Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Effects of meteorology on PM_{10} concentrations: a comparative assessment of machine learning methods

Relatore:
Prof. Gastone Castellani

Correlatrice:
Prof.ssa Claudia Sala

Presentata da:
Davide Ferraresi

Anno Accademico 2018/2019

Contents

1 Introduction **7**

1.1 Air quality and pollution . . . 9
1.1.1 Air pollutants and main sources of emissions . . . 9
1.1.2 Geographical variability and concentration monitoring . . . 10
1.1.3 International policies . . . 12
1.2 Particulate matter (PM) . . . 13
1.2.1 Effects on human health . . . 15
1.2.2 Legal thresholds and guideline values for concentration . . . 16
1.2.3 Emission and concentration monitoring . . . 17
1.3 Context overview: Emilia-Romagna . . . 23
1.3.1 Geographical and meteorological elements . . . 23
1.3.2 Anthropic pression and emission sources . . . 24
1.3.3 Air quality and meteorology . . . 25
1.3.4 Regional concentration monitoring . . . 25
1.4 Models for PM_{10} prediction . . . 29
1.4.1 Previous works on PM_{10} forecasting . . . 32
1.4.2 Previous works on other pollutants' forecasting . . . 32
1.4.3 Some remarks . . . 35

2 Materials and methods **36**

2.1 Data overview and exploratory analysis . . . 36
2.1.1 EDA techniques . . . 38
2.1.2 Particulate Matter concentration . . . 42
2.1.3 Temperature . . . 51
2.1.4 Precipitation . . . 55
2.1.5 Wind intensity and direction . . . 58
2.1.6 Radiant exposure . . . 65
2.1.7 Atmospheric pressure . . . 67
2.1.8 Mixed layer height . . . 69
2.2 Missing data treatment and imputation . . . 72
2.2.1 Listwise deletion of missing data . . . 72

2.2.2 Multiple imputation of missing data . . . 73
2.3 Linear regression models . . . 74
2.3.1 Standard linear regression . . . 75
2.3.2 Ridge regression . . . 76
2.3.3 Lasso regression . . . 77
2.4 Regression tree models . . . 77
2.4.1 Bagging and random forests . . . 79
2.4.2 Boosting . . . 80
2.5 Model assessment and selection . . . 81
2.5.1 Measuring the error . . . 81
2.5.2 Choosing the predictors . . . 82
2.5.3 Splitting the datasets . . . 82
2.5.4 Cross-validation procedure . . . 83
2.5.5 Model selection and assessment of the performances . . . 84
2.5.6 Classification task based on PM_{10} daily limit value . . . 86
2.5.7 Approach with MI datasets . . . 86
2.6 Implementation in R . . . 88
2.6.1 Missing data treatment and imputation . . . 88
2.6.2 Linear regression models . . . 91
2.6.3 Regression tree models . . . 92
2.6.4 Cross-validation, model selection and comparison . . . 92
2.6.5 Classification task . . . 93

3 Results **95**
3.1 Performance on LWD-*basic* datasets . . . 95
3.2 Comparison of performances on LWD-*basic* and MI-*basic* datasets . . . 98
3.3 Performance of models with non-meteorological predictors . . . 99
3.4 Comparing models for classification tasks . . . 101

4 Conclusions **103**
4.1 Further developments . . . 104

Abstract

Administrative decisions regarding the application of measures to address air quality issues have to rely both on present observation and future predictions of the concentration of various pollutants. Since PM_{10} is one of the most critical pollutants, the ability to provide accurate forecasts for its concentration, when required, is crucial in order to enforce the necessary measures at the right time.
Together with the pattern of emission sources which is present in a geographical area, meteorological conditions can significantly affect the concentration of pollutants in air, since they can favour the dispersion or, on the other hand, the build-up of those compounds. It is possible then to predict (at least partially) the concentration of PM_{10} in air using meteorological variables as predictors.
In fact, various statistical models have been proposed for accomplishing similar tasks on a number of geographical regions and urban areas, with varying results. The set of meteorological variables that have been considered in those cases included various predictors, measured both in the day of interest and in the previous ones. Sometimes also some non-meteorological descriptors (e.g. time-related variables) that are grossly related to the variation of the emission patterns have been considered as input variables for those models.

In this work an analysis of the relationship between meteorology-related variables and PM_{10} concentration levels in the capitals of the provinces of Emilia-Romagna has been performed in order to understand how the meteorological conditions affect PM_{10} concentration. Then the considered meteorological variables have been input as predictors to statistical regression models based on machine learning in order to obtain predictions for the daily mean value of PM_{10} concentration.
Taking a cue from a synthetic indicator defined by the regional agency ARPAE that links meteorological conditions to the building up of PM_{10}, a dataset containing time series of daily values of 10 meteorological variables and those of PM_{10} urban background concentration for the 10 cities, spanning a time interval of 2008 days (5 year and a half), has been initially created. Data have been obtained from the public database available on ARPAE websites and processed using R-based software *RStudio*.
Once the dataset has been built, it has been subjected to an exploratory data analysis

that has allowed to point out the main features of each variable and its relationship with PM_{10} concentration, evaluated on a daily basis.

After having adequately pre-processed the data, they have been used to train regression models with the aim of predicting PM_{10} daily mean concentration values starting from the same-day values of the meteorological variables. All the considered models, which include standard and regularized linear regressions and regression tree-based ones, have been trained separately with the data of each city, in order to reproduce specifically the patterns observed at a local level. At the beginning of the exam of those models, only the meteorological data from the same day for which the prediction has to be made have been fed into the model.
The results show that random forest and boosting models are generally better in the prediction tasks, for all the considered cities.

With respect to predictors, the level of model performance obtained with the chosen set of meteorological variables have been subsequently compared with the performances on the same dataset with the addition of non-meteorological variables such as the day of the week and the month related to each sample, and the previous-day PM_{10} mean concentration level, in order to improve the performance initially obtained. Statistical tests have shown that the performance improves significantly in a number of cases with time-related descriptors and in all the cases with the addition of the previous-day PM_{10} value.

Finally, an evaluation of the ability of the considered models to carry out a good "classification" with respect to the legal limit value for PM_{10} daily mean concentration has been made, obtaining good results for all the tested models.

Abstract

Le decisioni delle autorità relative all'applicazione di misure di contrasto al degrado della qualità dell'aria devono fondarsi sia su misurazioni effettuate, sia su previsioni per i valori futuri di concentrazione delle sostanze inquinanti. Il PM_{10} è uno degli inquinanti più controllati e la capacità di fornire previsioni accurate per la sua concentrazione, quando richiesto, è fondamentale per applicare le misure necessarie al momento giusto.
Oltre alla distribuzione e alle caratteristiche delle fonti di emissioni presenti in un'area geografica, un elemento che influenza in modo importante le concentrazioni di inquinanti è la meteorologia, dal momento che differenti condizioni meteo possono favorire la dispersione o, al contrario, l'accumulo di queste sostanze. È quindi possibile effettuare una previsione della concentrazione di PM_{10} in atmosfera impiegando come predittori alcune variabili meteorologiche.
In effetti diversi modelli statistici capaci di effettuare simili previsioni sono stati applicati su diverse regioni e aree urbane, con risultati di qualità variabile. Gli insiemi di variabili meteorologiche impiegati per quei modelli comprendevano diversi predittori, misurati sia nella giornata d'interesse sia in quelle precedenti. In tali modelli sono state talvolta considerate anche alcune variabili non legate al meteo (ad esempio le variabili temporali) che sono legate ai trend di variazione delle emissioni.

In questa attività è stata compiuta un'analisi della relazione fra variabili meteorologiche e concentrazioni di PM_{10} nei capoluoghi di provincia dell'Emilia-Romagna, con lo scopo di comprendere in che modo le condizioni meteo influenzano le concentrazioni di particolato. In seguito tali variabili sono state utilizzate come predittori all'interno di modelli statistici di regressione basati sul *machine learning* per effettuare previsioni del valore della concentrazione media giornaliera di PM_{10}.
Prendendo spunto da un indicatore sintetico elaborato dall'agenzia regionale ARPAE per identificare le condizioni meteorologiche che favoriscono l'innalzamento dei livelli di PM_{10} in atmosfera, è stato costruito un set di dati contenente le serie storiche dei valori giornalieri di 10 variabili meteorologiche e della concentrazione di fondo urbano di PM_{10} per ciascuna delle 10 città considerate; il set di dati copre un intervallo di 2008 giorni (circa 5 anni e mezzo). I dati sono stati estratti dai database pubblici di ARPAE disponibili sul sito dell'agenzia e sono stati elaborati mediante il software *RStudio* basato

sul linguaggio R.
Una volta costruito il set di dati, questo è stato sottoposto a un'analisi dati esplorativa per mettere in luce le caratteristiche principali di ciascuna variabile e la relazione fra ognuna di esse e la concentrazione di PM_{10}. In particolare si è evidenziata la relazione fra le condizioni meteo e la concentrazione dell'inquinante nella stessa giornata.

Una volta eseguite le necessarie operazioni di *preprocessing* per rendere fruibili i dati, essi sono stati utilizzati per addestrare modelli di regressione allo scopo di predire il valore della concentrazione media giornaliera di PM_{10} a partire dai valori delle variabili meteorologiche nello stesso giorno. Tutti i modelli considerati, che comprendevano modelli di regressione lineare standard, regolarizzata e modelli basati su alberi di regressione, sono stati addestrati separatamente con i dati di ciascuna città, in modo da poter riprodurre gli andamenti osservati nelle diverse località. Inizialmente la procedura di addestramento dei modelli ha incluso come dati di input solamente i valori delle variabili meteorologiche misurate nello stesso giorno per il quale il modello effettuava la previsione.
I risultati hanno dimostrato che, per quanto riguarda i modelli, le *random forest* e i *boosting models* si sono dimostrati i più efficaci in tutte le città considerate.

Sul fronte delle variabili utilizzate come predittori, le prestazioni dei modelli addestrati con le variabili meteorologiche sono state successivamente confrontate con le performance degli stessi modelli addestrati con i set di dati integrati con variabili temporali, quali il giorno della settimana e il mese di campionamento del dato, e il valore della concentrazione media giornaliera di PM_{10} del giorno precedente, allo scopo di valutare eventuali miglioramenti nell'accuratezza. Si sono osservati miglioramenti significativi in una parte dei modelli quando sono stati addestrati con l'integrazione delle variabili temporali; nel caso dei modelli addestrati con la variabile relativa alla concentrazione di PM_{10} del giorno precedente, l'incremento della qualità della predizione è risultato significativo in tutti i casi considerati.

Da ultimo, si è valutata l'abilità dei modelli considerati nel "classificare" correttamente un set di dati rispetto al valore limite legale della concentrazione media giornaliera di PM_{10}, ottenendo buoni risultati su tutti i modelli considerati.

Chapter 1

Introduction

In this work the issue of air pollution in a number of urban areas of Emilia-Romagna, an administrative region of Italy, has been addressed with respect to the effect of meteorology on the concentration of pollutants in air.
The considered areas are nowadays affected by serious problems concerning ambient air pollution due to PM_{10} concentrations. Therefore, administrative measures are routinely enforced in the winter months to try to tackle what is still called a "PM_{10} emergency". Applying those measures for harm reduction is necessary to protect the health of people and administrative bodies are responsible for the decisional process.[16] Measures are generally applied whenever both pollutant's concentration in the recent past and forecasts for the following days are above the thresholds that have been defined by the law. So the use of forecasting systems is common and a number of methods has been developed to address this need.

A distinction can be made [5] between numerical and statistical methods. The former ones (both bi- and tridimensional) are able to calculate the concentration of pollutant compounds in a selected area by performing a geospatial simulation, i.e. by geographically determining the sources of emissions, the presence of boundaries and the behaviour of the lower layers of the atmosphere and simulating the chemical and physical processes that happen in the air at different spatial and temporal scales. On the contrary, statistical methods are independent on those processes and analyse statistically the relationship between descriptors that approximate the various elements of the context (e.g. meteorological trends, geographical patterns of sources, ...). This latter kind of methods is the one that has been considered for the present work.

In this work the focus is on the relationship between the behaviour of meteorological variables and the trends in PM_{10} concentration. As pollutants build up in the atmosphere, some meteorological events can favour the increase or the reduction of the concentrations: for this reason, meteorology-related variables (such as temperature, precipitation

intensity and others) can be used as predictors in the aforementioned statistical models together with other variables that are related to other drivers (e.g. the periodical trend of emissions in a year cycle or in a week).

In the following chapters an analysis of the patterns of meteorological variables measured in all the capitals of the provinces of the region (Piacenza, Parma, Reggio Emilia, Modena, Bologna, Ferrara, Ravenna, Forlì, Cesena, Rimini) during the period of time between the 1^{st} of October, 2012 and the 31^{st} of March, 2018, and the relationship between those variables and the measured values of urban background PM_{10} daily mean concentration in the same cities is presented.
This analysis is followed by the evaluation of the performances of a number of statistical regression models based on machine learning techniques that take the meteorological variables measured in each city as input in order to predict the PM_{10} daily mean concentration in the same city.

The present chapter examines the issue of air pollution with particular focus on PM_{10}, the reasons why it is a current problem during winter months and the monitoring procedures that are currently implemented. Then a description of the geographical context in which measures considered in this work have been taken is made and a review of past works that involved the use of statistical models for similar tasks is made.

In chapter 2 an extensive exploratory data analysis on the considered data is presented, in which the patterns of each meteorological variable and the correlation between those and PM_{10} concentration is performed in order to understand how different meteorological conditions favour the building-up PM_{10} for each considered cities.
Then a number of the regression models that have been considered is presented: these models have been evaluated in order to select the one that best performs in predicting the value of PM_{10} daily mean concentration for each city starting from the values of the meteorological variables measured in the same day.

The results of the comparative assessment among the models are presented in chapter 3, where the performances of the same models on datasets integrated with non-meteorological variables are also shown.

Conclusions are made in chapter 4, where considerations are made about possible future developments of this work.

1.1 Air quality and pollution

The state of air quality is a current matter of concern, not only for decision makers but for society in general: a European Commission survey [14] has found it is considered as *the second biggest environmental concern* for European people after climate change.

Air pollution is responsible for a number of severe diseases, including premature death: it has been found that more than each year 500000 premature deaths in the European Union [31] can be attributed to ambient air pollution. Both *particulate matter* (PM) and air pollution in general have been classified as carcinogenic by the International Agency for Research on Cancer [26].

The environment in its entirety is also affected by air pollution: for example nitrogen compounds can cause eutrophication of ecosystems; NO_x and SO_2 produce acidification of soil and surface water; and O_3 has a negative impact on vegetation.
Some pollutants are also considered *climate forcers*, in the sense that they affect global warming in a positive (warming) or negative (cooling) way. Conversely, climate change affects the emission and the diffusion of air pollutants in the atmosphere (e.g. increasing temperatures intensifies the emission of volatile organic compounds and provides better conditions for the spread of wildfires, that are natural sources of pollutants).
Air pollution is also responsible for damaging properties and buildings, including cultural heritage ones. Finally, it has also economic consequences, that can be distinguished in *market costs* (e.g. reduced productivity, crop losses, ...) and *non-market* ones (increased mortality, degradation of water and soils, ...).

1.1.1 Air pollutants and main sources of emissions

Pollutants are generally distinguished into two categories [1]:

- primary pollutants are directly emitted to the atmosphere, both from natural and anthropogenic sources: primary particulate matter (PM), black carbon (BC), nitrogen oxides (NO_x, that includes NO and NO_2), sulphur oxides (SO_x), methane (CH_4), ammonia (NH_3) and volatile organic compounds (VOCs) are included in this class;

- secondary pollutants are produced from chemical reactions (sometimes favoured by the presence of sunlight) between precursor gases, i.e. primary or secondary pollutants as well, in the atmosphere: secondary PM, ozone (O_3), secondary NO_2 and other compounds are part of this category.

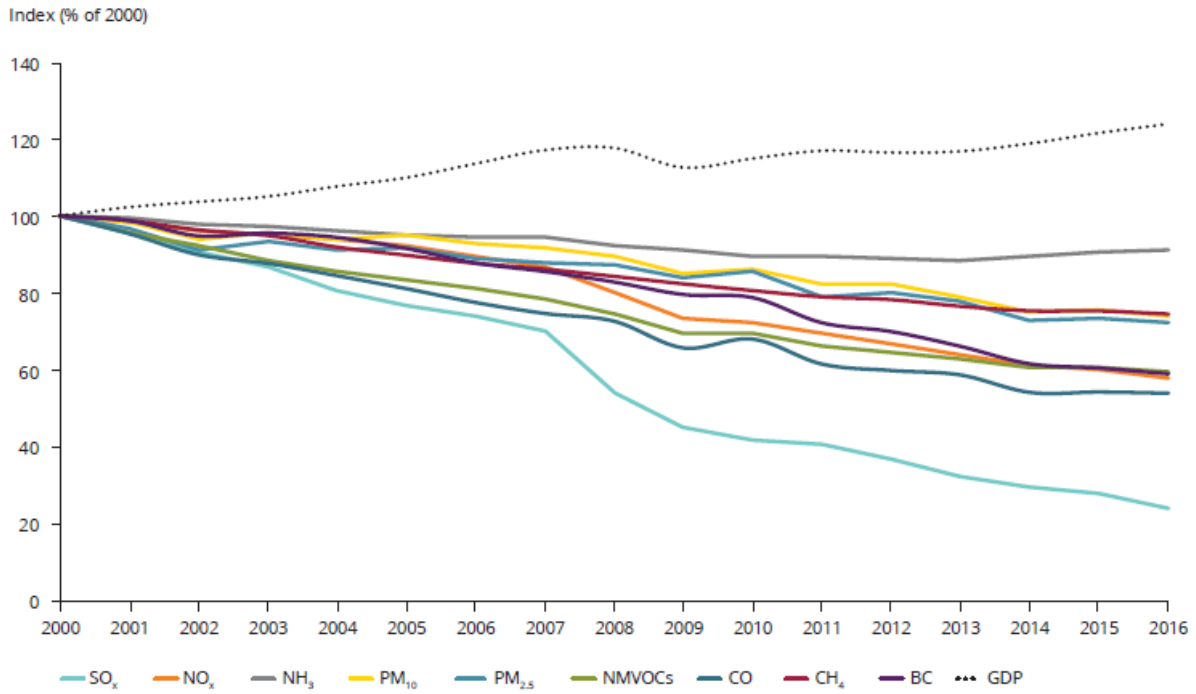


Figure 1.1: Development in EU-28 emissions with respect to 2000 levels (from [16]).

Time series of emission of primary pollutants starting from year 2000 show a general decrease, as can be seen in Figure 1.1.
Considering only primary pollutants, it is possible to analyse the contributions of the various anthropogenic sources to air pollution: emission inventories, i.e. analyses that estimate the quantities of compounds emitted in the atmosphere by the various economic sectors, are common tools for reviewing and analysing these data. Table 1.1 summarizes the shares of contribution by each economic sector in the European Union member States in 2016, with respect to pollutants ([16]; only principal pollutants have been considered).

1.1.2 Geographical variability and concentration monitoring

The variety and the density of sources of pollutants in anthropized areas are responsible for generally significant emissions.
Nonetheless, pollutants concentration in a specific place is the result of transport and dispersion phenomena from the surrounding sources, as well as chemical reactions that produce secondary pollutants. Both processes are influenced by several meteorological variables and the morphology of the region (e.g. orography).

This leads to a general behaviour in spatial concentrations that is represented in Figure

Source (sector)	SO_x	NO_x	PM_{10}	$PM_{2.5}$	NH_3	VOC	CH_4
Road transport	/	39	10	11	2	9	/
Non-road transport	3	9	2	2	/	1	/
Commercial and households	17	14	39	56	2	17	4
Energy	51	17	5	4	/	9	14
Industry	29	14	25	18	2	50	1
Agriculture	/	6	15	4	92	13	53
Waste	/	1	4	5	1	1	28
Other	/	/	/	/	1	/	/
	100	100	100	100	100	100	100

Table 1.1: Share contribution of principal air pollutants per sector in EU-28 (2016).

1.2.
Indeed, depending on the location of the monitoring station, the concentration of a pollutant is given by:

- a *regional background* that can be detected anywhere in the considered area and is produced by the transport of pollutants in the atmosphere;
- a *urban background* that adds up to the previous one in urban contexts, where more sources are densely packed;
- the contribution from specific *hotspots* such as trafficked road and industries.

So the regional background of a certain pollutant is related to a level of exposure of the population to that substance which is shared by all the inhabitants of the considered region, while the urban background is specific for residents in a certain urban area and hotspot levels can help assess the exposure of people who live by, work by or commute through particularly critical places.

Thus, sampling operations for analysing the concentration of pollutants are performed by a system of monitoring station placed at fixed locations in order to get a good representation of the considered area. In the EU framework, these locations are classified into:

- *traffic*: near trafficked roads, where concentration of various pollutants (NO_x, PM_{10},...) are generally high in specific periods of time during the day;
- *urban* and *suburban background*: sites within a urban context;
- *other*: specific locations near industries or other sources of specific pollutants;

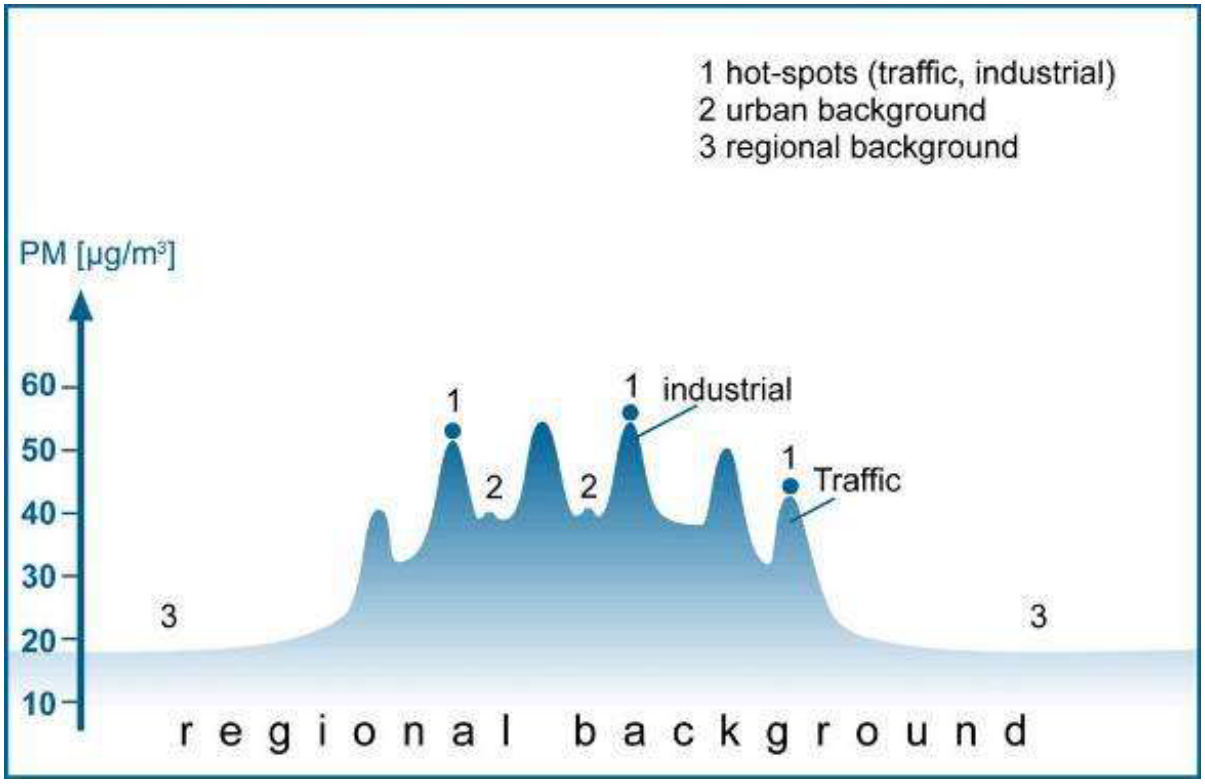


Figure 1.2: Spatial distribution of a pollutant concentration (from [12]).

- *rural background* sites: far from the largest cities, in order to measure the regional background levels.

1.1.3 International policies

The huge number of effects and economic costs of air pollution has encouraged local, national and international authorities, along with other organizations, to take measures in order to limit the emission of pollutants.

The United Nations Environment Programme (UNEP) and the World Health Organization (WHO) have guided the action at international level, calling for protection and promotion of people's health and well-being, along with offering support to governments for undertaking monitoring and assessment of air quality issues, and taking measures to prevent and reduce air pollution.

At the European level, the framework of the EU's air quality policy has been outlined in the 2018 Communication "A Europe that protects: Clean air for all" [15]. It is composed

of three *pillars*:

- the ambient air quality standards defined in the Ambient Air Quality Directives (2004/107/EC [17] and 2008/50/EC [18]);
- the national emission reduction targets defined in the National Emissions Ceiling (NEC) Directive (2001/81/EC as replaced by Dir. 2016/2284/EU) with limits that have to be complied with starting from 2020 and 2030;
- the emissions standards for key sources of pollution, defined in a number of Directives addressed to industries, power plants, vehicles, transport fuels and goods production.

At the national and local level, air quality plans to protect human health and environment are requested by the Ambient Air Quality Directives, while National Air Pollution Control Programmes are expected by 2019 following the NEC Directive.

1.2 Particulate matter (PM)

The term *particulate matter* (PM) refers to the microscopic solid and liquid matter suspended in air; it has been defined by WHO [40] as "a complex mixture with components having diverse chemical and physical characteristics", while the mixture composed of particulate and air is generally called *aerosol*.

Pollutants that belong to this category are generated through heterogeneous phase reactions, i.e. chemical reactions that involve compounds that are present in different phases in the atmosphere. Acid gases like sulfuric acid (H_2SO_4), nitric acid (HNO_3) and hydrochloric acid (HCl) are often part of these reactions. An example of chemical reactions that lead to the formation of these compounds is given in the following formulas:

$$\begin{cases} SO_{2(g)} + OH^{\bullet} + O_2 + H_2O \longrightarrow H_2SO_{4(l)} + HO_2^{\bullet} \\ {}_2NH_{3(g)} + H_2SO_{4(l)} \longrightarrow (NH_4)_2SO_{4(s)} \\ NH_{3(g)} + H_2SO_{4(l)} \longrightarrow (NH_4)HSO_{4(s)} \end{cases} \tag{1.1}$$

$$\begin{cases} NO_{2(g)} + OH^{\bullet} \longrightarrow HNO_{3(g)} \\ NH_{3(g)} + HNO_{3(g)} \longrightarrow NH_4NO_{3(s)} \end{cases} \tag{1.2}$$

$$NH_{3(g)} + HCl_{(g)} \longrightarrow NH_4Cl_{(s)} \tag{1.3}$$

As can be seen, the interaction of this gaseous molecules with radicals and ammonia (NH_3) produces solid salts that are incorporated by the aerosol.

The particles that make up particulate matter are generally classified by their *aerodynamic diameter*, a property that determines how particles are transported in the atmosphere and deposited in the environment, and how they interacts with the human respiratory system (the diameter influences the likelihood of deposition in the different sites). Aerodynamic diameter is approximately linked to the source or the processes that generate particles; it can consequently be exploited in the sampling operation to separate the different classes of particles and identifying the most probable sources.
Following this criteria, PM particles can be classified in:

- coarse particles (PM_{10}, or more correctly "PM_{10} - $PM_{2.5}$"), i.e. whose aerodynamic diameter is smaller than $10\,\mu$m and larger than $2.5\,\mu$m: these particles are generally produced breaking up larger solid particles, and include resuspended dust (by roads, industrial activities, agricultural processes, uncovered soil, ...), pollen grains, bacterial fragments, sea-spray particles and ashes;

- fine particles ($PM_{2.5}$), whose aerodynamic diameter is smaller than $2.5\,\mu$m: they are formed from gases, where ultrafine particles are generated through the processes of nucleation, coagulation (combination of two or more nuclei) and condensation (of gas or vapour molecules on the surface of nuclei), that can produce fine particles up to $1\,\mu$m in terms of aerodynamical diameter; fine particles are also produced in combustion processes, where vaporized metals and organic compounds can condense: SO_2, NO_x, NH_3 and VOCs are identified as $PM_{2.5}$'s main precursors;

- ultrafine particles, whose aerodynamic diameter is smaller than $0.1\,\mu$m.

