. The Sentinel 4 mission measures the key air pollutants NO_2, O_3, SO_2, formaldehyde, glyoxal, and other aerosols. Complementarily, the Low Earth Orbiting missions S5 and S5p provide additional data about CO, CH_4, and stratospheric O_3 [3].

[1] https://terra.nasa.gov/about/terra-instruments/modis

[2] https://terra.nasa.gov/about/terra-instruments/mopitt

[3] https://sentinel.esa.int/web/sentinel/missions/sentinel-4

3. Statistical models

This chapter is dedicated to the description of statistical models used to forecast univariate time series. In the first section we provide definitions for basic terms and methods used in the construction of statistical models as well as definitions of accuracy measures which evaluate forecast quality. The following section explains different approaches of time series decomposition, namely classical decomposition into the trend, seasonal and residual terms and the Box-Jenkins method of analysing time series. In further sections we discuss selected statistical models, which were used in modeling PM$_x$ pollution.

3.1 Introduction to time series analysis

Our main object of interest in this section will be a time series and the analysis of its behaviour in order to determine the mechanism, which generates the observed data. Understanding of this mechanism then enables us to model and forecast future behaviour of a given process. For purposes of this thesis, let us define a univariate (with one time-dependent variable) time series as follows.

Definition 1. *A* ***time series (TS)*** *of a length $n \in \mathbb{N}$, given as*

$$\{X_t\}_{t=1...n} = \{X_1, X_2, ...X_n\} \tag{3.1}$$

is a set of n equally spaced discrete data points in chronological order. Further we recognise

- ***Deterministic time series*** *- is fully described by mathematical formula*
- ***Stochastic time series*** *- contains random elements*

In the following text we will consider only stochastic TS.

Generally, each element of a stochastic TS is random variable with its own expected value $\mu_t = E(X_t)$ and variance $\sigma_t^2 = var(X_t)$. We assume standard definitions of these operators as in [13, 14]. TS is characterised by the covariance function (CF) of order the k defined as

$$\gamma_{k,t} = cov(X_t, X_{t+k}) = E[(X_t - \mu_t)(X_{t+k} - \mu_{t+k})] \tag{3.2}$$

and the autocorrelation function (ACF) of the order k defined as

$$\rho_{k,t} = \frac{cov(X_t, X_{t+k})}{\sigma_t \sigma_{t+k}}. \tag{3.3}$$

Definition 2. *Let $\{\varepsilon_t\}_{t=1...n}$ be a stochastic TS with*

$$E(\varepsilon_t) = 0, \quad var(\varepsilon_t) = \sigma^2, \quad cov(\varepsilon_t, \varepsilon_{t+k}) = 0, \qquad \text{for } k \neq 0 \ \& \ \forall t = 1...n$$

that is a TS with unrelated random variables and limited variance. Then $\{\varepsilon_t\}_{t=1...n}$ is called ***White Noise (WN)*** [15].

3.1.1 Weighted least square method

In the following chapters we often face the problem of approximating parameters of a given TS. For this purpose we generally use a weighted least square method described bellow.

Consider a regression model $F(x, \vec{\beta})$, where $\vec{\beta} = (\beta_1, \beta_2, \dots, \beta_n)$ is a vector of free parameters, which approximates a process $y = f(x)$. Let us choose one arbitrary set of values $\vec{\beta}$, then for the i-th element of a given process we can write

$$y_i = F(x_i, \vec{\beta}) + r_i, \tag{3.4}$$

where r_i is the residual i.e. the difference between the estimated and real values. Our goal will be to choose such $\vec{\beta} = \vec{b}$, that sum of weighted residuals for function $F(x, \vec{\beta})$,

$$S(\vec{\beta}) = \sum_{i=1}^{n} r_i^2 w_i = \sum_{i=1}^{n} \left[y_i - F(x_i, \vec{\beta})\right]^2 w_i, \tag{3.5}$$

with weight $w_i = \sigma_i^{-2}$ will be minimal. That gives us a necessary condition

$$\operatorname{grad} S(\vec{b}) = \vec{0}. \tag{3.6}$$

This equation can be solved analytically only for certain regression models (linear, polynomial, ...) and in most cases numerical methods are used instead.

3.1.2 Measures used to evaluate forecast accuracy

The accuracy of a forecast is determined by its ability to correctly predict new data, which were not used for the construction of a given model. To this end, we usually split available data into two portions, training and test data as shown in figure 3.1. The training data are used as an input for estimation of model parameters and the resulting forecast is then compared to the test data. To evaluate forecast accuracy, we use the following error metrics [16]:

a) **Mean absolute error (MAE):**

$$\text{MAE} = E\left(\left|\hat{X}_t - X_t\right|\right), \tag{3.7}$$

b) **Root mean square error (RMSE):**

$$\text{RMSE} = \sqrt{E\left((\hat{X}_t - X_t)^2\right)}, \tag{3.8}$$

c) **Mean absolute percentage error (MAPE):**

$$\text{MAPE} = E\left(\left|100\% \cdot \frac{\hat{X}_t - X_t}{X_t}\right|\right), \tag{3.9}$$

where $\hat{X}_t$ is the approximated value and X_t is the observed value.

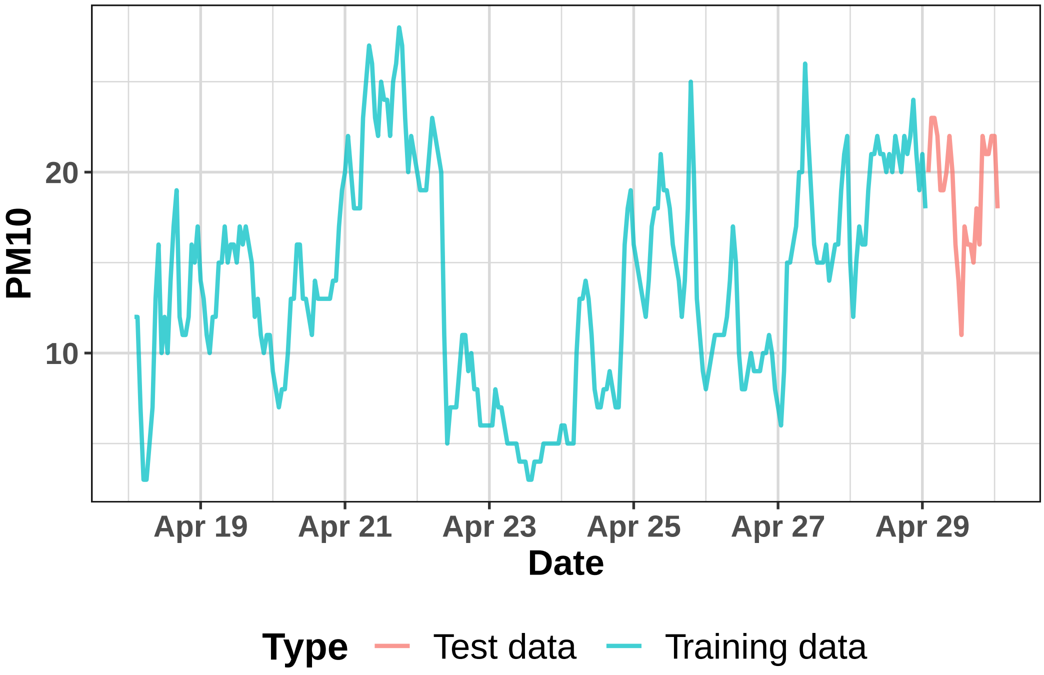

Figure 3.1: An example of data split into training and test portions. Only training data are used in model creation.

3.2 Decomposition of time series

In order to analyse a given TS in the time domain, we must decompose it into its underlying components. According to [15], this can be done in two ways:

1. **Classical decomposition** - based on regression analysis
2. **Box-Jenkins method** - based on correlation analysis.

3.2.1 Classical decomposition

Let us assume that a random process which generates our TS is only a function of time. We can then decompose the TS into a **deterministic part** and a **random part**. The deterministic part consists of a **trend term** T_t, which reflects long term behaviour of the TS and a **seasonal term** S_t, which describes periodic changes of the TS [15]. Further non-periodic fluctuations of varying frequency are included in a **cyclic term**, which is usually merged with the trend term into a single **trend-cycle component** [17]. A random part of the TS is characterised by a residual term ε_t given by random inconsistencies of the TS and can generally be described as white noise. Using these elements we define two models, which can be used for forecasting future behaviour of a given TS:

Definition 3. *Let $\{X_t\}_{t=1...n}$ be a TS decomposed to its trend term $\{T_t\}_{t=1...n}$, seasonal term $\{S_t\}_{t=1...n}$ and residual term $\{\varepsilon_t\}_{t=1...n}$. Then **additive model** of TS defined as:*

$$X_t = T_t + S_t + \varepsilon_t \qquad \text{for } \forall t = 1...n$$

*and its **multiplicative model** as:*

$$X_t = T_t \cdot S_t \cdot \varepsilon_t \qquad \text{for } \forall t = 1...n.$$

Examples of decomposed time series using an additive and a multiplicative model are shown in figures 3.2 and 3.3.

Trend and Cycle

The trend term reflects long term changes in the average behaviour of a TS i.e. its general tendency in a long time [20]. It is usually caused by factors which consistently affect the observed process in one way e.g. rising number of cars in an area causing a rising trend in air pollution measurements. In a broader sense, a trend may change its character in time and therefore it could contain cyclical elements with longer periods. Classical decomposition approximates trend term via regression models, with the most commonly used model being linear model given by equation

$$T_t = \beta_1 + \beta_2 t + u_t, \tag{3.10}$$

where u_t is the uncertainty of the model and parameters β_1, β_2 are obtained by the least square method [20].

In reality, the trend is typically not constant over long periods of time and based on the character of the TS it fluctuates with no fixed frequency. These irregular changes are described by a cyclic term C_t which is often added to the trend term. An example of a varying trend is shown in figure 3.2.

Seasonality

Seasonality is a periodic fluctuation of a TS which occurs with a fixed frequency and causes predictable changes in the TS. This term is usually observed in TS that are influenced by change of weather during different seasons, is dependent on the day of the week or changes during day and night. If a TS has a sufficient number of observations with high density (e.g. hourly data), we can observe multiple seasonalities (daily, weekly, monthly, . . .). To account for this behaviour, we can include trigonometric or other periodic terms to the model. In the case of long seasonal periods, Fourier terms may be added to estimate seasonality. These models are then often called harmonic regression [17].

STL decomposition

In further chapters we use STL as a main algorithm for decomposition. The STL, abbreviation of ”Seasonal and Trend decomposition using Loess”, is a method of classical decomposition, which uses locally estimated scatterplot smoothing or ”LOESS” for non-linear regression. This method was developed by Robert Cleveland et al. [21] in 1990 as a simple way to extract trend-cycle and seasonal terms from the TS. For the purpose of this thesis, describing the STL algorithm in detail is not crucial – for further details see the aforementioned article.

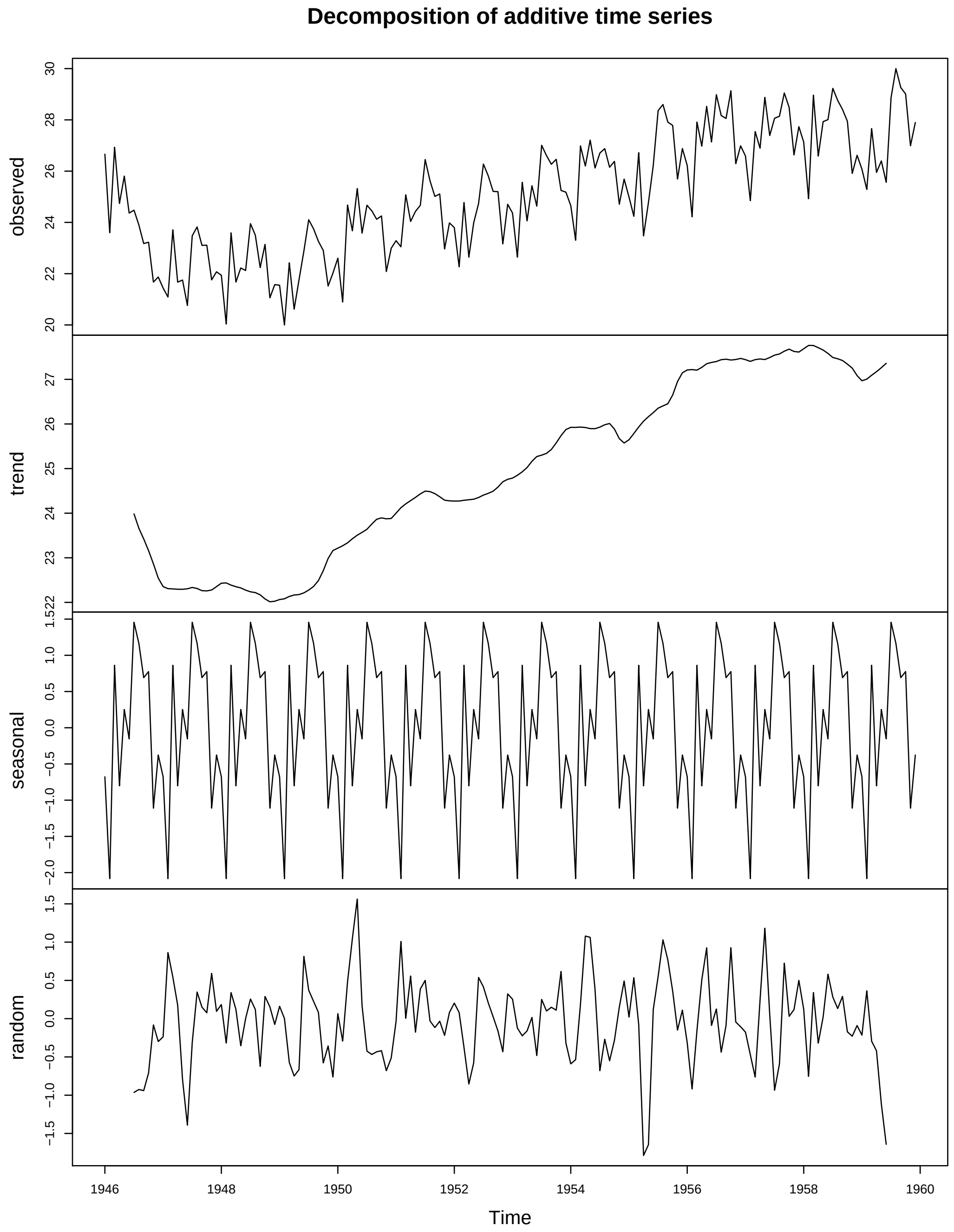

Figure 3.2: An example of the decomposition of additive TS. Data used for this plot come from an arbitrary data source [18].

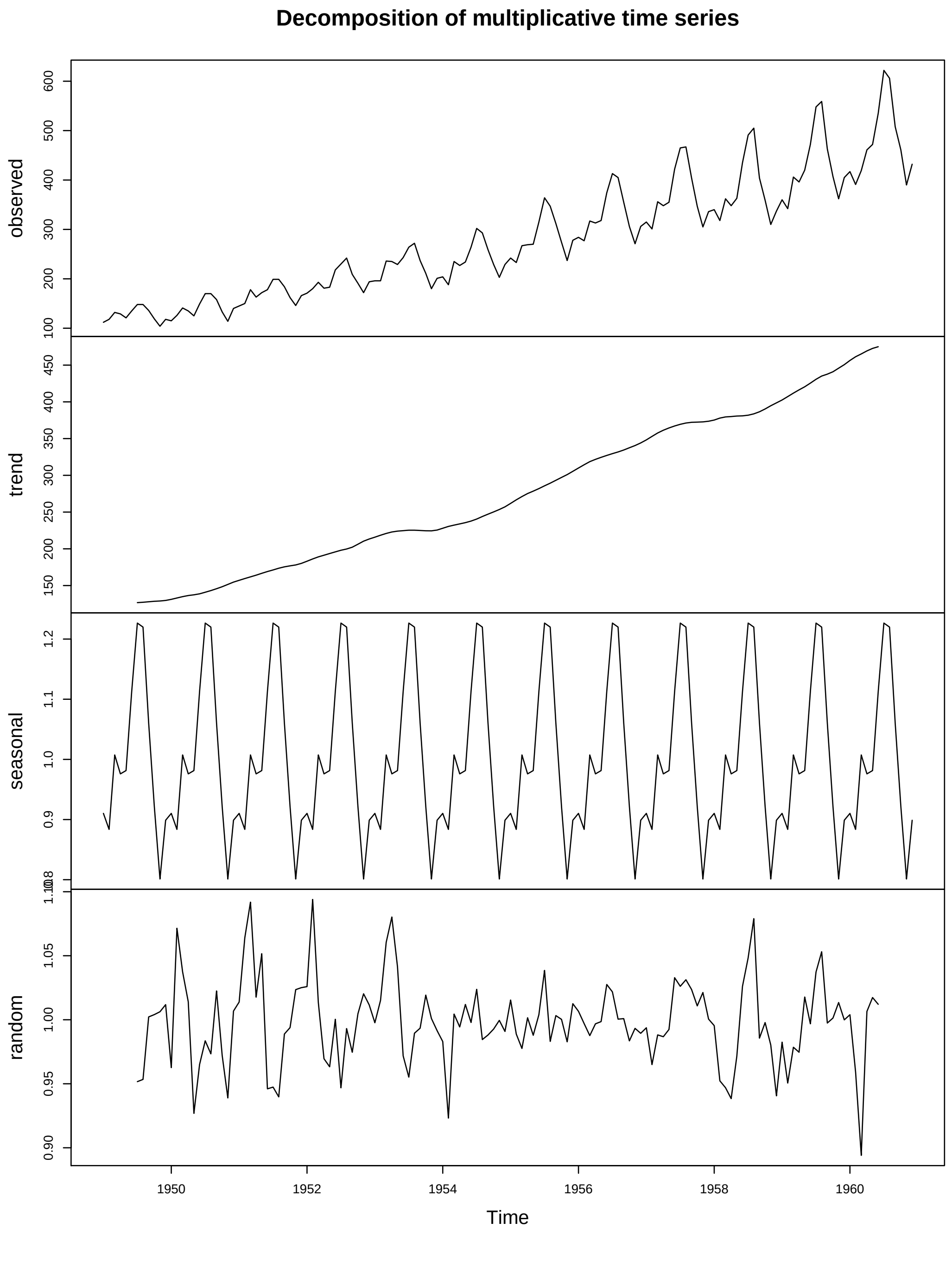

Figure 3.3: An example of the decomposition of multiplicative TS. Data used for this plot come from an arbitrary data source [19].

3.2.2 Box-Jenkins method

In contrast to classical decomposition, the Box-Jenkins method assumes a random character of each component of the decomposed TS and uses autoregressive (integrated) moving average models to predict future values. These models are usually very flexible and quickly adapt to changes in TS behaviour. However, in comparison with classical decomposition, the forecast produced by the Box-Jenkins method is harder to interpret.

The Box-Jenkins method is usually executed in the following steps [22]:

1. **Data preparation:** involves transformations and differencing. Data are transformed via a suitable function (logarithm, square roots, ...) to stabilise variance in the TS. Differencing is used to remove trend and seasonality from the data in order to make the TS easier to model.

2. **Model selection:** by using various graphs based on differenced and transformed TS we try to identify a potential model with a good fit to the data.

3. **Parameter estimation:** fine-tuning values of coefficients used in the selected model to achieve best fit.

4. **Model checking:** testing the assumptions of a model to find any areas where it may be inadequate. If the model is not accurate enough, we return to step 2.

5. **Forecasting:** the selected model is used to produce future values of the analysed TS.

In the following sections we will define basic elements used in constructing models via the Box-Jenkins method. All models built this way assume a stationary TS – if this condition is not met, we need to difference TS in order to reduce or eliminate trend and seasonal patterns and thus make it stationary. Next, we can apply either an autoregressive model, moving averages or their combination to fit and forecast the transformed TS.

Stationary time series

A time series is considered stationary when its characteristics are not changing in time (e.g. constant variance and trend). A more precise definition follows.

Definition 4. *Let $\{X_t\}_{t=1\dots n}$ be a stochastic TS. $\{X_t\}_{t=1\dots n}$ is said to be* ***stationary*** *if its expected value μ_t and variance σ_t^2 are constant and its CF and ACF is dependent only on time distance between two points [20]:*

$$\gamma_{k,t} = cov(X_t, X_{t+k}) = cov(X_t, X_{t-k}) = \gamma_{-k,t}$$
$$\rho_{k,t} = \frac{cov(X_t, X_{t+k})}{\sigma_t \sigma_{t+k}} = \frac{cov(X_t, X_{t-k})}{\sigma_t \sigma_{t-k}} = \rho_{-k,t}$$

Differencing and integrated model of the order d

To difference TS means to simply subtract its lagged version from the original TS. This can be easily denoted by using backshift and difference operators.

Definition 5. *Let $\{X_t\}_{t=1\dots n}$ be a stochastic TS. The* ***backshift operator*** *B is as [23]:*

$$BX_t = X_{t-1},$$

and its j^{th} application as

$$B^j X_t = X_{t-j}.$$

Definition 6. *Let X_t be a stochastic TS. The* ***difference operator*** *∇ is defined as:*

$$\nabla X_t = X_t - X_{t-1} = X_t - BX_t = (1-B)X_t,$$

and the difference operator of the order d as:

$$\nabla^d X_t = (1-B)^d X_t.$$

A TS is integrated of order d, denoted as $I(d)$, when its d^{th} difference is stationary.

Autoregressive model of the order p

A TS $\{X_t\}_{t=1\dots n}$ fitted by an autoregressive model of the order p, denoted as AR(p), can be written as [20]:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_t, \tag{3.11}$$

where $\phi_1, \phi_2, \dots, \phi_p \in \mathbb{R}$ are coefficients of the model, ε_t is a WN and

$$\phi_0 = \mu \left(1 - \sum_{i=1}^{p} \phi_i\right),$$

with μ describing the mean of X_t. We can rewrite this by using a backshift operator as

$$X_t = \phi_0 + \sum_{i=1}^{p} \phi_i B^i X_t + \varepsilon_t, \tag{3.12}$$

which is often shortened to

$$\phi_p(B) X_t = \phi_0 + \varepsilon_t. \tag{3.13}$$

Moving average model of the order q

A TS $\{X_t\}_{t=1\dots n}$ fitted by a moving average model of the order q, denoted as MA(q), can be written as [20]:

$$X_t = \phi_0 + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q}, \tag{3.14}$$

where $\phi_0 = \mu$ is a mean of the series, $\theta_1, \theta_2, \dots, \theta_q \in \mathbb{R}$ are coefficients of the model and $\varepsilon_t, \varepsilon_{t-1}, \dots, \varepsilon_{t-q}$ are WN error terms. By using a backshift operator we can write:

$$X_t = \phi_0 + (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)\varepsilon_t = \phi_0 + \theta_q(B)\varepsilon_t. \tag{3.15}$$

If we can rewrite a MA(q) model as a convergent AR(∞) model, that is when $\phi(B) = \theta_q^{-1}(B)$, we call the MA($q$) model **invertible**. This imposes constraints on MA parameters, which are complicated to compute for large q – statistical libraries in R are used to calculate them.

3.3 Exponential smoothing models

The term ”exponential smoothing” is derived from the exponentially decaying weights used in these models – forecasted values depend on weighted averages of past observations with an emphasis on the more recent data. In this chapter we will define basic exponential smoothing methods and their applications in forecasting.

3.3.1 Simple Exponential Smoothing (SES)

One of the most basic exponential smoothing methods is SES, which is suitable for forecasting data with no clear trend or seasonal pattern [17]. The forecasting equation is given as:

$$\begin{aligned}\hat{X}_{t+1} &= \alpha X_t + \alpha(1-\alpha)X_{t-1} + \alpha(1-\alpha)^2 X_{t-2} + \dots \\ &= \sum_{j=0}^{t-1} \alpha(1-\alpha)^j X_{t-j}, \qquad (3.16)\end{aligned}$$

where $0 \leq \alpha \leq 1$ is a smoothing parameter, which controls the rate of weight decay.