As said, the analysis of the chemical composition of PM allows to identify different sources, either natural or anthropogenic. Studies suggest that, in developed countries, more than 2/3 of $PM_{2.5}$ can be attributed to anthropogenic sources.
Considering natural sources, the main contributors to PM concentration are wildfires and desert dust. It has been found [33] that their contributions to PM_{10} and $PM_{2.5}$ in Europe are estimated in $4 \div 8\,\mu g/m^3$ and $1 \div 2\,\mu g/m^3$ respectively.
On the other side, the most important anthropogenic sources of PM are fossil fuel combustion (for energy production, heating and transport), biomass burning (in residential heating, agricultural burning and wildfires) and agricultural NH_3 emissions, as seen in the previous section. Part of these sources emit larger quantities of pollutants in winter: then this behaviour influences differently the concentrations of PM_{10} depending on the season of the year.
It is worth noting that, along with the reduction of anthropogenic emissions expected at the EU level, natural ones are destined to increase their relative importance in the whole. Some works [28] have also predict that wildfires will increase due to climate change: in the future, their effect on PM concentration could then approach (even exceed in some cases) that of anthropogenic emissions.

1.2.1 Effects on human health

The main reason for considering PM a matter of concern is the negative effects it has on human health.
Epidemiological and clinical studies have connected PM exposure to a number of negative health outcomes [40], from lung and respiratory system inflammation to increased risk for myocardial infarction, atherosclerosis, hospital admission and mortality in patients with a variety of diseases (chronic obstructive pulmonary disease, cardiovascular diseases) and respiratory cancer.

Due to the heterogeneity of PM composition, the potential of particles to produce adverse health effects depends not only on the size, but also on the composition and consequently on the sources that produced the various parts of the mixture.
Thus, being each epidemiological study related to specific places and periods of time, measurements of the whole PM mass concentration allow to infer only approximately the presence of the various components of PM in the air. In order to understand the different effects of PM component on human health, specific sampling of each component is required.[1]

Studies regarding personal PM exposure, which is related to the overall average concentration a person is subjected to in a certain period of time, can't take into account only outdoor PM concentration: having it been demonstrated that indoor PM levels can be significantly higher than outdoor concentrations at a certain time, a full characterization of the personal exposure is required in order to completely analyse the causal relationship between exposure itself and a range of diseases that could be observed.
Another factor affecting personal exposure is the variability of PM concentrations, also within the same city, as demonstrated by differences of daily values between roadside and urban background monitoring stations: people who live in a district can be exposed to a different outdoor concentration than those residing in another. Furthermore, living near busy streets generally means being exposed to higher concentrations.
Commuting is another important factor that affects personal exposure: PM concentration inside a car can reach values that are $5 \div 10$ times higher than those of a roadside monitoring station placed nearby.[40]
Personal exposure is also generally higher for children and people that exercise outdoor, due to the increased minute ventilation per unit mass with respect to an average person.

[1] It is worth mentioning that the improvement in sampling technologies has increasingly allowed to start routine and continuous measurements of single PM components and measurements of particle number concentration (particularly useful for ultrafine particles), along with the continuous measurements of PM_{10} and $PM_{2.5}$ mass concentration: these data can positively influence the depth of studies on the effect of these components on human health.

It is necessary to take into consideration also the biopersistence that characterizes mainly insoluble particles (black carbon, in particular).
In fact, about one third of insoluble/biopersistent particles are retained in the lungs for long periods of time (even years).

Toxicological studies are trying to understand which characteristics of PM produce negative consequences for people health.[40]
PM particles are inhaled and deposited throughout the respiratory system, and then deposited selectively depending on their size. The effects can be directly related with the respiratory system (inflammatory response, aggravation of existing diseases, weakening of defence mechanisms, e.g. against bacteria), worsening asthma and even allergies. Furthermore, these effects can induce secondary reactions in other systems, or particles can get through the respiratory tract and enter the body.
Toxicity arises from the interaction of PM particles with biological tissues: in particular, each component of PM interacts with different biological systems. A number of features of the particles (size fraction, mass concentration, number concentration, acidity, constituent chemicals, water solubility, ...) can favour those interactions and are currently under examination.

Critically, particle size is partially correlated with a number of features (e.g. chemical composition), making it difficult to understand which are the actual effect-related features. The identification of particle-size-dependent effect independent of chemical composition (i.e. caused by the mere presence of PM in the affected tissue) can help in these kind of analysis. In the case of ultrafine particles, for example, the small size itself produces a more significant response by pulmonary tissues with respect to larger particles with the same chemical composition and mass concentration.[2] On the other hand, particles can catalyse chemical reaction on their surfaces and also act as carrier of toxic chemical compounds, allowing them to reach the inner regions of the respiratory system.

1.2.2 Legal thresholds and guideline values for concentration

In the European Union, the reference values for PM_{10} and $PM_{2.5}$ concentration have been established by:

- the air quality standards for the protection of health, as defined in the Ambient Air Quality Directives ([17], [18]), that set legal thresholds which must be respected;

[2] Actually, studies have found that ultrafine particles can get translocated from the respiratory system to the brain, the nervous system and the liver, aside from being able to affect cellular organelles. The evaluation of the relation with brain tumour is currently under assessment [38].

- the WHO Air Quality Guidelines, updated in 2005 ([39], [40]).

Table 1.2 and 1.3 summarize the reference values for both pollutants.

It must be noted that two additional targets for EU member States have been sets by EU Directives for $PM_{2.5}$, based on the Average Exposure Indicator (AEI), i.e. the 3-year average of concentration values for a set of urban background stations selected on purpose by every national authority. These targets are:

- the Exposure Concentration Obligation, i.e. the target AEI value to be reached in 2015 (averaging 2013-2015 concentration values), set at $20\,\mu g/m^3$;
- the National Exposure Reduction Target (NERT), i.e. the percentage of reduction in AEI (comprised between 0% and 20%, with different values for each EU member State) to be met in 2020.

Reference value name	Averaging period	Value	Max n. of exceedances
Daily limit value (EU)	1 day	$50\,\mu g/m^3$	35 days per year
Annual limit value (EU)	Calendar year	$40\,\mu g/m^3$	/
Daily AQG value (WHO)	1 day	$50\,\mu g/m^3$	3 days per year
Annual AQG value (WHO)	Calendar year	$20\,\mu g/m^3$	/

Table 1.2: Reference values for PM_{10} concentration limits.

Reference value name	Averaging period	Value	Max n. of exceedances
Annual limit value (EU)	Calendar year	$25\,\mu g/m^3$	/
Daily AQG value (WHO)	1 day	$25\,\mu g/m^3$	3 days per year
Annual AQG value (WHO)	Calendar year	$10\,\mu g/m^3$	/

Table 1.3: Reference values for $PM_{2.5}$ concentration limits.

1.2.3 Emission and concentration monitoring

As already mentioned, PM emission comes from both natural and anthropogenic sources.

Natural sources (like desert dust transport events and wildfires) can't be controlled and usually contribute to background concentration, but also to high-concentration PM pollution events.[3]
As concerns anthropogenic sources, both primary PM and precursors emissions must be considered for assessing the contribution from the different sources.

In the European Union (see Figure 1.1), the periodic monitoring lead to observe that the emissions of secondary PM precursors (NO_x, SO_x and VOCs, with the important exception of NH_3) have been reduced more remarkably than primary PM emissions ([16]).

As said above, particulate matter concentration in the atmosphere is a critical subject of monitoring for multiple reasons, going from human health protection to assessment of air quality improvement measures. PM concentration is connected both to primary PM emissions and secondary PM produced by chemical reaction in atmosphere and to meteorological and geomorphological conditions that facilitate translocation, dispersion and deposition of pollutants far from their sources; secondary PM generation itself is linked to various environmental factors that influence the mentioned reactions.

The European Environment Agency periodically monitors the reference values for PM_{10} and $PM_{2.5}$ concentration in 39 countries (28 member States and 11 other reporting countries), assessing the compliance with EU legal thresholds.[16]

Figure 1.3 shows the results of 2016 monitoring on about 2900 stations for the EU daily limit value of $50\,\mu g/m^3$: in particular, each point shows the 36^{th} highest daily mean concentration value in the year. Since the threshold can be exceeded for 35 days at maximum, a point in the map that go beyond that reference value actually corresponds to a non-complying site.
In 2016 19% of the reporting stations exceeded the limit value. Notably, 97% of these stations where either urban or suburban sites.

The annual mean concentration values for PM_{10} in 2016 with respect to the EU limit value of $40\,\mu g/m^3$ are reported in Figure 1.4.
In this case, only 6% of the considered station reported a value above the threshold.
It is also interesting to mention that 48% of the considered stations exceeded the WHO guideline values for the mean annual concentration ($20\,\mu g/m^3$). As can be seen from the

[3]Specific environmental conditions can lead to the so-called *PM pollution episodes* [16], defined as large-scale events of widespread high values of concentration for PM_{10} or $PM_{2.5}$. Such events generally happen in autumn, winter or spring, when favourable meteorological conditions combine with large anthropogenic emissions (mainly in the agricultural or residential sector) and the eventual addition of contribution from natural sources (e.g. dust transport). In these situations, the level of PM concentration can exceed the reference values (guideline values or even legal thresholds) for several days.

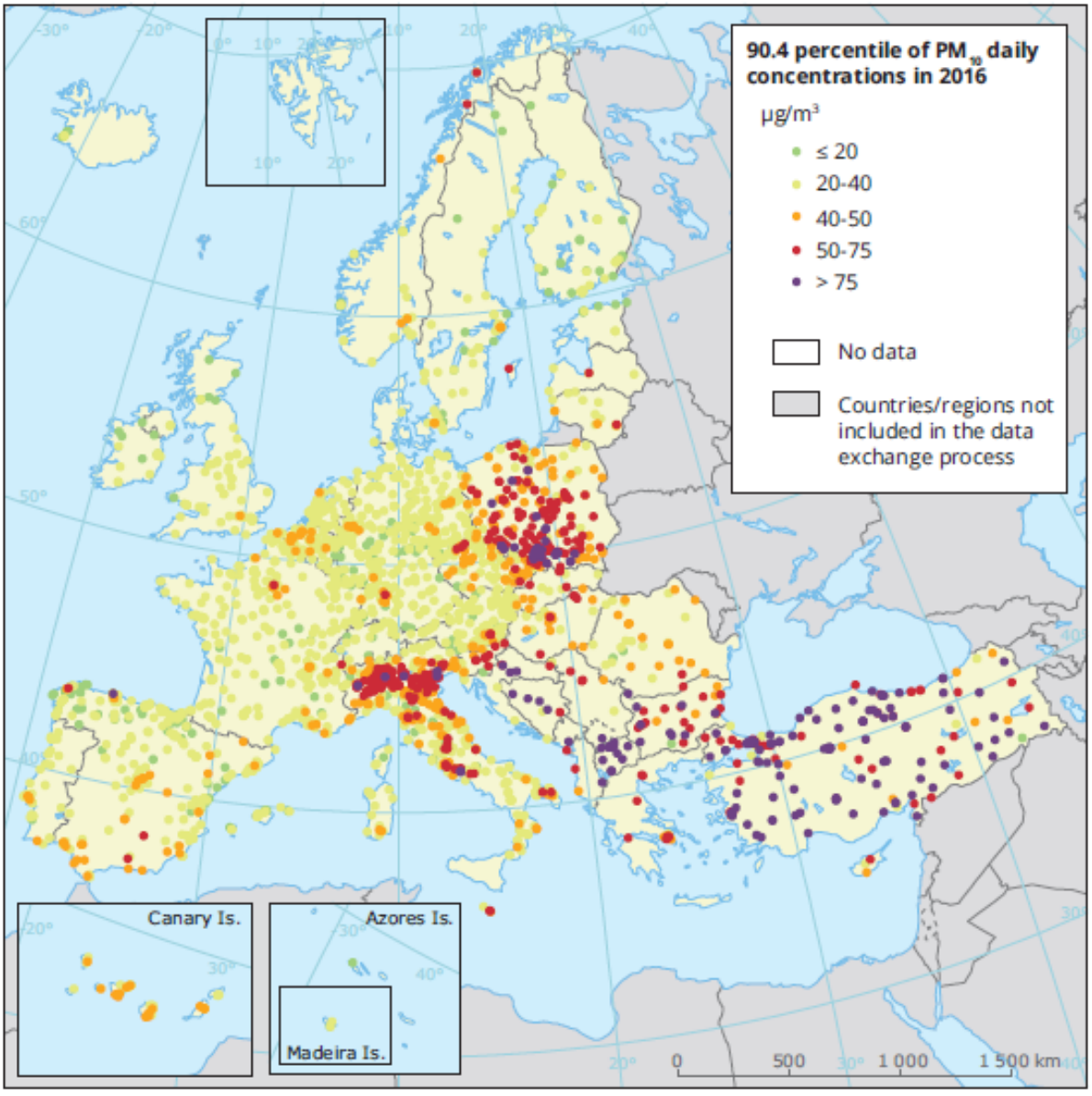


Figure 1.3: PM_{10} daily mean concentrations in Europe in 2016 - 36^{th} highest value (from [16]).

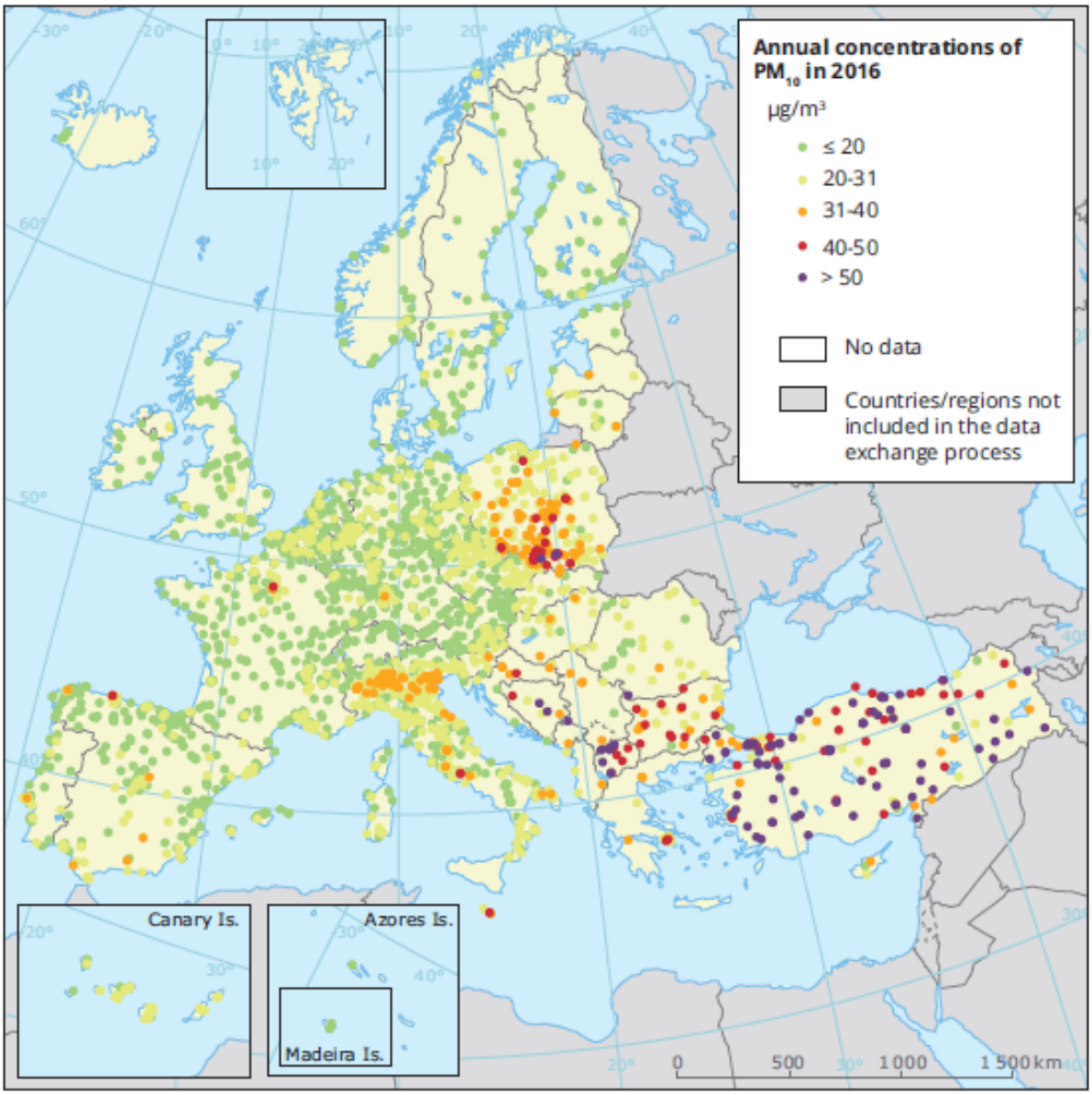


Figure 1.4: PM_{10} annual mean concentrations in Europe in 2016 (from [16]).

map, only four countries (Estonia, Iceland, Ireland and Switzerland) have all the stations annual mean below this value.

The last two maps show large areas where PM_{10} concentration values are generally high: the Po valley, eastern Europe, the Balkans and Turkey. Apart from the first region, the others show exceeding values both in the daily and the annual limit values.

Figure 1.5 shows the annual mean of $PM_{2.5}$ concentration values for 1327 monitoring stations in 2016, with respect to the EU limit value of $25\,\mu g/m^3$.
In this case, values above concentration threshold were reported from 5% of the monitoring station, mainly (97%) in urban areas.
Concerning WHO annual guideline value ($10\,\mu g/m^3$), it was exceeded at 68% of the stations; five countries (Estonia, Finland, Hungary, Norway and Switzerland) reported only values below the guideline.

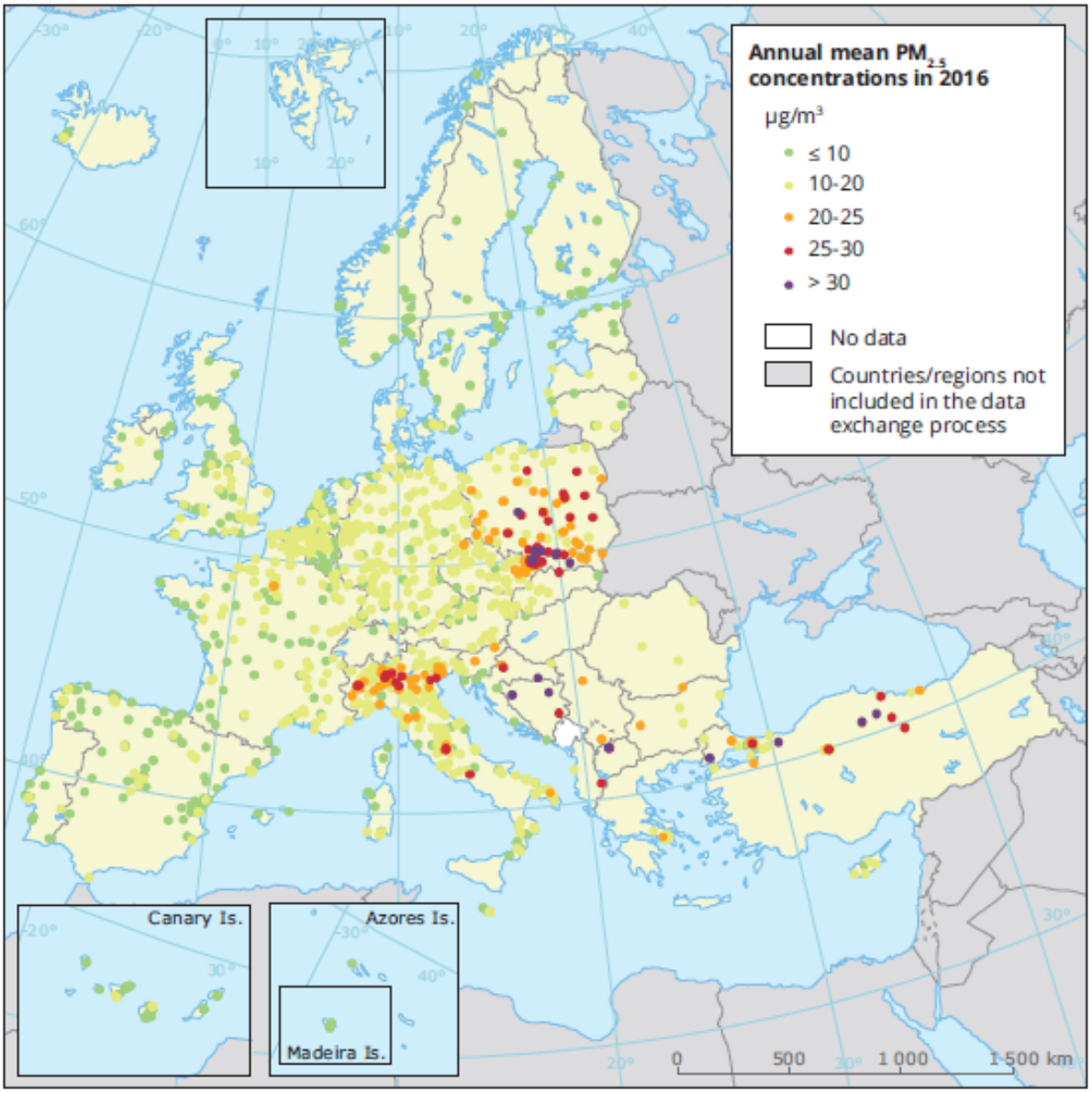


Figure 1.5: $PM_{2.5}$ annual mean concentrations in Europe in 2016 (from [16]).

1.3 Context overview: Emilia-Romagna

Various characteristics of Emilia-Romagna, the Italian administrative region whose area is the research field of this work, make it a place where pollutant concentrations (PM, O_3 and NO_2 in particular) regularly exceed the EU legal thresholds.
In this region, the Regional Agency for Prevention, Environment and Energy (ARPAE, *Agenzia Regionale per la Protezione dell'Ambiente e l'Energia*) is the public agency responsible for monitoring air quality, drafting emisson inventories and assessing the impact of measures addressed to air pollution issues in the region.

1.3.1 Geographical and meteorological elements

Emilia-Romagna is located in the Po-Adriatic basin: half of the region (the north-north-east part) corresponds to a flat strip on the southern side of the Po river, while the other half is characterized by a portion of the Apennine chain of mountains. The whole Po-Adriatic basin is sorrounded by the Apennines (southern edge) and the Alps (western and northern edges), while the Adriatic sea closes the area on the east side.

Although westerly winds are prevalent at these latitudes, the enclosing orography of the basin determines unfavourable conditions for the dispersion of air: in fact, calms charaterizes the wind regime of the region, as air circulation between northern Italy and the rest of the continental Europe is hindered by mountains, and dispersion processes in absence of significant wind require days in order to remove pollutants from the air. In the low Po valley wind speed does not exceed 2.5 m/s in general (4 m/s can be reached on the coast) and the mixing of air is due primarily to the thermal component of the wind, which depends on the solar radiation: as radiation increases during summer months, this leads to a reduction in concentrations for various pollutants (including PM and NO_2).
Therefore a useful meteorological parameter is the mixing height, which describes the vertical depth (above surface) which is available for air mixing processes such as convection. As it will be seen, its value is particularly low during winter months since it is correlated with the presence of the thermal component of the wind.
In the same period it is common to see temperature inversion: this term refers to the situation in which a warmer air mass is found above a colder one that is immediately near the surface. In this condition, convection is hindered so that air stagnates and pollution tends to build up, leading to very high concentrations that are quite homogeneous in the whole area.

On the other hand, during summer months photochemical pollution (that involves O_3) is enhanced by solar radiation and higher temperature: for this reason a variety of secondary pollutants show higher concentrations during this period. As mentioned above, the stronger thermal component of the wind during summer allows a more effective

mixing in the atmosphere, a condition that produces an approximately homogeneous distribution of these pollutants in the basin.

1.3.2 Anthropic pression and emission sources

The Po-Adriatic basin is home to over 23 million people (nearly 40% of the Italian population). This macroregion contributes to over the 50% of the national GDP.
In this context, Emilia-Romagna is characterized by a high population density in its flat part (198 inhabitants/km^2) and a total land consumption of about 10% of its total surface area.
Cities and industrial areas are concentrated along the main communication routes. The rest of the flatland is occupied by intensive agriculture and animal farming.
The urban polycentrism that characterizes the area is the cause of the great demand of mobility, that in turn produces huge emissions in the transport sector; this adds to transit traffic, due to the central position of the region with respect to the main national routes.

The strong anthropization of the region, with the concentration of economic activities (industrial, agricultural and farming sector), vast residential areas and the aforementioned high levels of traffic, is responsible for the important emissions of air pollutants. As said, ARPAE is responsible for drafting emission inventories for the region. In Table 1.4 the contributions to the emission of pollutants from the different economic sectors for the year 2015 are reported, as provided in the last report available [3].

Sector	SO_2 (t)	NO_x (t)	PM_{10} (t)	$PM_{2.5}$ (t)	NH_3 (t)	VOC (t)	CH_4 (t)
MS1	387	4057	44	43	17	146	133
MS2	216	6238	5606	5548	107	6505	3804
MS3	8112	10915	469	366	25	504	315
MS4	2614	1892	723	468	131	4428	1594
MS5	2	2	0	0	0	2902	35723
MS6	0	0	302	255	0	30392	0
MS7	60	47229	2859	2189	424	16891	998
MS8	81	9491	423	422	2	974	14
MS9	23	674	8	8	164	54	44476
MS10	0	503	532	241	47565	41192	69322
MS11	/	/	/	/	/	34940	/
Total	11495	81001	10966	9540	48435	133988	156379

Table 1.4: Contribution of principal air pollutants per sector in EU-28 (2016).

Sectors are classified following the SNAP (Selected Nomenclature for sources of Air

Pollution) coding, required by the CORINAIR metodology; the economic sectors for each class are specified in Table 1.5.

Class	Economic sectors	Class	Economic sectors
MS1	Energy production and fuel reprocessing	MS7	Road transport
MS2	Non-industrial burning	MS8	Other mobile sources and machineries
MS3	Industrial burning	MS9	Waste processing and disposal
MS4	Production processes	MS10	Agriculture
MS5	Fuel extraction and distribution	MS11	Other sources and absorptions
MS6	Use of solvents		

Table 1.5: Sectors included in SNAP coding.

1.3.3 Air quality and meteorology

The load of anthropogenic primary emissions in the basin, together with meteorological factors that determine air stagnation, leads to high concentrations of various pollutants that facilitate the formation of secondary pollutants. Significant levels of concentration tend to persist at ground level until major meteorological events (winds, rains) of sufficient strenght allow the removal of pollutants through transport or deposition.

As already said, emissions are only the starting element in the process of air pollution, while meteorological factors are the main responsibles for the processes of dispersion, translocation and catalysis of chemical reactions that produce secondary pollutants.
Thus a statistical assessment of the influence of meteorological factors on the concentration of pollutants is of particular interest in order to evaluate the conditions that favour the building up and the dispersion of pollutants. Such analysis is the particular focus of the present work.

1.3.4 Regional concentration monitoring

In order to measure regularly pollutants concentration, a monitoring network has been established by ARPAE within the regional borders: it is composed of 47 monitoring station equipped with automatic analysers for different compounds (nitrous oxides, particulate matter, ozone,...; each station can have a different combination of these instruments).
Following the national legislation (*dlgs.* 155/2010, art. 3), the region has been divided into four zones: *Agglomerato* (that includes the regional capital and its neighbouring towns), *Appennino*, *Pianura ovest* and *Pianura est*.

Similarly to the geographical classification of sites given in a previous section, ARPAE monitoring stations are classified into 4 categories according to their location. Each category is characterized by a specific combination of measure instruments, as follows:

- 12 *Traffico Urbano* (TU; traffic) stations, equipped with PM_{10} and NO_2 analysers (sometimes also with CO and C_6H_6 (benzene) analysers);
- 21 *Fondo Urbano* (FU) and *Fondo Suburbano* (FS) (urban background) stations, equipped with PM_{10}, $PM_{2.5}$, O_3 and NO_2 analysers; some FU stations are also used to collect samples of PM in order to periodically determine some metals (Pb, As, Ni, Cd) concentrations;
- 14 *Fondo Rurale* (FR; regional background) stations, equipped with PM_{10}, $PM_{2.5}$, O_3 and NO_2 analysers;

Figure 1.6 shows the locations of the monitoring stations on the regional area; the number of station per category and zone is summarized in Table 1.6.

Zone	TU	FU	FS	FR	Total per zone
Agglomerato	2	1	1	0	4
Appennino	0	0	0	5	5
Pianura ovest	5	5	4	4	18
Pianura est	5	6	4	5	20
Total per class	12	12	9	14	47

Table 1.6: Number of monitoring station per zone and category.

The results of monitoring on air pollutants reported by ARPAE for 2017 are obtained applying a geographical model based on data recorded by urban and rural background stations.[1]

As can be understood by Figure 1.7, in that year the plain part of the region has been affected by several exceedences of EU daily limit value of $50\,\mu g/m^3$ for PM_{10}: in particular, the northern area that encloses the cities of Piacenza, Parma, Reggio Emilia, Modena and Ferrara has illegally passed the threshold of 35 daily exceedances in the year.
The map does not allow to figure out if the WHO guideline objective for exceedances (3 days per year at maximum) of the same daily mean concentration value was met in any part of the region.

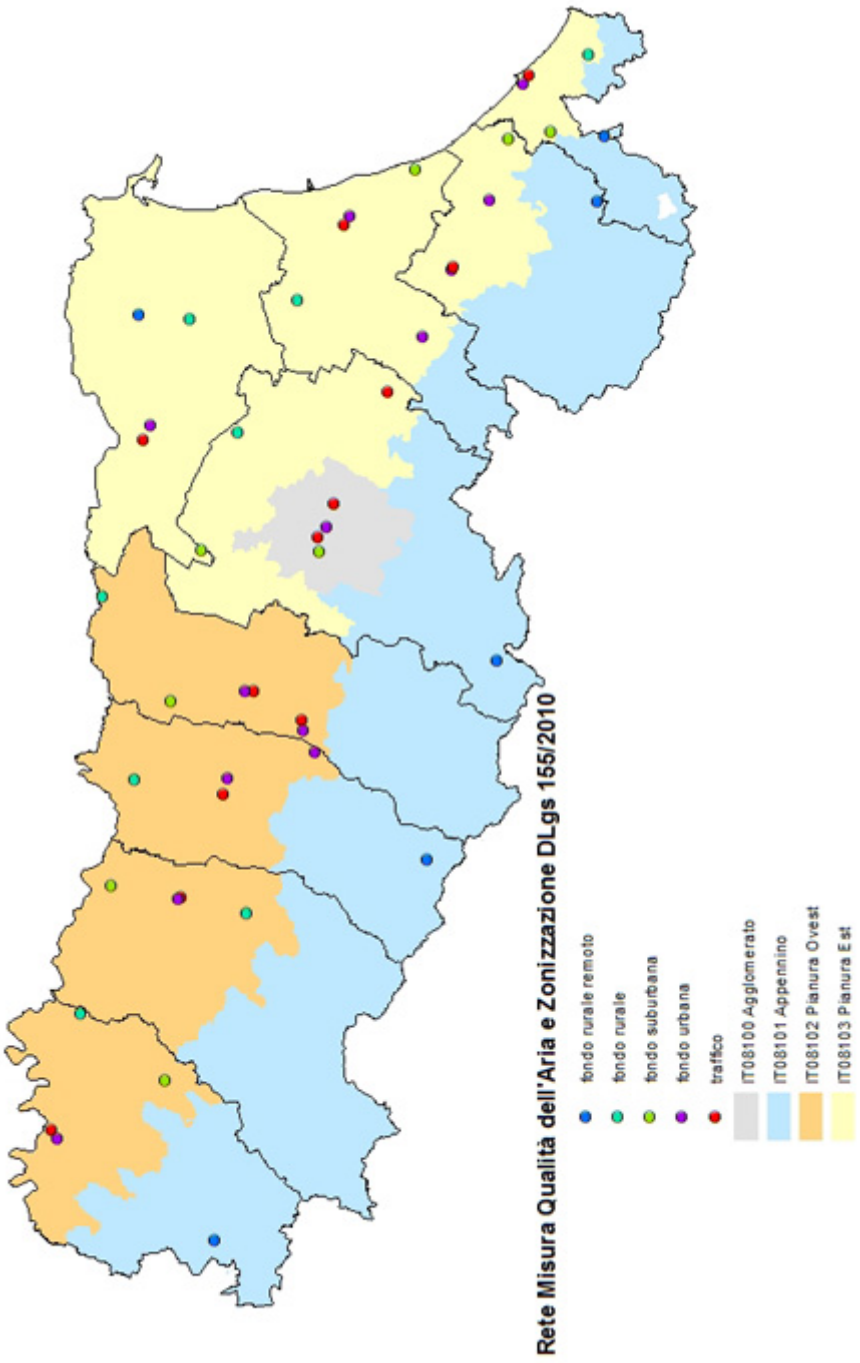


Figure 1.6: ARPAE monitoring network (from [2]).

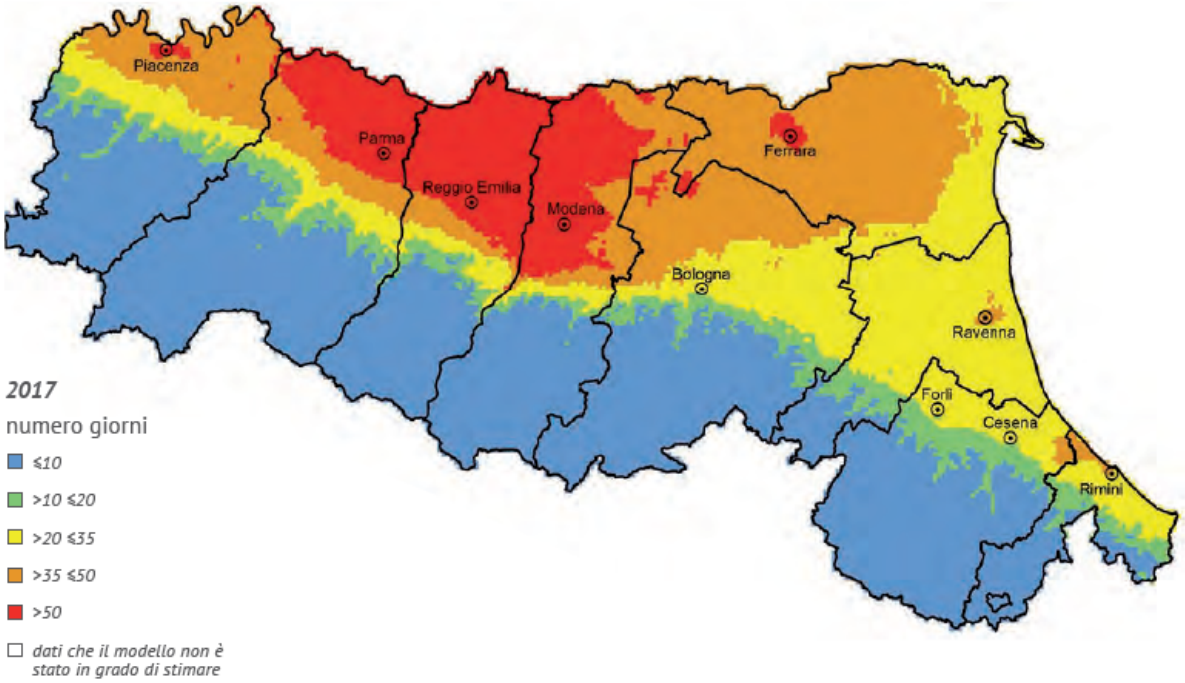


Figure 1.7: PM_{10} daily mean concentrations in Emilia-Romagna in 2017 - Number of exceedances of the limit value (estimate) (from [1]).

Talking about the annual mean concentration of PM_{10}, the map in Figure 1.8 shows that the EU limit value of $40\,\mu g/m^3$ has not been exceeded in any part of the region.
On the other hand, it must be noticed that the WHO guideline for the same reference value ($20\,\mu g/m^3$) was exceeded in the whole plain part of the region.

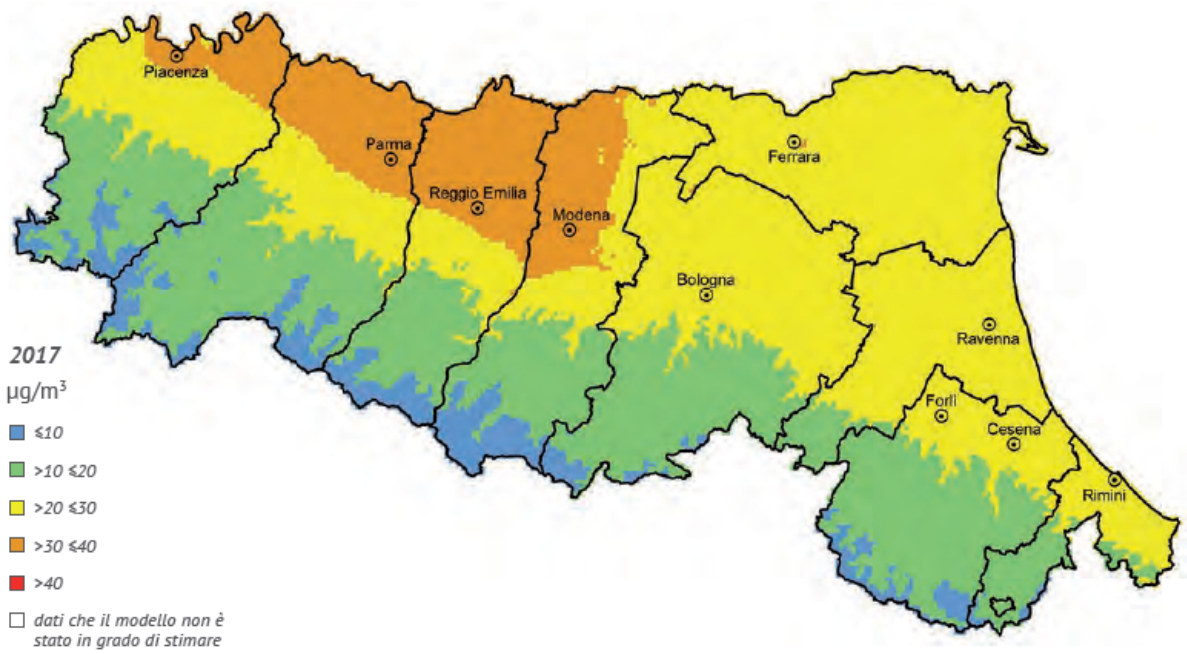


Figure 1.8: PM_{10} annual mean concentrations in Emilia-Romagna in 2017 (estimate) (from [1]).

The annual mean concentration estimate for $PM_{2.5}$ is shown in Figure 1.9. Exceedances of the EU limit value ($25\,\mu g/m^3$) are confined to small rural areas in the northern part of the region.
At the same time, the WHO guideline value ($10\,\mu g/m^3$) is exceeded in most of the regional area, including in the Apennine valleys.

1.4 Models for PM_{10} prediction

The final aim of this work is to develop a statistical model based on machine learning able to predict PM_{10} concentration levels by exploiting the relationship between the pollutant's concentration and meteorological conditions observed in the capital cities of the provinces of Emilia-Romagna.
A similar classification task has been previously performed for the regional area, as outlined in the 2018 report on regional air quality published by ARPAE.[1] In that

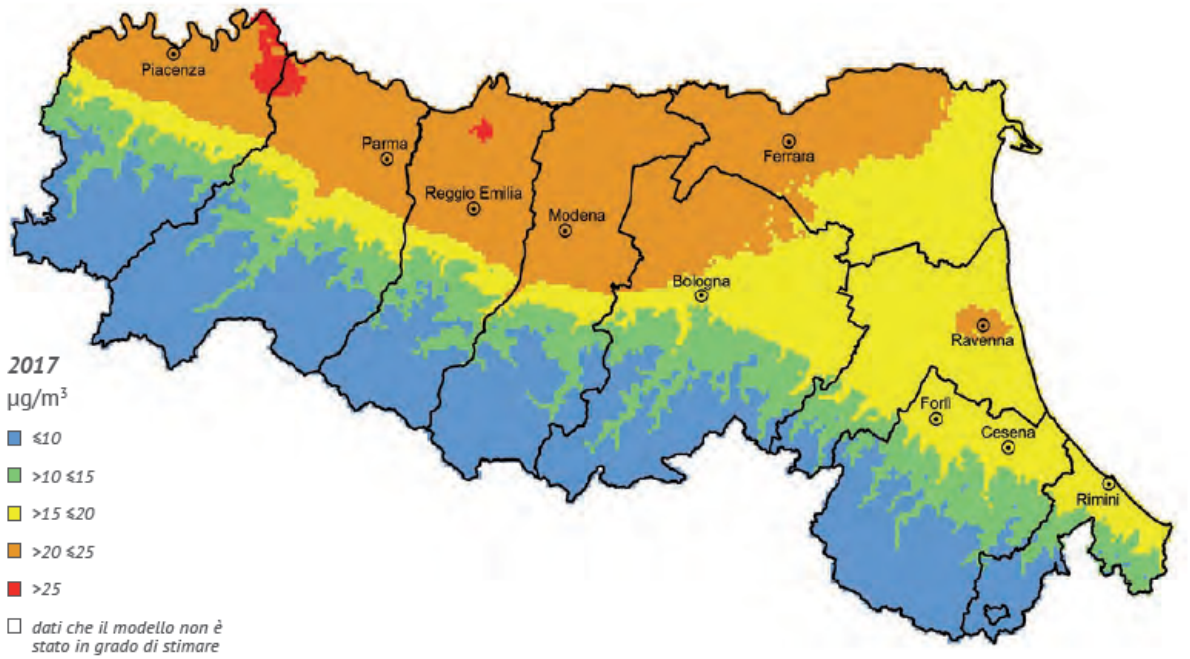


Figure 1.9: $PM_{2.5}$ annual mean concentrations in Emilia-Romagna in 2017 (estimate) (from [1]).

report two syntetic indicators for the evaluation of the meteorological condition that can favour the building up of pollutants (the *number of favourable days for PM_{10} build-up* and the *number of favourable days for O_3 build-up*) are presented: these indexes represent the days in which meteorological conditions are "good" for the generation and build-up of the pollutant, i.e. the days characterized by a high probability that legal threshold would be exceeded.
The meteorological parameter values that define the *favourable conditions* have been obtained [4] using classification tree technique on a dataset of significant meteorological variables corresponding to one year of observations in Bologna, recorded in the *Giardini Margherita* monitoring station. The model has been used in order to make predictions for the whole regional area, despite being trained on local data.
The analysis performed by ARPAE on the meteorological data for years 2008-2017 has lead to the results shown in Figure 1.10, that describes the percentage of days (with respect to the considered seasons) in which the model predicts that the concentration of the pollutant will exceed the legal threshold.

As the task was similar to the one of the present work (in particular for the geographical area that was considered), it must be reported that no information has been obtained from the author of [4] apart from those aforementioned: in particular, no precise tempo-

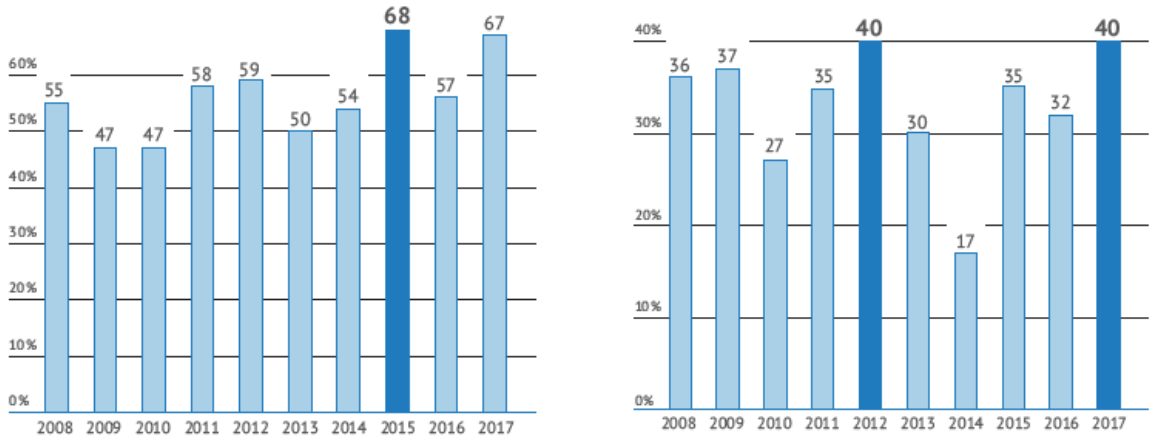


Figure 1.10: Share of favourable days for PM_{10} (during autumn and winter; on the left) and O_3 (during spring and summer; on the right) build-up in 2008-2017 (from [1]).

ral information on the considered set of data, nor the dataset itself, nor the reasons for which a classification tree was chosen as model have been made available.

Since the lack of information made it impossible to use the results obtained by the classification tree applied on data from Bologna as a benchmark, while on the other hand literature provides a large number of works in which quantitative predictions have been performed starting from similars sets of meteorological data, a completely independent analysis has been performed starting from a newly assembled dataset on the same kind of data, geographically widened in order to comprehend all the provincial capital cities of Emilia-Romagna, so to achieve a quantitative result similar to the one obtained in [4] for each city separately.

In the next paragraphs, an overview of past works is made. The main focus of the research is that of stastical models based on machine learning which have been applied to tasks similar to the one of interest fot the present work.[35]
As different kinds of models can be applied for the task of predicting a pollutant's concentration [5], here only statistical methods have been considered. Other kinds of models, such as numerical ones, have not. There are in fact important differences between these two categories: numerical models are based on a simulation of the concentration of pollutants on spatial and temporal scales, based on the chemical and physical processes that take place in the atmosphere, where the geography of the considered area, the atmospheric behaviour in a layer of defined height, the position and the characteristics of the sources of emissions and other inputs of the model must be characterized on a defined temporal scale. Characterizations of this kind can be performed using suitable

simulation models that starts from information given on a discrete spatial and temporal scale and determines numerically the most likely behaviour for each considered variable. These models can be characterised by a significant precision both in the spatial and temporal scales; on the other hand, the volume of data and the computational costs that are necessary in order to run them are relevant. On the other hand, a different kind of results can be obtained by using statistical models: these does not take into account the physical and chemical processed given a certain geospatial region of interest, but evaluate the behaviour of a number of input variables in a statistical way in order to predict output descriptors that are related to the considered predictors. As the geographical range of those models can be reduced to a point-level prediction, the quantity of necessary data is reduced and in general they are more easily available thanks to the presence of various open databases.

1.4.1 Previous works on PM_{10} forecasting

A study performed on a 1999-2001 dataset of 5 meteorological variables collected in the metropolitan area of greater Athens [9] presents a comparison between multilinear regression techniques and feed-forward multilayer perceptrons. The best performer, a multilayer perceptron trained only on meteorological data, was characterized respectively by $R^2 = 0.47\,(0.03)$ and $RMSE = 21.19\,(0.95)\,\mu g/m^3$. Adding the previous-day PM_{10} concentration values to the set of predictors improved the performance of that model, leading to $R^2 = 0.65\,(0.03)$ and $RMSE = 16.94\,(0.76)\,\mu g/m^3$.

Similar works have also involved PM_{10} hourly concentration prediction: in [25] multilinear regressions and multilayer perceptrons were trained on data taken from four different locations in Greater Athens; the best RMSE ($12.16\,(0.67)\,\mu g/m^3$) was obtained with a multilayer perceptron coupled with a genetic algorithm for variable selection.

Another work on PM_{10} hourly concentration prediction was performed on data (both meteorological and descriptive, as "day of the week") collected in 2005 in Phoenix, Arizona [22]: a comparison between a deterministic model (Community Air Quality Modelling System, or $CMAQ$) and a statistical one (3-layer neural network) was made, resulting in RMSE values of $25 \div 40\,\mu g/m^3$ with the statistical model performing better.

1.4.2 Previous works on other pollutants' forecasting

In order to get a wider knowledge of previous efforts in the statistical modelling field applied to the task of pollutant concentration prediction on a meteorological basis, other works have been considered.

A study on 8 USA cities [11] comparing the performance of multilinear regressions and neural networks in predicting O_3 daily maximum 1-hour concentrations starting from a dataset of four meteorological variables showed that neural network generally outperforms multilinear regressions. The best result for both algorithms was obtained considering also the previous-day value of O_3 concentration: in the best case (neural network) the regression reached $R^2 = 0.69\,(0.04)$ and $RMSE = 9.24\,(0.54)$.

In Canada, a classification tree analysis on maximum surface O_3 daily observation has been performed on a 1985-92 dataset concerning the areas of Vancouver and the lower Fraser River valley [6]: the study, which has been considered useful for the wide range (57) of variables chosen as predictors, describes a categorical approach in considering the response variable (i.e. predictions and measured values were classified with respect to 2 thresholds and matching between classes has been evaluated).

A comparison between multilinear regression, ARIMA model and neural networks was performed on a 1993-1994 dataset containing temperature- and wind-related variables, along with emissions of a number of gaseous compounds, in order to evaluate the best result in predicting hourly daily maximum ozone level O_3 in Dallas, Texas [42]. The best result was achieved by the neural network, with a MAD of $6.4 ppb \approx 12.8\,\mu g/m^3$.

An attempt to model hourly concentrations of NO_x and NO_2 in Central London starting from a dataset of 6 meteorological variables, which was supplemented with sinusoidal indexes accounting for the hour of the day or with the previous-hour concentration of either NO_x or NO_2, was made using multilayer perceptrons [23]. The best model, which used the lagged concentration for NO_x prediction, reached a best R^2 value of $0.92\,(0.02)$ corresponding to $RMSE = 33.8\,(2.3)$.
Furthermore, a comparative study in 5 UK cities [24] aimed to analyse the performances of multilinear regression, regression trees and multilayer perceptrons on a dataset of 7 meteorological variables from the period 1993-1997 in order to predict hourly O_3 concentrations; seasonal sinusoidal indexes representing the day of the year and the time of the day were also added in order to account for periodical variation of emissions of precursors. The best test R^2 value was $0.68\,(0.01)$, corresponding to $RMSE = 6.60\,(0.13)$.

The Greater Athens Area hosted further analysis. One [43] was performed on 1987-1990 meteorological data (3 variables), along with previous-day maximum hour concentration of NO_x and day of the week index, which were used to predict both the the increase or decrease of the concentration of the same pollutant and a numerical forecast of that concentration. In the second task the best achieved test RMSEs were of $45\mu g/m^3$ (in case of increasing concentration) and $35\mu g/m^3$ (in case of decreasing concentration).
Again in Athens, a dataset collected in the summer months of 1987-1993 has been used to evaluate regression models for the prediction of the daily maximum of O_3 hourly

concentrations [7]. The works involved both meteorological variables, previous-day O_3 hourly maximum and the same-day concentrations of precursor gases NO_2 and CO, which were feeded into multiple linear regression and ARIMA model. The best performed showed a RMSE of $47.99\,\mu g/m^3$.
Another work [10] focused again on maximum hourly concentration of O_3, starting from summer data collected in Athens during 1992-1999, in order to compare multilinear regressions and neural networks. A set of 8 meteorological variables was considered, while O_3 measures were taken from 4 nearby monitoring stations. In this case the best reported R^2 was $0.59\,(0.01)$, corresponding to a RMSE of $21.7\,(0.7)\,\mu g/m^3$, obtained using a neural network.
In the same area, an analysis concerning O_3 and NO_2 daily concentration levels employed feed-forward neural networks that were trained using four meteorological variables, previous-day concentration and an index representing the day of the week [32]. Best model R^2 of 0.802 for O_3 and 0.690 for NO_2,corresponding to RMSE values of $27.4\,\mu g/m^3$ and $39.3\,\mu g/m^3$ respectively.

In Beijing, data of three meteorological variables and UV radiation taken during a single summer period were used as predictors for ozone concentration. The results were produced by neural networks of different kind [21]: the best R^2 of $0.76\,(0.01)$ (corresponding to a RMSE of $36.56\,(5.15)\,\mu g/m^3$) was obtained from a neural network trained with genetic algorithm coupled with a SVM classifier.

A work on Besiktas district in Istanbul [30] was focused on forecasting SO_2, CO and PM_{10} levels using neural networks feeded on the basis of spatial criteria: this was made possible by the presence of a network of monitoring stations located in neighboring districts. A non-geographical model (which used only data from Besiktas) with 9 meteorological variables and the same-day level of the considered pollutant was compared with 1-, 2- and 3-neighborhood-trained models in which the set of input variables was expanded on a geographical basis; in the last case, the same-day level of pollutant was feeded to the model as a weighted sum of the concentration levels measured in the considered neighboring districts. The results were reported as error rates based on a error grid: the comparison showed that the 3-neighboring-districts model gave the best results.

A comparison between standard multilinear regression on 14 meteorological variables measured in New Delhi in 2000-2006 and the same model feeded with principal components of the dataset (the overall algorithm is called "principal component regression") was made in order to predict the values of the Indian Air Quality Index [29] (a continuous numerical value in the interval $0 \div 500$, related to the presence of pollutants in the air). The results of the analysis, performed separately for the four seasons, shown a generally better beahviour of the coupled model, with the best R^2 of 0.5767 (RMSE of 30.90) achieved for the winter season.

1.4.3 Some remarks

Being aware of the results presented in the described works, the aim of the present one is to develop a similar quantitative approach starting from the classification model proposed in [4] and widening the geographical range of the analysis to the capitals of the provinces of Emilia-Romagna.
As seen in the previous paragraphs, a number of machine learning methods have been used in tasks similar to the one that is presented in this work. In particular, city-level approach is common since the urban areas are generally more affected by air pollution than rural ones. Noticeably, none of the articles that have been found concerns the geographical region that is analysed in the present work.

The prediction quality significantly varies depending on the work: this can be related to the representativeness of the meteorological quantities in describing the conditions in which the data from PM_{10} monitoring stations have been collected, the kind of source of pollution that are present in the area and other context-related issues that can affect the modelling results. Obviously also the kind of models that have been applied strongly influence the performance.
It must also be noticed that, in a number of cases, the chosen task concerns forecasting the concentration of pollutants for the day following the one in which the measurements have been performed. This purpose can be more interesting from a administrative point of view, since it can provide information in advance to decision-makers so that they can enforce restrictions and other measures in order to tackle a potential critical situation. However, in accordance with the previously described classification model built by ARPAE, it has been chosen not to consider previous-day values of meterological variables for the regression models that have been trained and to limit the predictors to the same-day quantities.

So, concerning the present work, the assessment has been made on two kind of regression models: standard and regularized linear regression models and regression tree-based models. The first have been chosen since they are widely used as a basis point for a large number of works in the cited literature, while the second ones are the counterparts of the classification tree model whose results have been used by ARPAE in [1] to assess the number of days with favourable meteorological conditions for the build-up process of pollutants.
The previous literature review is also useful, apart from providing a framework of models and sets of meteorological and non-meteorological variables that have been used for similar modelling tasks, to give suggestions for further developments of the present analysis (see Chapter 4).

Chapter 2

Materials and methods

In this chapter a description of the data and the models that have been used in this work is given. Starting from an exploratory analysis of the dataset, the chapter continues with a presentation of the algorithms that have been used to address the problem of missing data and to perform modelling tasks: as already explained, the aim is to find the regression model that gives the best prediction of PM_{10} concentration values starting from the values of the meteorological variables measured in the same day.

In section 2.1 the dataset of the considered variables is presented and analysed using exploratory data analysis (EDA) techiniques, with a focus on the relationship between each variable and PM_{10} concentration. The aim is to provide a first overview on the meteorological variables, their trends and distributions in the considered period of time. In section 2.2 the problem of missing data is addressed and the methods that have been applied on the considered dataset is presented. Subsequently the regression models are presented: in section 2.3 the considered linear models for data regression and prediction are described; in section 2.4 the applied regression tree methods are explained. Then, in section 2.5, the procedure of cross-validation which has been used to assess the performance of the chosen regression models is outlined.

Finally, the implementation of the models and the assessment procedures using the R-based software *RStudio* are described in section 2.6.