3.3.2 Holt’s linear trend

We can extend the SES method to also apply for TS with trend by using two smoothing equations. This approach was proposed by Charles Holt and is known as Holt’s linear trend method [17]. The forecasting equation is given as:

$$\hat{X}_{t+h} = \ell_t + hb_t, \tag{3.17}$$

where ℓ_t is the equation estimating level of the TS and b_t is the equation estimating trend. These equations are defined as:

$$\ell_t = \alpha X_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{3.18}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1}, \tag{3.19}$$

where $0 \leq \alpha \leq 1$ is a smoothing parameter for the level and $0 \leq \beta \leq 1$ is a smoothing parameter for the trend. The h-step-ahead forecast is equal to the last estimated level plus h times the last estimated trend value – forecasts are linear functions of h [17].

3.3.3 Holt-Winters’ seasonal method

To account for seasonality, we need to modify the previous model by adding a third smoothing equation s_t, with a corresponding smoothing parameter γ, for the seasonal component. This can be done in two ways, additive or multiplicative, depending on the nature of seasonality – the additive method is suitable for constant seasonal patterns while the multiplicative method preforms better when seasonal variations are dependent on the level of the TS. Both methods were developed by Charles Holt and his student Peter Winters and are together known as Holt-Winters' seasonal method [17]. Forecasting and smoothing equations for the additive method are:

$$\hat{X}_{t+h} = \ell_t + hb_t + s_{t+h-m(k+1)} \tag{3.20}$$
$$\ell_t = \alpha(X_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{3.21}$$
$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \tag{3.22}$$
$$s_t = \gamma(X_t - \ell_t) + (1-\gamma)s_{t-m}, \tag{3.23}$$

where m is the frequency of seasonality, $k \in \mathbb{N}$ is the integer part of $(h-1)/m$, which ensures that the seasonal pattern uses final period of available data as an input and $0 \leq \alpha; \beta; \gamma \leq 1$ are smoothing parameters [17].The forecasting and smoothing equations for multiplicative the method are:

$$\hat{X}_{t+h} = (\ell_t + hb_t)s_{t+h-m(k+1)} \tag{3.24}$$
$$\ell_t = \alpha\frac{X_t}{s_{t-m}} + (1-\alpha)(\ell_{t-1} + b_{t-1}) \tag{3.25}$$
$$b_t = \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} \tag{3.26}$$
$$s_t = \gamma\frac{X_t}{\ell_{t-1} + b_{t-1}} + (1-\gamma)s_{t-m}. \tag{3.27}$$

3.3.4 ETS model framework

In this section we will discuss a common classification of exponential smoothing models, the ETS framework, as described in [17]. Methods described in the previous sections are not the only exponential smoothing models available. If we account for two common ways of modeling the trend term (additive and damped additive) and two ways of modeling the seasonal term (additive and multiplicative), we can sort each exponential smoothing model into a category based on the methods it uses. This gives us 9 distinct models labelled by ordered pairs of letters (Trend, Seasonal) – possible combinations are shown in table 3.1.

Seasonal Trend	None	Additive	Multiplicative
None	(N,N)	(N,A)	(N,M)
Additive	(A,N)	(A,A)	(A,M)
Additive damped	(A$_d$,N)	(A$_d$,A)	(A$_d$,M)

Table 3.1: Classification of exponential smoothing models based on trend and seasonal components [17].

Using this framework, we can assign previous models to categories:

Simple exponential smoothing	(N,N)
Holt's linear trend	(A,N)
Holt-Winters' additive method	(A,A)
Holt-Winters' multiplicative method	(A,A)

Models from the ETS framework generate forecasts with prediction intervals, which are obtained by adding error terms to smoothing equations. We can use

ADDITIVE ERROR MODELS

Trend	**Seasonal**		
	N	**A**	**M**
N	$y_t = \ell_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$	$y_t = \ell_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/\ell_{t-1}$
A	$y_t = \ell_{t-1} + b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t$ $b_t = b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + b_{t-1})$
$\mathbf{A_d}$	$y_t = \ell_{t-1} + \phi b_{t-1} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$	$y_t = \ell_{t-1} + \phi b_{t-1} + s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t$ $b_t = \phi b_{t-1} + \beta\varepsilon_t$ $s_t = s_{t-m} + \gamma\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m} + \varepsilon_t$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha\varepsilon_t/s_{t-m}$ $b_t = \phi b_{t-1} + \beta\varepsilon_t/s_{t-m}$ $s_t = s_{t-m} + \gamma\varepsilon_t/(\ell_{t-1} + \phi b_{t-1})$

MULTIPLICATIVE ERROR MODELS

Trend	**Seasonal**		
	N	**A**	**M**
N	$y_t = \ell_{t-1}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$	$y_t = (\ell_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \alpha(\ell_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + s_{t-m})\varepsilon_t$	$y_t = \ell_{t-1}s_{t-m}(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1}(1 + \alpha\varepsilon_t)$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
A	$y_t = (\ell_{t-1} + b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + b_{t-1} + \alpha(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = b_{t-1} + \beta(\ell_{t-1} + b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$
$\mathbf{A_d}$	$y_t = (\ell_{t-1} + \phi b_{t-1})(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1} + s_{t-m})(1 + \varepsilon_t)$ $\ell_t = \ell_{t-1} + \phi b_{t-1} + \alpha(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$ $s_t = s_{t-m} + \gamma(\ell_{t-1} + \phi b_{t-1} + s_{t-m})\varepsilon_t$	$y_t = (\ell_{t-1} + \phi b_{t-1})s_{t-m}(1 + \varepsilon_t)$ $\ell_t = (\ell_{t-1} + \phi b_{t-1})(1 + \alpha\varepsilon_t)$ $b_t = \phi b_{t-1} + \beta(\ell_{t-1} + \phi b_{t-1})\varepsilon_t$ $s_t = s_{t-m}(1 + \gamma\varepsilon_t)$

Figure 3.4: Forecasting and smoothing equations used by models in the ETS framework [17].

two different kinds of error terms – additive or multiplicative. To distinguish between the type of error, we add a third letter into the previously described ordered set of letters (Error, Trend, Seasonal), which completes the ETS framework. The table in the figure 3.4 shows forecasting and smoothing equations for all available ETS models.

3.4 ARIMA models

As mentioned in chapter 3.2.2, the Box-Jenkins method uses a combination of autoregression and moving averages for forecasting TS. Both of these methods require a stationary TS, therefore we need to integrate any non-stationary TS in order to apply them. In this chapter we will describe the combined autoregressive integrated moving average (ARIMA) model and its usage in forecasting both non-seasonal and seasonal TS.

3.4.1 Non-seasonal ARIMA model

By combining equations (3.13), (3.15) and thr difference operator from definition 6 we obtain non-seasonal ARIMA(p, d, q) model [20]:

$$\phi_p(B)(1-B)^d X_t = \phi_0 + \theta_q(B)A_t, \tag{3.28}$$

where p is the order of the AR, d is the degree of differencing and q is the order of the MA. The constant ϕ_0 determines behaviour of the long term forecast produced by this model (e.g. for ϕ_0=0 and d=0 long term forecast will tend to 0 and for $\phi_0 \neq 0$ and d=0 long term forecast will tend to the mean of the TS).

Determining values of parameters (p, d, q) is not trivial and usually can not be done from the time plot – for better insight we need to examine relations between the observed TS and its lagged version. To do this, we plot values of ACF for the sufficient number of lags and evaluate its behaviour. Additionally, we also need to consider that two lags of TS can show correlation only because they both correlate with the third lag. We avoid this by using the partial autocorrelation function (PACF), which removes the effect of other lags. As a rule of thumb, we can use the following statements to identify parametes for the ARIMA model [17]:

1. If the ACF of a TS is exponentially decaying or shows sinusoidal behaviour, and there is a significant peak at lag p in PACF but none beyond, then the data may follow the ARIMA$(p, d, 0)$ model.

2. If the a PACF of TS is exponentially decaying or shows sinusoidal behaviour, and there is a significant peak at lag q in ACF but none beyond, then the data may follow the ARIMA$(0, d, q)$ model.

An example of ACF and PACF plots for 24 lags is shown in figure 3.5.

3.4.2 Seasonal ARIMA model

The non-seasonal ARIMA(p, d, q) model can be extended to account for seasonality by adding seasonal terms with parameters $(P, D, Q)_m$, which plays a similar role as their non-seasonal counterparts. The constant m represents the seasonal frequency (e.g. for monthly data $m = 12$). The resulting seasonal ARIMA$(p, d, q)(P, D, Q)_m$ model is obtained by multiplying each term in equation (3.28) by its seasonal equivalent, where the backshift operator is applied with order m [20]:

$$\phi_p(B)\Phi_P(B^m)(1-B)^d(1-B^m)^D X_t = \theta_q(B)\Theta_Q(B^m)A_t. \tag{3.29}$$

3.4.3 Dynamic regression models

In a regression model, we assume that the modelled TS has a linear relationship to other predictor TS (e.g. air pollution is dependent on temperature, humidity, ...) [17]. Let Y_t be a forecasted TS and $X_{1,t}, X_{2,t}, ..., X_{k,t}$ the TS of its predictor variables. The regression model is then defined as:

$$Y_t = \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \varepsilon_t, \tag{3.30}$$

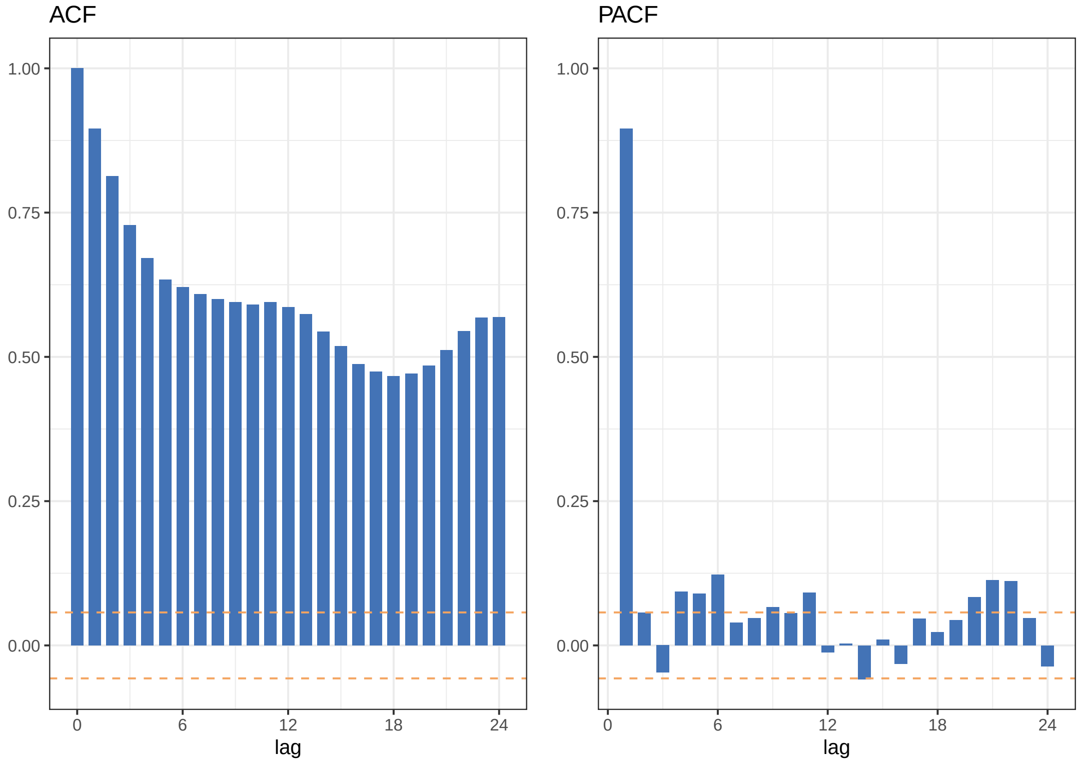

Figure 3.5: Values of the autocorrelation function and the partial autocorrelation function of time series of PM_{10} concentration for 24 lags.

where $\beta_0, \beta_1, \ldots, \beta_k$ are parameters of the model, which are usually estimated by a least square method, and ε_t is the WN error term. The main advantage of this method is its ability to include external variables into the equation and thus enrich the model, however, to use predictors in a forecast we need to know their forecasted values as well. The main disadvantage is that these models do not allow forecasts as dynamic as ARIMA. To improve this, we can use ARIMA to model the error term ε_t in the regression model, which results in a dynamic regression given by equations [17]:

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_k X_{k,t} + \eta_t \\ \phi_p(B)(1-B)^d \eta_t &= \theta_q(B)\varepsilon_t, \end{aligned} \tag{3.31}$$

where η_t is an error series following the ARIMA(p,d,q) model and ε_t is WN.

Dynamic harmonic regression (DHR)

As was briefly mentioned in chapter 3.2.1, if there is a long seasonal period or multiple seasonalities (e.g. in hourly data with daily, weekly and monthly seasonality), Fourier terms may be added to the model to estimate seasonal terms.

For the seasonal frequency m, Fourier terms are given as:

$$\begin{aligned}
X_{1,t} &= \sin\left(\frac{2\pi t}{m}\right), \quad X_{2,t} = \cos\left(\frac{2\pi t}{m}\right) \\
X_{3,t} &= \sin\left(\frac{4\pi t}{m}\right), \quad X_{4,t} = \cos\left(\frac{4\pi t}{m}\right) \\
X_{5,t} &= \sin\left(\frac{6\pi t}{m}\right), \quad X_{6,t} = \cos\left(\frac{6\pi t}{m}\right) \\
&\vdots
\end{aligned} \tag{3.32}$$

Using these TS as predictors in (3.31) we obtain a dynamic harmonic regression model, where the seasonal term is modelled by Fourier terms and short-term dynamics are handled by ARIMA [17]. The smoothness of the seasonal pattern can be regulated by a number of Fourier terms, which we include into the model (less terms result in a smoother pattern). For TS with multiple seasonalities, we can include multiple sets of Fourier terms.

3.5 BATS and TBATS models

The DHR model mentioned in chapter 3.4.3 can process long seasonal periods very well but it can not handle non-integer values of seasonal frequency and changes in period length. To overcome these issues, A. M. De Livera et al. proposed a framework of models, which adds Box-Cox transformation, Fourier representations with time varying coefficients and ARMA error correction to exponential smoothing models. Their approach also includes an evaluation of best model based on the Akaike information criterion (AIC) due to a significant number of parameters, which these models use as an input. In this chapter we will describe basic principles used in this method.

3.5.1 Box-Cox transformations

The Box-Cox transformations are a group of logarithm and power transformations, which are used to reduce or eliminate changes in the variation of the TS. Generally, these can be defined for any series, but we will focus only on non-negative TS.

Definition 7. *Let $\{X_t\}_{t=1\ldots n}$ be a TS, $X_t \geq 0$ for every $t \in \mathbb{N}$. The* ***Box-Cox transformations*** *from X_t to $X_t^{(\lambda)}$ is given as:*

$$X_t^{(\lambda)} = \begin{cases} \dfrac{X_t^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\[2ex] \ln X_t & \text{for } \lambda = 0 \end{cases}$$

where λ is a parameter defining a particular transformation. Back-transform from $X_t^{(\lambda)}$ to X_t is [24]:

$$X_t = \begin{cases} (\lambda X_t^{(\lambda)} + 1)^{1/\lambda} & \text{for } \lambda \neq 0 \\ \exp(X_t^{(\lambda)}) & \text{for } \lambda = 0. \end{cases}$$

3.5.2 BATS model

The name BATS is an acronym for the key features of the model: Box-Cox transform, ARMA errors, Trend, and Seasonal components. It is described by a set of parameters $(\lambda, \phi, p, q, m_1, m_2, \ldots, m_T)$ to indicate the Box-Cox parameter, damping parameter, ARMA parameters (p and q), and the seasonal periods $(m_1, m_2, \ldots, m_T)$ [25]. This model is an extension of the Holt-Winters' additive method from section 3.3.3 with T seasonal patterns (represented by T seasonal equations) and damped trend. Modelling and smoothing equations are:

$$\begin{aligned}
X_t^{(\lambda)} &= \begin{cases} \dfrac{X_t^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \\ \ln X_t & \text{for } \lambda = 0 \end{cases} \\
X_t^{(\lambda)} &= \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t \\
\ell_t &= \ell_{t-1} + \phi b_{t-1} + \alpha d_t \\
b_t &= (1-\phi)b + \phi b_{t-1} + \beta d_t \\
s_t^{(i)} &= s_{t-m_i}^{(i)} + \gamma_i d_t \\
d_t &= \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t,
\end{aligned} \tag{3.33}$$

where $X_t^{(\lambda)}$ is a TS after the Box-Cox transformation, $(m_1, m_2, \ldots, m_k)$ are seasonal periods, ℓ_t is the local level in a period t, b is the long-run trend, b_t is a short-run trend in period t, $s_t^{(i)}$ is the i-th seasonal component, d_t denotes ARMA(p, q) process and ε_t is a WN error term [25].

3.5.3 TBATS model

To further improve the previous model, we can replace seasonal equations $s_t^{(i)}$ in (3.33) by trigonometric seasonal formulation, based on the Fourier series [25]:

$$\begin{aligned}
s_t^{(i)} &= \sum_{j=1}^{k_i} s_{j,t}^{(i)} \\
s_{j,t}^{(i)} &= s_{j,t-1}^{(i)} \cos \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin \lambda_j^{(i)} + \gamma_1^{(i)} d_t \\
s_{j,t}^{*(i)} &= -s_{j,t-1}^{(i)} \sin \lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos \lambda_j^{(i)} + \gamma_2^{(i)} d_t,
\end{aligned} \tag{3.34}$$

where $\gamma_1^{(i)}, \gamma_2^{(i)}$ are smoothing parameters and $\lambda_j^{(i)} = 2\pi j / m_i$. The modelling equation is:

$$X_t^{(\lambda)} = \ell_{t-1} + \phi b_{t-1} + \sum_{i=1}^{k} s_{t-1}^{(i)} + d_t. \tag{3.35}$$

This gives us the Trigonometric BATS model (TBATS), which is described by a set of parameters $(\lambda, \phi, p, q, \{m_1, k_1\}, \{m_2, k_2\}, \ldots, \{m_t, k_t\})$.

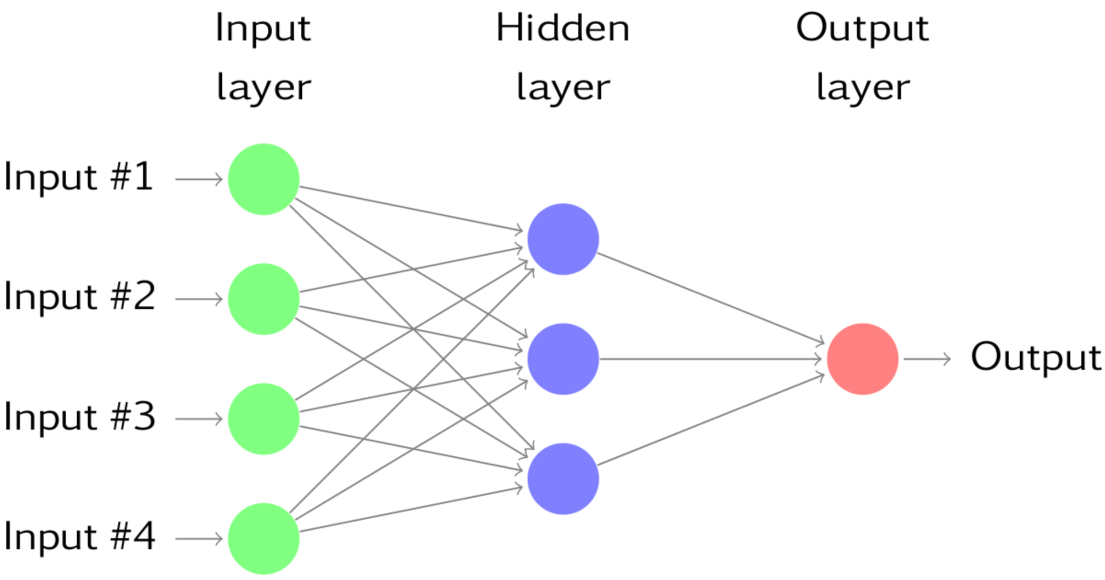

Figure 3.6: A feed-forward neural network consisting of 3 layers. [17]

3.6 Neural Network Auto-Regressive model

A multilayer feed-forward neural network is a type of neural network where information flows only in the forward direction - i.e. from input nodes, through hidden nodes and to the output nodes, without any cycles. We can utilise this kind of system to create an autoregressive model by feeding lagged values of a TS as an input to the network.

3.6.1 Neural network architecture

In its simplest form, a neural network consists only of an input layer and an output layer of nodes, which is equivalent to linear regression. More complicated networks have a number of hidden layers, which causes a non-linearity. Each layer of nodes uses modified outputs from the previous layer as its input - e.g. the hidden layer of neural network in figure 3.6 uses weighted linear combinations of the first layer nodes as its input, then modifies them by a nonlinear function and outputs them to the next layer[17].

3.6.2 NNAR model

The NNAR(p, k) model is a feed-forward neural network with 1 hidden layer of k nodes, which uses a linear combination function and an activation function to produce forecasts, based on p lagged values of the TS. The modelling equation is [26]:

$$X_t = w_0 + \sum_{j=1}^{k} w_j g\left(w_{0,j} + \sum_{i=1}^{p} w_{i,j} X_{t-j}\right) + \varepsilon_t, \tag{3.36}$$

where $w_{i,j}$ and w_j are connection weights. The transfer function $g(x)$ is defined as a logistic function:

$$g(x) = \frac{1}{1+e^{-x}}. \tag{3.37}$$

Given a TS with seasonal a pattern, it is useful to add the last observed values from same season as input – this gives us a NNAR$(p, P, k)_m$ model, where m is a seasonal frequency. It uses $(X_{t-1}, X_{t-2}, \ldots, X_{t-p}, X_{t-m}, X_{t-2m}, \ldots, X_{t-Pm})$ as input values and has k nodes in its hidden layer [17].

3.7 Classification and regression trees

Another approach to TS forecasting is to use a decision tree as a predictive model. Depending on the kind of the target variable, we distinguish two kinds of decision trees – a **classification tree** when the target variable has discreet values and a **regression tree** when the target variable is continuous. In this chapter we will discuss basic principles of using this method.

3.7.1 Decision tree

A decision tree is a structure that is used to represent classifying examples – its nodes represent a test of a certain attribute and its branches show the outcome of this test. In order to use a decision tree in modelling, we need to provide a set of features (individual measurable properties of the observed TS), which will be used in its construction. For TS analysis the most useful features are lagged versions of the TS itself and, in the case that seasonal patterns are present, Fourier coefficients. One disadvantage of using this approach is the fact, that it can not handle a trend term as it uses only rules made on training data. Therefore, the original TS must be detrended by other methods and the trend term must be modelled separately (e.g. by the ARIMA model) [27], [28]. An example of a decision tree constructed on a TS of PM_{10} concentration is depicted in figure 3.7.

3.7.2 Recursive partitioning

Recursive partitioning is a statistical method used to produce a decision tree and classify the input population by splitting it into sub-populations. These can be then split again until the maximal depth of the tree is reached (based on a predetermined criterion or when all the leaf nodes are homogeneous). At each step, the split is made based on the independent feature that results in the largest possible reduction in heterogeneity of the dependent (predicted) variable.

3.7.3 Conditional inference trees

Recursive partitioning can lead to overfitting – a state, where the model corresponds very closely or exactly to the training data. Another way to use recursive partitioning, which avoids this problem, is a conditional inference tree. This method takes into account the distributional properties of the data and uses statistical theory (permutation-based significance tests) to determine which features to select instead of minimising heterogeneity. If no significant association between any of the features and the response is found, recursion is stopped [28].