2.1 Data overview and exploratory analysis

In this section a general overview of the considered set of data and basic statistical analysis techniques that have been used in order to make an exploratory data analysis are presented. Each variable is described and analysed separately, highlighting its relationship with PM_{10} concentrations.

Concerning the dataset in general, a group of daily-measured meteorological variables has been included in it along with the values of PM_{10} daily mean concentrations.
Meteorological variables have been downloaded by the applet *Dext3r*[1], that allows the user to define the set of variables , the time window and the location of interest. PM_{10} daily mean concentrations are available on the ARPAE website, in the thematic section "Aria"[2], in which a page for each monitoring station is present.
Each considered variable corresponds to daily-measured or calculated values in the period of time between the 1st of October, 2012 and the 31st of March, 2018. The measuring processes have taken place in all the provincial capital cities of Emilia-Romagna: Piacenza, Parma, Reggio Emilia, Modena, Bologna, Ferrara, Ravenna, Forlì, Cesena and Rimini. The overall number of considered days is 2008.
PM_{10} concentrations have been measured in the monitoring stations classified as *Fondo Urbano*, while the meteorological variables have been measured by urban meteorological stations that are part of ARPAE meteorological network[3].

In Table 2.1 the lists of the considered monitoring stations for PM_{10} and of the meteorological stations are presented.

City	PM_{10} monitoring station			Meteorological station		
	Name	Latitude	Longitude	Name	Latitude	Longitude
Piacenza	*Parco Montecucco*	45.04	9.67	*Piacenza urbana*	45.05	9.67
Parma	*Cittadella*	44.79	10.33	*Parma urbana*	44.30	10.32
Reggio Emilia	*S. Lazzaro*	44.69	10.66	*Reggio Emilia urbana*	44.68	10.63
Modena	*Parco Ferrari*	44.65	10.91	*Modena urbana*	44.65	10.92
Bologna	*Giardini Margherita*	44.48	11.36	*Bologna urbana*	44.50	11.32
Ferrara	*Villa Fulvia*	44.82	11.65	*Ferrara urbana*	44.82	11.62
Ravenna	*Caorle*	44.42	12.22	*Ravenna urbana*	44.40	12.18
Forlì	*Parco Resistenza*	44.22	12.04	*Forlì urbana*	44.22	12.03
Cesena	*Franchini Angeloni*	44.14	12.24	*Cesena urbana*	44.13	12.23
Rimini	*Marecchia*	44.06	12.55	*Rimini urbana*	44.05	12.57

Table 2.1: Position of PM_{10} monitoring stations and meteorological stations in which data has been measured.

The resulting dataset contains daily-calculated values for the variables summarized in Table 2.2. The properties of each variable are described in the following sections.
The choice of the variables has been performed both by asking experts' opinion and by reviewing literature[4] (see "References" in Table 2.2).

As it is not unusual in the case of automated measuring devices, some missing values are present in the set. In Table 2.3 the absolute number of missing values by variable

[1]Available at `https://simc.arpae.it/dext3r/`
[2]Available at `https://www.arpae.it/dettaglio_generale.asp?id=2921&idlivello=1637`
[3]The chosen stations contain the term *Urbana* in their name and are generally located within the

Name	Description	Unit of measure	References
T_{mean}	Daily mean temperature	°C	[8], [9], [21], [25], [29]
T_{min}	Daily minimum temperature	°C	[9], [29]
T_{max}	Daily maximum temperature	°C	[6], [7], [8], [9], [11], [29], [42]
T_{range}	Daily temperature range ($T_{max} - T_{min}$)	°C	[10], [29]
P	Daily precipitation amount	kg/m^2	[6], [25], [29], [32]
W_{int}	Daily mean wind intensity	m/s	[6], [7], [8], [9], [11], [21], [23], [24], [25], [29], [30], [32], [42], [43]
RE	Daily radiant exposure	J/m^2	[6], [10], [11], [25], [29], [32]
p	Daily mean atmospheric pression	Pa	[23], [24], [25], [29], [30]
$max(H_{mix})$	Daily maximum mixing height	m	[6]
W_{dir}	Daily dominant wind direction	◦	[6], [7], [8], [9], [10], [24], [25] [29], [30], [32], [42], [43]

Table 2.2: List of considered meteorological variables.

and by city is shown, along with the corresponding percentage with respect to all the observation (2008 values for each variable, for each city).
The total number of missing values (considering also temperature range, defined as $T_{range} = T_{max} - T_{min}$) amounts to 2591. Samples can be grouped with respect to the number of missing values contained in them. Table 2.4 provides the distribution of missing values in the incomplete samples.
The total number of incomplete samples (i.e. those for which at least the value of one variable is missing) is 1772. The ways in which the problem of missing data values has been approached in this work will be discussed in Section 2.2.

2.1.1 EDA techniques

In order to perform a so called *Exploratory Data Analysis* (or EDA) [41] on the considered dataset, some quantities have to be defined in advance.

The main summary measures provided in order to numerically represent the distribution of data in the considered datasets are the *sample quantiles* q_p, i.e. numerical values that have the same units of measure as the data and exceed a proportion of the data (which are considered as arranged in ascending order) corresponding to the subscript p, with $0 < p < 1$. From a statistical point of view, a quantile can be seen as the value that is expected to exceed with a certain probability p a randomly chosen member of the data set.

urban area.

[4] As seen in section 1.4, in order to consider a wider collection of articles, the topics considered in the performed literature research have been not only the assessment of the effects of meteorological variables on PM_{10} concentration, but also in the case of other pollutants.

City	T_{mean}	T_{min}, T_{max}	P	W_{int}	RE	p	$max(H_{mix})$	W_{dir}	$[PM_{10}]$
Piacenza	0 (0%)	1 (0.05%)	2 (0.1%)	0 (0%)	1 (0.05%)	26 (1.29%)	0 (0%)	1 (0.05%)	109 (5.43%)
Parma	0 (0%)	1 (0.05%)	89 (4.43%)	0 (0%)	1 (0.05%)	74 (3.69%)	0 (0%)	0 (0%)	54 (2.69%)
Reggio Emilia	19 (0.95%)	105 (5.23%)	116 (5.78%)	32 (1.59%)	54 (2.69%)	235 (11.70%)	0 (0%)	32 (1.59%)	62 (3.09%)
Modena	1 (0.05%)	1 (0.05%)	5 (0.25%)	3 (0.15%)	3 (0.15%)	30 (1.49%)	0 (0%)	3 (0.15%)	37 (1.84%)
Bologna	0 (0%)	0 (0%)	0 (0%)	5 (0.25%)	1 (0.05%)	25 (1.25%)	0 (0%)	9 (0.45%)	142 (7.07%)
Ferrara	0 (0%)	2 (0.10%)	17 (0.85%)	0 (0%)	1 (0.05%)	49 (2.44%)	0 (0%)	5 (0.25%)	51 (2.54%)
Ravenna	0 (0%)	1 (0.05%)	20 (1.00%)	0 (0%)	1 (0.05%)	17 (0.85%)	0 (0%)	3 (0.15%)	54 (2.69%)
Forlì	6 (0.30%)	7 (0.35%)	28 (1.39%)	6 (0.30%)	7 (0.35%)	43 (2.14%)	0 (0%)	86 (4.28%)	93 (4.63%)
Cesena	6 (0.30%)	6 (0.30%)	18 (0.90%)	6 (0.30%)	7 (0.35%)	35 (1.74%)	0 (0%)	8 (0.40%)	67 (3.34%)
Rimini	4 (0.20%)	4 (0.20%)	6 (0.30%)	156 (7.77%)	5 (0.25%)	33 (1.64%)	0 (0%)	101 (5.03%)	97 (4.83%)
Total	36	128 (x 2)	301	208	81	567	0	248	766

Table 2.3: Number of missing values in the dataset (percentages are calculated with respect to 2008, i.e. the number of samples taken for each variable in each city).

Missing values per sample	Number of samples	Missing values per sample	Number of samples
0	18308	5	86
1	1496	6	3
2	132	7	4
3	16	8	8
4	0	9	27

Table 2.4: Distribution of missing values in the incomplete samples.

The q_p quantile corresponds to the $(p * 100)th$ percentile.

The *median*, which corresponds to $q_{0.5}$ (or the 50th percentile), is a common measure of *location* of the distribution of tha data. Given a set of n data, it is defined as:

$$q_{0.5} = \begin{cases} x_{(n+1)/2} & (n \text{ odd}) \\ \frac{1}{2}\left(x_{n/2} + x_{[n/2]+1}\right) & (n \text{ even}) \end{cases} \tag{2.1}$$

Another measure of location which is used in this chapter is the *mean*, or *sample average*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.2}$$

Because its lack of robustness and resistance, the mean is generally replaced by the median. In this chapter the former statistic is used exclusively to quantify long-period (e.g. monthy, annual) values of considered quantities, while the latter is preferred for characterising the distribution of data and for graphical representations.

The *spread* of a distribution of data is given as the *interquartile range* (IQR), computed starting from the values of the lower quartile ($q_{0.25}$) and the upper quartile ($q_{0.75}$). The IQR corresponds to the difference between the upper and the lower quartile:

$$IQR = q_{0.75} - q_{0.25} \tag{2.3}$$

In order to assess the *association degree* of a pair of variable, a common measure to compute is the *correlation coefficient*. Since the relationship between the variables considered in this work are hardly linear, as it will be seen, a non-parametric coefficient has been chosen, specifically the *Spearman rank correlation coefficient*. This coefficient describes the strenght in monotone relationship between a pair of variables (x_1, x_2) contained in the dataset. The computation of this coefficient takes into account only the ranks (i.e. the position of each value in the ascending-ordered sequence of all the values of the considered variables) of the values of the two variables. In the case of the two variables x_1 and x_2, being $R(x_1), R(x_2)$ the rank sequences corresponding to the values of the variables, the Spearman coefficient is computed as:

$$\rho_{12} = r_{rank_{12}} = \frac{cov(R(x_1), R(x_2))}{\sigma_{R(x_1)}\,\sigma_{R(x_2)}} \tag{2.4}$$

where $cov(R(x_1), R(x_2))$ is the covariance between the rank sequences of the two variables and $\sigma_{R(x_1)}$ is the standard deviation of the rank sequence of x_1; the Spearman coefficient is defined by analogy with the definition of Pearson correlation coefficient r_{12} for the same

variables. Defining the difference between two ranks of a point as $D_i = R(x_1^i) - R(x_2^i)$, the computation can be simplified to:

$$\rho_{12} = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \tag{2.5}$$

Concerning the way of graphically representing data, some kind of plots are used in this section.

- *Time series plots* represent the changes in the values of a variable during an interval of time.

  In order to provide a better understanding of the pattern underlying the collected data, in some cases a n-days running mean transformation has been applied: this means that each point of the transformed graph represents the average of the daily values measured among the corresponding day and the the following $n - 1$ days:

  $$\tilde{x}_i = \frac{1}{n} \sum_{j=0}^{n-1} x_{i+j} \tag{2.6}$$

  It is necessary to notice that this operation implies a forward time shift with respect to the actual distribution of values.

- *Barplots* are a common way of comparing quantities. In this chapter they will be used to evaluate the trends of some variables evaluated on an annual scale.

- *Boxplots* (in this case, more precisely, *schematic plots*) are a common way to sintetically represent the distribution of the data. The box is confined between the two quartiles and contains the median, while its notches identify the interval $\left[q_{0.5} - (1.58 * \frac{IQR}{\sqrt{n}}); q_{0.5} - (1.58 * \frac{IQR}{\sqrt{n}})\right]$, that can be interpreted as an error on the median value.

  The "whiskers" (linear segments that extends from the box) reach the minimum and maximum value of the data, except when one or both these values exceed respectively $q_{0.25} - (1.5 * IQR)$ and $q_{0.75} + (1.5 * IQR)$; in this case, the *outliers* (i.e. values outside the range of the whiskers) are printed as isolated points.

- *Scatterplots* are generally used to compare relationships between pair of variables and to assess a functional dependence. Here their main use is related to graphically evaluating the relationship between PM_{10} concentration and meteorological variables.

2.1.2 Particulate Matter concentration

The physical quantity this work is focused on is PM_{10} concentration in outdoor air. As explained in section 1.3, the processes of dispersion of pollutants in the Po basin are hindered by meteorological conditions typical of winter. So it's not surprising that PM_{10} daily mean concentrations in the considered cities, as shown in Figure 2.1, are higher in the winter periods and lower in summertime.

More precisely, the oscillatory pattern of this variable appears quite the same in all the considered locations: the oscillations are generally stronger in the autumn and winter months, when the concentration can reach very high values indeed, thanks to the environmental conditions of these seasons. At the same time, some deviations from the collective behaviour of the trends are present: these deviations appear stronger when they are positive, i.e. when a peak is formed, than when they are negative.

Analysing separately the annual distribution of PM_{10} daily mean concentration values for each city using boxplots allows to understand better the evolution of the indicators of central tendency and spread, and also the frequency of extreme values. In the case of Figures 2.2 and 2.3 only the data measured between 2013 and 2017 are considered, as data from an entire year are needed in order to correctly evaluate the quantile values for that year.

As median values appear to be quite stable during time for the considered cities, all the considered distributions have long tails towards high values of concentration. While all the cities show a considerable spread of outlier values, in some cases very high values of concentration with respect to the median have been measured (e.g. in 2017). Table 2.5 summarizes the most important statistics for each distribution of PM_{10}.

City	Median					IQR					# outliers				
	2013	2014	2015	2016	2017	2013	2014	2015	2016	2017	2013	2014	2015	2016	2017
Piacenza	25.0	22.0	28.0	23.0	26.0	18.0	15.0	19.0	14.0	24.0	12	17	16	22	15
Parma	28.0	26.0	29.0	24.5	29.0	20.25	20.25	19.0	16.25	26.00	10	16	16	18	19
Reggio Emilia	24.0	21.0	24.0	24.0	25.0	17.0	14.0	18.0	19.0	23.0	13	22	18	14	22
Modena	23.0	21.0	25.0	22.0	25.0	19.0	16.75	20.25	18.0	26.0	14	20	14	16	12
Bologna	16.0	17.0	22.0	19.0	19.0	12.25	14.0	19.0	15.0	16.0	25	15	9	21	24
Ferrara	22.5	19.0	23.0	21.0	22.0	19.75	17.75	21.00	18.00	21.50	16	21	21	18	32
Ravenna	22.5	20.0	24.0	21.0	22.0	17.75	17.00	19.00	15.75	20.00	20	20	25	22	16
Forlì	18.0	16.0	20.0	18.0	18.0	15.00	13.00	17.00	16.00	16.75	16	16	18	18	17
Cesena	19.0	18.0	22.0	19.0	20.0	15.00	13.00	15.75	16.00	15.00	14	24	16	10	20
Rimini	23.0	22.0	26.0	22.0	23.0	15.00	18.00	19.50	16.00	18.00	25	13	23	23	16

Table 2.5: Annual statistics for PM_{10} daily mean concentration in the considered cities.

As some peaks in Figure 2.1 appear extremely high, a comparison of the number of

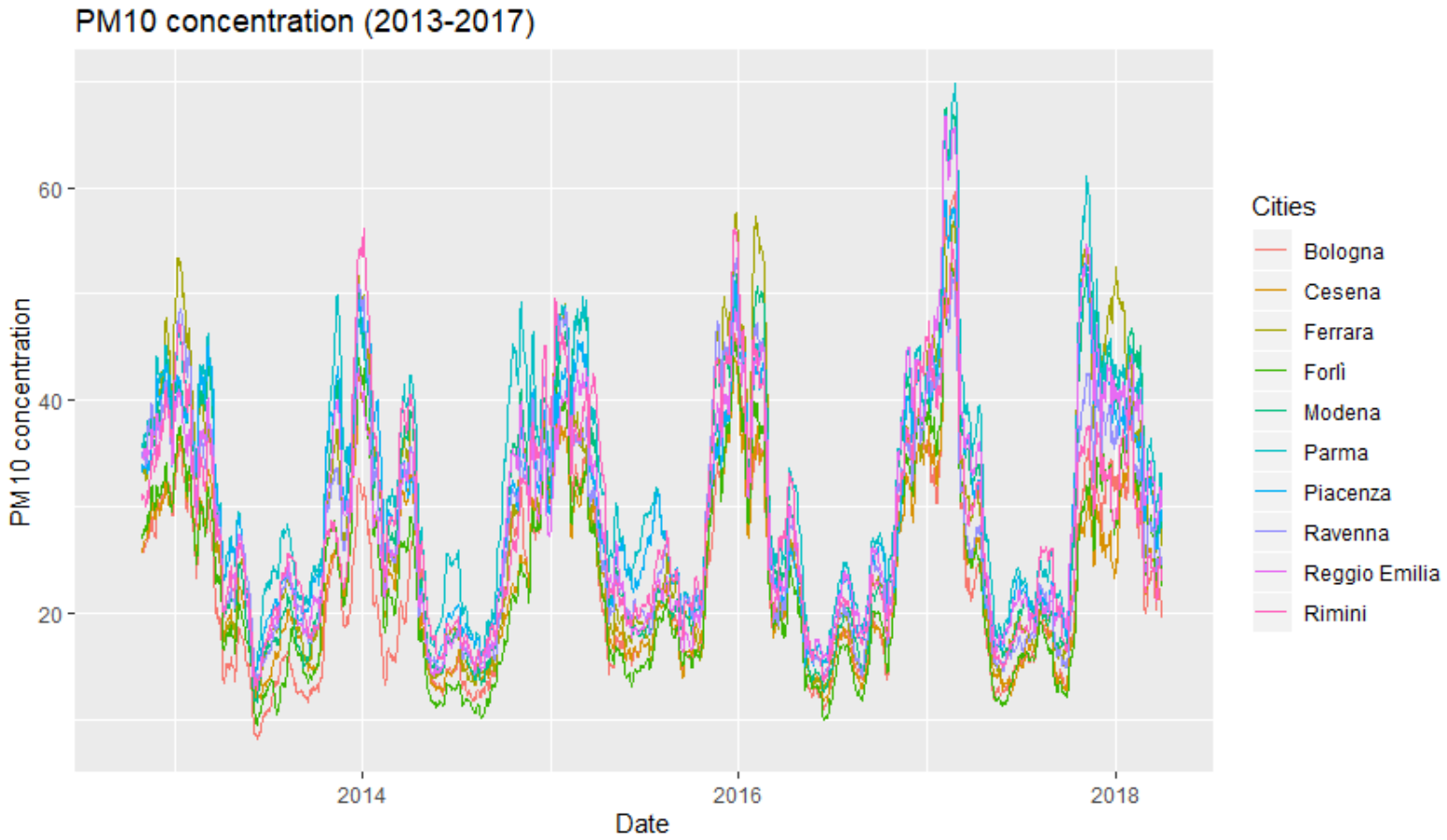
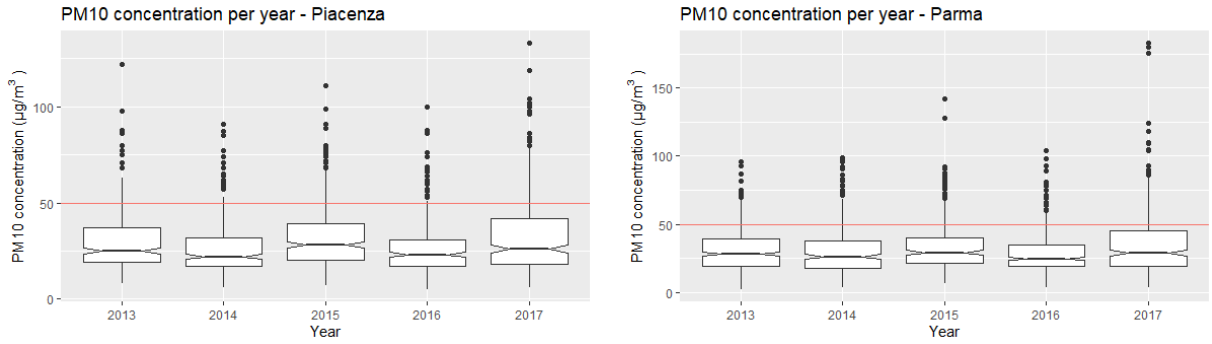
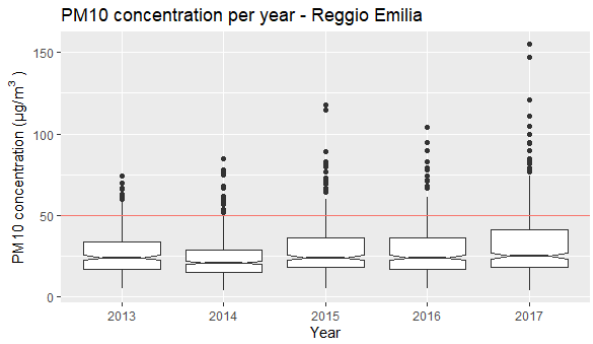


Figure 2.1: PM_{10} concentration values (in $\mu g/m^3$) in the considered time interval (represented as 30-days mobile mean on daily values).

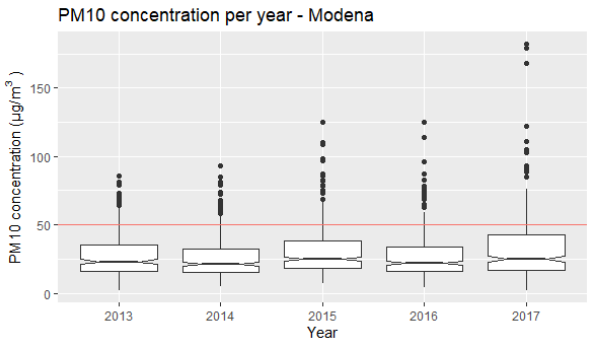


(a) Piacenza

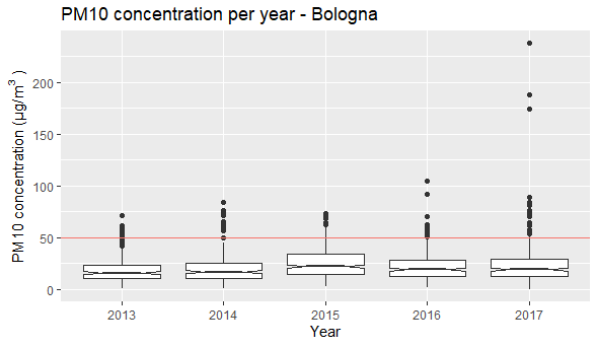
(b) Parma



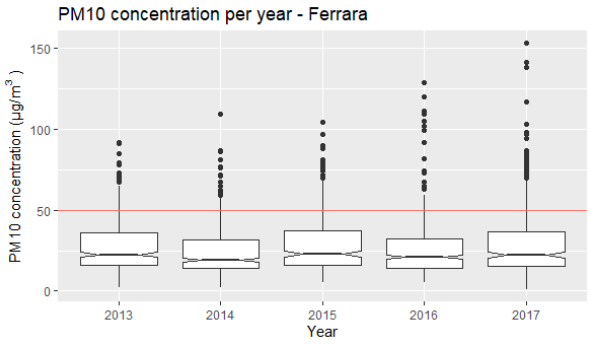
(c) Reggio Emilia



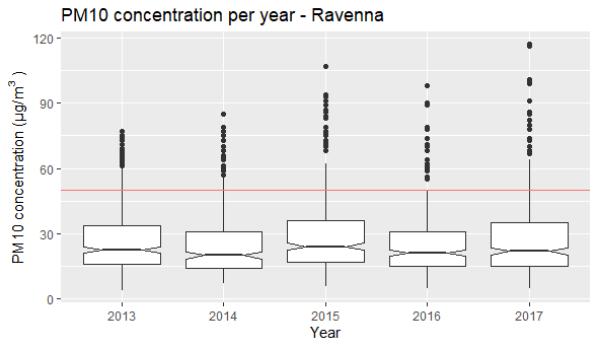
(d) Modena



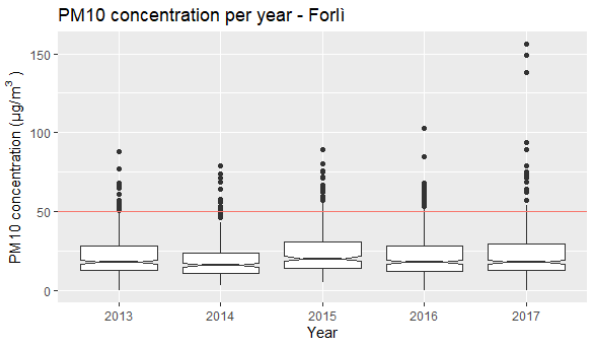
(e) Bologna



(f) Ferrara

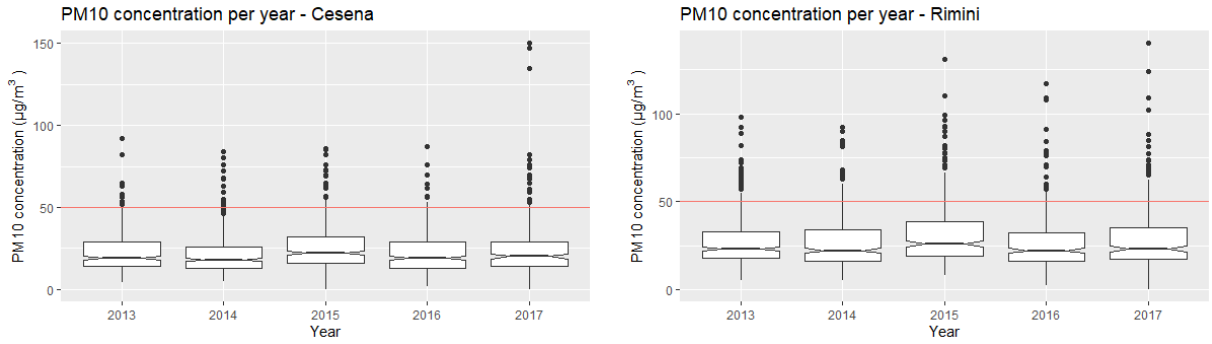


(g) Ravenna



(h) Forlì

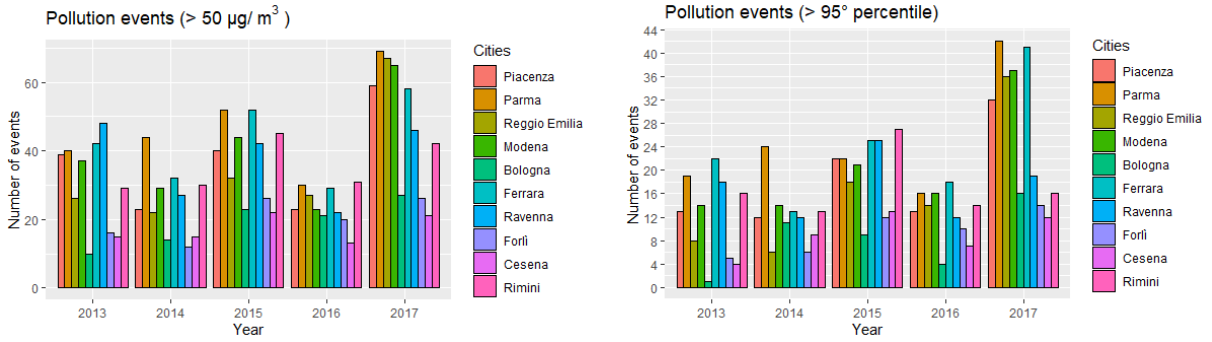
Figure 2.2: Boxplot distributions of PM_{10} concentration values (1).



(a) Cesena (b) Rimini

Figure 2.3: Boxplot distributions of PM_{10} concentration values (2).

pollution episodes (i.e. the days in which high PM_{10} daily mean concentration values are measured) can be used to evaluate the number of these events in each city for the considered interval of time. Such a comparison is performed in Figure 2.4.



(a) Concentration superior than $50\mu g/m^3$ (b) Concentration superior than 95° percentile ($62\mu g/m^3$)

Figure 2.4: PM_{10} pollution events.

In Figure 2.4(a) the exceedances of the EU daily threshold are shown: the general behaviour does not contain any specific trend, although 2017 appears to be the year with the highest number of events for all cities, except Cesena.
Regarding Figure 2.4(b), where a threshold corresponding to the 95° percentile calculated over the entire dataset of 20080 samples has been fixed in order to considered a subset of "high pollution events", it can be seen that these events are concentrated in

the second half of the considered time interval[5]. It would be of interest, for a different kind of analysis, to assess the general trend of high pollution events on a longer (e.g. 20 years) time window.

A useful insight on the differences between local trends on a larger timescale is given by the distribution of the monthly PM_{10} concentration average in the period of interest: the graph in Figure 2.5 allows some deeper understanding of the seasonal trend, while accounting for significant differences between cities.

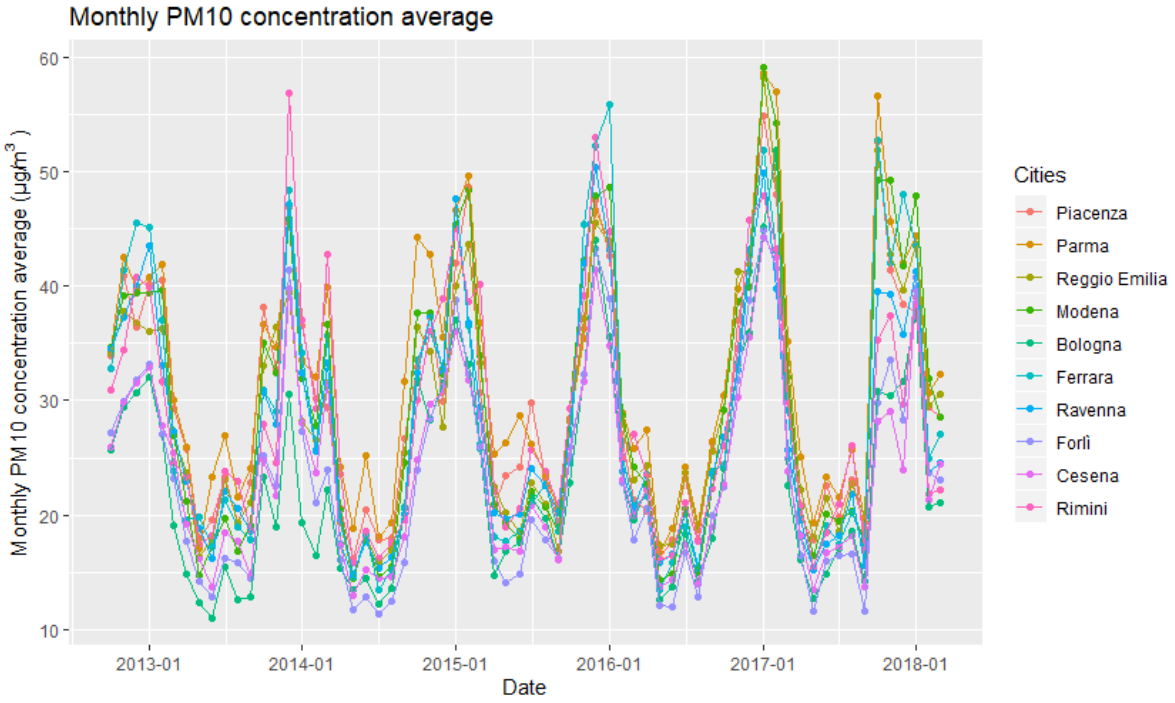


Figure 2.5: Monthly average of PM_{10} daily mean concentrations.

As some patterns can be easily motivated by geographical reasons (e.g. lowest values of PM_{10} concentration for cities on the boundary of the basin, such as those closer to the coast or the Apennines, and higher values for the innermost ones), others (e.g. peaks of concentration in Rimini during the winter of 2013-2014) should be attentively considered in case an analytical analysis had to be undertaken.

[5]It is useful to note that pollution events with significant levels of PM_{10} concentration (e.g. superior to $100\mu g/m^3$) for more than one day have happened: for example in year 2017, when an important pollution event took place in the entire regional area during three consecutive days (31 January, 1 and 2 February). Events of this kind can produce an important signal in the considered distributions. Obviously these events do not undermine the significance of the observed trend: on the contrary, such events should be explicitly considered, whenever they took place, in order to evaluate their impact on the observed trend.

It is worth remembering, however, that the present analysis is going to focus exclusively on meteorological parameters: in fact, while some of these patterns are mostly influenced by geographical features (e.g. wind speed and direction) that explain part of the discrepancies between the considered cities, others depend on causes that are not related to the physical characteristics of the considered area (e.g. presence of specific sources of emissions).

The annual average of daily mean values provides information on the trend of PM_{10} concentration using a much wider time window that averages over seasonal variations. The values for the considered cities are shown in Figure 2.6.

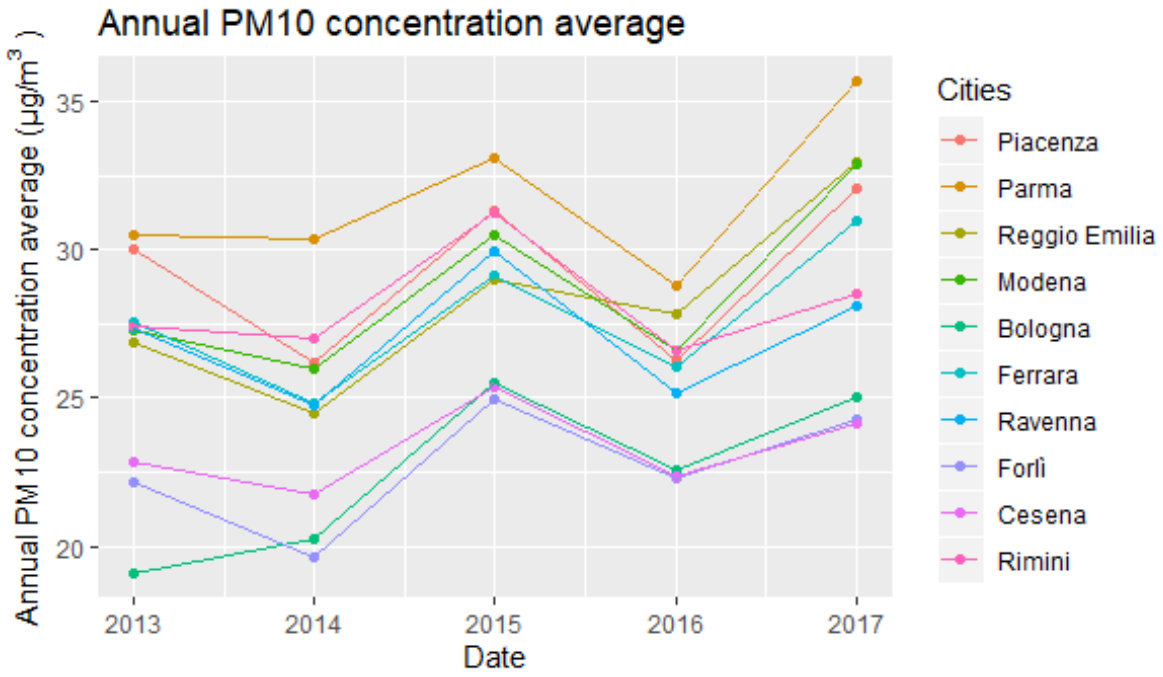


Figure 2.6: Annual average of PM_{10} daily mean concentrations.

Although a general oscillatory trend is common to all the cities (probably related to major variations in the meteorological situations), some common features are clear:

- Parma has the highest mean concentration for all the 5 years considered;
- Bologna, Cesena and Forlì have lower concentrations for the whole time period, suggesting some kind of structural (possibly geographical) reason;
- despite being a coastal city, Rimini (as well as Ravenna, which is near the Adriatic coast too) shares average values similar to the cities in the inner part of the region.

From a legal point of view, the EU annual limit value ($40\,\mu g/m^3$) has never been exceeded, while the WHO AQG value ($20\,\mu g/m^3$) has been respected only in Bologna (in 2013) and in Forlì (in 2014).

PM persistence and seasonality

A large number of atmosphere-related quantities are statistically dependent on their past values (i.e. weather tends to be similar in a relatively short time window): in particular, this dependence appears when the measurement interval is shorter than the timescale of the physical processes that influence that quantity. This behaviour is called *persistence*, or positive serial dependence.[41]
Also in the case of PM_{10} concentrations, since build-up processes develop in a period of time of some days, the daily mean value shows this kind of relationship. On the other hand, meteorological events that cause dispersion (e.g. rain) happen relatively quickly: this affects negatively the positive serial dependence of concentration values.

A quantitative measure of persistence is the *serial correlation*, also called *temporal autocorrelation*. Given a sequence of values $(x_i)_{i=1,\dots,n}$, the temporal autocorrelation is computed with respect to a temporal lag (or for a set of values for the lag itself). Being k a value for the lag, the quantity corresponds to the Pearson correlation coefficient calculated for the $n-k$ data pairs (x_i, x_{i-k}), i.e.

$$r_k = \frac{\sum_{i=1}^{n-k}\left[(x_i - \overline{x_i})(x_{i+k} - \overline{x_{i+k}})\right]}{\left[\sum_{i=1}^{n-k}(x_i - \overline{x_i})^2\right]^{1/2}\left[\sum_{i=k+1}^{n}(x_i - \overline{x_{i+k}})^2\right]^{1/2}} \tag{2.7}$$

where

$$\overline{x_i} = \frac{\sum_{i=1}^{n-k} x_i}{n-k} \qquad \overline{x_{i+k}} = \frac{\sum_{i=k+1}^{n} x_i}{n-k} \tag{2.8}$$

In this way, a sequence of values of r_k can be obtained for $k = 1, 2, \dots$, representing the correlation between values of the considered variable using a time lag of k units of time.

In the case of PM_{10} concentrations, since the daily mean value has been used, the time lag corresponds to a number of days.
Figure 2.7 shows the autocorrelation function for the 10 considered cities for a maximum lag of 90 days.
Table 2.6 reports the values of the autocorrelation function for a lag $k = 1$.

Apart from persistence, some papers (see section 1.4) describe models that make use of variables accounting for periodicity in the sequence of PM_{10} daily mean concentrations.

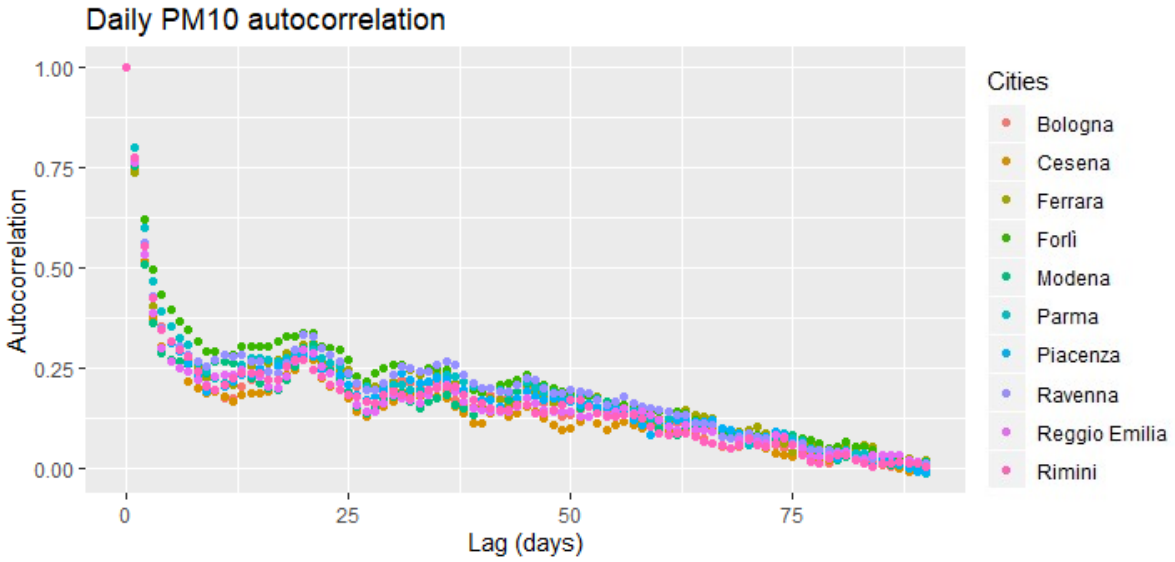


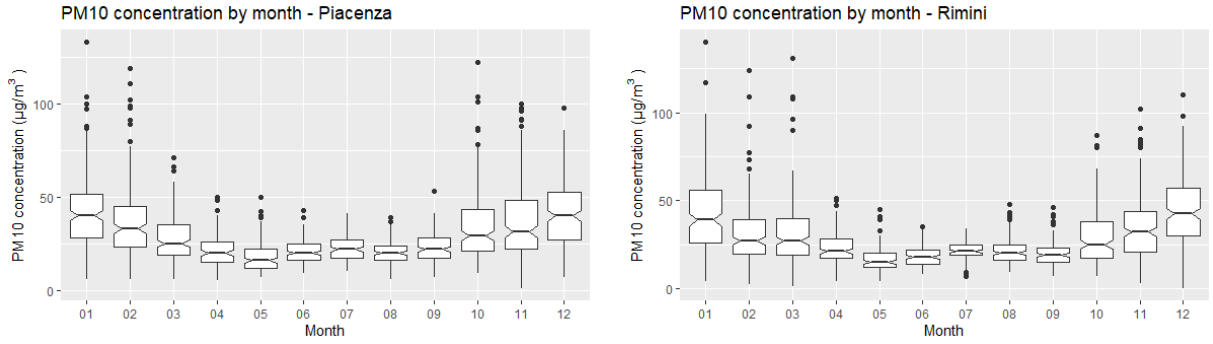
Figure 2.7: Autocorrelation of PM_{10} values.

Variable	PM_{10} temporal autoconcentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
r_1 ($[PM_{10}]$)	0.765	0.750	0.738	0.801	0.756	0.802	0.763	0.777	0.761	0.775

Table 2.6: Temporal autocorrelation of PM_{10} concentration with lag $k = 1$.

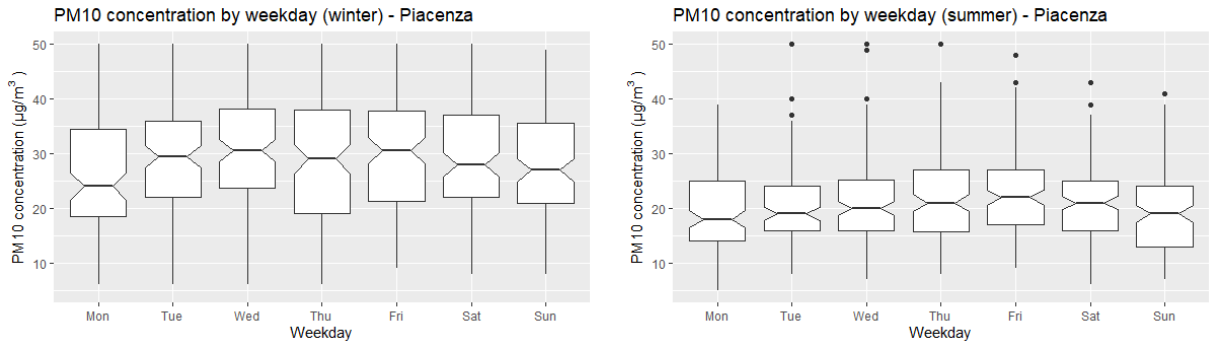
As concerns the annual cycle of the seasons, it is well-established that meteorological conditions are (or should be) similar during the same period of the year. As PM_{10} concentrations are affected by meteorological conditions (see paragraph 1.3), a corresponding behaviour is understandable. Also the pattern of emissions varies during the year, with different amounts of primary PM_{10} and precursors emitted in different periods of the year. The resulting pattern, shown in Figure 2.5, presents some regularities and can be better understood by Figure 2.8, where the distribution of PM_{10} daily mean values are grouped by month and reported as boxplots.

Another kind of periodicity is associated with the variation of the intensities of the emissions from some sources during the time interval of a week. So a weekly variability can be considered and graphically assessed. Figure 2.9 reports the boxplots of PM_{10} concentration by day of the week in Piacenza considering a restricted interval of concentration values. Summer and winter periods are compared: variations of concentration values with respect to the day of the week are present in both cases, despite the obvious shift in the absolute values due to the seasonal variation.



(a) Piacenza (b) Rimini

Figure 2.8: Boxplot distributions of PM_{10} concentration values by month.



(a) Winter (b) Summer

Figure 2.9: Boxplot distributions of PM_{10} concentration values by weekday (winter vs summer).

2.1.3 Temperature

The first meteorological quantity to be considered in order to assess its relationship with PM_{10} concentration is temperature. For this work, a set of temperature-related variables has been considered: daily mean temperature, daily minimum temperature, daily maximum temperature and daily range of temperatures (i.e. the variation of temperature in a single day).

In a calendar year, daily mean temperatures in the cities of Emilia-Romagna varies from lowest values in winter to highest ones in summer. A summary of the extreme values of the temperature variables in the considered period is shown in Table 2.7.

City	Mean temperatures		Minimum temperatures		Maximum temperatures	
	Min. value	Max. value	Min. value	Max. value	Min. value	Max. value
Piacenza	-2.45 (9/12/2012)	31.19 (5/8/2017)	-8.8 (9/12/2012)	26.0 (7/7/2015)	-0.9 (1/3/2018)	37.8 (22/7/2015)
Parma	-2.55 (27/2/2018)	32.93 (8/4/2017)	-7.4 (9/2/2015)	26.8 (8/7/2015, 4/8/2017)	-1.6 (1/3/2018)	39.3 (4/8/2017)
Reggio Emilia	-2.74 (27/2/2018)	33.02 (4/8/2017)	-6.9 (28/2/2018)	27 (4/8/2017)	-1.0 (1/3/2018)	39.7 (3/8/2017)
Modena	-3.32 (27/2/2018)	32.86 (4/8/2017)	-6.9 (28/2/2018)	26.5 (4/8/2017)	-1.8 (1/3/2018)	39.5 (3/8/2017)
Bologna	-3.58 (27/2/2018)	34.02 (4/8/2017)	-7.8 (28/2/2018)	27.3 (4/8/2017)	-1.7 (26/2/2018, 1/3/2018)	39.7 (4/8/2017)
Ferrara	-2.84 (27/2/2018)	32.08 (4/8/2017)	-6.1 (28/2/2018)	26.4 (22/7/2015)	-1.7 (1/3/2018)	38.8 (4/8/2017)
Ravenna	-2.52 (27/2/2018)	32.32 (4/8/2017)	-6.8 (28/2/2018	26.3 (4/8/2017)	-0.4 (26/2/2018, 27/2/2018)	39.4 (4/8/2017)
Forlì	-4.2 (27/2/2018)	34.23 (4/8/2017)	-7.4 (28/2/2018)	28.5 (4/8/2017)	-1.8 (27/2/2018)	41.7 (4/8/2017)
Cesena	-4.42 (27/2/2018)	33.64 (4/8/2017)	-7.8 (28/2/2018)	27.2 (4/8/2017)	-1.9 (27/2/2018)	38.6 (4/8/2017)
Rimini	-3.67 (27/2/2018)	31.24 (4/8/2017)	-5.8 (28/2/2018)	26.4 (5/8/2017)	-1.9 (27/2/2018)	38.4 (8/8/2013)

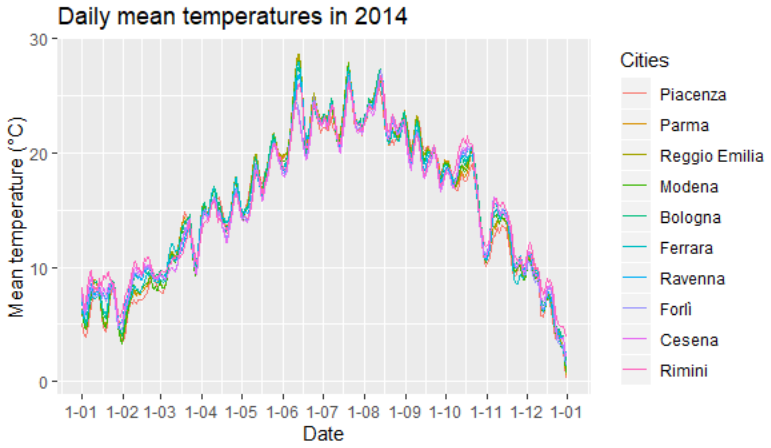
Table 2.7: Extreme values of temperature in the time interval [1/10/2012 − 31/3/2018].

An example of the typical trend is shown in Figure 2.10(a) for the year 2014: the oscillatory behaviour is caused by meteorological perturbations on the regional scale, while smaller variations are present between cities.

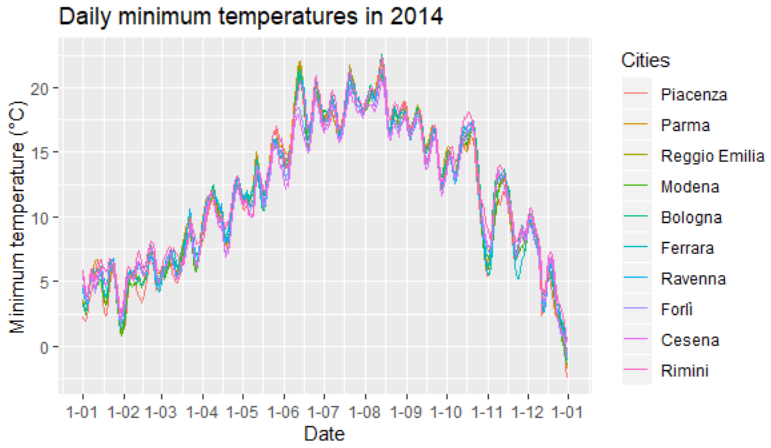
Similar trends are present in the distribution of daily minimum and maximum temperatures in the same year (Figures 2.10(b) and 2.10(c)): apart from the range of values, which is obviously different, the oscillations in the considered cities are mostly the same.

Regarding the daily range of temperatures shown in Figure 2.10(d), some information can be inferred on the seasonal variation of temperature range during the year:

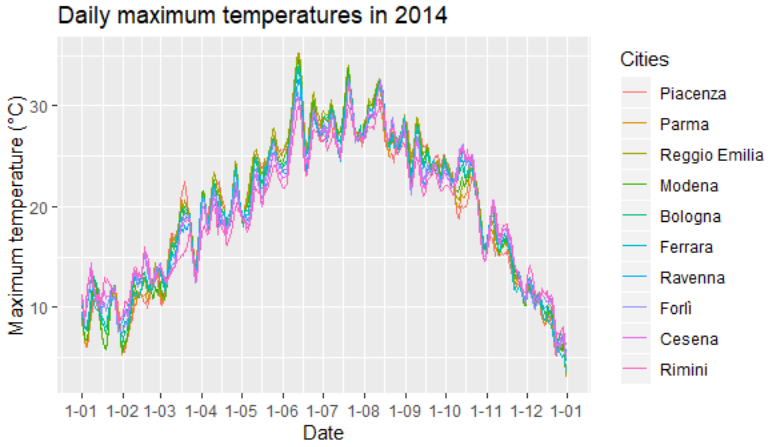
- the daily oscillation of temperatures (i.e. the temperature range) is generally stronger in summer and weaker in winter;
- the variation of the temperature range in a year is small in a coastal city such Rimini with respect to more internal cities such as Reggio Emilia and Piacenza.



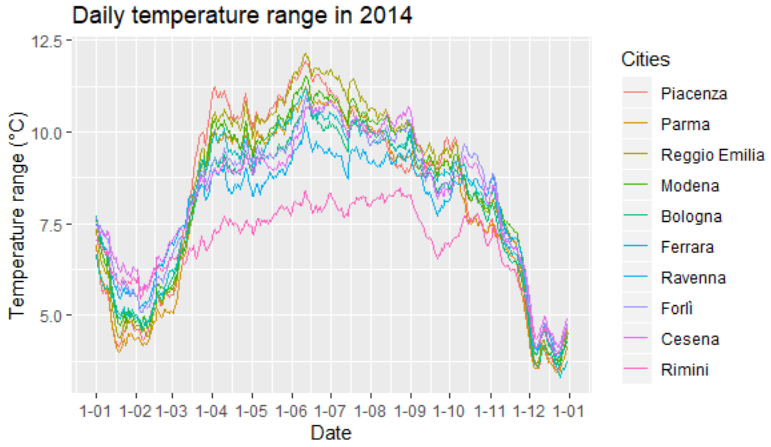
(a) Daily mean temperature ($°C$) (5-days mobile mean)



(b) Daily minimum temperature ($°C$) (5-days mobile mean)



(c) Daily maximum temperature ($°C$) (5-days mobile mean)



(d) Daily temperature range ($°C$) (30-days mobile mean)

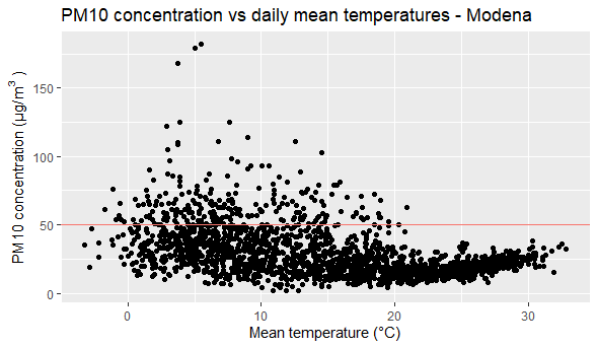
Figure 2.10: Series of temperature-related quantities in 2014.

In order to assess the correlation between these quantities and PM_{10} concentration, the Spearman correlation has been computed for each variable and each city using the whole time interval available for this work. The resulting values are presented in Table 2.8.

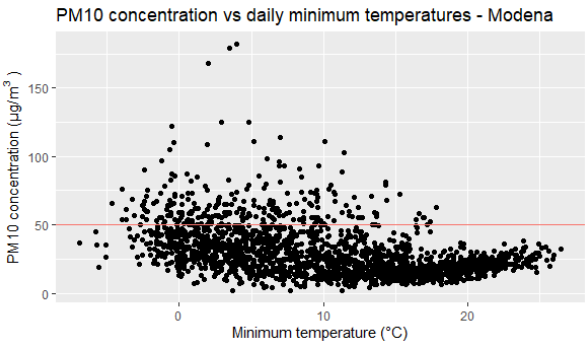
Variable	Correlation with PM_{10} concentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
T_{mean}	-0.390	-0.352	-0.388	-0.492	-0.396	-0.422	-0.370	-0.471	-0.410	-0.407
T_{min}	-0.365	-0.335	-0.381	-0.479	-0.373	-0.428	-0.386	-0.461	-0.389	-0.384
T_{max}	-0.388	-0.352	-0.385	-0.477	-0.391	-0.392	-0.369	-0.448	-0.397	-0.405
T_{range}	-0.259	-0.249	-0.262	-0.288	-0.265	-0.136	-0.116	-0.218	-0.229	-0.200

Table 2.8: Correlations between temperature and PM_{10} concentration.

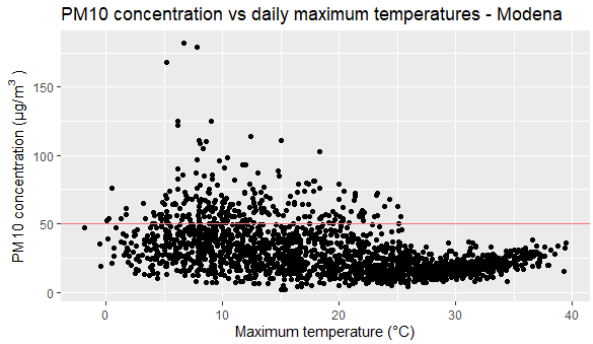
A graphical representation of the actual relationships between these quantities is given by Figure 2.11, where scatter plots regarding Modena are reported. As can be easily understood there's not a strong relationship between temperatures and PM_{10} concentration, although high values of pollution tend to concentrate in days with lower-than-average temperatures and a limited temperature variability.