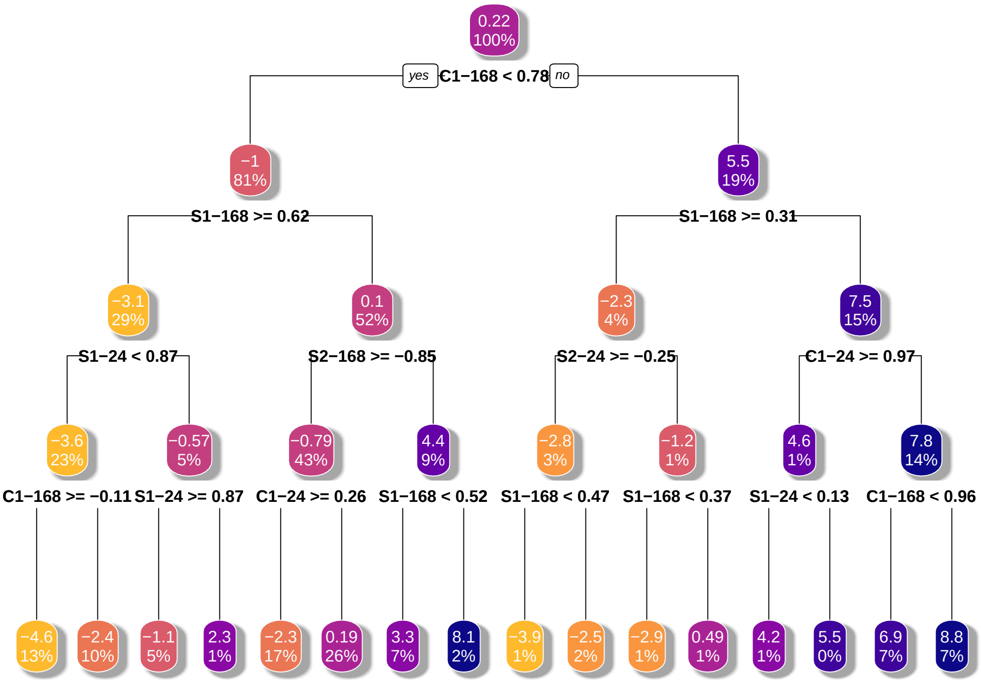

Figure 3.7: An example of a decision tree constructed using recursive partitioning with a lag feature and two sets of Fourier terms on a TS of PM_{10} concentration.

4. Modelling PMx pollution

In this chapter we use statistical libraries in the programming language R to model and forecast PM_{10} pollution, using the statistical methods described in the previous chapter. Data used for this analysis were collected from a network of air quality sensors in Prague, which are operated by the Czech Hydrometeorological Institute (CHMI)[1]. The collection of data was realised via application that was designed by J. Zídek and is described in [29].

4.1 Introduction

In the analysis bellow, we followed principles described in the book 'Forecasting: Principles and Practice' by Rob J. Hyndman [17], who is one of the leading authorities in forecasting and statistics. For the forecasting via regression trees we followed the method described by Peter Laurinec in [28].

4.1.1 Analysed data

We used data from 5 different sensors in Prague, which are operated by CHMI – IDs and locations of these sensors sre shown in table 4.1. We were working only with PM_{10} concentration as not all of these sensors measure $PM_{2.5}$, however, all models are applicable to $PM_{2.5}$ as well. Data were collected between 25/02/2017 and 09/05/2019 in hourly measurements as an actual PM_{10} concentration as well as a 24 hour average concentration. These two sets of data differ only in averaging used by the latter, which after a few tests was evaluated as redundant – we let the models handle this task. To further improve the quality of data, we clean outlier values by replacing them with linear interpolation. The length of the collected data is too large for some forecasting methods, according to [17] it is unrealistic for a model to stay the same for long periods of time. Due to this, we chose to trim the data to a length sufficient to cover several seasonal cycles. We split the data into training and testing sets, constructed model using the training data and then evaluated the accuracy against the test data. In order to quantify the accuracy of our model, we used metrics described in chapter 3.1.2. In the following chapters we used figures of models from the ASMIA sensor for demonstration, models for the remaining sensors are shown in figures in appendix A. An example of analysed data from the ASMIA sensor before and after cleaning the outliers are shown in figure 4.1.

4.1.2 Challenges in working with hourly data

The main problem we encountered when working with hourly data was multiple seasonalities. For this reason we could not use any of the basic, single-seasonal models and had to work mainly with STL, dynamic harmonic regression and other methods that allow at least double seasonality. For our data we assumed daily ($m_1 = 24$) and weekly ($m_2 = 24 \cdot 7 = 168$) seasonalities – a monthly seasonality

[1] http://portal.chmi.cz/files/portal/docs/uoco/web_generator/actual_hour_data_CZ.html

Sensor ID	Location
AKALA	Praha 8, Karlín
ALEGA	Praha 2, Legerova
AREPA	Praha 1, n. Republiky
ARIEA	Praha 2, Riegrovy sady
ASMIA	Praha 5, Smíchov

Table 4.1: ID and location of selected CHMI sensors which were used for data collection.

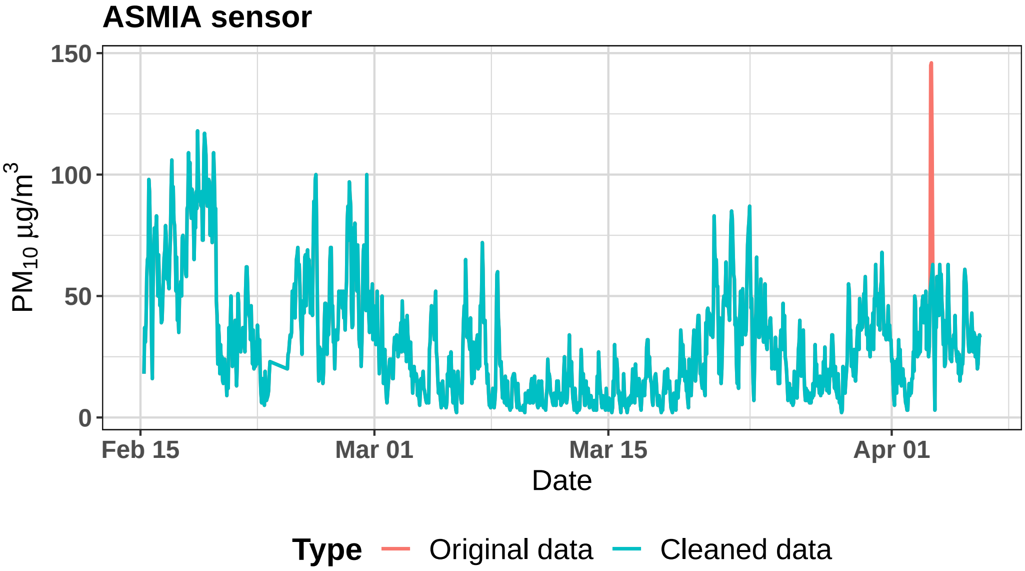

Figure 4.1: Hourly concentration of PM_{10} collected by the ASMIA sensor between March and May 2019 and data with cleaned outliers.

was omitted mainly due to the need to normalise data for months with different number of days and an increased length of data.

4.1.3 Forecasting with R

Most methods and models described in chapter 3 are already implemented in the programming language R. We converted our data to R's time series objects, which consist of measured value and their associated time stamps, and used the following libraries to create models and forecasts:

```
library(tseries)    # TS analysis and computational finance
library(dplyr)      # grammar of data manipulation
library(forecast)   # tools for displaying and forecasting TS
library(xts)        # handling of R's time-based data classes
library(rpart)      # recursive partitioning decision tree method
library(rpart.plot) # decision tree plots
library(party)      # conditional inference decision tree method
```

4.2 Exponential smoothing and STL

For the first experiments we chose exponential smoothing models, namely Holt's linear trend and the Holt-Winters' method. As expected, Holt's model yielded only straight lines capturing the trend of the TS – this model does not contain any seasonal terms, therefore it does not perform well for short term forecasts. However, it may provide a reasonable approximation for a long period forecasting as it models the mean value of the series. An example forecast produced by this method is shown in figure 4.2 for ASMIA sensor (see appendix A, figures A.2, A.13, A.24 and A.35 for the rest of the sensors).

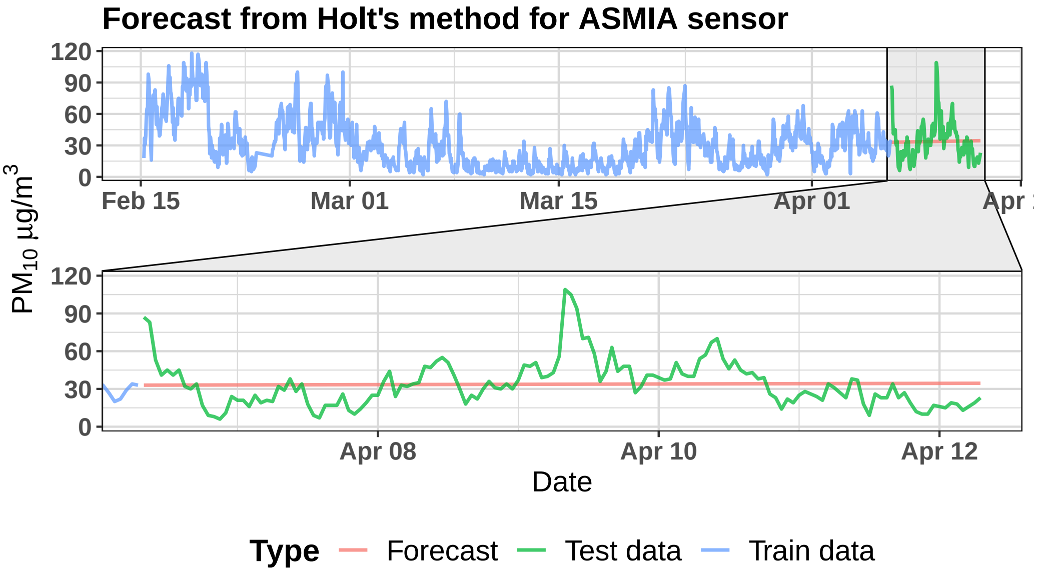

Figure 4.2: Forecast obtained by Holt's linear model (ETS(A,A,N)) from data collected from the ASMIA sensor.

Standard Holt-Winter's method did not work for any available data – the forecast library in R simply failed to estimate any valid parameters. To overcome this, we used the double-seasonal method, which adds a second seasonal equation to the expression in equation 3.20. This solved the problem and proved to work surprisingly well. Forecast produced by Holt-Winters' method is shown in figure 4.3 for the ASMIA sensor (see appendix A, figures A.3, A.14, A.25 and A.36 for the rest of the sensors).

The next method we used was STL in combination with the exponential smoothing model. Again, because of the multiple seasonalities present in the data, we used a multi-seasonal version of this function (MSTL). This method decomposes the TS into its seasonal and trend-cycle terms and then forecasts each of them separately via the exponential smoothing method – in our case the SES model. A forecast produced this way is shown in figure 4.4 for the ASMIA sensor (see appendix A, figures A.4, A.15, A.26 and A.37 for the rest of the sensors).

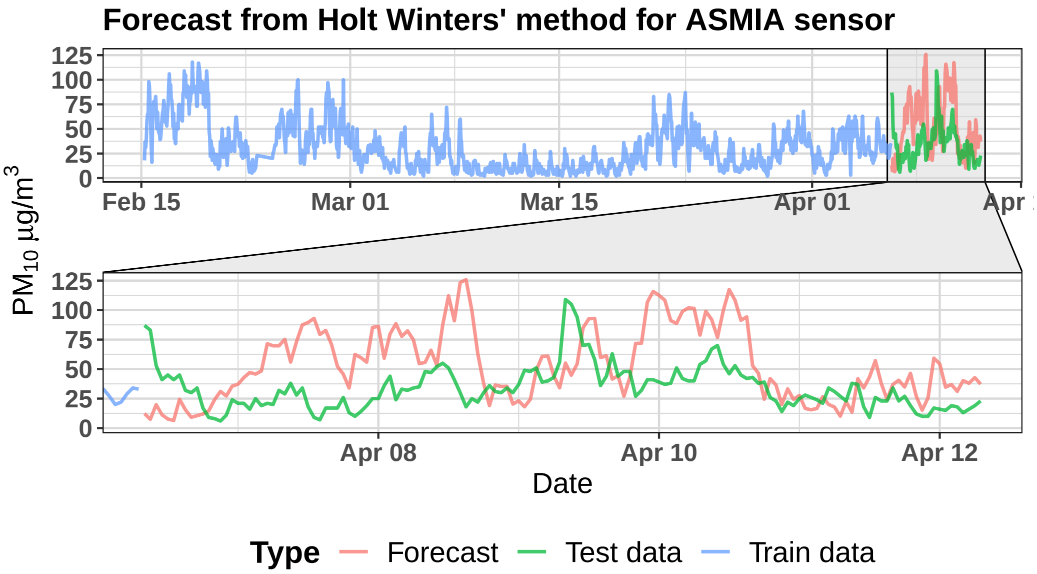

Figure 4.3: Forecast obtained by the Holt-Winters' double-seasonal method (ETS(A,A,N)) from data collected from the ASMIA sensor.

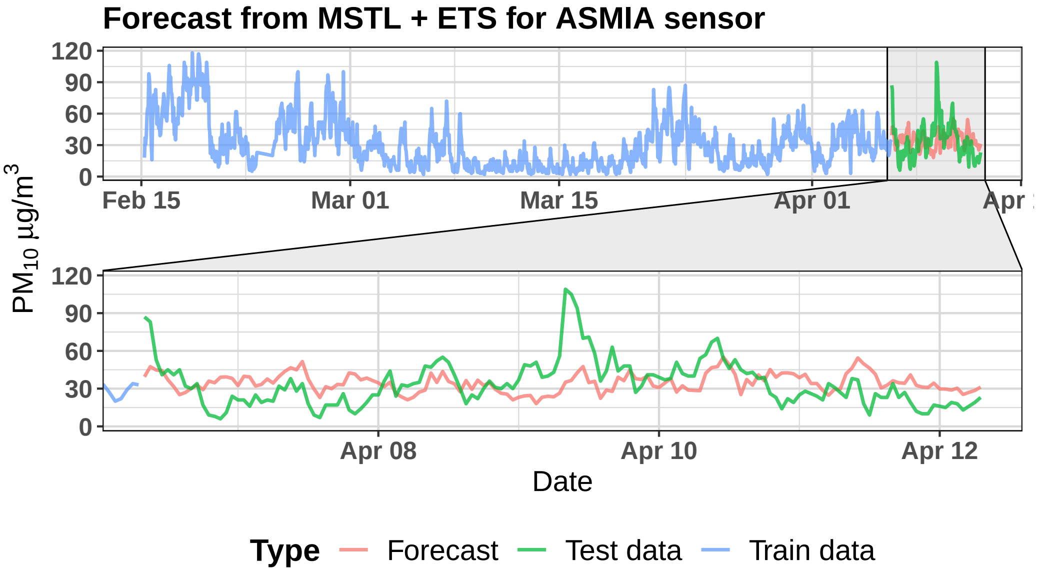

Figure 4.4: Forecast obtained by MSTL (ETS(A,N,N)) from data collected from the ASMIA sensor.

4.3 ARIMA and DHR

The ARIMA model encountered similar issues to those encountered by the simple Holt-Winters' method mentioned above – multiple seasonalities prevented the model from capturing single seasonal patterns. Due to this fact, only non-seasonal ARIMA models could be fitted to data which again resulted in mostly straight line forecasts. A forecast from ARIMA is shown in figure 4.5 for the ASMIA sensor (see appendix A, figures A.5, A.16, A.27 and A.38 for the rest of the sensors).

To better model seasonalities, we tried to improve our results by using dynamic harmonic regression with Fourier terms as predictors. We used 10 sine and cosine pairs of Fourier terms for $m_1 = 24$ and other 10 pairs for $m_2 = 168$. The error term was then modelled by ARIMA. The DHR forecast for the ASMIA sensor is shown in figure 4.6 (see appendix A, figures A.6, A.17, A.28 and A.39 for the rest of the sensors).

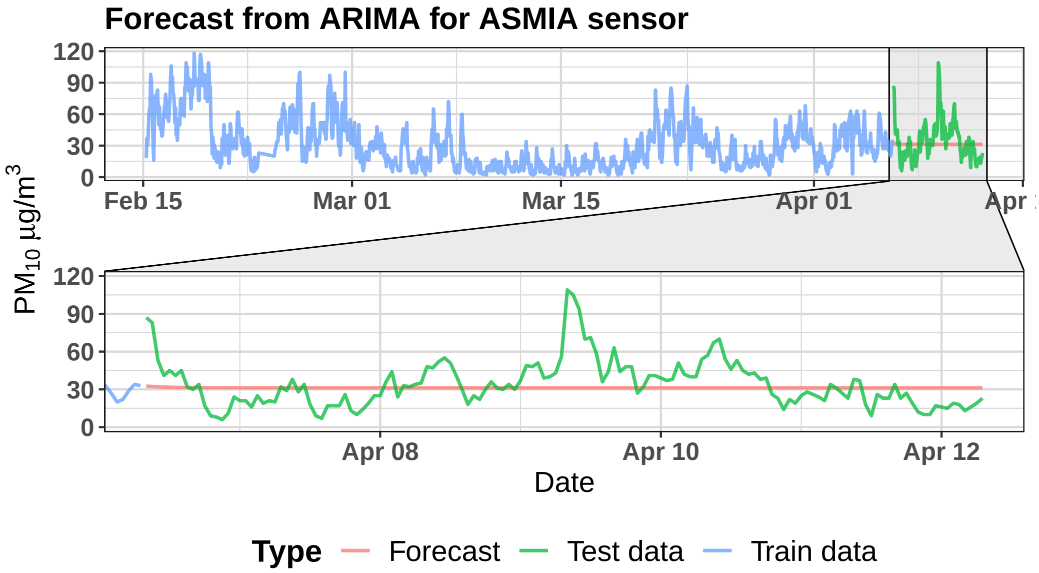

Figure 4.5: Forecast obtained by ARIMA (3,1,1) from data collected from the ASMIA sensor.

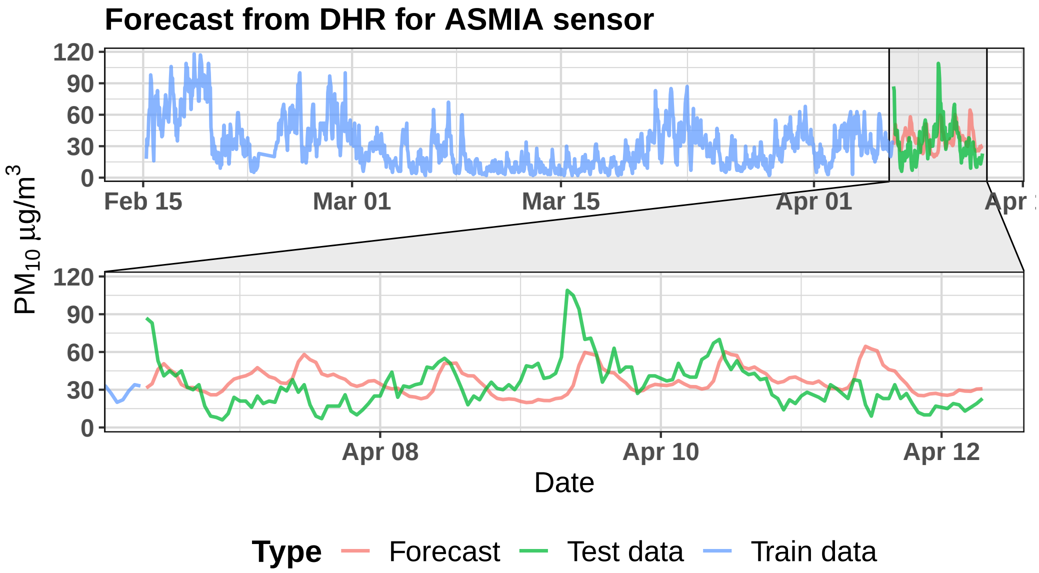

Figure 4.6: Forecast obtained by DHR (with ARIMA(1,1,2) errors) model from data collected from the ASMIA sensor.

4.4 BATS and TBATS

After the success with Fourier terms in DHR, we expected TBATS models to perform well. These models are generated automatically by a library in R with a particular model being selected according to AIC. For our data and forecasting conditions, models failed to identify the need for trigonometric seasonal patterns and kept producing BATS models, which once more resulted in mostly straight lines. TBATS model has been chosen only for data from the ASMIA sensor, which is shown in figure 4.7. Models for the rest of the sensors are shown in appendix A (figures A.7, A.18, A.29 and A.40).

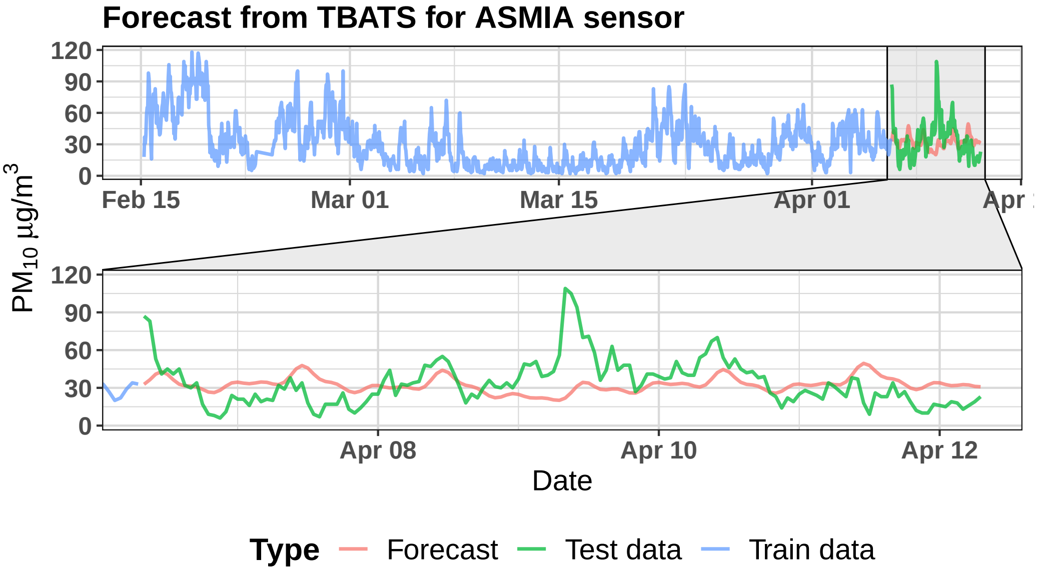

Figure 4.7: Forecast obtained by the BATS(0,399; {0,0}; 0,836; {[24,4]; [168,5]}) model from data collected from the ASMIA sensor.

4.5 NNAR

As with the previous method, parameters for NNAR are chosen automatically based on AIC. For the majority of these models only a set with length of about a day has been used as an input, which often resulted in forecasts diverging from the observed series after a few points. Despite that, for the very short horizon, this method managed to produce a reasonable approximation. A forecast from NNAR is shown in figure 4.8 for the ASMIA sensor (see appendix A, figures A.8, A.19, A.30 and A.41for the rest of the sensors).

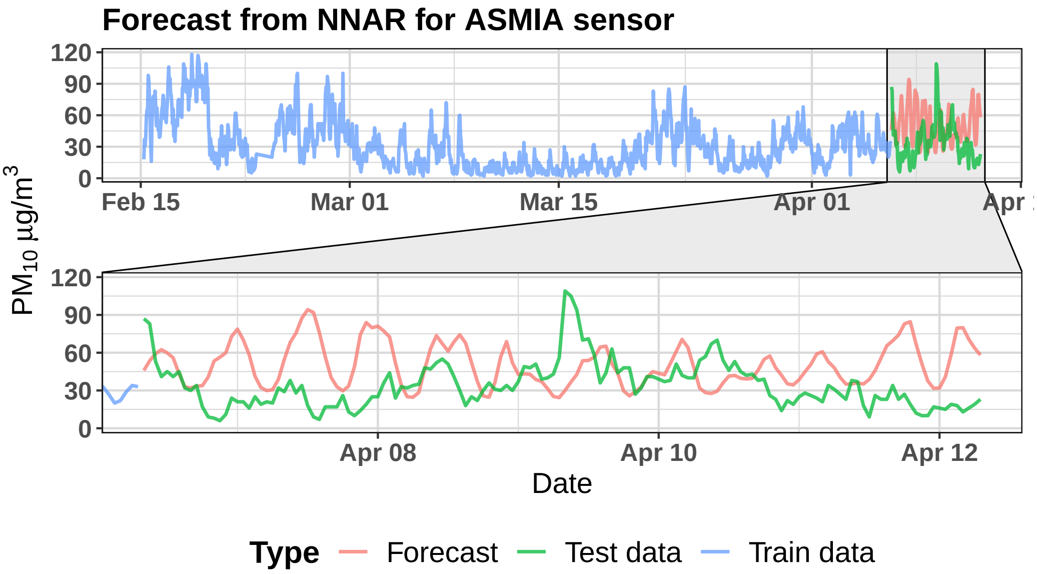

Figure 4.8: Forecast obtained by NNAR(26,1,114)$_{168}$ model from data collected from ASMIA sensor.