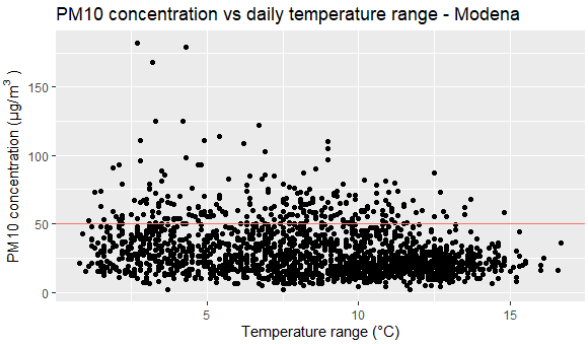
(a) Daily mean temperature vs PM_{10} concentration



(b) Daily minimum temperature vs PM_{10} concentration



(c) Daily maximum temperature vs PM_{10} concentration

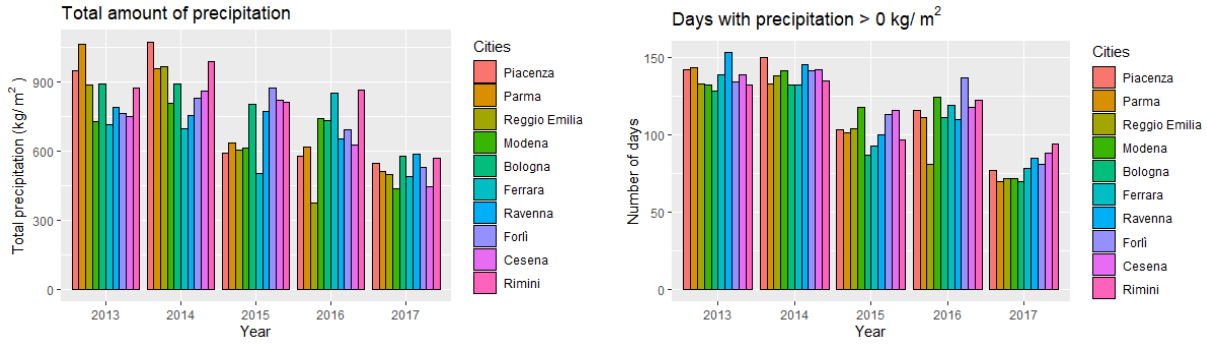


(d) Daily temperature range vs PM_{10} concentration

Figure 2.11: PM_{10} concentration vs temperature-related quantities in Modena.

2.1.4 Precipitation

A second meteorological variable to be considered is the daily amount of precipitation. In Figure 2.12(a) the annual amount of precipitation (per unit area) measured in each city in the year is shown, while in Figure 2.12(b) the number of days with precipitation are reported. A negative trend can be observed in both cases.



(a) Total amount of precipitation

(b) Number of days with precipitation

Figure 2.12: Annual statistics for precipitation in the considered cities (2013-2017).

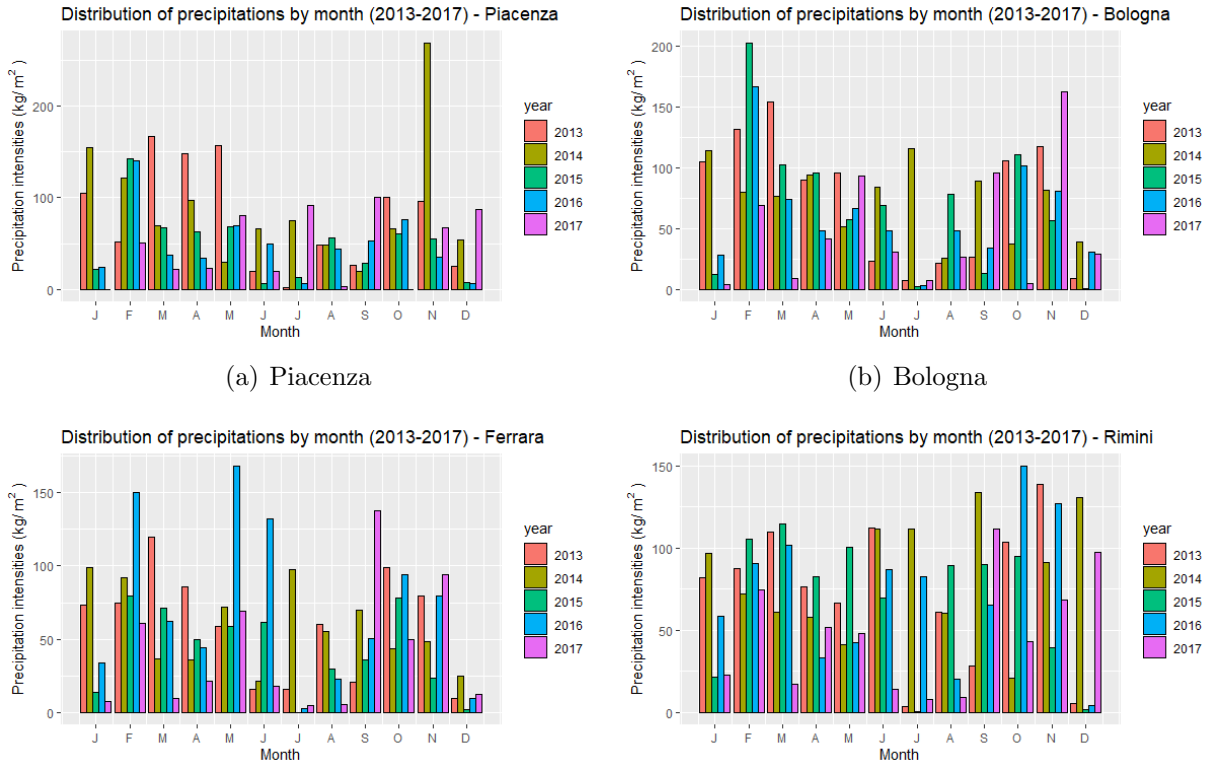
Comparing the monthly distribution of precipitations in the considered period for the four cities in Figure 2.13, some differences in the patterns can be noticed: in summer the internal areas are drier than the coastal ones. In general, precipitations are not distributed homogeneously during the year, and the distributions change depending on the year.

Turning to the relationship between daily precipitation and PM_{10} concentration, Table 2.9 summarizes the values of the Spearman correlation between the two variables.

Variable	Correlation with PM_{10} concentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
P	-0.094	-0.168	-0.174	-0.111	-0.186	-0.227	-0.191	-0.201	-0.213	-0.268

Table 2.9: Correlations between precipitation and PM_{10} concentration.

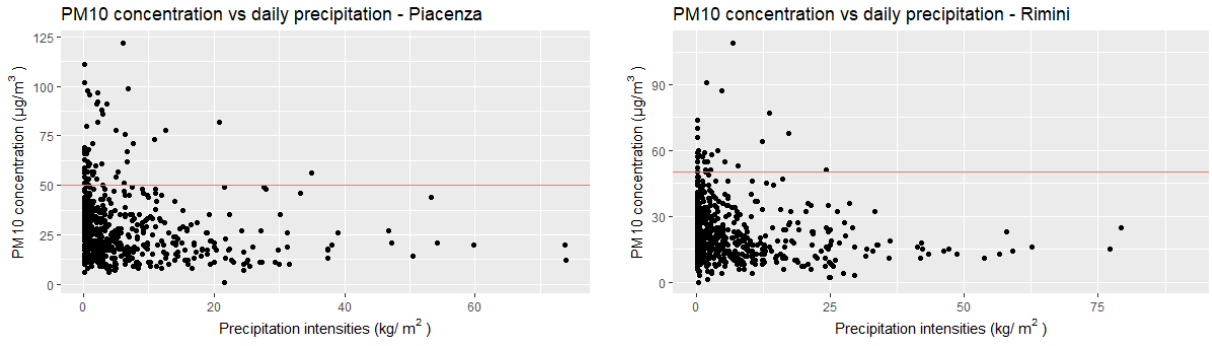
The scatterplots for Piacenza and Rimini are presented in Figure 2.14: precipitation are generally associated with lower PM_{10} concentration. No significant differences can be seen between the cities (the other locations are characterized by similar plots).



(a) Piacenza (b) Bologna

(c) Ferrara (d) Rimini

Figure 2.13: Monthly distribution of precipitation (2013-2017).



(a) Piacenza (b) Rimini

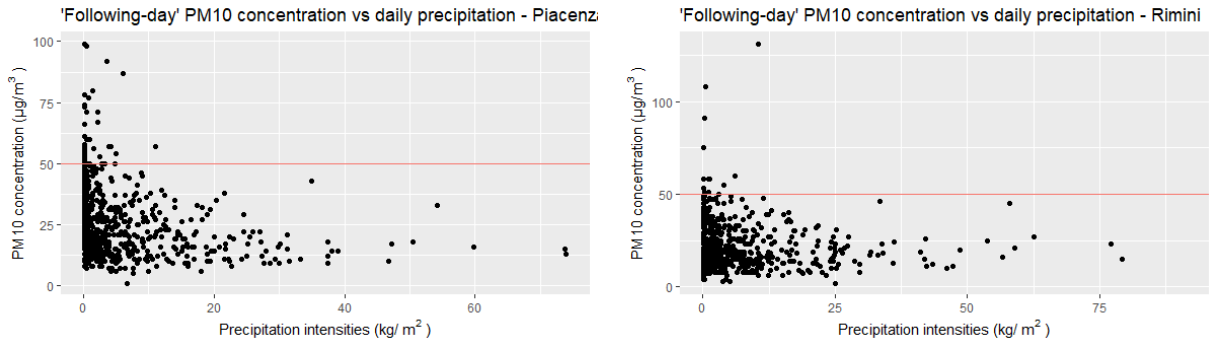
Figure 2.14: PM_{10} concentration vs precipitation (2013-2017).

In order to assess the presence of effects of precipitation on PM_{10} concentration in the day following the precipitation event, Spearman correlation has been computed also between these two quantities. Results are summarized in Table 2.10.

Variable	Correlation with PM_{10} concentration ($PM_{10}(d+1)$)									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
P	-0.226	-0.300	-0.310	-0.221	-0.267	-0.288	-0.258	-0.260	-0.281	-0.289

Table 2.10: Correlations between precipitation and PM_{10} concentration on the following day.

While a small improvement can be seen in the correlation values, no particular differences can be seen in the features of the scatterplots for the same cities considered before (Figure 2.15).



(a) Piacenza

(b) Rimini

Figure 2.15: PM_{10} concentration of the following day vs precipitation (2013-2017).

2.1.5 Wind intensity and direction

Concerning wind description in a data set, it can be described both as a scalar variable (i.e. considering only its intensity) or as a vectorial one (considering both its intensity and direction); in the reviewed literature both approaches have been applied. A characterization of wind intensity is firstly given, then wind direction is considered.

Wind intensity

Annual sets of values for wind intensity (those corresponding to year 2013 are presented in Figure 2.16) shows that eastern cities are generally windier than western one: a possible explanation for this behaviour is the geographical position, with coastal cities being favoured by the presence of sea and the absence of mountain ranges (excluding the Apennines in the south-west), while the western cities are located in the inner part of the basin and surrounded by mountains (even if the distance is great, the effects are recognisable).

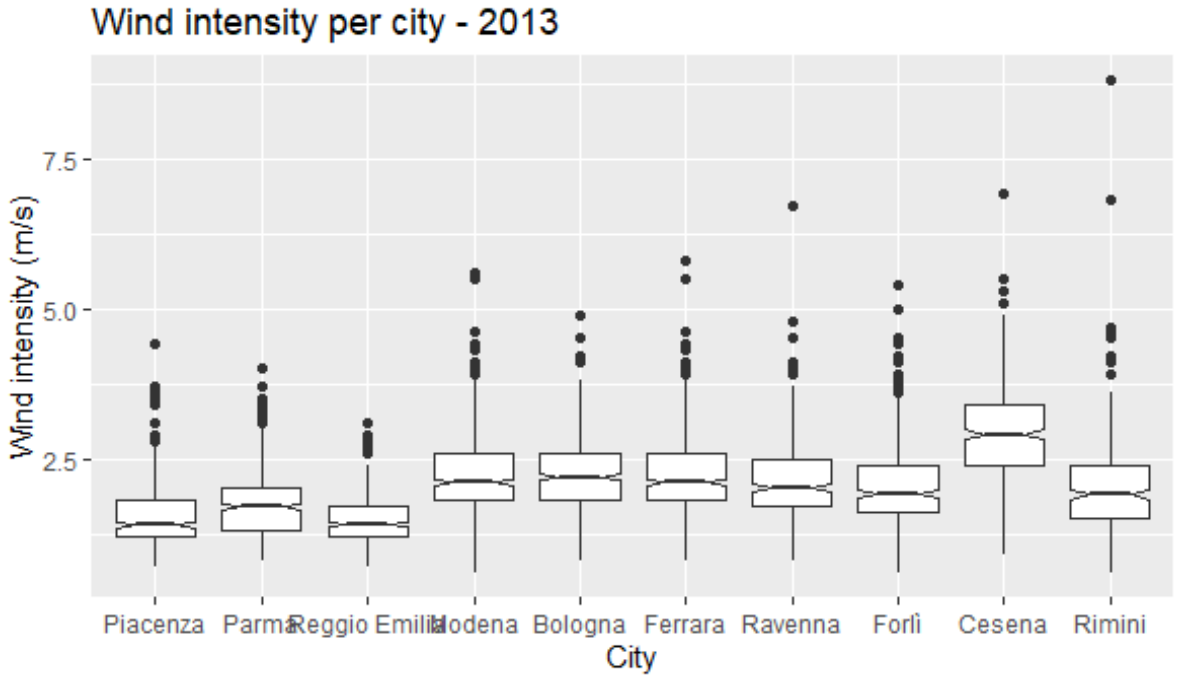


Figure 2.16: Boxplot distributions of wind intensity values (in m/s) in 2013.

Observing the distribution of the monthly mean values of wind intensity (Figure 2.17), despite a strong noise, a common feature can be spotted: all distributions have a peak during summer months, while lowest values have been measured in winter months. This can be linked to the geographical feature of the Po basin: as already said in section 1.3,

the considered area is characterised by low winds that tend to strengthen in summer due to increased thermal convection, because of higher surface temperatures that produce vertical motion of the air. This can justify the observed trend.

Furthermore, with regard to the general behaviour of wind intensity trends, three "clusters" of cities can be detected:

- the western cities (Piacenza, Parma, Reggio Emilia);
- the central-eastern cities (Modena, Bologna, Ferrara, Forlì, Rimini);
- Cesena, which appears to be the windiest city.

The relationship between wind intensity and PM_{10} concentration as been assessed using Spearman correlation as before. Results are shown in Table 2.11.

Variable	Correlation with PM_{10} concentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
W_{int}	-0.532	-0.554	-0.592	-0.473	-0.535	-0.468	-0.585	-0.510	-0.393	-0.437

Table 2.11: Correlations between wind intensity and PM_{10} concentration.

The minimum and maximum values of correlation, which is generally significant, are found for Rimini and Ravenna respectively, both belonging to the central-eastern cluster of cities. There's no clear pattern that links the previously detected clusters with the values of correlation.

The corresponding scatterplots are shown in Figure 2.18, along with the ones for Piacenza and Cesena, chosen as representatives for the other clusters.
Features of these plots appear quite similar, with highest values of PM_{10} concentration associated to days characterised by weak wind.

Wind direction

In order to obtain a descriptor for wind direction (i.e. the direction from which the wind originates) on a daily basis, the quantity called "daily dominant wind direction" has been chosen from ARPAE database. This quantity corresponds to the most frequently reported value for istantaneous wind direction (which is calculated as the mean direction during time intervals of 10 minutes) during the considered day. There are 8 possible angular values for this variable, corresponding to the four cardinal directions (North, East, South and West) and their four intermediate directions (NE, SE, SW, NW); North

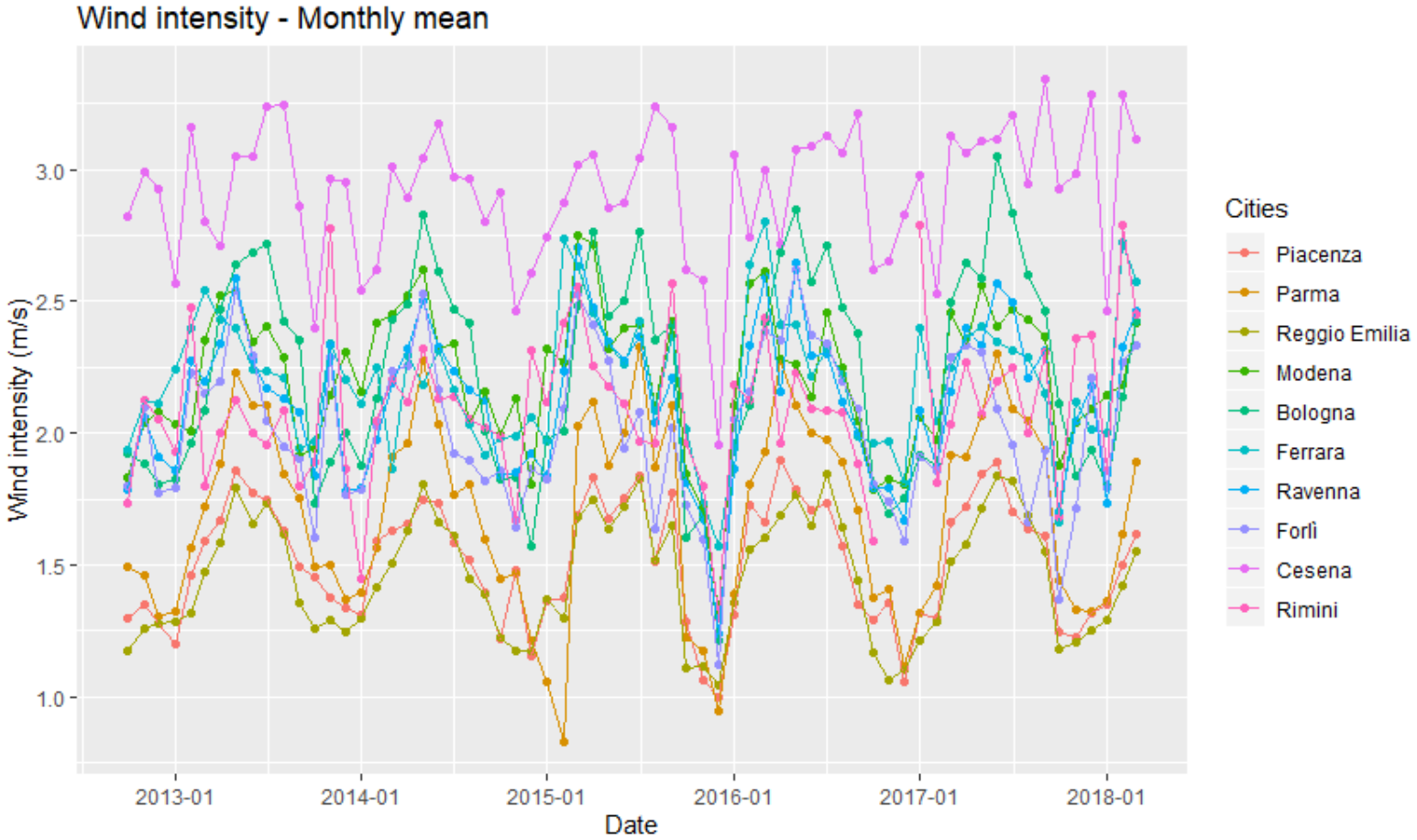
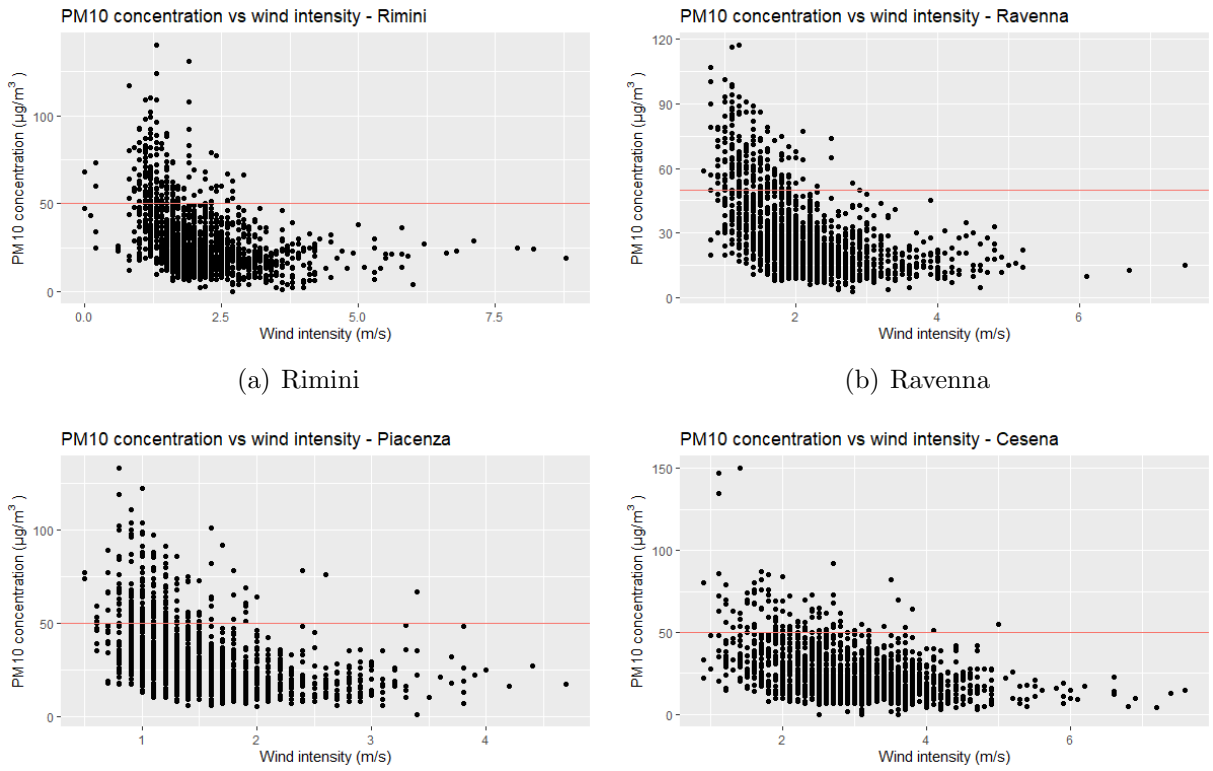


Figure 2.17: Series of monthly mean values of wind intensity in m/s.



(a) Rimini

(b) Ravenna

(c) Piacenza

(d) Cesena

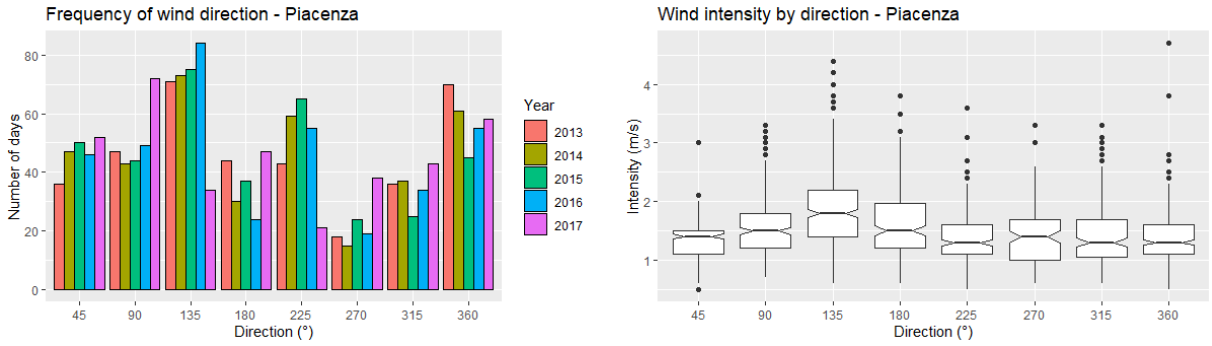
Figure 2.18: PM_{10} concentration of the following day vs wind intensity (2013-2017).

corresponds to 0° or 360°. Because of this way of reporting, this variable has been considered a *categorical* variable.

The annual behaviour of wind direction for each city can be descrived in two ways:

- the number of days characterized by a certain wind direction;
- the wind intensity distribution for each considered direction.

An example of these two distribution is given for the city of Piacenza, Bologna, Ferrara and Rimini respectively in Figures 2.19, 2.20, 2.21 and 2.22.
It can be seen that, regarding the distribution of daily predominant wind direction, while westerly winds prevail in the case of Bologna, in other locations the directions are more homogeneously distributed (Piacenza) or are characterised by more than one modal value (Ferrara, Rimini).

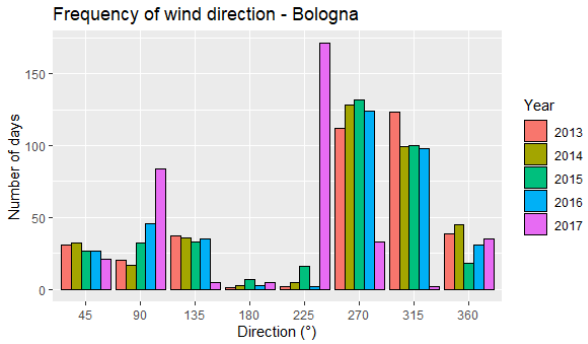


(a) Number of days per direction

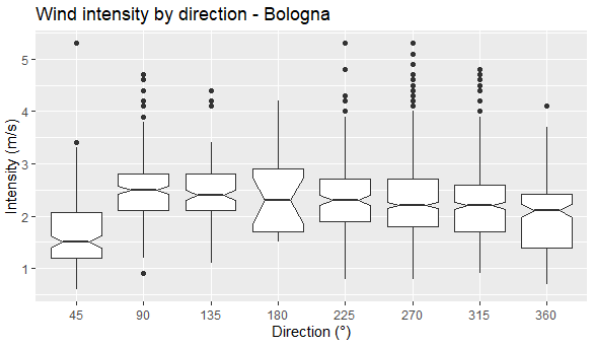
(b) Intensity distribution per direction

Figure 2.19: Wind characteristics per direction in Piacenza (2013-2017).

Concerning the relationship between wind direction and PM_{10} concentration, the assessment can be performed for each city by computing the distribution of concentration values for each value of wind direction and evaluating the descriptive statistics for the distribution. The results are summarizable in boxplot graphs, that are provided in Figure 2.23 for the four cities considered in Figures 2.19, 2.20, 2.21 and 2.22.
PM_{10} concentration values seem not to depend strongly on wind direction. Slightly higher values of concentration are measured for westerly winds in 3 cities out of 4: this can be linked to the fact that the innermost part of the Po basin, the one enclosed between the Alps and the Apennines, is located westerly with respect to the considered cities, while the same cities necessarily influence one another because of transport phenomena.

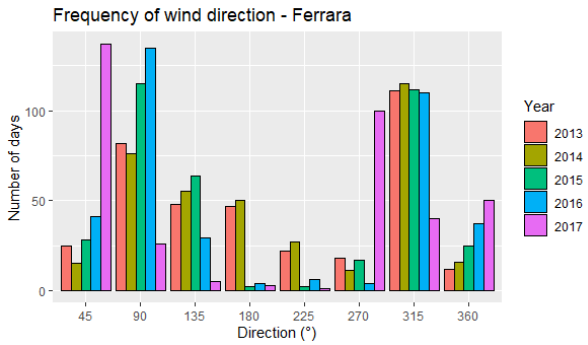


(a) Number of days per direction

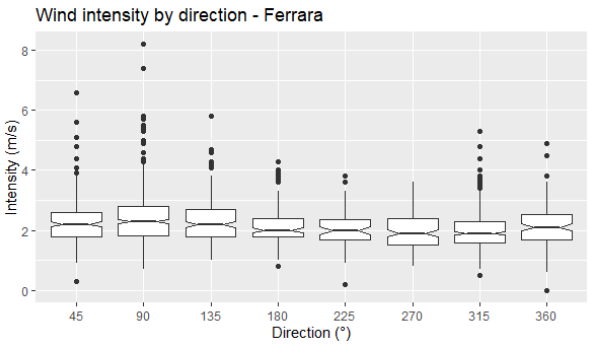


(b) Intensity distribution per direction

Figure 2.20: Wind characteristics per direction in Bologna (2013-2017).