4.6 RPART and CTREE

The last experiment we tried was to use decision trees using both recursive partitioning and conditional inference. Both methods showed promising results for some sensors, but in the case of ALEGA sensor they slightly diverged from the observed data. A forecast from RPART is shown in figure 4.9 and a forecast from CTREE in figure 4.10 for the ASMIA sensor. For the rest of the sensors see appendix A (figures A.9, A.20, A.31 and A.42 for RPART and figures A.10, A.21, A.32 and A.43 for CTREE).

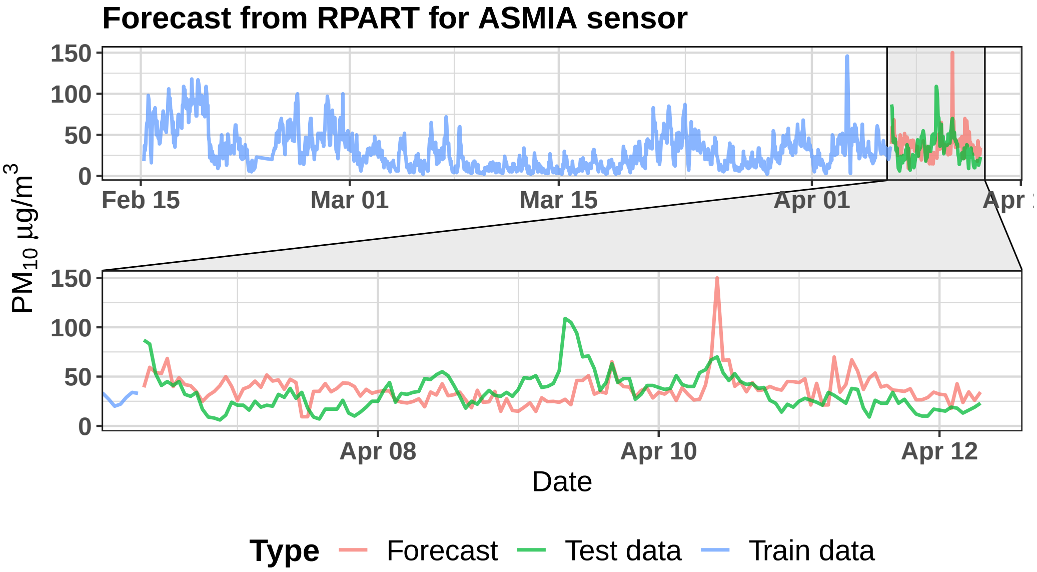

Figure 4.9: Forecast obtained by the RPART from data collected from the ASMIA sensor.

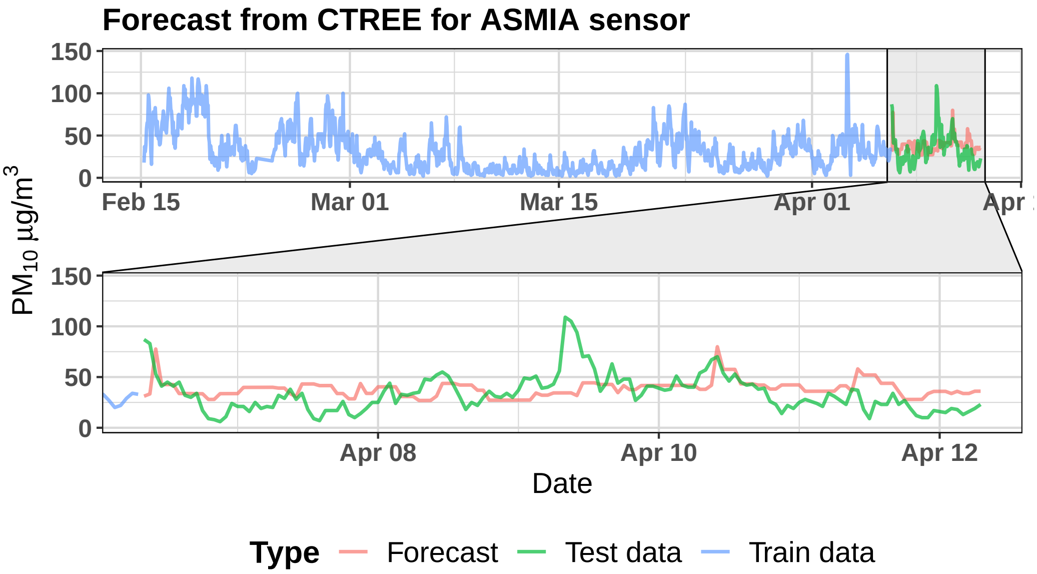

Figure 4.10: Forecast obtained by CTREE from data collected from the ASMIA sensor.

4.7 Comparing used models

All the metrics, which we considered for measuring accuracy, are dependent on the length of the compared time series. Due to the fact that the forecasting error increases with the length of the prediction period, we had chosen 3 different horizons for which we evaluated accuracy - short horizon (6h), medium horizon (24h) and long horizon (144h). Both MAE and RMSE are scale dependent measures, therefore they can be used only to compare models of the same TS – for the comparison between different TS we use MAPE. We have to consider that MAPE is an asymmetric measure, which means that it penalises under-forecasting more than over-forecasting, as well as the danger of division by zero when the observed value is 0. Lastly, it is also important to observe the shape of the forecasted series – sometimes the straight line is evaluated as the best model by error measures even though it fails to capture significant behaviour features of the observed data.

Accuracy measures for the short horizon forecasts for the ASMIA sensor are shown in table 4.2, for the rest of the sensors see appendix B (tables B.1, B.4, B.7 and B.10). For better clarity, we color-coded columns in each table to highlight the best models - greener values mean lower errors and redder values indicate higher errors. In case of the ASMIA sensor we can observe that all models except Holt-Winters' method managed to forecast the short horizon with relatively low errors. Across the sensors, MSTL and DHR models provide good results in every case while both decision trees produced higher errors.

The medium horizon measures are shown in table 4.3 for the ASMIA sensor, for the rest of the sensors see appendix B (tables B.2, B.5, B.8 and B.11). In this case the models that forecasted straight lines, namely Holt's linear trend, ARIMA and TBATS, generated low errors for all sensors. As in the previous scenario, MSTL and DHR managed to create solid forecasts and decision trees (RPART and CTREE) are among the worst models in every case.

	MAE	RMSE	MAPE
HOLT	25.29	25.29	37.54
HW	47.48	47.48	79.82
ARIMA	26.32	26.32	39.66
MSTL	18.65	17.54	26.91
DHR	20.54	16.35	27.56
TBATS	20.59	19.96	27.74
NNAR	15.20	13.03	20.60
RPART	18.22	5.98	28.34
CTREE	22.63	13.34	30.62

Table 4.2: Forecast accuracy metrics for the 6h horizon for data from the ASMIA sensor.

	MAE	RMSE	MAPE
HOLT	15.97	2.26	87.38
HW	29.30	1.64	118.47
ARIMA	15.24	0.64	80.64
MSTL	16.38	5.17	98.58
DHR	16.80	6.28	88.34
TBATS	14.00	2.54	74.13
NNAR	20.49	13.32	131.08
RPART	18.23	12.18	109.68
CTREE	16.89	6.92	90.48

Table 4.3: Forecast accuracy metrics for the 24h horizon for data from the ASMIA sensor.

	MAE	RMSE	MAPE
HOLT	13.98	0.28	59.32
HW	30.73	19.56	122.49
ARIMA	13.82	2.71	54.61
MSTL	15.14	0.66	65.14
DHR	15.52	2.83	68.25
TBATS	14.64	1.57	59.24
NNAR	25.52	16.25	121.79
RPART	16.60	3.02	70.63
CTREE	14.81	4.31	67.05

Table 4.4: Forecast accuracy metrics for the 144h horizon for data from the ASMIA sensor.

The long horizon accuracy measures are shown in table 4.4 for the ASMIA sensor, for the rest of the sensors see appendix B (tables B.3, B.6, B.9 and B.12). DHR is again among the best models for long-term forecasting joined with Holt's linear model and ARIMA. Regression trees (RPART and CTREE) performed slightly better than in the short horizon and the medium horizon. The NNAR model diverged significantly from the observed data for all sensors and produced higher errors.

In order to better visualise differences between the performance of different models, we created a plot of their MAPE for all three horizons. This plot for the ASMIA sensor is shown in figure 4.11 (see appendix A, figures A.11, A.22, A.33 and A.44 for the rest of the models).

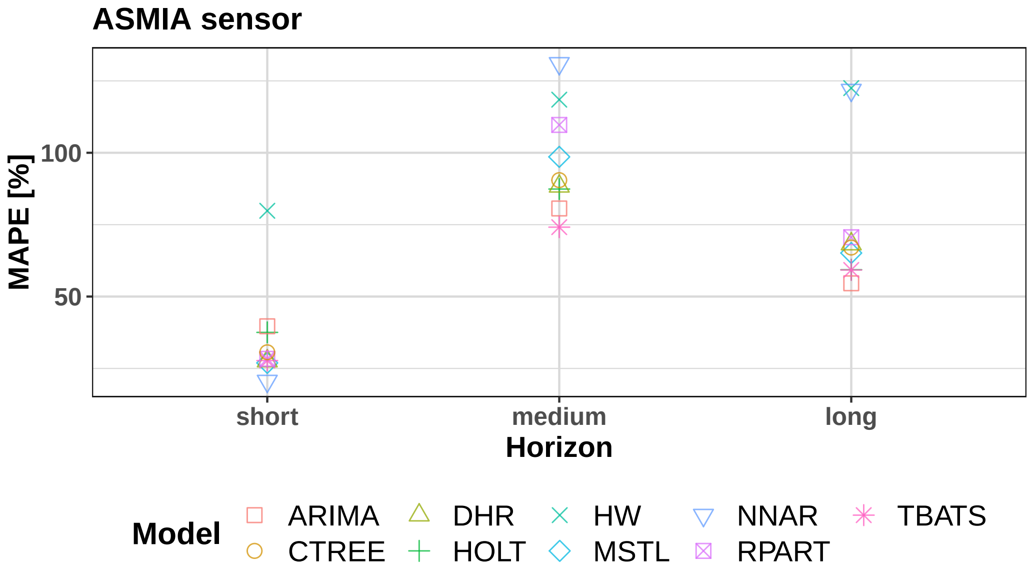

Figure 4.11: Values of the MAPE for all models created from data collected from the ASMIA sensor for 3 different horizons.

The best average MAPE (see table 4.5) was achieved by the Holt's linear method. This was caused mainly by the fact, that this model had low MAPE in both medium and long horizons for each sensor. However, Holt's linear model does not capture any seasonal patterns and fluctuations in TS – if we want to reflect these, we need to choose another model. The next best models were MSTL and DHR with close average MAPE. For these two methods we created a plot of their MAPE across the sensors for all three horizons. These plots are shown in the figure for MSTL and in the figure for DHR.

Sensor	MAPE [%]
Holt	45.65
HW	79.19
ARIMA	57.48
MSTL	55.44
DHR	49.03
TBATS	58.62
NNAR	103.24
RPART	99.16
CTREE	104.94

Table 4.5: The average value of MAPE calculated across all sensors and all 3 horizons.

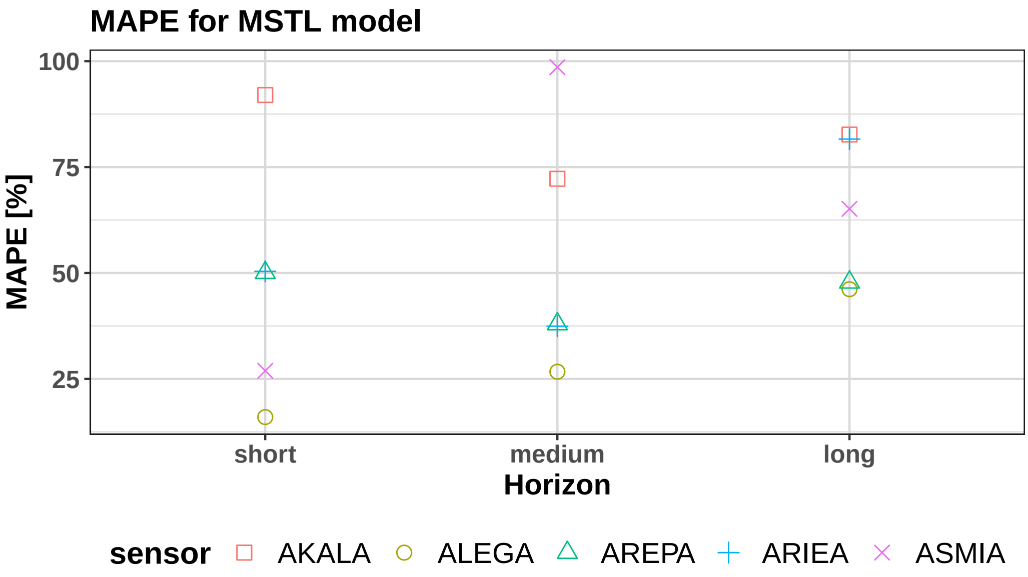

Figure 4.12: Value of the MAPE of MSTL models for 3 horizons across the sensors.

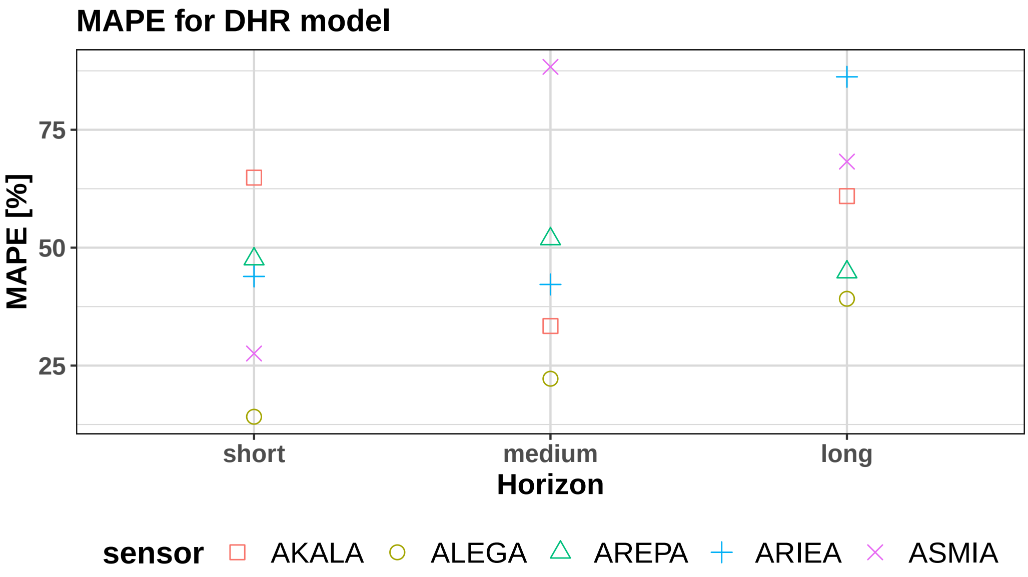

Figure 4.13: Value of the MAPE of DHR models for 3 horizons across the sensors.

Conclusion

In this thesis, we outlined the main issues regarding air pollution, its measurements and prediction. We named the main substances responsible for air pollution, sources of their emission and their effects on humans and the environment. Further, we researched emerging low-cost sensors used for detection of the pollutants as well as explored larger projects focused on monitoring the climate. In chapter 3, we explained statistical principles of forecasting univariate time series and how we can leverage these for prediction of the evolution of air pollution.

In the final part, we constructed several models of the concentration of PM_{10} pollutant based on data collected from 5 different measuring stations in Prague. These models were then used to produce forecasts of pollution for the next 6, 24 and 144 hours and their accuracy was calculated. After the evaluation, we concluded that all models except for RPART and CTREE were capable to predict the 6-hour horizon very well. For the longer horizons, the best method based on MAPE is Holt's linear trend which models the mean of the time series. However, this model does not include any information about seasonal patterns and random fluctuations and fails to predict peaks in the concentration of PM_{10}. Due to this, we recommend using MSTL or DHR models, which consistently produced forecasts with relatively small errors. These observations were made for the Prague area – depending on the location and severity of pollution other models could perform better. Although we forecasted only the PM_{10}, methods that we used are not constricted to this pollutant and can be used for any univariate discrete set of data.

We hope that this thesis will contribute to efforts to provide better information about air pollution and climate change. It provides foundations for the development of automatic an modelling system, that will be used to predict potentially dangerous air pollution situations.

Bibliography

[1] WHO. Ambient (outdoor) air pollution. *World Health Organization fact sheets*, 2018. URL `https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health`. accessed on 30/07/2020.

[2] Cedric D. Koolen and Gadi Rothenberg. Air pollution in europe. *ChemSusChem*, 12(1):164–172, dec 2018. doi: 10.1002/cssc.201802292. accessed on 30/07/2020.

[3] MŽP. Amoniak. *Integrovaný Registr Znečišťování*, 2019. URL `https://www.irz.cz/sites/default/files/latky/Amoniak_Karta_latky_11012019.pdf`. accessed on 30/07/2020.

[4] EEA. Nitrogen oxides (nox) emissions. *European Environment Agency*, 2018. URL `https://www.eea.europa.eu/data-and-maps/indicators/eea-32-nitrogen-oxides-nox-emissions-1`. accessed on 30/07/2020.

[5] EEA. Non-methane volatile organic compounds (nmvoc) emissions. *European Environment Agency*, 2015. URL `https://www.eea.europa.eu/data-and-maps/indicators/eea-32-non-methane-volatile-1`. accessed on 30/07/2020.

[6] Sef van den Elshout, Karine Léger, and Hermann Heich. CAQI common air quality index — update with PM2.5 and sensitivity analysis. *Science of The Total Environment*, 488-489:461–468, aug 2014. doi: 10.1016/j.scitotenv.2013.10.060. accessed on 30/07/2020.

[7] Petra Bauerová and Josef Keder. Hodnocení testovacího měření různých typů malých senzorů kvality ovzduší na observatoři tušimice. *ČHMÚ*, 2019. URL `http://portal.chmi.cz/files/portal/docs/tiskove_zpravy/2019/Testovani_malych_senzoru_OBT_zprava_MZP_FINAL.pdf`. accessed on 30/07/2020.

[8] C. Borrego, A.M. Costa, J. Ginja, M. Amorim, M. Coutinho, K. Karatzas, Th. Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, P. Smith, N. André, P. Gérard, L.A. Francis, N. Castell, P. Schneider, M. Viana, M.C. Minguillón, W. Reimringer, R.P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, and M. Penza. Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. *Atmospheric Environment*, 147:246–263, dec 2016. doi: 10.1016/j.atmosenv.2016.09.050.

[9] Michel Gerboles, Laurent Spinelle, and Annette Borowiak. Measuring air pollution with low-cost sensors. *The European Commission's science and knowledge service*, 2017. URL `https://ec.europa.eu/jrc/en/publication/brochures-leaflets/measuring-air-pollution-low-cost-sensors`. accessed on 30/07/2020.

[10] M.I. Mead, O.A.M. Popoola, G.B. Stewart, P. Landshoff, M. Calleja, M. Hayes, J.J. Baldovi, M.W. McLeod, T.F. Hodgson, J. Dicks, A. Lewis, J. Cohen, R. Baron, J.R. Saffell, and R.L. Jones. The use of electrochemical sensors for monitoring urban air quality in low-cost, high-density networks. *Atmospheric Environment*, 70:186–203, may 2013. doi: 10.1016/j.atmosenv.2012.11.060.

[11] Philip Peterson, Amrita Aujla, Kirsty Grant, Alex Brundle, Martin Thompson, Josh Vande Hey, and Roland Leigh. Practical use of metal oxide semiconductor gas sensors for measuring nitrogen dioxide and ozone in urban environments. *Sensors*, 17(7):1653, jul 2017. doi: 10.3390/s17071653.

[12] Alastair C. Lewis, Erika von Schneidemesser, and Richard E. Peltier. Low-cost sensors for the measurement of atmospheric composition: overview of topic and future applications. *World Meteorological Organization*, 2018. URL `https://library.wmo.int/doc_num.php?explnum_id=9881`. accessed on 30/07/2020.

[13] Jiří Anděl. *Statistické metody.* Matfyzpress, Praha, 2007. ISBN 8073780038.

[14] Milan Meloun. *Kompendium statistického zpracování dat.* Karolinum, Praha, 2012. ISBN 9788024621968.

[15] Marie Forbelská. *Stochastické modelování jednorozměrných časových řad.* Masarykova univerzita, Brno, 2009. ISBN 9788021048126.

[16] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 2006. URL `https://www.sciencedirect.com/science/article/abs/pii/S0169207006000239?via=ihub`. accessed on 18/07/2020.

[17] Rob J. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice.* OTexts: Melbourne, Australia, 2018. URL `https://otexts.com/fpp2/`. accessed on 18/07/2020.

[18] Additive TS dataset, Monthly births in NY. `https://robjhyndman.com/tsdldata/data/nybirths.dat`.

[19] Multiplicative TS dataset, Airline passengers. `https://raw.githubusercontent.com/jbrownlee/Datasets/master/airline-passengers.csv`.

[20] Josef Arlt and Markéta Arltová. *Ekonomické časové řady.* Professional Publishing, Praha, 2009. ISBN 9788086946856.

[21] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 1990. URL `https://www.wessa.net/download/stl.pdf`. accessed on 19/07/2020.

[22] R. J. Hyndman. Box-Jenkins modelling. *Informed Student Guide to Management Science*, 2001. URL `https://robjhyndman.com/papers/BoxJenkins.pdf`. accessed on 18/07/2020.

[23] Marie Forbelská. Náhodné procesy II. *Skriptum*, 2005. URL `https://is.muni.cz/el/1431/jaro2006/M0122/M0122.pdf`. accessed on 18/07/2020.

[24] G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 1964. URL `http://links.jstor.org/sici?sici=0035-9246(1964)26:2<211:AAOT>2.0.CO;2-6`. accessed on 25/07/2020.

[25] A. M. De Livera, R. J. Hyndman, and R. D: Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Department of Econometrics and Business Statistics Working Paper*, 2010. URL `https://robjhyndman.com/papers/ComplexSeasonality.pdf`. accessed on 25/07/2020.

[26] A. Maleki, S. Nasseri, M. S. Aminabad, and M. Hadi. Comparison of ARIMA and NNAR models for forecasting water treatment plant's influent characteristics. *KSCE Journal of Civil Engineering*, 2018. doi: 10.1007/s12205-018-1195-z. URL `https://www.researchgate.net/publication/324525859_Comparison_of_ARIMA_and_NNAR_Models_for_Forecasting_Water_Treatment_Plant's_Influent_Characteristics`. accessed on 26/07/2020.

[27] X. Wu and V. Kumar et al. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 2007. doi: 10.1007/s10115-007-0114-2. URL `https://www.researchgate.net/publication/29467751_Top_10_algorithms_in_data_mining`. accessed on 26/07/2020.

[28] P. Laurinec. Using regression trees for forecasting double-seasonal time series with trend in r. 2017. URL `https://petolau.github.io/Regression-trees-for-forecasting-time-series-in-R/`. accessed on 26/07/2020.

[29] J. Zídek. Dolování znalostí z bezdrátových senzorových sítí. Master's thesis, České vysoké učení technické v Praze, 2019.

List of Abbreviations

ACF	autocorrelation function
ARIMA	autoregressive integrated moving average
BATS	Box-Cox transform, ARMA errors, trend, and seasonal
CAQI	common air quality index
CF	covariance function
CHMI	Czech Hydrometeorological Institute
CTREE	conditional inference trees
DHR	dynamic harmonic regression
ETS	error, trend, seasonal
EU	European Union
HW	Holt-Winters' seasonal method
LED	light-emitting diode
LOESS	locally estimated scatterplot smoothing
MAE	mean absolute error
MAPE	mean absolute percentage error
MODIS	Moderate Resolution Imaging Spectroradiometer
MOPITT	Measurement of Pollution in the Troposphere
MOS	metal oxide semiconductor
MSTL	multiseasonal and trend decomposition using LOESS
NMVOC	non-methane volatile organic compounds
NNAR	neural network auto-regressive
PACF	partial autocorrelation function
RMSE	root mean square error
RPART	recursive partitioning
SES	simple exponential smoothing
STL	seasonal and trend decomposition using LOESS
TBATS	trigonometric BATS
TS	time series
UV	ultra violet
WHO	World Health Organization
WN	white noise

A. Figures

A.1 Models for AKALA sensor

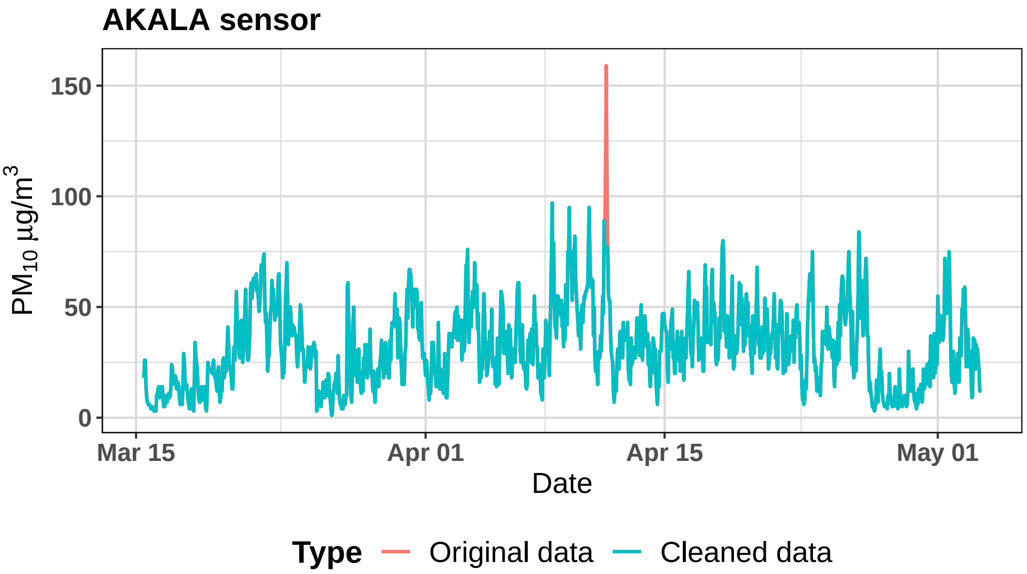

Figure A.1: Hourly concentration of PM$_{10}$ collected by AKALA sensor between March and May 2019 and data with cleaned outliers.

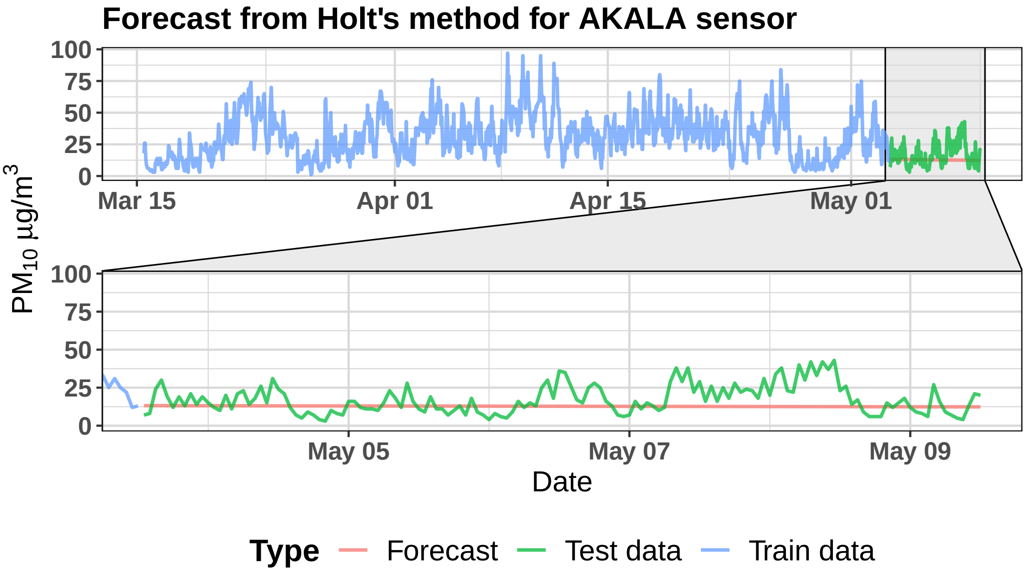

Figure A.2: Forecast obtained by Holt's linear model (ETS(A,A,N)) from data collected from AKALA sensor.

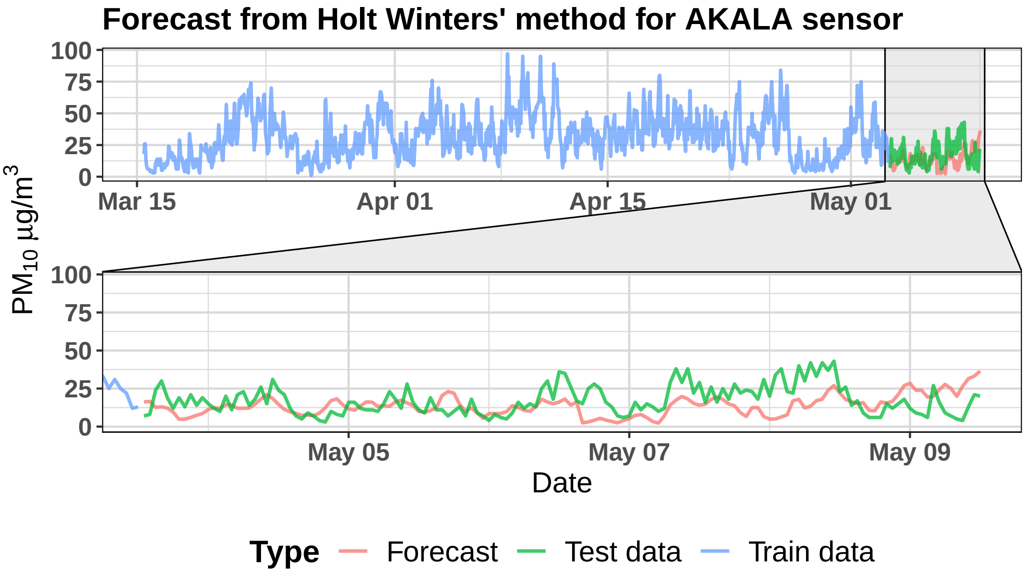

Figure A.3: Forecast obtained by Holt-Winters' double seasonal method (ETS(A,A,N)) from data collected from AKALA sensor.

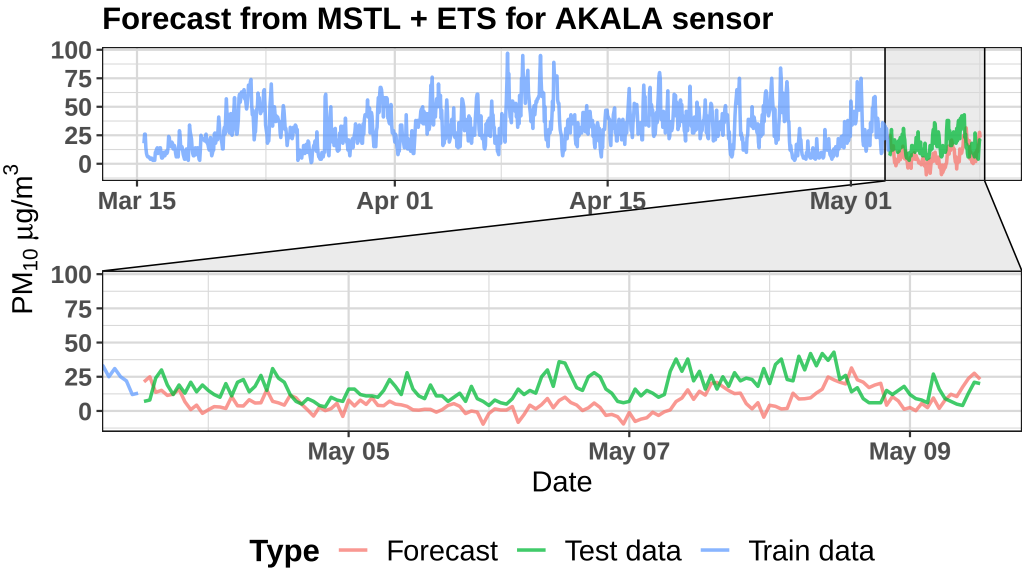

Figure A.4: Forecast obtained by MSTL (ETS(A,N,N)) model from data collected from AKALA sensor.

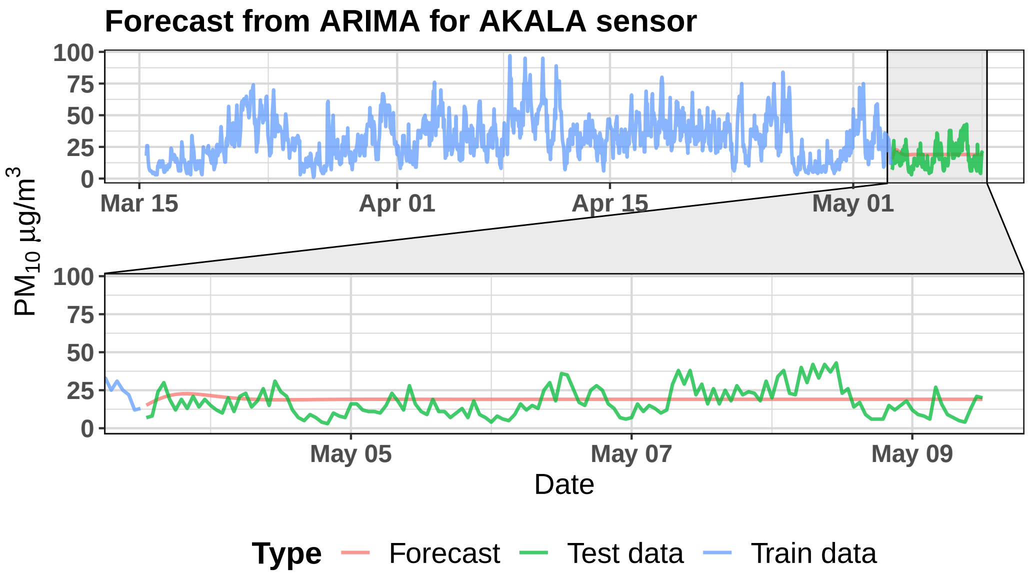

Figure A.5: Forecast obtained by ARIMA(2,1,2) model from data collected from AKALA sensor.

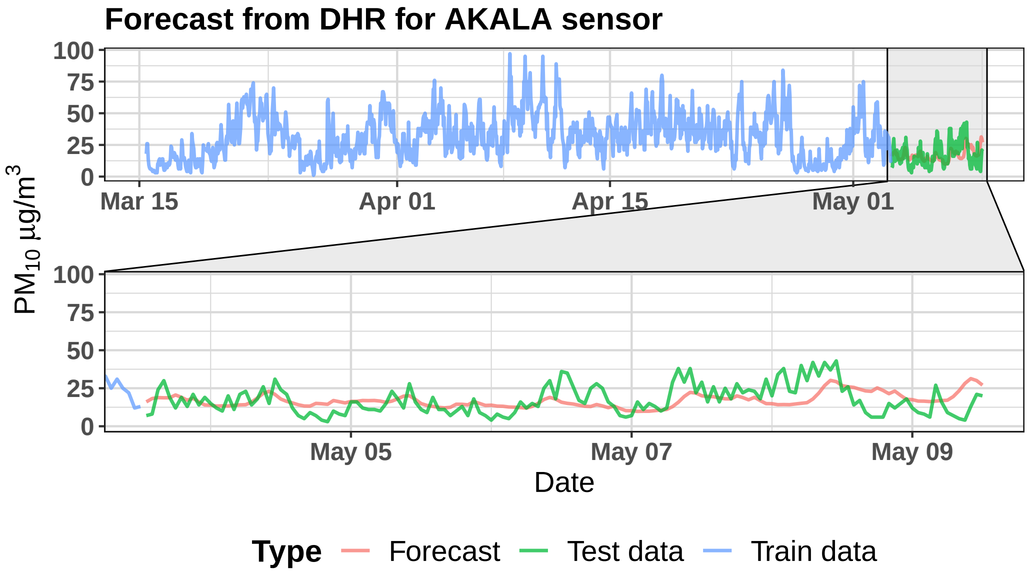

Figure A.6: Forecast obtained by DHR (with ARIMA(3,1,1) errors) model from data collected from AKALA sensor.

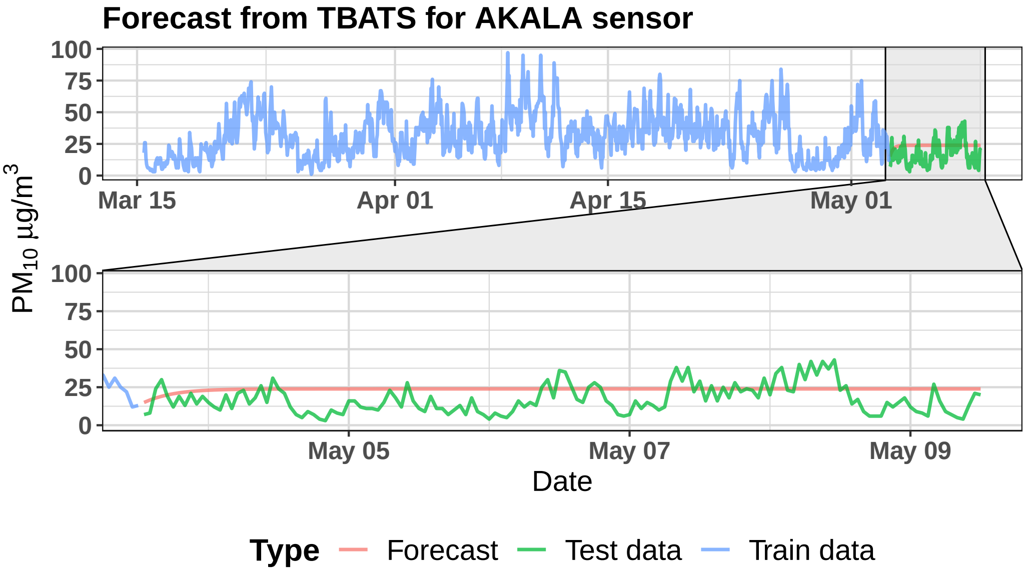

Figure A.7: Forecast obtained by BATS(0,416; {0,0}; 0,8; -) model from data collected from AKALA sensor.

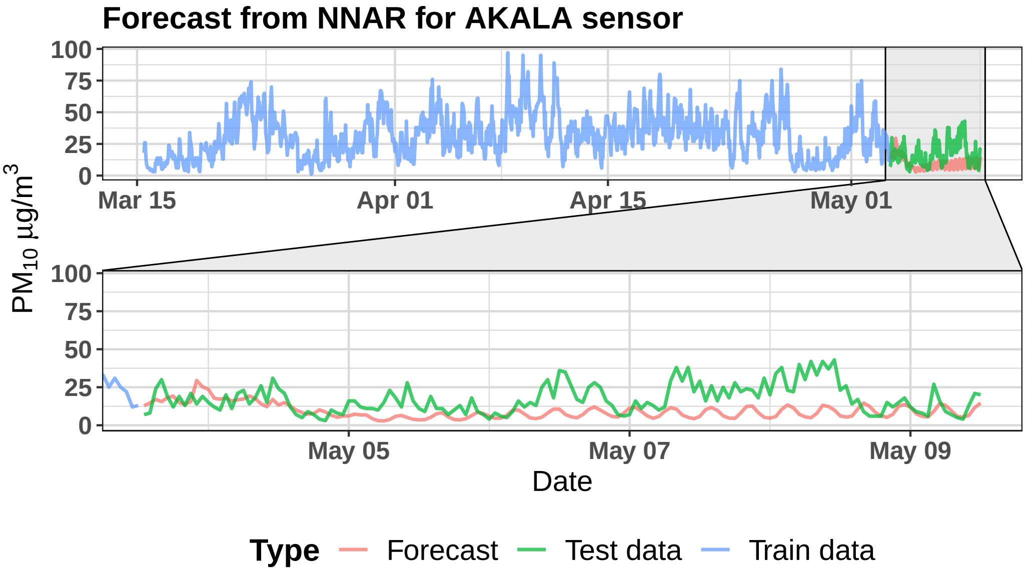

Figure A.8: Forecast obtained by NNAR(30,1,16)$_{168}$ model from data collected from AKALA sensor.

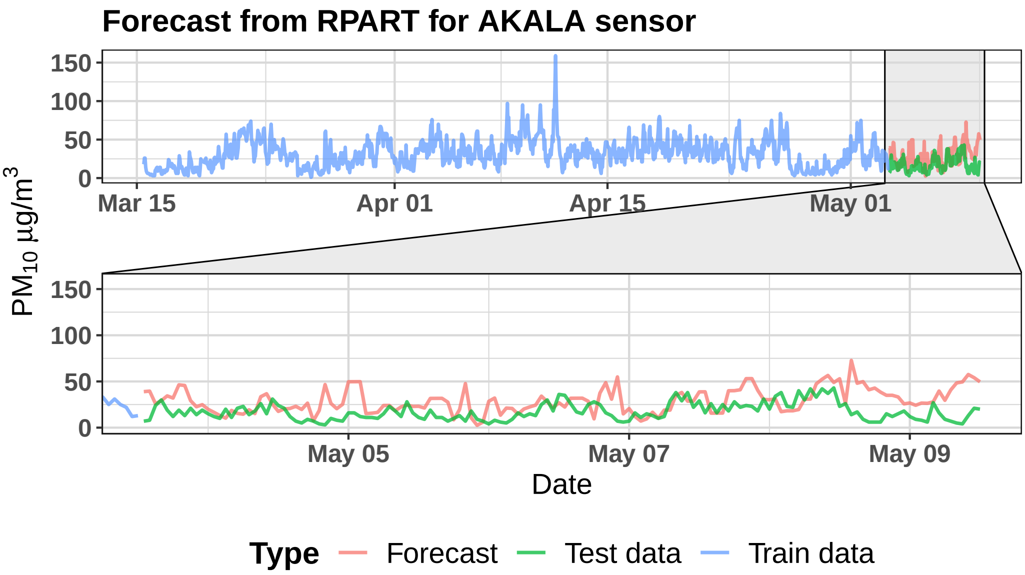

Figure A.9: Forecast obtained by recursive partitioning tree from data collected from AKALA sensor.

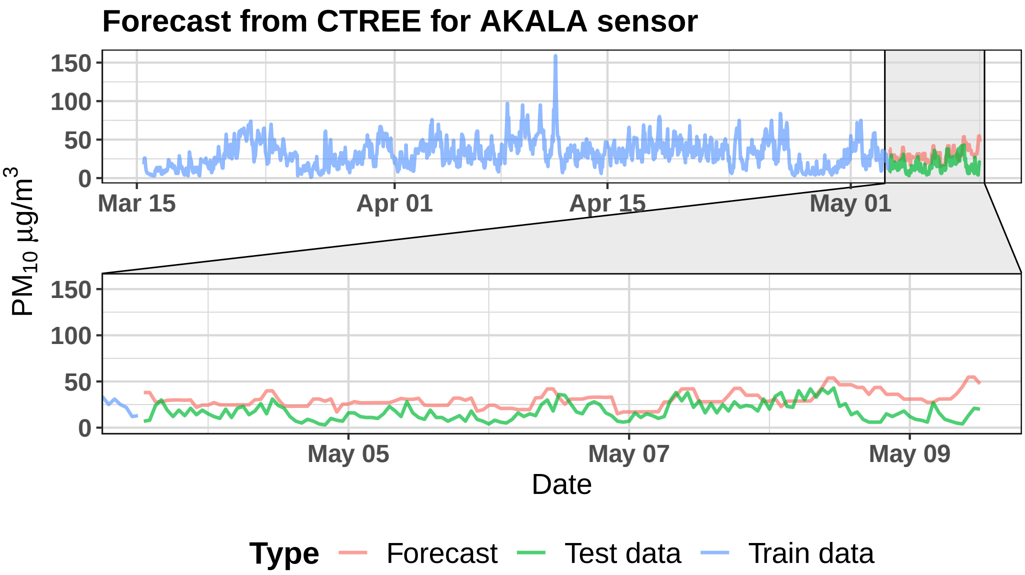

Figure A.10: Forecast obtained by conditional inference tree from data collected from AKALA sensor.

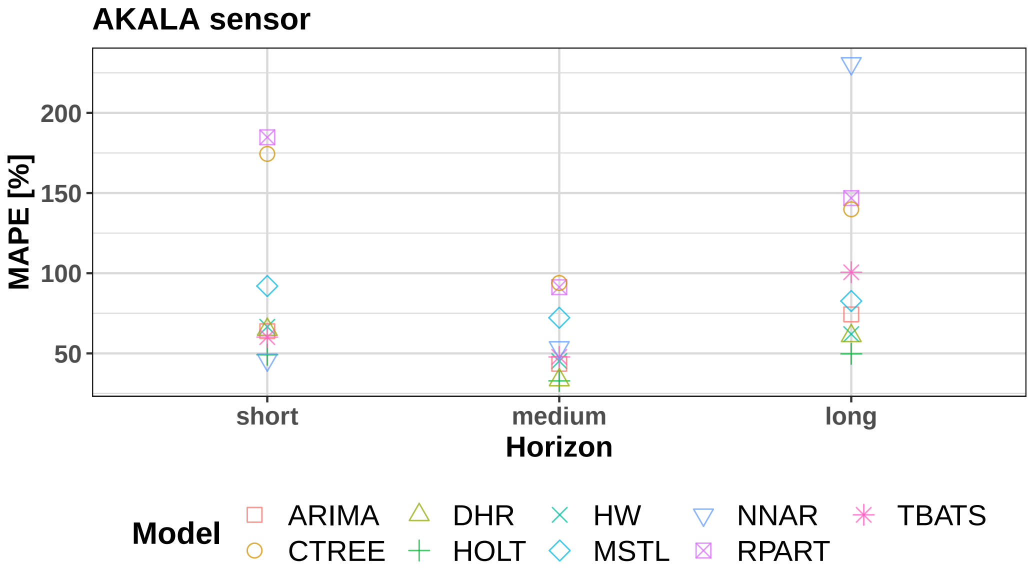

Figure A.11: Values of the MAPE for all models created from data collected from the AKALA sensor for 3 different horizons.