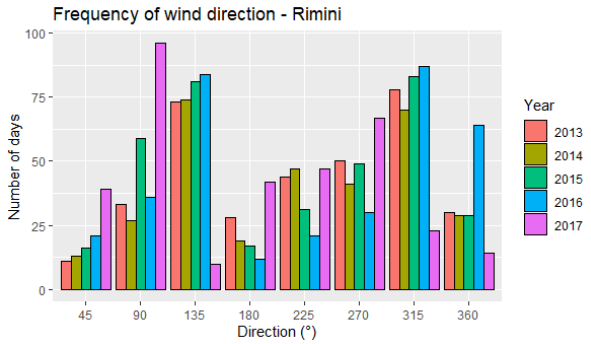


(a) Number of days per direction

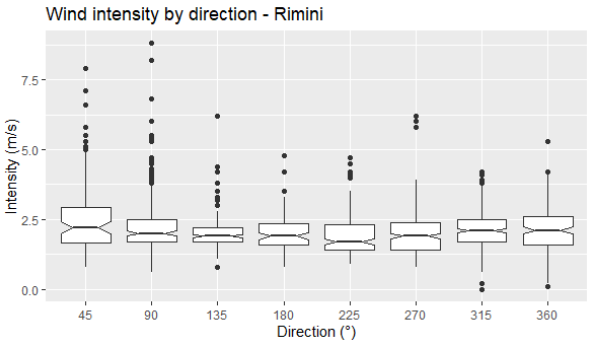


(b) Intensity distribution per direction

Figure 2.21: Wind characteristics per direction in Ferrara (2013-2017).

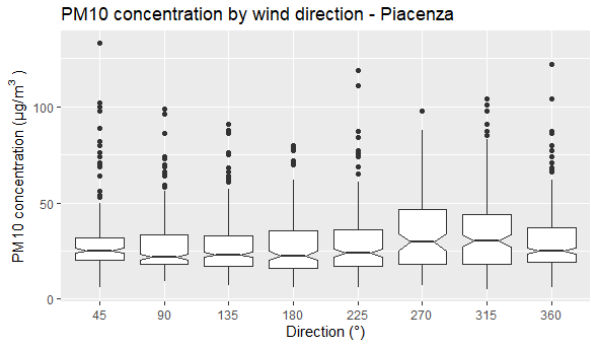


(a) Number of days per direction

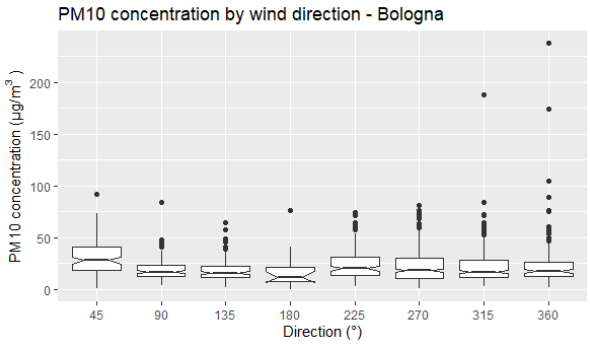


(b) Intensity distribution per direction

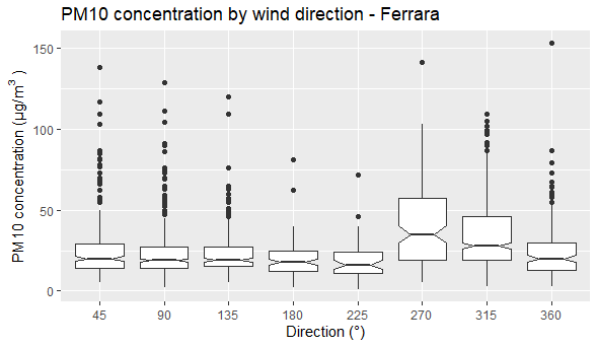
Figure 2.22: Wind characteristics per direction in Rimini (2013-2017).



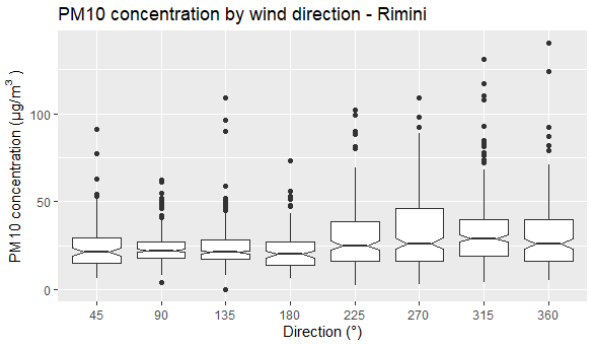
(a) Piacenza



(b) Bologna



(c) Ferrara



(d) Rimini

Figure 2.23: PM_{10} concentration vs wind direction (2013-2017).

2.1.6 Radiant exposure

In order to quantify the amount of solar radiation at the surface, ARPAE performs measures of radiant exposure, i.e. the radiant energy received by a surface per unit area (the corresponding unit of measure is J/m^2). As can be easily predicted, the exposure varies seasonally with higher values in summer and lower ones in winter. The trend for the year 2016 is shown in Figure 2.24, where the predicted behaviour can be observed. The presence of oscillations with respect to a regular annual oscillation is due to perturbations caused by cloud cover and other phenomena related to sunlight dimming: this justifies the variability betweeen different places in the same period of time.

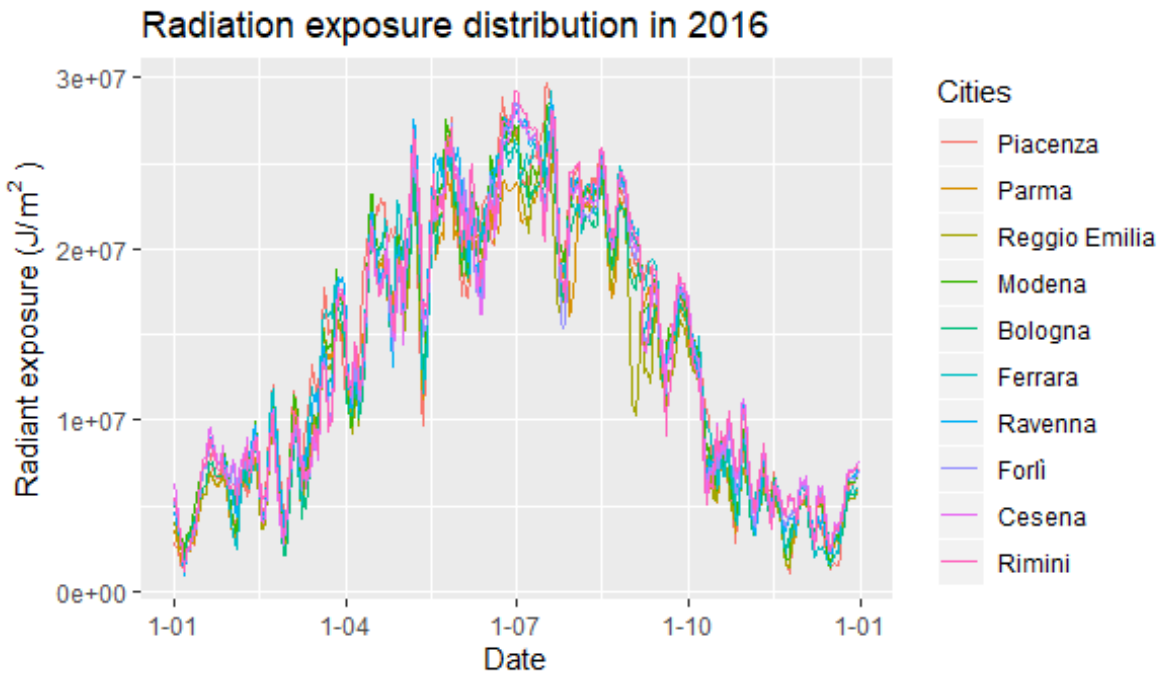


Figure 2.24: Radiant exposure values (in m/s) in 2016 (represented as 5-days mobile mean on daily values).

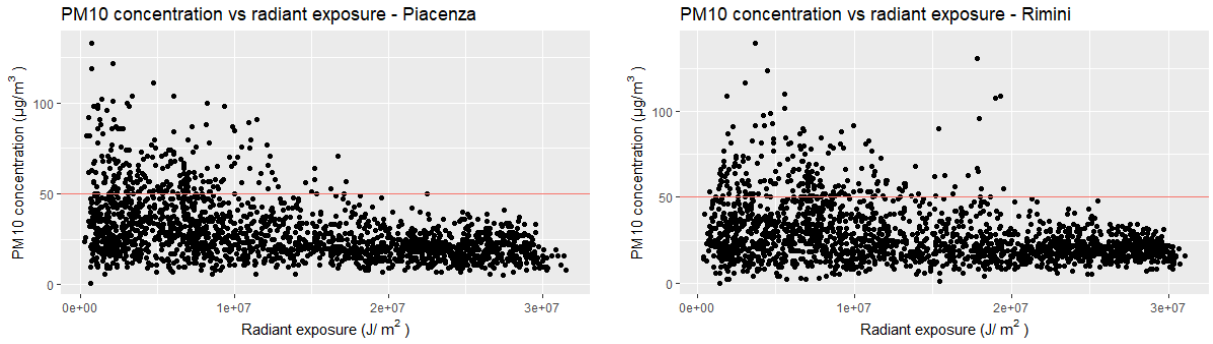
In order to evaluate the relationship between radiant exposure and PM_{10} concentration, the Pearson correlation values have been computed for all the cities considered. The results are shown in Table 2.12.

Variable	Correlation with PM_{10} concentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
RE	-0.440	-0.395	-0.445	-0.509	-0.404	-0.388	-0.400	-0.423	-0.380	-0.343

Table 2.12: Correlations between radiant exposure and PM_{10} concentration.

As days with higher radiant exposure are typical of summer, high values of radiant exposure correlates with lower levels of PM_{10} concentration: this implies that the correlation is negative.

Scatterplots for Piacenza and Rimini are shown in Figure 2.25.



(a) Piacenza

(b) Rimini

Figure 2.25: PM_{10} concentration vs radiant exposure (2013-2017).

While the behaviour in Figure 2.25(a) is common to the other cities (not shown) in the western part of the region, Rimini is characterized by some anomalies (days with a significant radiant exposure and high level of pollution at the same time).

2.1.7 Atmospheric pressure

Atmospheric pressure, which describes the pressure exerted by the atmosphere on the surface (or the point of the atmosphere where the quantity is measured) is generally linked to other quantities, such as wind speed and density variations of the air (which can be caused by alterations of temperature or composition of the atmosphere).
Figure 2.26 shows the distribution of atmospheric pressure in 2017: the quantity fluctuates simultaneously in all the considered cities, while some differences persist between different locations for the whole period.

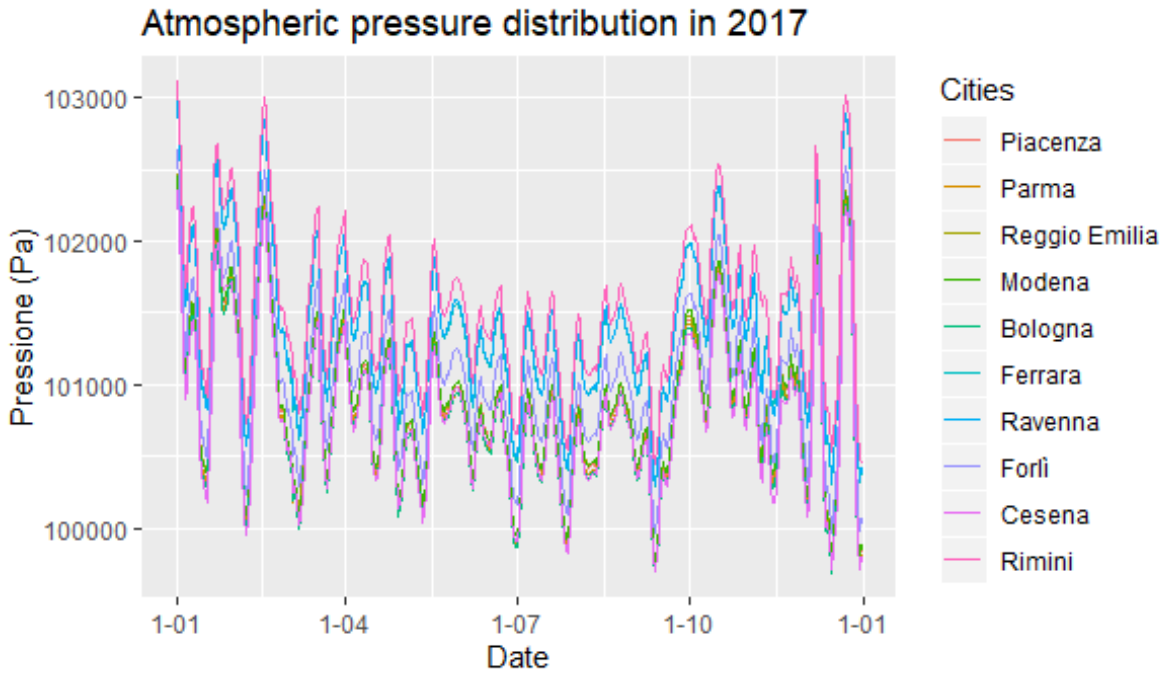


Figure 2.26: Atmospheric pressure values (in Pa) in 2017 (represented as 5-days mobile mean on daily values).

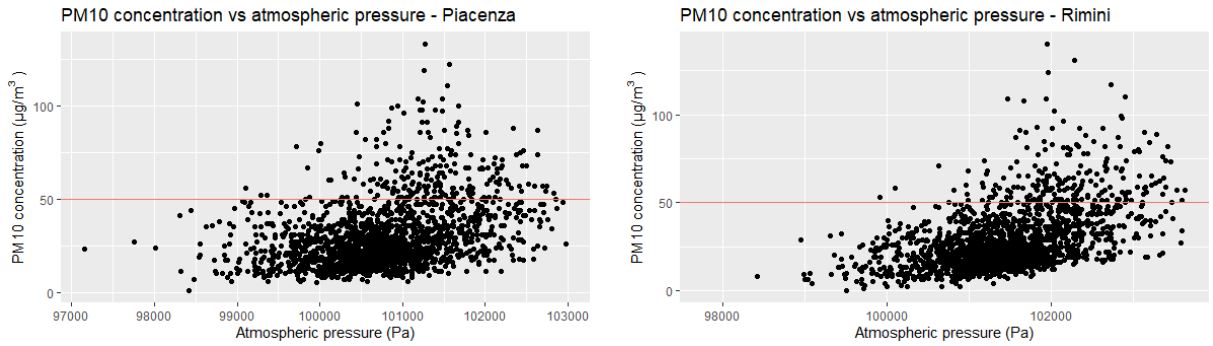
The Spearman correlation values are presented in Table 2.13.

Variable	Correlation with PM_{10} concentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
p	0.376	0.437	0.416	0.422	0.439	0.440	0.394	0.438	0.409	0.487

Table 2.13: Correlations between atmospheric pressure and PM_{10} concentration.

Atmospheric pressure is positively (although not strongly) correlated with PM_{10} concentration.

Plots in Figure 2.27 show the kind of relationship which is shared by all cities: higher pollution levels are associated with high pressure values, despite low PM_{10} concentration values can be associate with both low and high pressure.



(a) Piacenza

(b) Rimini

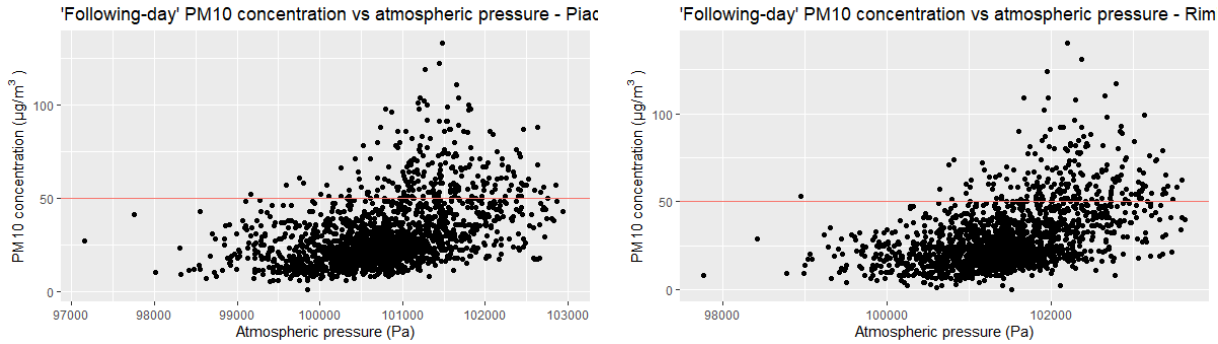
Figure 2.27: PM_{10} concentration vs atmospheric pressure (2013-2017).

Since the variation of atmospheric pressure can have effect on PM_{10} concentration on a longer timescale, Spearman correlation has been computed between pressure values and PM_{10} levels with a lag of one day. Results are summarized in Table 2.14.
It is interesting to notice that, while the correlation improves for the western cities, it worsen for Rimini.

Variable	Correlation with PM_{10} concentration ($PM_{10}(d+1)$)									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
p	0.484	0.518	0.523	0.472	0.475	0.495	0.470	0.465	0.454	0.483

Table 2.14: Correlations between atmospheric pressure and PM_{10} concentration on the following day.

Scatterplots in Figure 2.28 don't show different patterns with respect to the previous plots (the correlation is not strong in either cases).



(a) Piacenza

(b) Rimini

Figure 2.28: PM_{10} concentration of the following day vs atmospheric pressure (2013-2017).

2.1.8 Mixed layer height

The mixed layer identifies the lower part of the atmosphere where the motion of the air is directly affected by the presence of the surface: in this layer fluctuations of various physical variables are common and quite rapid, while the motion of the winds is influenced by surface drag.

The height of the mixed layer, called *mixing height*, corresponds to the vertical depth available for air mixing processes. It is proportional to the atmospheric turbulence intensity, that depends both on convection processes due to daily surface warming and drag between wind and surface. Mixing height is also negatively affected by temperature inversion phenomena, which are common during wintertime.
Mixing height is therefore an important variable concerning PM_{10} concentration, because it identifies the amount of space where particulate matter can diffuse in absence of horizontal winds of adequate strenght: in these cases, pollutants can only be moved vertically by convection and so they tend to fill the available volume.

For the reasons explained above, the mixing height is expected to decrease during winter months, when PM_{10} concentration generally increases.
Figure 2.29 showing the annual distribution of daily maximum values for the mixing height in 2013 proves that behaviour. Rimini and Ravenna presents the most different behaviour for this variable, which is probably affected by their location near the sea.

The dependence of PM_{10} concentration on mixing height is quantified by the values of Spearman correlation shown in Table 2.15.

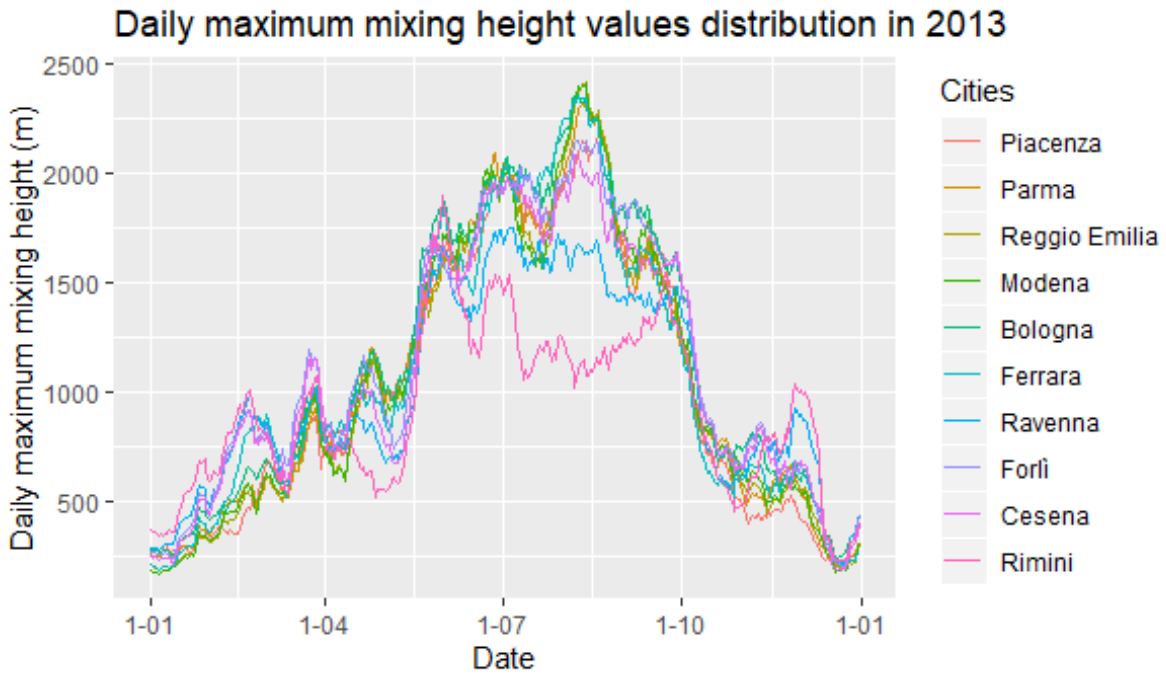


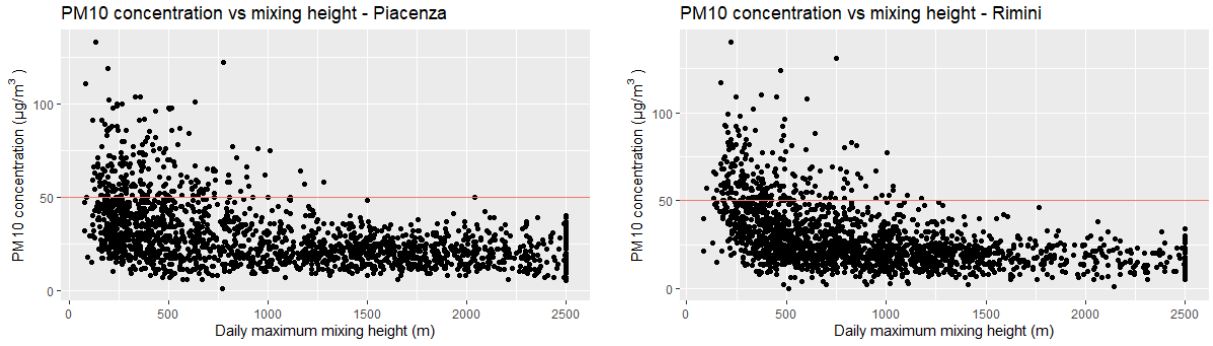
Figure 2.29: Daily maximum values of the mixing height (in m) in 2013 (represented as 15-days mobile mean on daily values).

Variable	Correlation with PM_{10} concentration									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
$max(H_{mix})$	-0.499	-0.484	-0.502	-0.577	-0.486	-0.502	-0.521	-0.516	-0.498	-0.535

Table 2.15: Correlations between daily maximum mixing height and PM_{10} concentration.

The correlation is significant, although there's no strong evidence for a linear dependence.

These values are reflected in the scatterplots reported in Figure 2.30: for high values of mixing height no high pollution episodes are present, while high concentrations can be found in days with low values of mixing height.



(a) Piacenza

(b) Rimini

Figure 2.30: PM_{10} concentration vs daily maximum mixing height (2013-2017).

2.2 Missing data treatment and imputation

As seen at the beginning of section 2.1, the dataset of meteorological variables contains 1772 samples with one or more missing values. Since all the variables had to be considered in the regression models, it has been necessary to address this issue in order to deal with a complete dataset.

In presence of an incomplete predictor Y_j ($j = 1, \ldots, p$, where p is the number of variables in the considered dataset), i.e. a variable in which there is at least one missing value in the considered set of data, a single observation where the value of Y_j is missing, i.e. an *incomplete sample*, can be written as $Y_{-j}^{(i)} = \left(Y_1^{(i)}, \ldots, Y_{j-1}^{(i)}, Y_{j+1}^{(i)}, \ldots, Y_p^{(i)}\right)$; similarly, the collection of predictors of the dataset without the j^{th} one is represented by $Y_{-j} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$.
In presence of missing data, the incomplete dataset composed of the values which have been actually observed for the measured variables can be denoted by Y^{obs}.

Since the estimation of a response variable starting from an evaluation of the values of its predictors can't be performed without making some kind of assumptions about the missing data, a number of methods have been developed in order to deal with this situation. In this work, two paths have been considered:

- the removal of all the incomplete samples prior to analysis, known as *listwise deletion* or *complete-case analysis*;
- the *multiple imputation* of missing data, i.e. the replacement of missing data with values drawn from an appropriate distribution: this task is iterative and is generally performed a number m of times.

2.2.1 Listwise deletion of missing data

Listwise deletion is a procedure that involves the removal of all the incomplete samples that are present in the considered dataset.
It is considered as a "standard approach" [37]. It also generally implies a considerable loss of information.

In the case of the present work, the use of listwise deletion approach has involved the deletion of the 1772 incomplete samples.
The cleaned dataset (called *LWD dataset*) contains 18308 samples: their distribution with respect to the city in which the sample has been taken is summarized in Table 2.16 (in the original dataset, each city corresponds to 2008 samples).

City	Number of samples	City	Number of samples
Piacenza	1872	Ferrara	1893
Parma	1793	Ravenna	1914
Reggio Emilia	1673	Forlì	1775
Modena	1933	Cesena	1891
Bologna	1831	Rimini	1733
TOTAL	18308		

Table 2.16: Number of samples after listwise deletion.

2.2.2 Multiple imputation of missing data

In order not to lose the information contained in the incomplete samples, imputation is used as a way to complete those samples by filling appropriately evaluated values for the missing variables.
Concerning the imputation of missing data in this work, the approach of *multiple imputation* has been taken.[36] This approach implies the creation of a number of imputed versions of the original dataset Y: in each version the missing data are replaced by plausible values drawn from a distribution specifically modelled for each missing entry; each version is generated independently from the others. So the imputed datasets are identical in the non-missing part Y^{obs} and differ in the imputed part; the difference between the values imputed for a certain missing entry is generally larger when the uncertainty on the distribution of the involved variable is higher.
The imputation model should consider the process that created the missing data, preserve the relation in the data and the uncertainty about these relations, as well as address a number of issues regarding the characteristics of the predictors [6], the interdependency (i.e. correlation) between variables, the different types of variable that can be present in the dataset (e.g. numeric continuous variables and categorical ones), the imputation of impossible or unlikely values and others [36].

The imputation model which has been used for this work involves the *technique of chained equations* in order to get a separate model for each variable of the dataset. The hypothetically complete dataset is considered as obtainable by random sampling from the multivariate distribution $P(Y|\theta)$, where $Y = Y_1, \ldots, Y_p$ and θ is a vector of unknown parameters that completely specifies the distribution. The posterior distribution of θ is obtained sampling iteratively from the conditional distributions $P(Y_i|Y_{-i}, \theta_i)$: the t^{th}

[6] In the imputation of variable Y_j, the maximum set of predictors that can be considered is the subset of all the other variables in the dataset, i.e. $(Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots, Y_p)$. Concretely, some or all of the predictors may be themselves incomplete variables: in these case, they (or their imputed values) may have to be ignored during a part of or the entire imputation procedure (this happens certainly at the beginning of the imputation process, when no imputed values has been assigned yet).

iteration of the process computes sequentially the values of

$$\theta_1^{*[t]} \sim P(\theta_1 | Y_1^{obs}, Y_2^{[t-1]}, \ldots, Y_p^{[t-1]}) \tag{2.9}$$
$$Y_1^{*[t]} \sim P(Y_1 | Y_1^{obs}, Y_2^{[t-1]}, \ldots, Y_p^{[t-1]}, \theta_1^{*[t]}) \tag{2.10}$$
$$\vdots \tag{2.11}$$
$$\theta_p^{*[t]} \sim P(\theta_p | Y_p^{obs}, Y_1^{[t]}, \ldots, Y_{p-1}^{[t]}) \tag{2.12}$$
$$Y_p^{*[t]} \sim P(Y_p | Y_p^{obs}, Y_1^{[t]}, \ldots, Y_{p-1}^{[t]}, \theta_p^{*[t]}) \tag{2.13}$$
$$\tag{2.14}$$

where $Y_j^{[t]} = \left(Y_j^{obs}, Y_j^{*[t]}\right)$ and $Y_j^{*[t]}$ are the imputed values at step t: in this way, previous imputations of Y_j do not influence directly the same variable at the successive steps, while they affect it indirectly (in fact, Y_j^* values are considered in the imputation of all the other variables $Y_i\,(i \neq j)$).