A.2 Models for ALEGA sensor

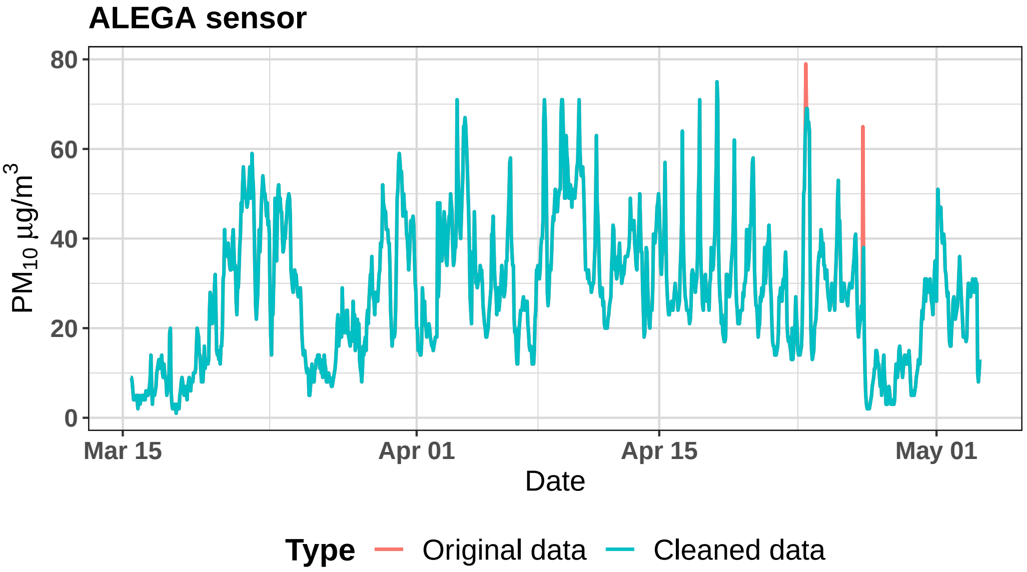

Figure A.12: Hourly concentration of PM_{10} collected by ALEGA sensor between March and May 2019 and data with cleaned outliers.

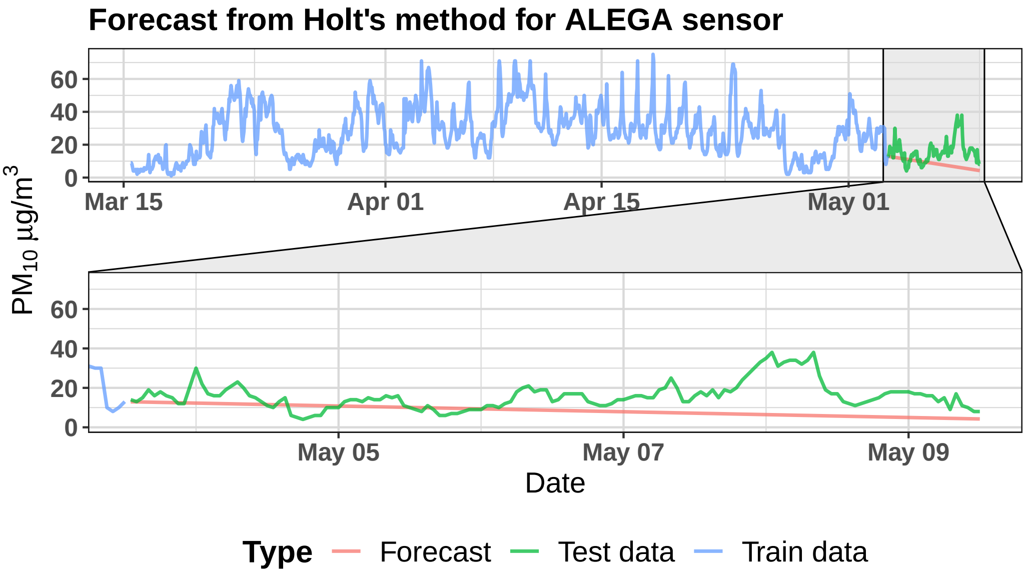

Figure A.13: Forecast obtained by Holt's linear model (ETS(A,A,N)) from data collected from ALEGA sensor.

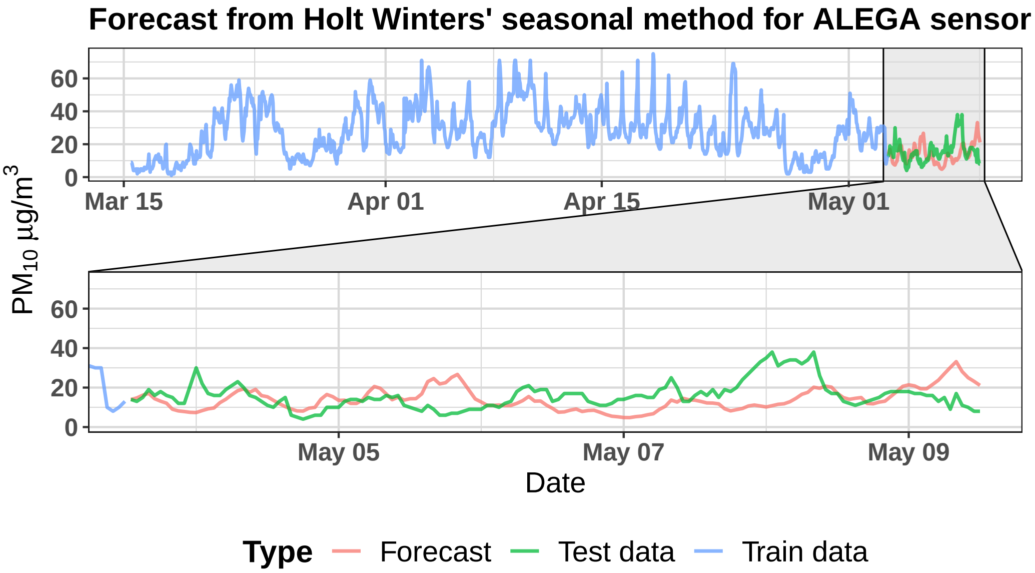

Figure A.14: Forecast obtained by Holt-Winters' double seasonal method (ETS(A,A,N)) from data collected from ALEGA sensor.

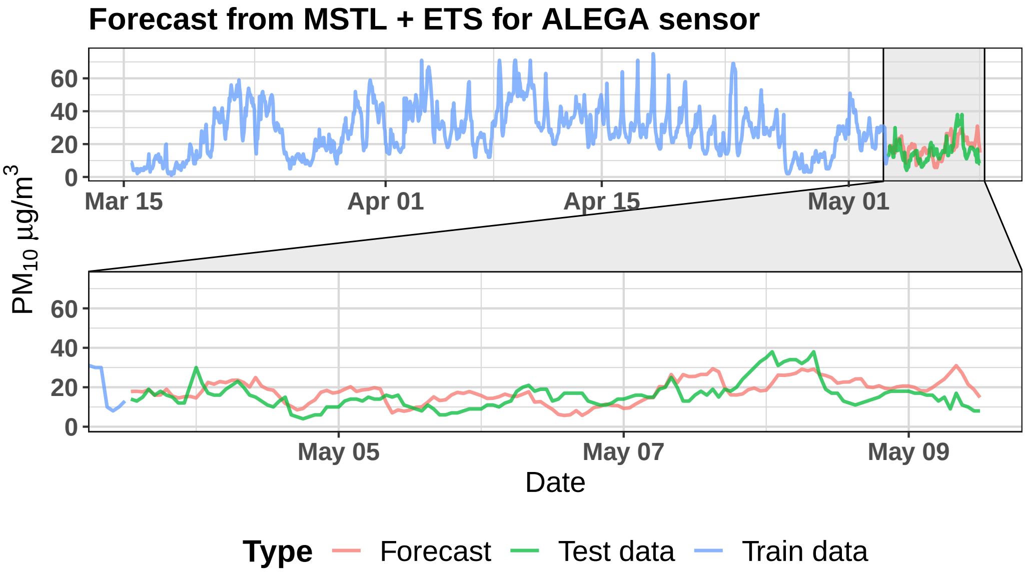

Figure A.15: Forecast obtained by MSTL (ETS(A,N,N)) model from data collected from ALEGA sensor.

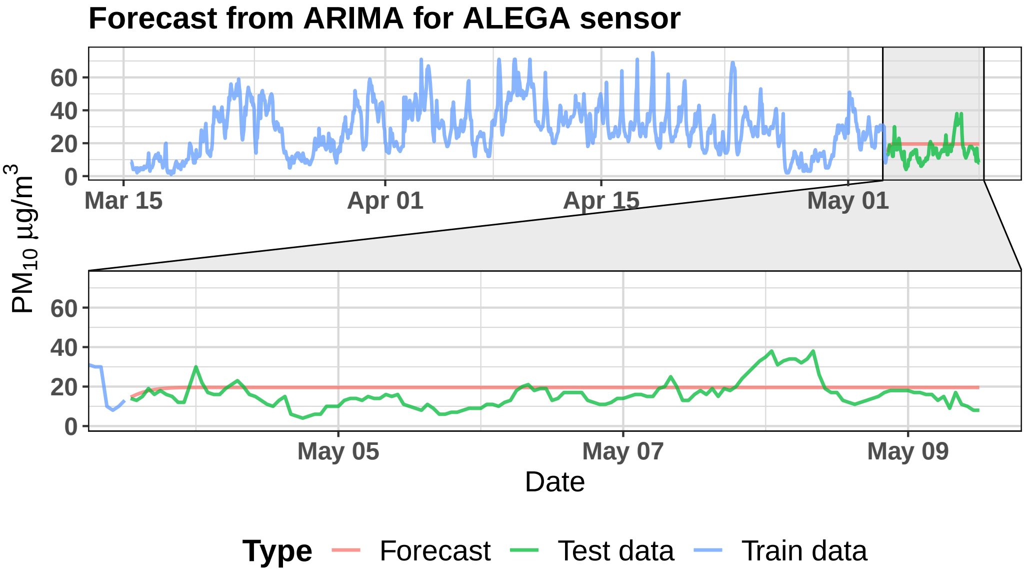

Figure A.16: Forecast obtained by ARIMA(2,1,2) model from data collected from ALEGA sensor.

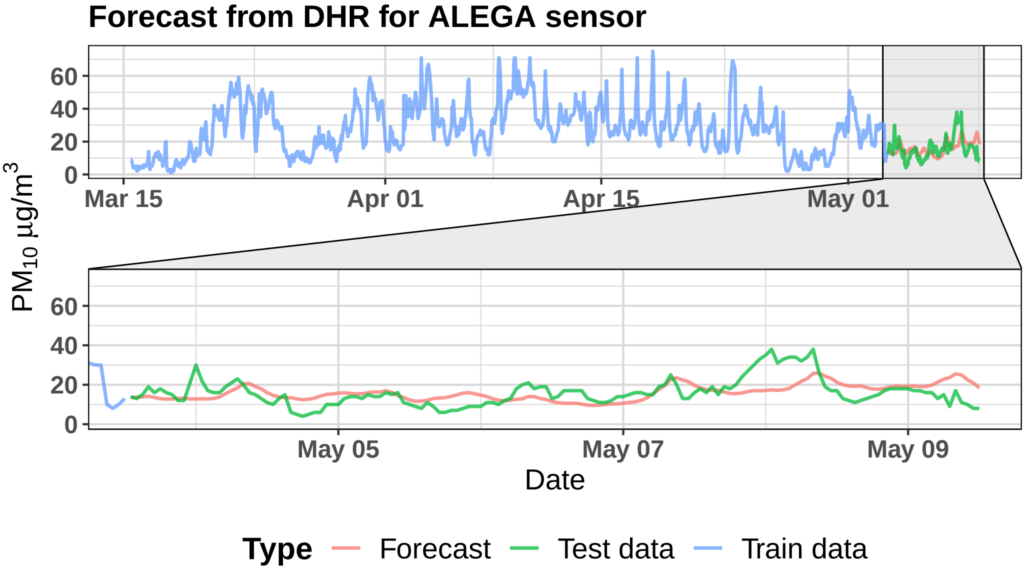

Figure A.17: Forecast obtained by DHR (with ARIMA(2,1,1) errors) model from data collected from ALEGA sensor.

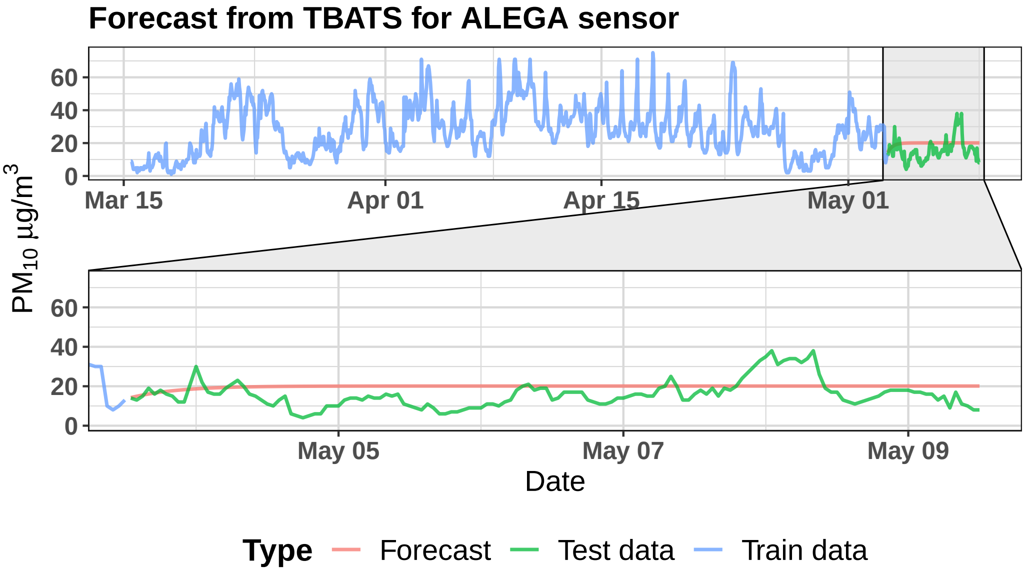

Figure A.18: Forecast obtained by BATS(0,388; {0,0}; 0,873; -) model from data collected from ALEGA sensor.

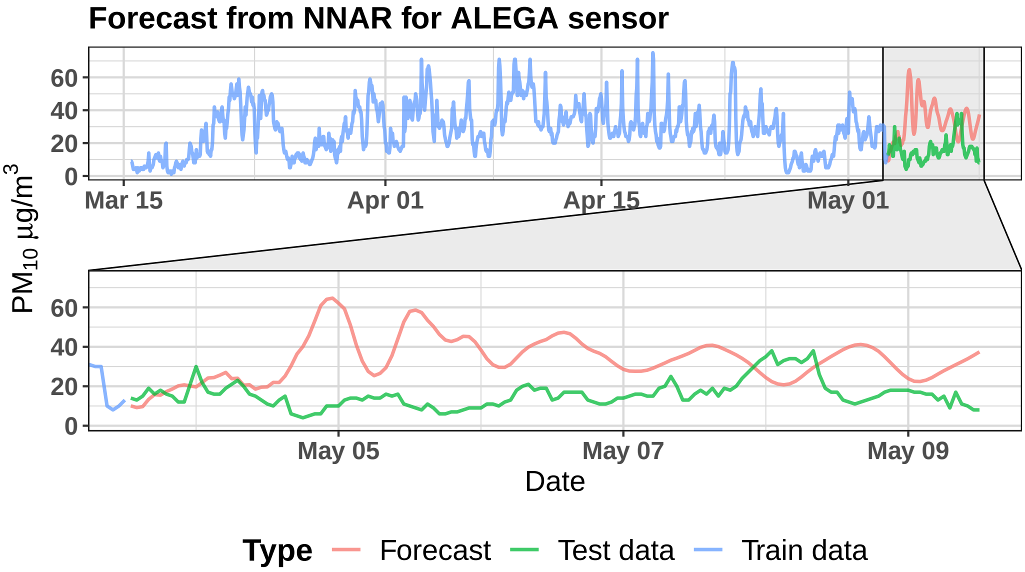

Figure A.19: Forecast obtained by NNAR(28,1,15)$_{168}$ model from data collected from ALEGA sensor.

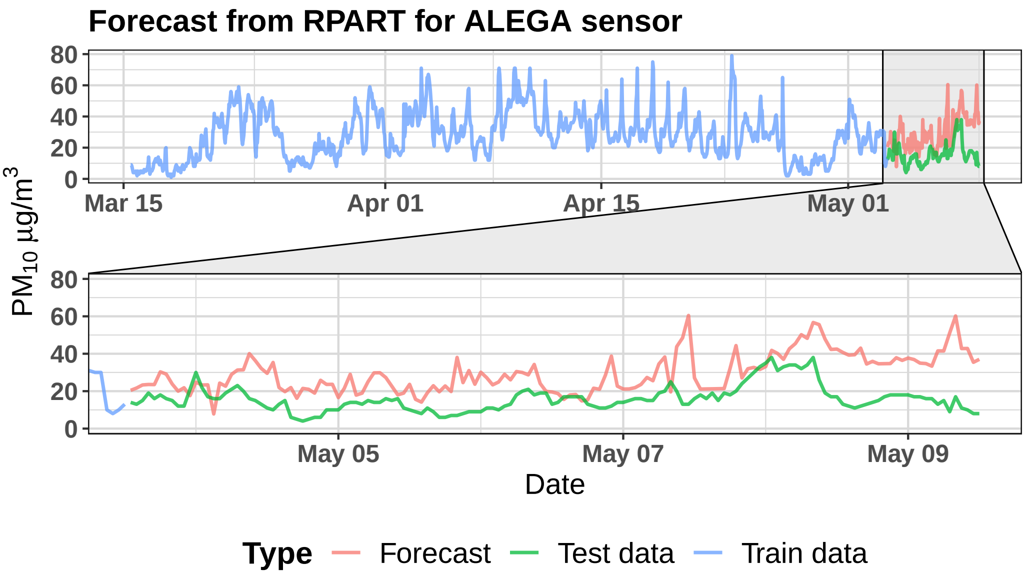

Figure A.20: Forecast obtained by recursive partitioning tree from data collected from ALEGA sensor.

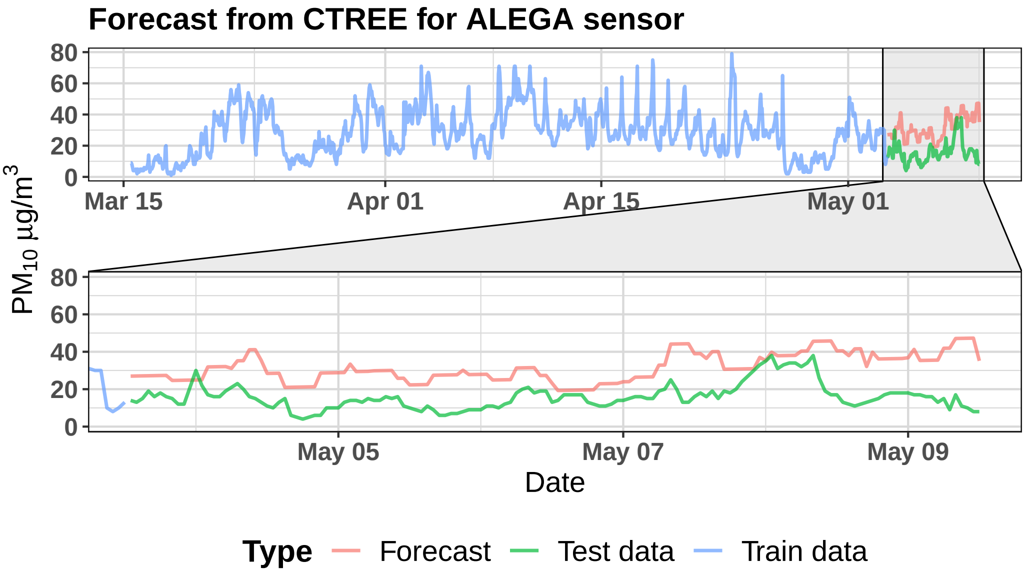

Figure A.21: Forecast obtained by conditional inference tree from data collected from ALEGA sensor.

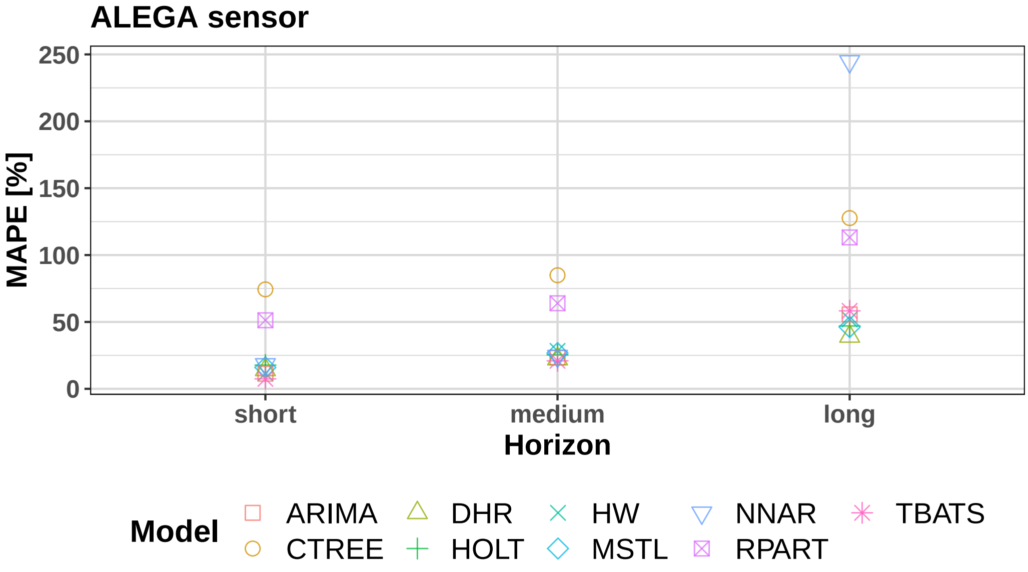

Figure A.22: Values of the MAPE for all models created from data collected from the ALEGA sensor for 3 different horizons.

A.3 Models for AREPA sensor

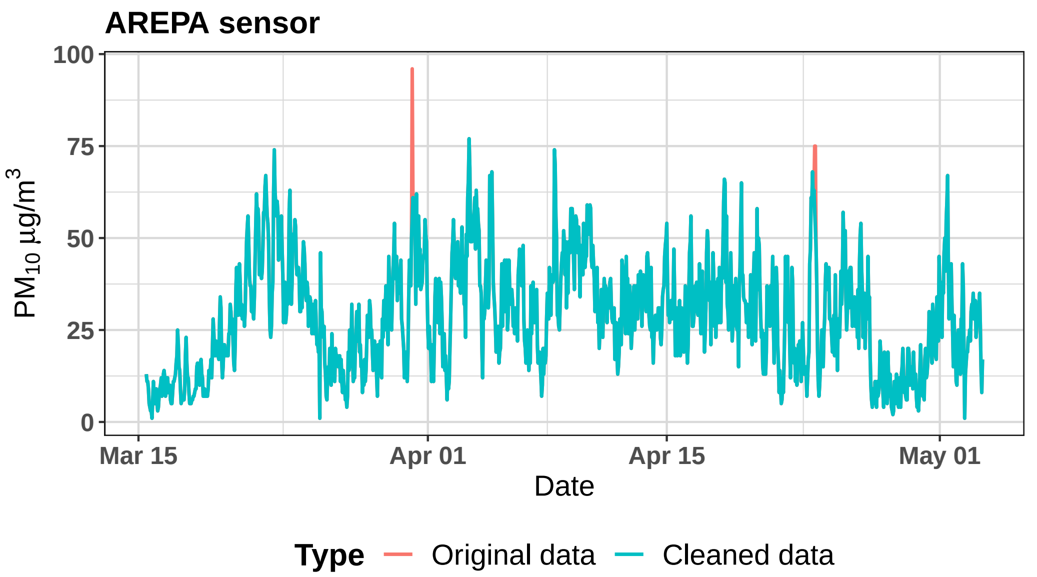

Figure A.23: Hourly concentration of PM$_{10}$ collected by AREPA sensor between March and May 2019 and data with cleaned outliers.

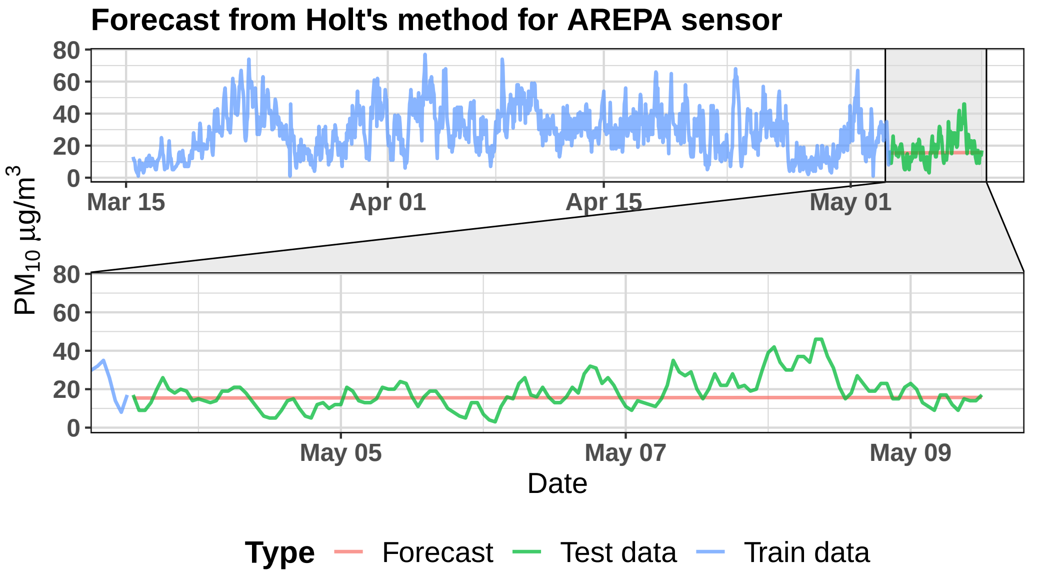

Figure A.24: Forecast obtained by Holt's linear model (ETS(A,A,N)) from data collected from AREPA sensor.

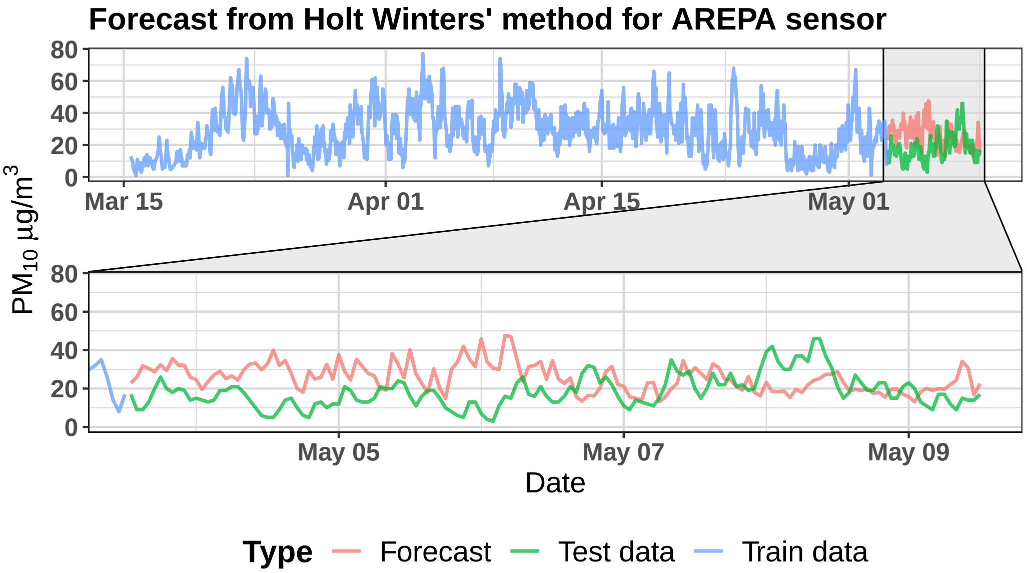

Figure A.25: Forecast obtained by Holt-Winters' double seasonal method (ETS(A,A,N)) from data collected from AREPA sensor.

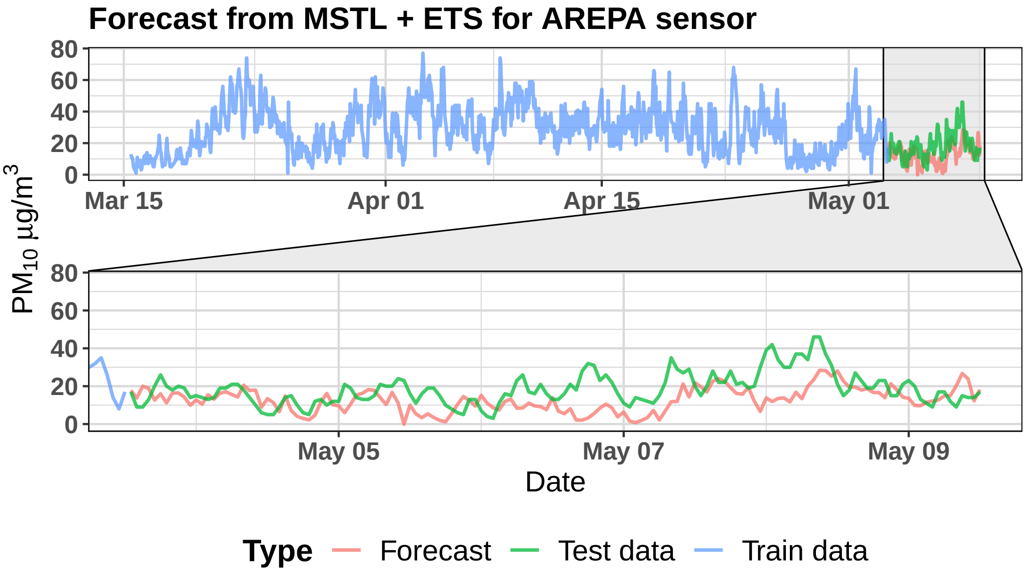

Figure A.26: Forecast obtained by MSTL (ETS(A,N,N)) model from data collected from AREPA sensor.

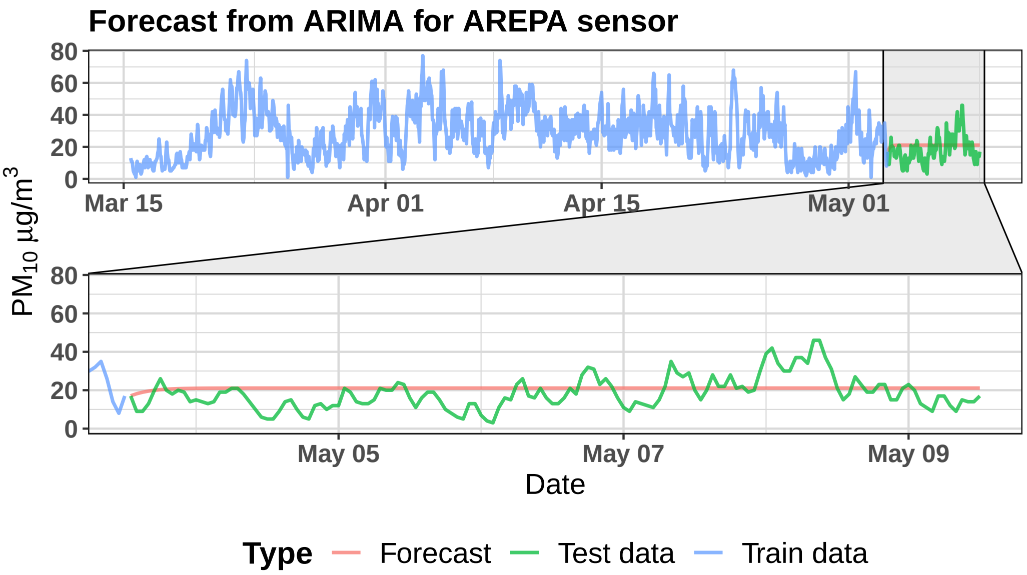

Figure A.27: Forecast obtained by ARIMA(1,1,2) model from data collected from AREPA sensor.

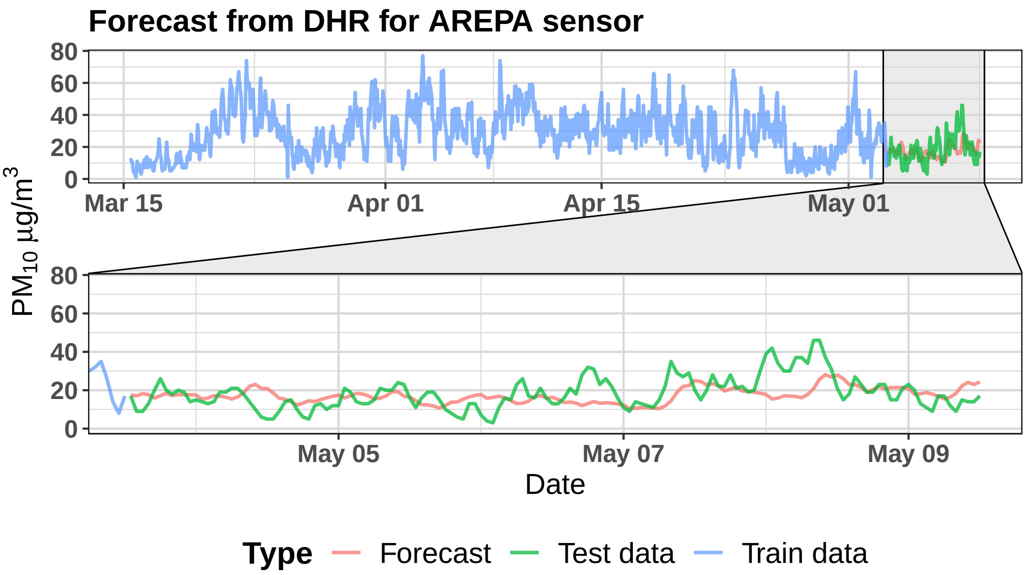

Figure A.28: Forecast obtained by DHR (with ARIMA(1,1,2) errors) model from data collected from AREPA sensor.

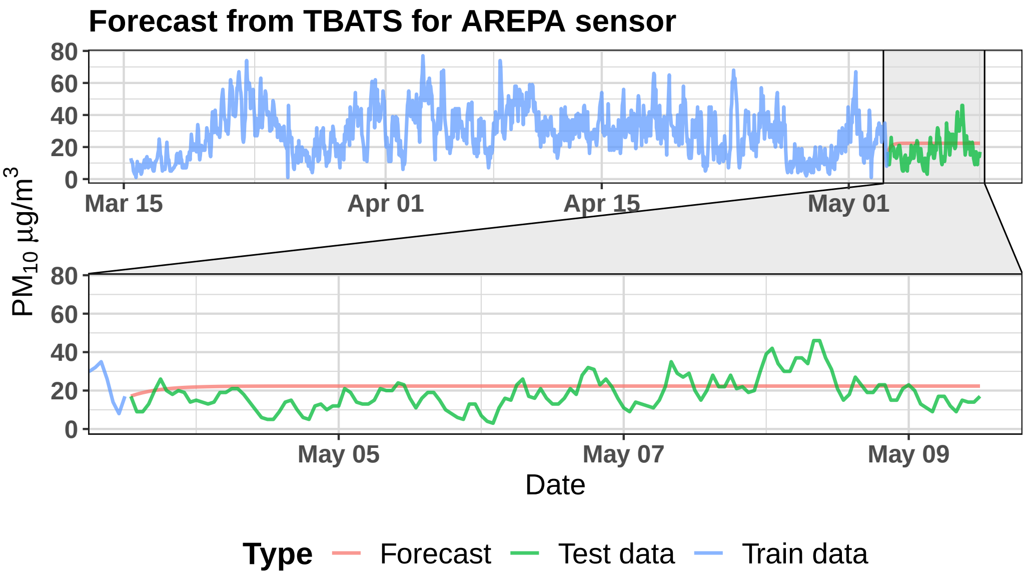

Figure A.29: Forecast obtained by BATS(0,508; {0,0}; 0,8; -) model from data collected from AREPA sensor.

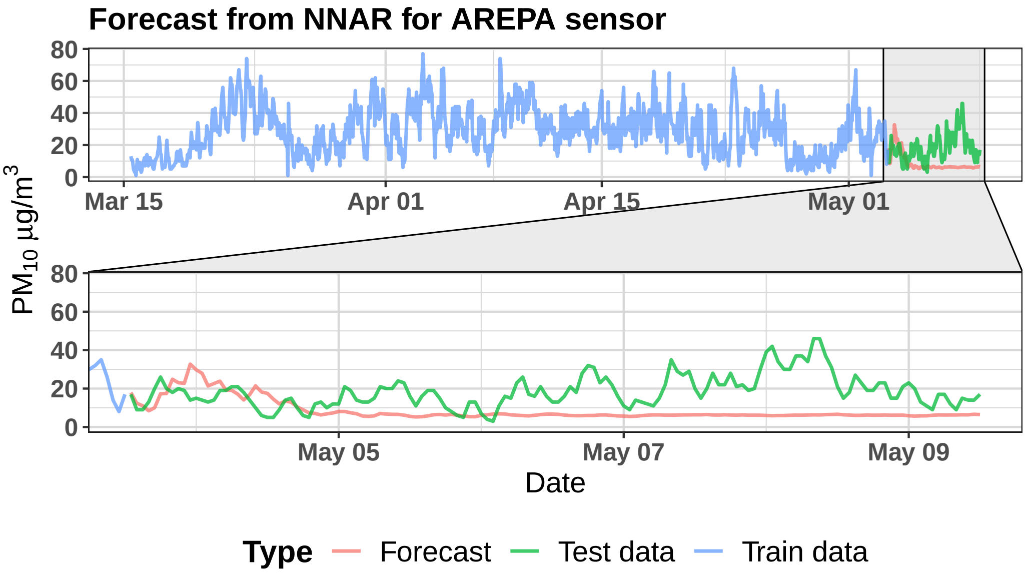

Figure A.30: Forecast obtained by NNAR(25,1,14)$_{168}$ model from data collected from AREPA sensor.

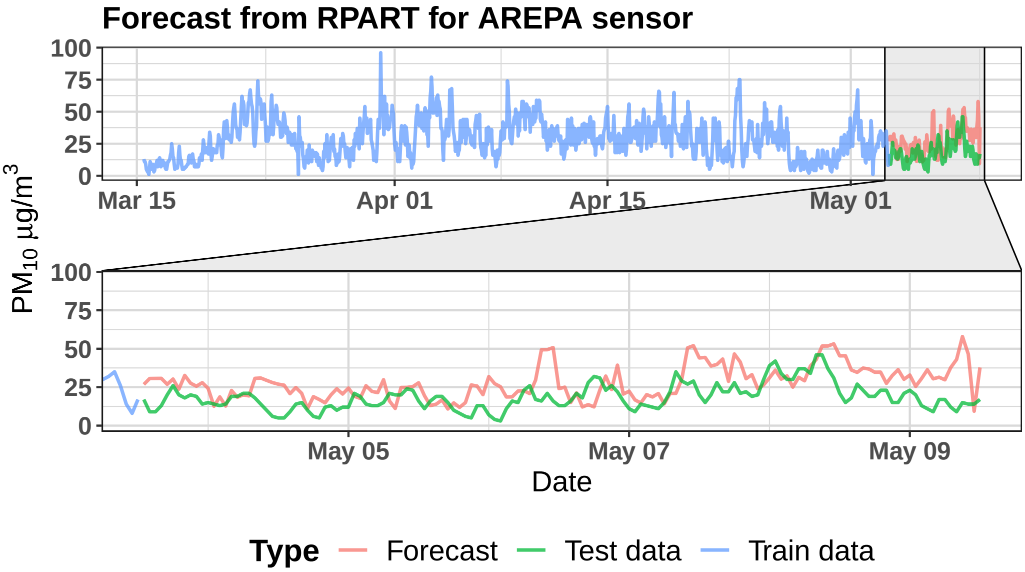

Figure A.31: Forecast obtained by recursive partitioning tree from data collected from AREPA sensor.

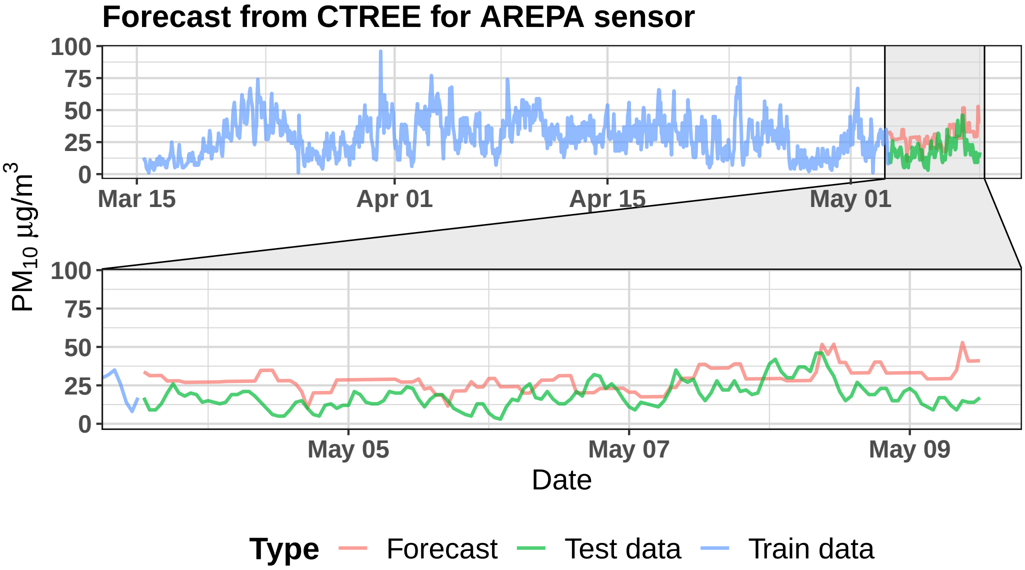

Figure A.32: Forecast obtained by conditional inference tree from data collected from AREPA sensor.

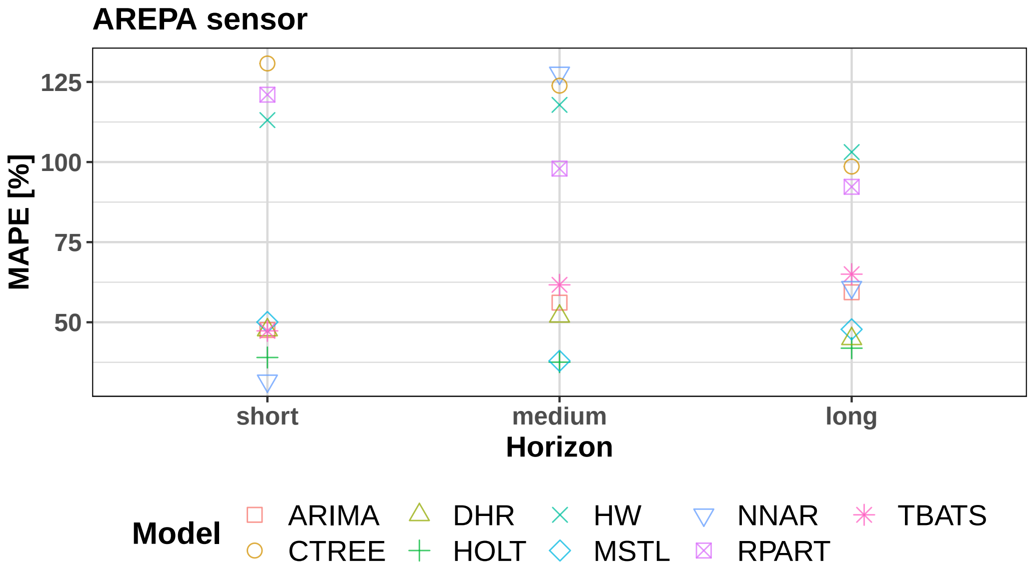

Figure A.33: Values of the MAPE for all models created from data collected from the AREPA sensor for 3 different horizons.

A.4 Models for ARIERA sensor

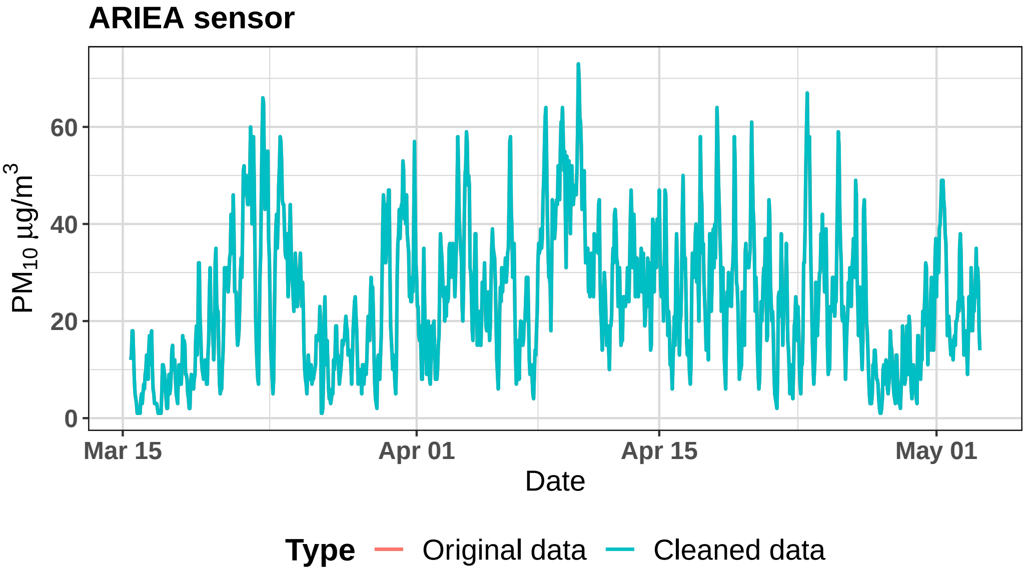

Figure A.34: Hourly concentration of PM$_{10}$ collected by ARIEA sensor between March and May 2019 and data with cleaned outliers.

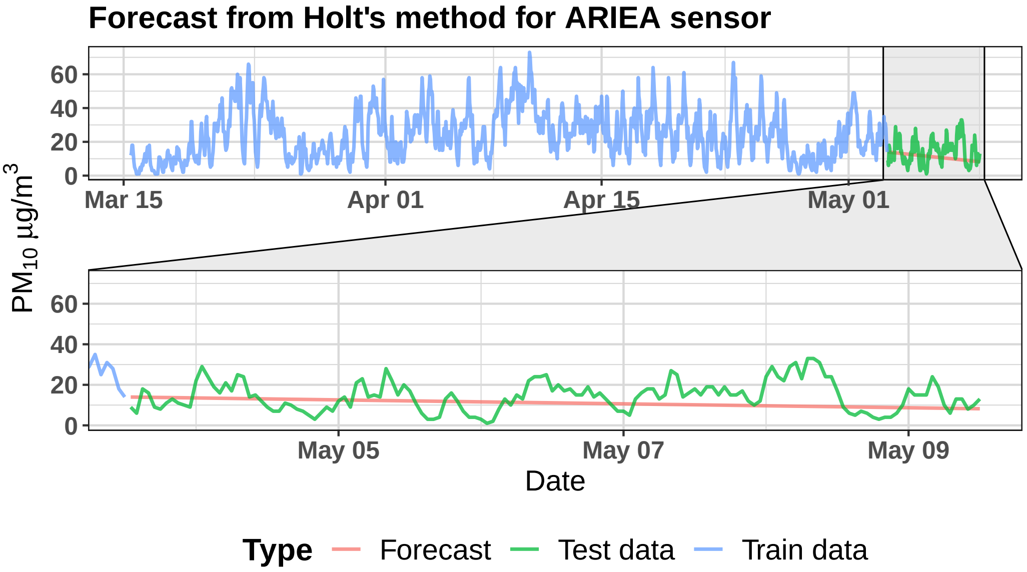

Figure A.35: Forecast obtained by Holt's linear model (ETS(A,A,N)) from data collected from ARIEA sensor.

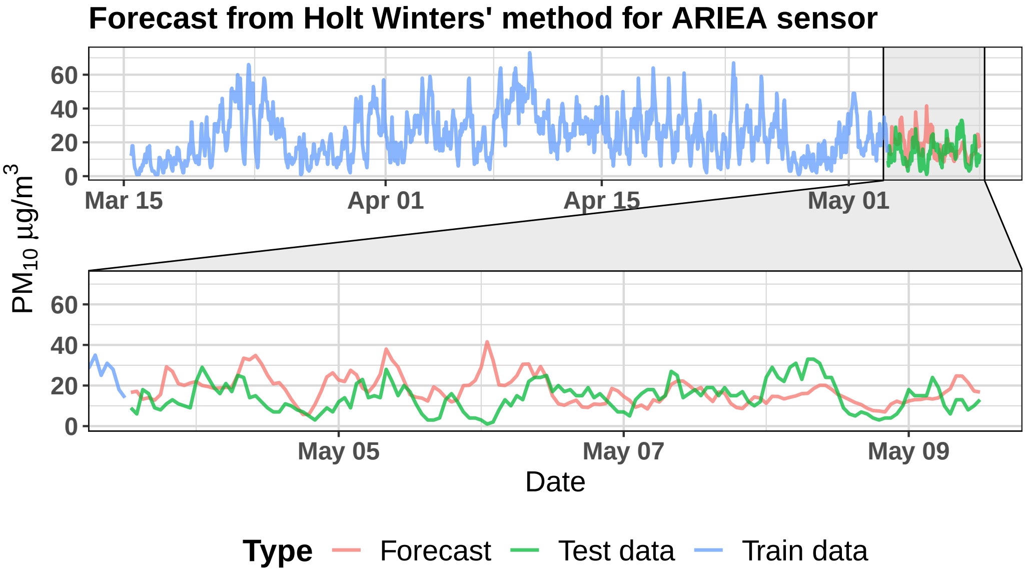

Figure A.36: Forecast obtained by Holt-Winters' double seasonal method (ETS(A,A,N)) from data collected from ARIEA sensor.

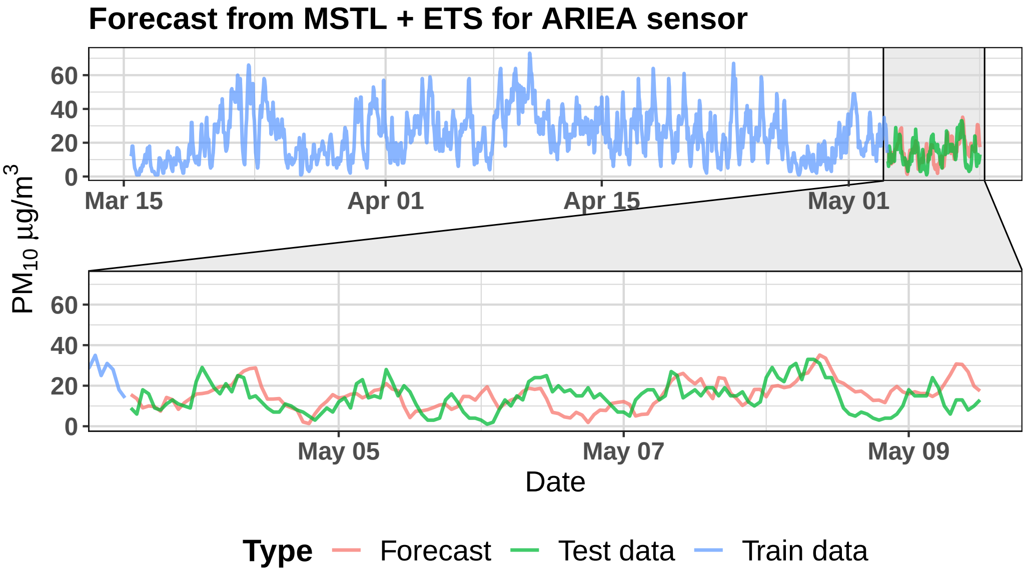

Figure A.37: Forecast obtained by MSTL (ETS(A,N,N)) model from data collected from ARIEA sensor.

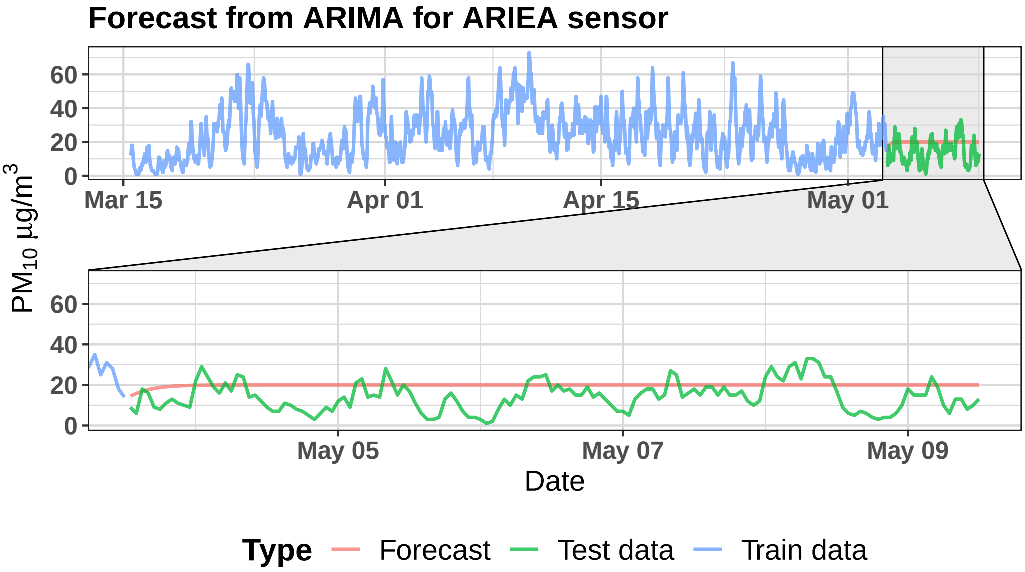

Figure A.38: Forecast obtained by ARIMA(1,1,2) model from data collected from ARIEA sensor.

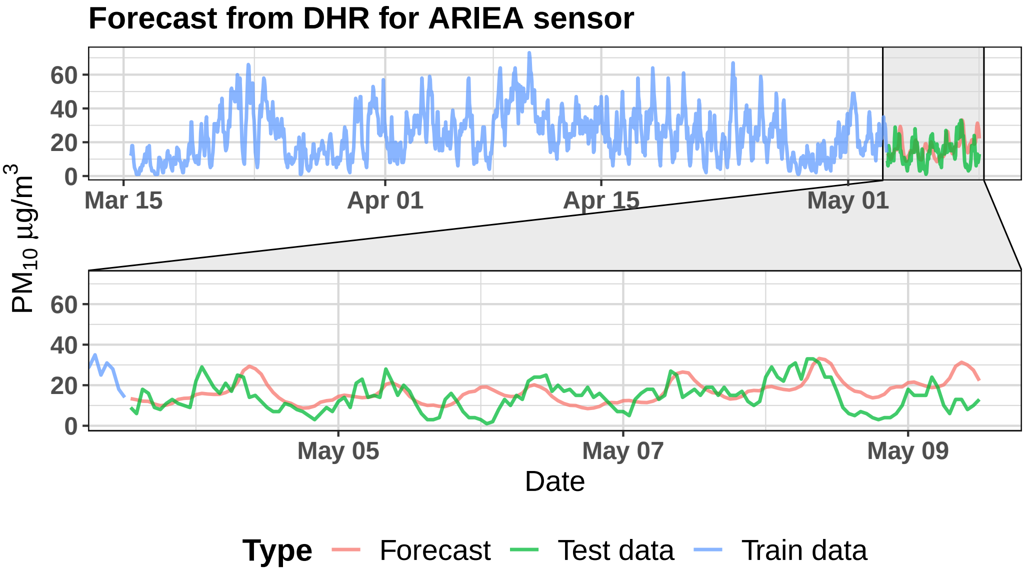

Figure A.39: Forecast obtained by DHR (with ARIMA(1,1,2) errors) model from data collected from ARIEA sensor.

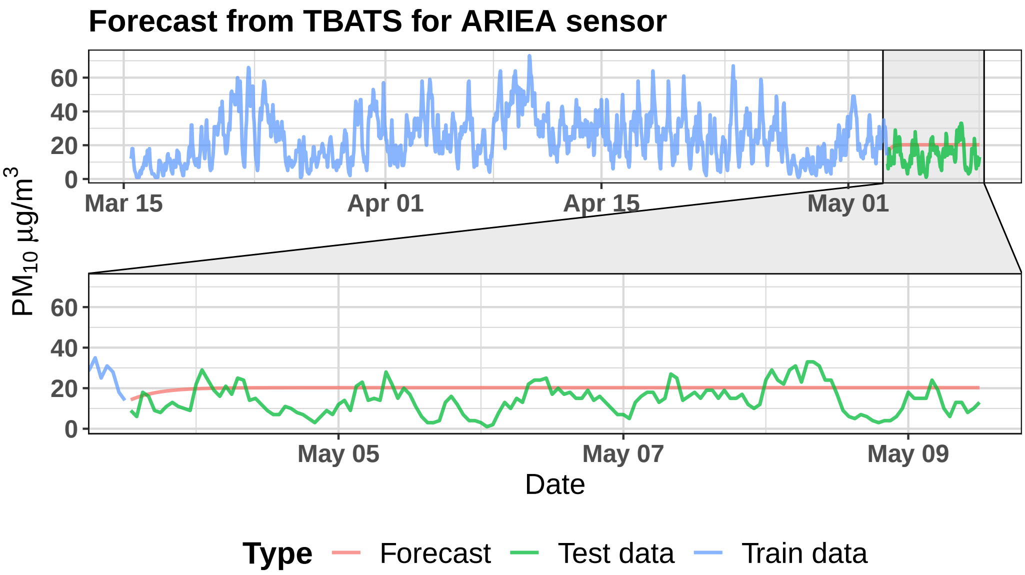

Figure A.40: Forecast obtained by BATS(0,534; {0,0}; 0,8; -) model from data collected from ARIEA sensor.

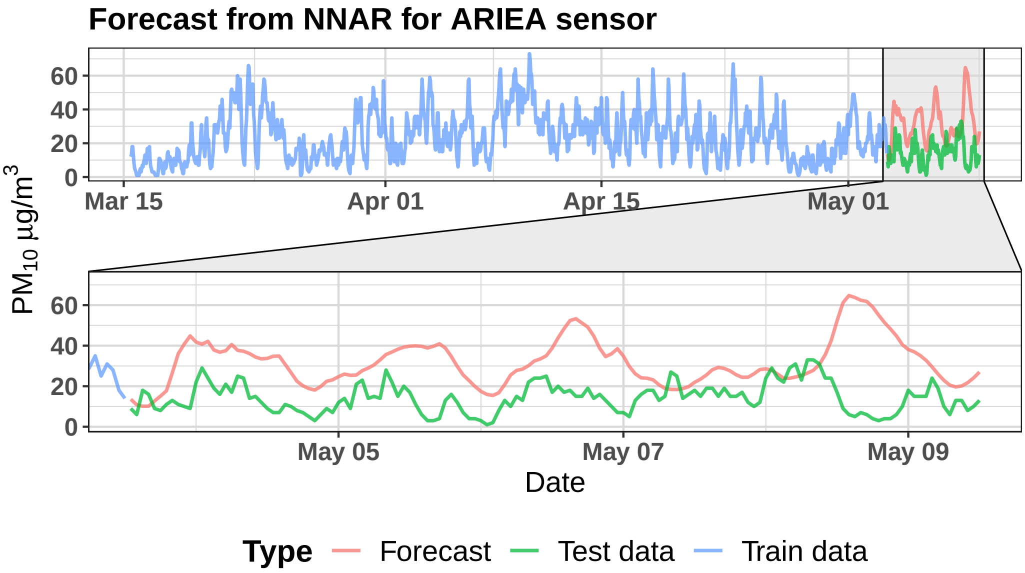

Figure A.41: Forecast obtained by NNAR$(27,1,14)_{168}$ model from data collected from ARIEA sensor.

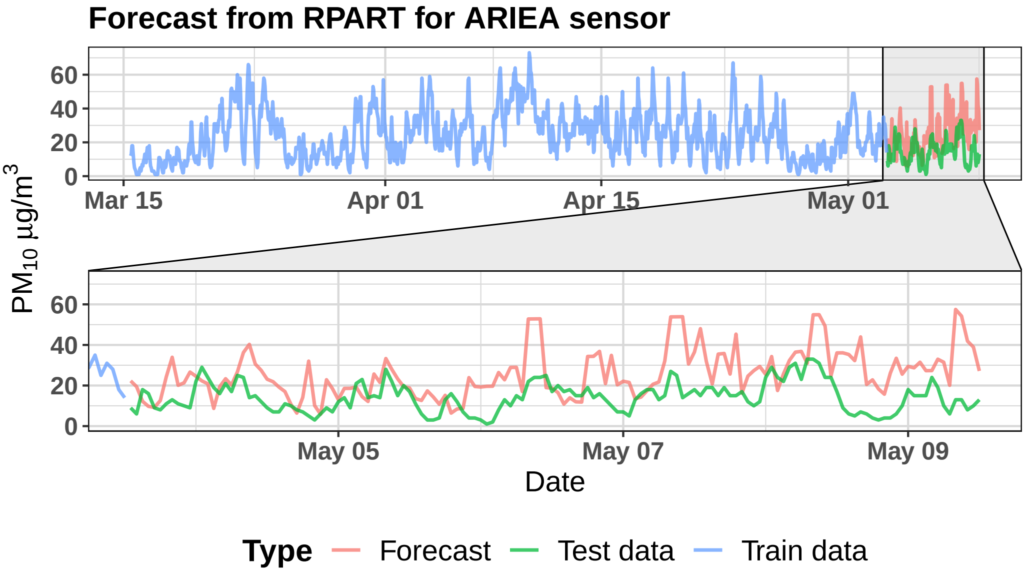

Figure A.42: Forecast obtained by recursive partitioning tree from data collected from ARIEA sensor.

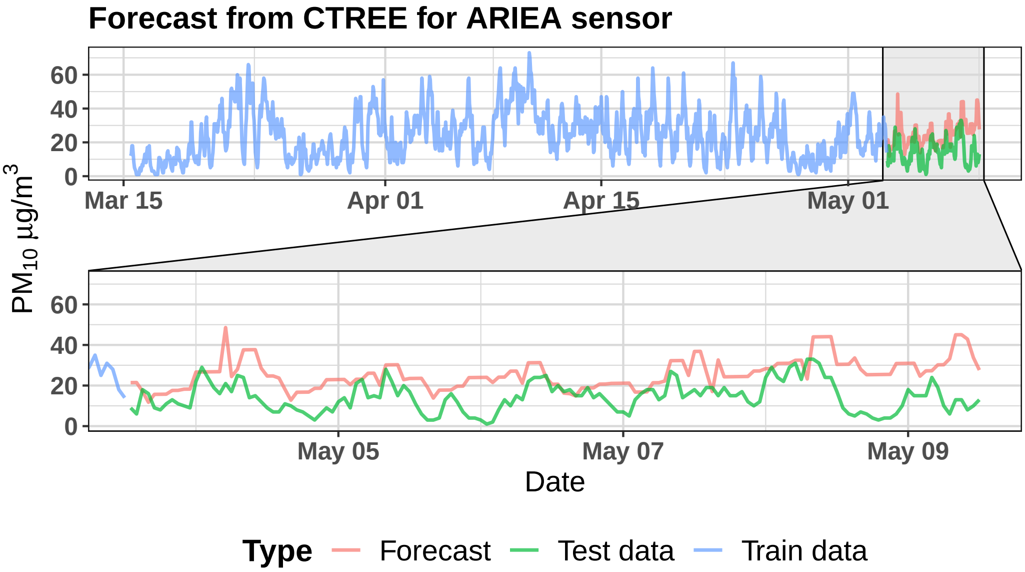

Figure A.43: Forecast obtained by conditional inference tree from data collected from ARIEA sensor.

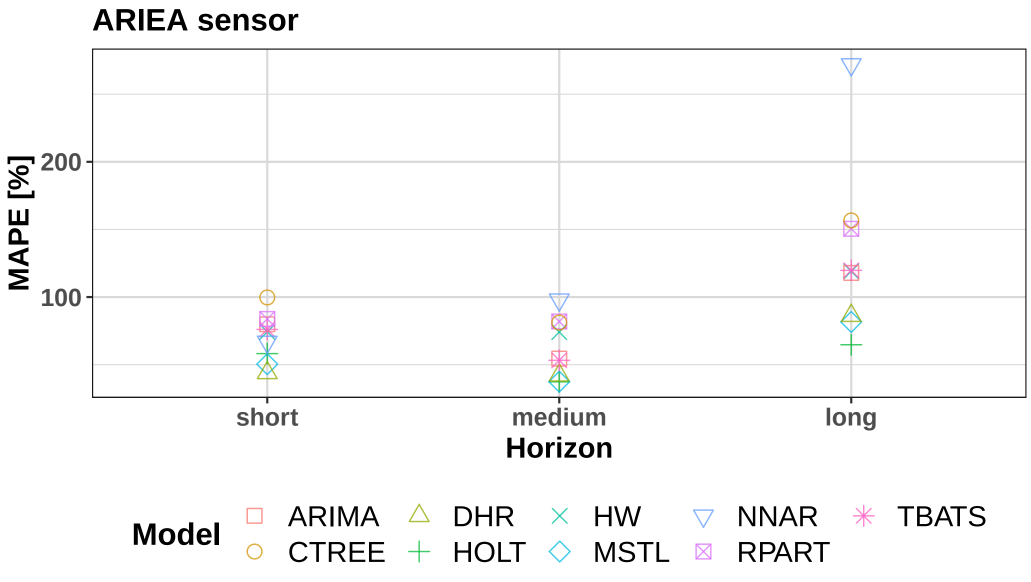

Figure A.44: Values of the MAPE for all models created from data collected from the ARIEA sensor for 3 different horizons.

B. Tables

B.1 Accuracy measures for AKALA sensor

	MAE	RMSE	MAPE
HOLT	7.67	3.49	49.14
HW	9.24	3.30	66.59
ARIMA	7.45	2.60	63.84
MSTL	10.75	0.23	92.01
DHR	7.43	1.85	64.85
TBATS	7.21	1.53	60.15
NNAR	8.11	5.11	45.98
RPART	16.93	16.66	184.79
CTREE	15.91	15.22	174.43

Table B.1: Forecast accuracy metrics for the 6h horizon for data from AKALA sensor.

	MAE	RMSE	MAPE
HOLT	6.21	4.62	32.85
HW	7.71	5.27	45.16
ARIMA	6.07	2.36	43.48
MSTL	12.10	9.43	72.24
DHR	5.02	0.68	33.38
TBATS	6.40	4.18	47.78
NNAR	7.77	3.70	53.67
RPART	11.57	8.28	91.29
CTREE	11.52	11.35	93.92

Table B.2: Forecast accuracy metrics for the 24h horizon for data from AKALA sensor.

	MAE	RMSE	MAPE
HOLT	7.85	4.54	49.71
HW	8.99	3.49	62.02
ARIMA	8.16	1.88	74.31
MSTL	12.89	10.70	82.67
DHR	7.28	0.28	60.93
TBATS	9.82	6.25	100.58
NNAR	27.32	25.82	230.83
RPART	14.93	11.82	146.85
CTREE	14.10	13.17	139.88

Table B.3: Forecast accuracy metrics for the 144h horizon for data from AKALA sensor.

B.2 Accuracy measures for ALEGA sensor

	MAE	RMSE	MAPE
HOLT	3.05	3.05	17.73
HW	1.91	0.94	11.55
ARIMA	1.70	1.34	11.31
MSTL	2.29	1.55	16.00
DHR	2.43	2.22	14.15
TBATS	1.20	0.11	7.44
NNAR	2.83	1.93	18.23
RPART	7.97	7.97	51.25
CTREE	11.30	11.30	74.37

Table B.4: Forecast accuracy metrics for the 6h horizon for data from ALEGA sensor.

	MAE	RMSE	MAPE
HOLT	4.96	4.84	25.36
HW	5.19	3.87	28.34
ARIMA	3.57	1.84	23.38
MSTL	4.36	2.09	26.70
DHR	4.06	2.22	22.22
TBATS	3.28	1.09	20.94
NNAR	4.46	2.93	24.46
RPART	9.78	8.38	63.97
CTREE	13.11	12.69	84.92

Table B.5: Forecast accuracy metrics for the 24h horizon for data from ALEGA sensor.

	MAE	RMSE	MAPE
HOLT	8.16	7.33	47.01
HW	7.29	1.85	51.90
ARIMA	6.35	3.50	55.66
MSTL	5.98	1.92	46.18
DHR	5.22	0.07	39.17
TBATS	6.62	3.85	58.29
NNAR	28.61	28.22	244.72
RPART	13.89	13.49	113.18
CTREE	15.55	15.48	127.65

Table B.6: Forecast accuracy metrics for the 144h horizon for data from ALEGA sensor.

B.3 Accuracy measures for AREPA sensor

	MAE	RMSE	MAPE
HOLT	5.33	0.26	39.00
HW	13.03	13.03	113.09
ARIMA	5.25	3.34	47.57
MSTL	6.63	0.84	50.05
DHR	5.90	1.60	47.58
TBATS	5.21	3.39	47.32
NNAR	5.13	3.32	31.70
RPART	13.67	13.67	121.02
CTREE	14.97	14.97	130.77

Table B.7: Forecast accuracy metrics for the 6h horizon for data from AREPA sensor.

	MAE	RMSE	MAPE
HOLT	4.25	0.15	37.55
HW	12.98	12.98	117.84
ARIMA	5.40	4.92	56.12
MSTL	4.75	0.55	37.95
DHR	5.09	2.30	51.85
TBATS	6.09	5.64	61.68
NNAR	15.02	13.25	127.81
RPART	10.26	9.77	97.95
CTREE	13.58	13.58	123.84

Table B.8: Forecast accuracy metrics for the 24h horizon for data from AREPA sensor.

	MAE	RMSE	MAPE
HOLT	6.47	2.86	41.91
HW	11.91	7.22	103.12
ARIMA	7.01	2.60	59.36
MSTL	8.27	5.66	47.79
DHR	6.30	1.04	44.82
TBATS	7.52	3.73	65.01
NNAR	12.47	11.89	60.92
RPART	11.64	9.46	92.28
CTREE	11.94	10.22	98.59

Table B.9: Forecast accuracy metrics for the 144h horizon for data from AREPA sensor.

B.4 Accuracy measures for ARIEA sensor

	MAE	RMSE	MAPE
HOLT	4.96	2.87	58.21
HW	6.13	3.90	74.28
ARIMA	6.36	5.99	79.88
MSTL	5.07	0.09	50.38
DHR	4.12	0.84	43.90
TBATS	6.07	5.51	76.06
NNAR	6.11	2.87	66.83
RPART	7.26	3.25	83.80
CTREE	7.91	6.20	99.72

Table B.10: Forecast accuracy metrics for the 6h horizon for data from ARIEA sensor.

	MAE	RMSE	MAPE
HOLT	5.36	1.78	37.29
HW	8.55	6.67	74.12
ARIMA	6.04	3.83	54.49
MSTL	5.06	1.15	37.44
DHR	5.64	1.23	42.20
TBATS	5.93	3.70	53.23
NNAR	12.32	10.87	98.02
RPART	9.72	7.03	81.93
CTREE	9.98	9.36	81.15

Table B.11: Forecast accuracy metrics for the 24h horizon for data from ARIEA sensor.

	MAE	RMSE	MAPE
HOLT	6.60	3.18	64.73
HW	7.89	3.50	119.05
ARIMA	7.79	5.59	117.87
MSTL	6.45	1.19	81.61
DHR	6.27	2.45	86.23
TBATS	7.91	5.79	119.73
NNAR	22.55	22.21	271.99
RPART	12.83	11.48	150.49
CTREE	11.62	11.27	156.68

Table B.12: Forecast accuracy metrics for the 144h horizon for data from ARIEA sensor.

C. Digital files

This thesis contains following digital files:

forecast_PM10.R program, which creates forecasts and exports their plots, model parameters and accuracy measures

measures.R program, which exports plots of accuracy measures from prepared data