The specific imputation method chosen for each variable to be imputed takes, at each step, the subset of considered predictors (which have been imputed themselves at that moment, if incomplete[7]) and computes a single imputed value for each missing entry using a specific function chosen for the variable.
The subset of predictors of a variable can be restricted when the number of these predictors is huge or when computational problems arise during the imputation process (e.g. when predictors are linearly dependent and the corresponding coefficient matrix is singular.).
The number of iterations to be performed has to be empirically chosen comparing the actual convergence of basic parameters (e.g. mean and variance) among the imputed datasets. Convergence is achieved when the values of the parameters appear to intermingle throughout the iterations.

Once the multiple imputation is performed, the imputed datasets (which will be called *MI datasets*) are obtained.

2.3 Linear regression models

In order to quantitatively model the relationship between meteorological variables (predictors) and PM_{10} concentration (response variable), a number of regression models suitable to this kind of task (generally called *supervised statistical learning*) has been

[7]The 0^{th} imputation is performed taking a random draw from the observed data.

considered.

Given the true relationship f between a set of predictors $X = (X_1, ..., X_p)$ and a response Y as $Y = f(X) + \epsilon$, where ϵ is a random error with zero mean, a model $\hat{f}$ is an estimate of that relationship and can be used in order to predict values $\hat{Y}$ of the response variable as $\hat{Y} = \hat{f}(X)$. A model is always affected by a reducible error (that depends on the chosen model, so that it can be reduced choosing the best statistical learning technique and parameters) and an irreducible error (due to the presence of the random error ϵ which is not known), that diminishes the accuracy of the prediction. The effects of these kinds of error can be summarized in the expression of the expected value of the squared difference between the modelled and the actual value of the response variable:

$$E(Y - \hat{Y})^2 = E\left[f(X) + \epsilon - \hat{f}(X)\right]^2 = E\left[f(X) - \hat{f}(X)\right]^2 + Var(\epsilon) \tag{2.15}$$

where the first term represents the reducible error and the second one corresponds to the irreducible error (i.e. the variance of the random error).

In the following paragraphs an overview of the standard and regularized linear regression model that have been applied to the meteorological dataset will be presented, while in section 2.4 the same will be made for regression tree-like models. In section 2.5 the chosen method to assess and compare the performances of the considered models will be described, while the results of the assessment will be exposed and commented in section 3.

2.3.1 Standard linear regression

Response estimation by linear regression is a common way of assessing the presence of a linear relationship between predictive and response variables in a dataset.
In order to simultaneously evaluate the contribution of the considered set of predictors, *multiple linear regression* has been used. This model can be represented by the relationship:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{2.16}$$

where β_0 is the intercept and β_i, $i = 1, \ldots, p$ are the coefficients (also called *slopes*) that model the association between the predictor values X_i and the response valus Y.
Since the coefficients are unknown, multiple linear regression allows to obtain an estimate of this relationship, which can be written as $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$. The estimate of the parameters $\hat{\beta}_0, \ldots, \hat{\beta}_p$ is obtained applying the *least squares criterion* on a set of samples $(x_{i1}, \ldots, x_{ip}, y_i)$, i.e. by minimizing the *residual sum of squares*

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip} \right)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right)^2 \tag{2.17}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$ $(i = 1, \ldots, n)$ are the predicted values of y_i . Because of this, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are called the *least squares coefficient estimates* for multiple linear regression.

2.3.2 Ridge regression

A modified version of linear regression involves the estimation of the model parameters $\hat{\beta}_0, \ldots, \hat{\beta}_p$ that minimize a *regularized* expression of the residual sum of squares

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.18}$$

where the second addend defines a *shrinkage penalty* and $\lambda\,(> 0)$ is the *tuning parameter* that determines the strength of the regularization (for $\lambda = 0$ the linear RSS is obtained). The process of regularization has the effect of shrinking the fitting coefficient towards zero without setting them to zero: they progressively decrease their squared values as λ increases. In this way, predictors whose contribution to the modelling process is "less important" see their coefficients reduced.

Regularization corresponds to a reduction of the variance and an increase of the bias of the fitted model: the strength of this effect grows with the magnitude of the tuning parameter λ.
Since the test mean squared error (or MSE, the metric used to estimate the error between the measured values of the response variable and their estimates; for the definition and the way it has been used in this work, see paragraph 2.5.1) is itself a function of both the squared bias and the variance, exploring a range of values for λ allows to find the best condition for the regularization, i.e. that in which the test MSE takes its lowest value. The process of selection of the best λ value is generally performed by means of a cross-validation procedure: for this work, the chosen path is explained in paragraph 2.5.

A notable element to highlight is that, as regularization of the fitting coefficients is affected by the scaling of the corresponding predictors, it is necessary to standardize their values: this requires the transformation of their values by means of the formula

$$\bar{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x_j}\right)^2}} \tag{2.19}$$

2.3.3 Lasso regression

As showed, Ridge regression shrinks the values of the fitting coefficients but does not set them to zero. Setting a coefficient to zero implies a reduction of the number of variables considered in the model, i.e. a process of *feature selection*.
A second type of regularized linear regression, called *Lasso regression*, is a slightly modified version of the Ridge model and is able to perform feature selection through the minimization of the expression

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \tag{2.20}$$

The second addend adds a penalty which is proportional to $|\beta_j|$, instead of the β_j^2 term which characterizes Ridge regression. Increasing the value of the tuning parameter λ, the number of predictors gets shrunk by progressively removing the least important ones; on the other hand, as for Ridge regression, $\lambda = 0$ corresponds to the standard linear regression.
As in the case of Ridge regression, standardization of the values of predictors is required.

2.4 Regression tree models

Decision trees algorithms allow to predict the value (in the regression case) or the class (in the classification case) corresponding to a set of values for the predictors by *segmentation* of the predictor space.[27]
A decision tree is trained using a set of data and is composed of a number of *nodes*, each one corresponding to a splitting rule, i.e. a function of one of the predictors that splits the variable space in two regions; each split is chosen in order to minimize the "distance" between samples that belongs to the same region of the space obtained in the splitting.
A *recursive binary splitting* approach (explained below) allows to create a set on nodes, starting from a *root* and finishing on the terminal nodes (*leaves*): nodes are generated by iteratively performing a region split at each step of the process. Each *internal node* generates two separate regions which can be depicted as *branches* in a graphical representation of the algorithm. The regions correspond to hyper-rectangles in the variable space and are defined in order to minimize the expression of the residual sum of squares

$$RSS = \sum_{m=1}^{M} \sum_{i \in R_m} (y_i - \hat{y}_{R_m})^2 \tag{2.21}$$

where y_i is the response value for the sample, $\hat{y}_i$ is the corresponding prediction and J is the number of regions R_j in which the space has been split; each sample (X_i, y_i) only belongs to one region.

The pseudo-algorithm can be sinthetically described as follows, starting from the whole variable space R:

1. for each predictor X_i $(i = 1, \ldots, p)$ find the cutpoints s_i, i.e. the numerical values that minimize the RSS expression

$$RSS(i) = \sum_{k: x_k \in R_1(i,s)} (y_k - \hat{y}_{R_1})^2 + \sum_{k: x_k \in R_2(i,s)} (y_k - \hat{y}_{R_2})^2 \tag{2.22}$$

   where $R_1(i, s) = \{X : X_i < s\}$ and $R_2(i, s) = \{X : X_i >= s\}$;

2. choose the cutpoint s_j corresponding to $RSS(j) = \min_i RSS(i)$ as next node of the tree and split the variable space correspondingly;

3. iterate from step 1, considering separately each region that has already been defined and choosing at each iteration the best cutpoint among all the regions and the predictors;

4. stop when a certain criterion is reached.

At the end of this algorithm, a partition of the variable space is obtained. It can be mathematically described by the corresponding model function

$$f(X) = \sum_{m=1}^{M} c_m \cdot 1_{(X \in R_m)} \tag{2.23}$$

where c_m is the predicted value to each sample X that is assigned to region R_m given the values of its predictors, and M is the number of regions.

Since such procedure can easily lead to overfitting (e.g. when the tree gets too complex, or when each terminal node corresponds to a very limited number of training samples), a common strategy involves *pruning* the tree and evaluating a cross-validation error on a number of subtrees.
In details, once the complete tree T_0 has been computed, the pruning approach involves:

- defining the cost function

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T| \tag{2.24}$$

  where $|T|$ is the number of terminal nodes of a subtree $T \subset T_0$ and α is the tuning parameter;

- evaluating the cost function for an interval of values of α, in order to get the sequence of "best" subtrees that minimize its value;

- using k-fold cross-validation (i.e. subsetting the training sample set and using it to evaluate a mean error value on a cycle of k training and validation tasks) on the set of obtained subtrees in order to choose the "best" value of α as the one corresponding to the subtree with the lowest mean error.

In the cost function, the tuning parameter α is responsible for determining the complexity of the trees.

2.4.1 Bagging and random forests

Even when pruning is performed, a regression tree is generally affected by high variability depending on the training data. In order to reduce variance, a useful approach is *bagging* (*bootstrap aggregation*) in which variance reduction is achieved by using more than one training set and then averaging. Bagging then implies

- the construction of B separate training sets by bootstrapping the original training set;

- the computation of a regression tree $\hat{f}^{*b}(x)$ for each set ($b = 1, \ldots, B$; the asterisk refers to the bootstrap process that has generated the subsets);

- the averaging of the obtained trees:

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{2.25}$$

Since variance is reduced by averaging, pruning is not necessary anymore.

An issue that can arise in bagging is related to the presence a limited number of strong predictors among the considered variables: this situation usually leads to the computation of very similar bagged trees, with a high correlation that preserve a high variability of the model.
In order to overcome this problem, a preliminary random selection of m predictors from the original set of p variables can be performed at each splitting step of the tree computation, for each tree, so modifying the original bagging approach.
The obtained set of trees is averaged just like in the bagging model (see Equation 2.25). This model is generally referred to as a *random forest*.

In this case, cross-validation is perform in order to find the best values for both hyperparameters: the number B of trees and the size m of the set of considered variables.

2.4.2 Boosting

The process of *boosting* involves the evaluation, for a given training set, of a sequence of B regression trees: each tree is grown until it reaches a thresholded value d of terminal nodes, then its contribution, reduced by a shrinkage parameter λ, is added to the previously computed trees. The result is an model which is iteratively built with a *slow learning* approach. Instead of regressing on the original response values, residuals are computed at each step so that the model progressively manages to improve the fitting in the response regions where it performs worse.
In the form of pseudo-algorithm, boosting can be summarized as follows:

1. set $\hat{f}(x) = 0$ and $r_i = y_i$ for each sample i of the training set;
2. for $b = 1, \dots, B$, given a maximum number d of splits, fit a regression tree $\hat{f}^b$ to the training set (X, r), then update both the boosting model, as $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$, and the residuals, as $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)$;
3. average the obtained models:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x) \tag{2.26}$$

Although this process implies that each step is built on the previous one, and so the choices at subsequent steps are necessarily interdependent, the slow learning allows to avoid overfitting.

The use of cross-validation allows to explore some ranges of values for the hyperparameters B, d and λ.

2.5 Model assessment and selection

Assessing the performance of a number of regression models, as in the case of this work, implies the evaluation of quantitative parameters that allow to understand how well each model is able to reproduce the values of the response variable given a sample of values of the predictors, i.e. the quality of the prediction. After the *selection* of the best performing model, *testing* is a subsequent task: it involves evaluating the performance of the chosen model again, using a completely new set of data.

2.5.1 Measuring the error

In the context of regression tasks, the evaluation of the *mean squared error* (MSE) is considered an index of quality for the fit both in the assessment phase (which is also called *validation*) and in the testing one.
Starting from a dataset of n samples (X_i, y_i) used to assess the model's prediction ability, this parameter is computed as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(X_i)\right)^2 \tag{2.27}$$

where y_i is the measured value of the response and $\hat{f}(X_i) = \hat{y}_i$ is the predicted value of the response for the same sample obtained from the model applied the values X_i of the predictors.
As said, MSE can be used as a cost function both in the validation of a group of algorithms (in this phase it is called *training mean squared error*) and in testing (when it is called *test mean squared error*), and so it will be used in both tasks in this work.

The MSE measures the "lack of fit" of the model [27] and is measured in the same units of y, thus it cannot be readily compared with analogous measures. In order to overcome this problem, another commonly used measure of the quality of the fit is the R^2 statistic, whose expression is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.28}$$

where $\hat{y}_i$ is the predicted value of y_i and $\bar{y}$ is the average of the measured values y_i ($y = 1, \ldots, n$).
It can be said that R^2 describes the amount of variability in the values of the response variable that can be explained by the model. R^2 assumes a value in the range $[0, 1]$, so its complement $(1 - R^2)$ can be seen as the amount of variability that is not explained

by the model.

In this work, the choice of the best model among a number of fitted ones has been performed by choosing the one corresponding to the highest value of R^2. However, the corresponding value of MSE has always been reported in order to quantify the average error in the prediction task.

2.5.2 Choosing the predictors

As previously declared in section 1.4, the aim of this work was to perform a comparative assessment among regression models in order to find the best one in predicting the value of PM_{10} daily mean concentration starting from the values of the meteorological variables measured in the same day, separately for each considered city.
The *basic* set of predictors that has been considered corresponds to the variables described in Table 2.2. It is necessary to report that W_{dir} has been treated as a categorical variable, since its range of values corresponds to 8 discrete values representing the cardinal points (as explained in paragraph 2.1.5).

As the review of literature showed that a significant number of works also considered time-related variables, a choice has been made in order to include non-meteorological variables among the set of predictors. This choice has been applied, in particular, in two subsequent steps:

- adding two categorical variables that represent the month and the day of the week in which each sample has been taken, in order to incorporate the periodicities analysed in paragraph 2.1.2; the resulting set of predictors has been identified as a *nonmet* set;
- adding to the previous set of predictors the value of PM_{10} daily mean concentration which has measured in the previous day with respect to the day in which the sample has been taken, so to provide to the model an information on persistence of the pollutant (again in paragraph 2.1.2); the resulting set of predictors has been identified as a ($nonmet$+$lag[PM_{10}]$) set.

2.5.3 Splitting the datasets

In order to perform a quantitative assessment of the performance of the best model in an ensemble, the datasets which have previously been described (the listwise-deleted or LWD dataset, and the multiple-imputed or MI datasets; for definitions see section 2.2, while for the use of MI datasets see paragraph 2.5.7) have to be split.

It has been chosen to divide both of them into a *traning set* containing a percentage of 80% of the total samples and a *test set* containing the remaining 20%. A so called *stratified sampling* operation has been used in order to maintain the proportion of the values of the response variable in both datasets. In Table 2.17 the size of each city-level dataset obtained from the original LWD dataset is reported; for MI-derived city-level datasets, each of them contains 2008 samples and the split created sets with the same number of samples (1608 for the training sets, 400 for the test sets).

City	Piacenza	Parma	Reggio Emilia	Modena	Bologna	
Training set	1498	1436	1341	1548	1467	
Test set	374	357	332	385	364	
City	Ferrara	Ravenna	Forlì	Cesena	Rimini	Total
Training set	1517	1498	1421	1515	1388	14648
Test set	376	374	354	376	345	3660

Table 2.17: Number of samples in each city-level dataset obtained from LWD dataset.

Each training set has been used in order to cross-validate the models that have been described in sections 2.3 and 2.4, while the test set has been used to assess the performance of the best model chosen after cross-validation. The next paragraph describes these tasks in greater detail.

2.5.4 Cross-validation procedure

In order to select a model, a cross-validation procedure has been implemented. Such a procedure has been necessarily performed in order to evaluate the performance of models that have to be provided with a number of hyperparameters: a grid of values for each hyperparameter has then been selected. Table 2.18 summarizes the hyperparameters subjected to cross-validation for each considered model.

The procedure of cross-validation that has been followed, i.e. a standard *10-fold cross-validation* procedure, involved the following steps:

- the training set has been divided, performing a random split, into $K = 10$ uniquely identified folds of (approximately) equal size;
- a *for* loop with 10 iterations has been implemented: at each of these, a single fold (which changed in every iteration) has been used for evaluating the MSE and R^2 values for that iteration, while the remaining 9 folds have been used for training the model before the measurement of the regression error;

Algorithm	Hyperparameter
Standard linear regression	(none)
Ridge regression	Penalty λ
Lasso regression	Penalty λ
Single regression tree	Size (after pruning)
Bagging / Random forests	Number of tree B Number of predictors m (in range $1 \div p$)
Boosting	Number of tree B Depth of growth d Shrinkage λ

Table 2.18: Hyperparameter values considered in cross-validation.

- once all the iterations has been performed, the best estimates and standard errors for MSE and R^2 have been computed as:

$$\text{MSE} = \frac{1}{K}\sum_{k=1}^{K}\text{MSE}_k \qquad R^2 = \frac{1}{K}\sum_{k=1}^{K}R_k^2 \tag{2.29}$$

$$\text{SE}_{\text{MSE}} = \frac{\sigma(\text{MSE}_k)}{K} \qquad \text{SE}_{R^2} = \frac{\sigma\left(R_k^2\right)}{K} \tag{2.30}$$

  where MSE_k and R_k^2 are the values calculated on the k^{th} fold.

2.5.5 Model selection and assessment of the performances

At the end of the cross-validation procedure, the best performing model for each city has been selected in order to be trained again using the whole training set and evaluate the values of MSE and R^2 on the test set, as an unbiased measure of the performance.

Whenever two or more models have to be compared in order to determine which one performs better, a 2-sample statistical test has been performed.[13] In these cases, the null hypothesis that is considered is that the two samples come from the same distribution.
In order to perform a statistical test of this kind, a sample of values for the quantity that represents the performance of each involved model has to be used. In this work the values of R_k^2 ($k = 1, \ldots, 10$) obtained from cross-validation of each compared model have been used for these tasks.
Since some of the comparisons involve R^2 samples obtained performing 10-fold cross-validation on models trained with dataset which have been fold-split in the same way (e.g. each fold contains the same samples from the dataset; the number of used predictors may vary from one model to the other), those samples have to be considered *paired*.

In other cases, the comparison involves R^2 samples obtained from dataset with different fold-spitting (e.g. fold obtained by dataset with different numbers of samples): these are considered as *non-paired* samples.

A common parametric test such as Student's t-test, which can be used for similar tasks, requires the distribution of the differences between the two considered samples to be normally distributed: however, in the present case all the distribution contains only 10 samples, making it difficult to assume such an hypothesis.
So it has been chosen to use non-parametric tests that do not require the hypothesis of normality. In particular, two tests have been used[8]:

- *Wilcoxon signed-rank test* for paired samples, that assumes that the distribution of the differences is symmetric around zero as the null hypothesis; the calculation of the test statistic

$$W = \sum_{i=1}^{N_r} [sgn(x_{2,i} - x_{1,i}) \cdot R_i] \tag{2.31}$$

  where R_i is the ascending rank of the pair $(x_{1,i}, x_{2,i})$ based on its absolute difference $|x_{2,i} - x_{1,i}|$, is followed by the determination of the corresponding significance level on the basis of the W distribution (characterized by $\mu_W = 0$ and $\sigma_W = \sqrt{N(N+1)(2N+1)/6}$; the use of significance table in order to identify the significance is suggested);

- *Mann-Whitney-Wilcoxon test* for non-paired samples, that assumes the null hypothesis that both distribution share the same location value; the test statistic is given by

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}, \qquad U_2 = R_2 - \frac{n_2(n_2+1)}{2} \tag{2.32}$$

  where n_1 (n_2) is the number of data in the first (second) sample and R_1 (R_2) is the sum of the ranks of the data in the first (second) sample (the ranks are attributed by putting data from both samples in the same set and assigning ranks in ascending order); the smaller value between R_1 and R_2 is used to obtain the corresponding p-value from the significance table.

[8]It must be noticed that both the considered tests, as the Student's t-test, assume that all the data are independent from each other. This is certainly false in the case of R^2 values considered in this work, since they have been calculated using overlapped folds of the same dataset. However, using the non-parametric tests avoid the necessity of assuming the normality of the distribution of the samples: in this, they appear better than Student's t-test.

2.5.6 Classification task based on PM_{10} daily limit value

Since each model provides predictions for the values of the response variable, it is possible to perform a further assessment of the performance of each model on a classification task: in particular, the task of interest for the present work is the one of predicting if each sample of predictors corresponds to a value of PM_{10} concentration that exceeds the daily limit or not.

In order to do so, the measured values for the response variables have to be binarized so that

$$y_{bin}^{(i)} = \begin{cases} 0 & y^{(i)} \leq 50\,\mu\text{g/m}^3 \\ 1 & y^{(i)} > 50\,\mu\text{g/m}^3 \end{cases}$$

Using the predicted values $\hat{y}_i$ and the corresponding binarized true ones obtained in the model testing task, a Receiver Operating Characteristics (ROC) curve [19] has been created in order to evaluate the performance of the models based on a binary threshold value used to assign smaller models to the negative class (the one of *non-exceeding* values) and larger ones to the positive class (of *exceeding* values).

The performance in this case has been assessed using the Area Under the Curve (AUC), whose value is always between 0 (for a model that classifies all the samples in the wrong way) and 1 (for a classifier that is always right); 0.5 is the value of AUC for a classifier that guesses randomly. The AUC can be considered a quality parameter for evaluating the performance of classifiers because it is "equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance".[9] So models whose prediction correspond to higher AUC have been considered better in the classification task than those with lower AUC.

2.5.7 Approach with MI datasets

As concerns the datasets obtained following the multiple imputation method described in paragraph 2.2, considering that a number of datasets is generally obtained from the multiple imputation process, a method has to be defined in order to treat them systematically and obtain an analogous result with respect to a standard model selection procedure on a single dataset.

As concerns this work, following one of the possible approaches to this problem [20], it has been decided to perform cross-validation separately on the considered regression models for each one of the 5 MI dataset with the same grid of hyperparameters used

[9] It must be noticed, however, that a classifier with a larger value of AUC could perform worse than a second classifier with lower AUC in a specific region of ROC space.

for LWD dataset, and then to determine, for each combination of the hyperparameters, the average validation MSE and validation R^2 calculated for the models resulting from the training processes: the best model for each city has been chosen as the one with the highest average R^2 value.
Once the best model for each city has been selected, it has been tested with the corresponding MI test sets: the conclusive values for MSE and R^2 have been again obtained by averaging the ones separately obtained by each test set.

Since this approach implies a great computational cost, it has been chosen to use the MI datasets only for a regression task on the *basic* set of predictors in order to compare the results with those obtained with the LWD dataset for the same predictors.

2.6 Implementation in R

In this section, the details of the R implementation of the methods explained so far will be provided.

2.6.1 Missing data treatment and imputation

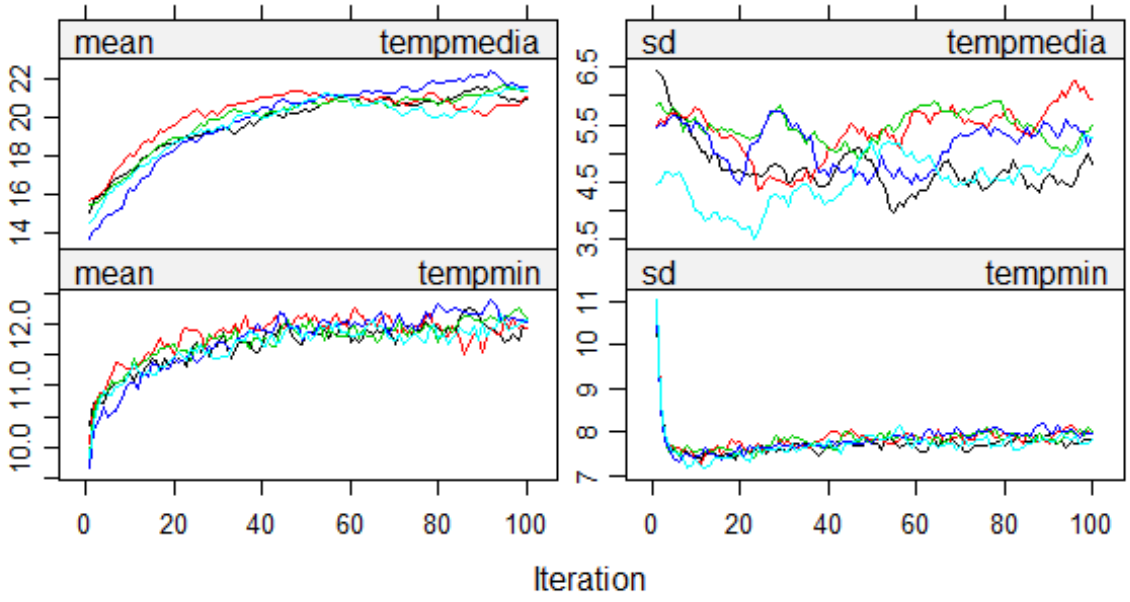
As concerns listwise deletion, the operation has been performed directly on the original dataset using the R function `na.omit`.

On the other hand, in order to practically perform the tasks involved in multiple imputation procedure, the package called `mice` has been used.
The function `mice()` from this package performs the imputation creating the selected number m of imputed datasets (here a value of 5 has been set), which are stored in a *multiply imputed datasets* (or `mids`) class object. After initializing a `mids` object without actually perfoming any imputation (parameter `maxit` has been set to 0), the imputation methods for different kind of variables have been chosen:

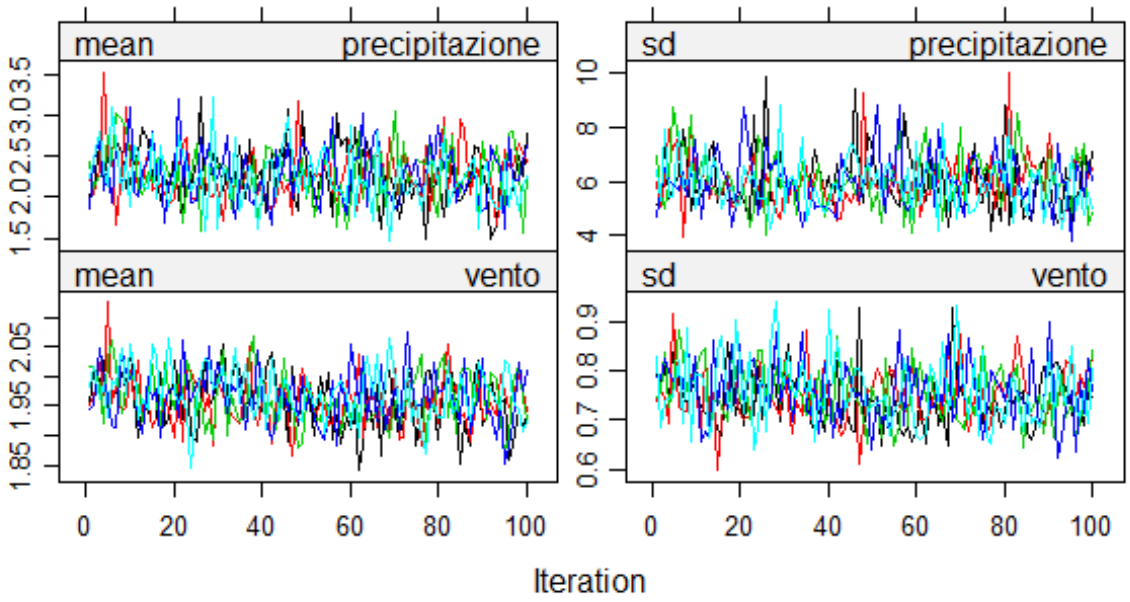
- for numeric variables, *predictive mean matching* function has been used: the imputed values which can be possibly assigned are restricted to observed ones for the same variable; it is considered a generally robust method;
- for factor-like variables (only W_{dir}), *multinomial logit model* has been used.

In order to select a useful subset of predictor for each variable, the `quickpred` function has been used. The function computes two correlation values for each pair of variables (the first using the values of target and predictor, the second using the binary response indicator of the target and the values of the predictor) and rejects the predictor if both values are below a chosen threshold (here the value of 0.1 has been set).
Furthermore, both T_{range} and $[PM_{10}]$ have been excluded from the predictors in the imputation process: for the former variable, it has been specified that the value of the variable must correspond to $T_{max} - T_{min}$; for the latter, it has been chosen not to use the variable as a predictor for meteorological conditions, since it is considered the response variable in this context.

As it is usual for imputation tasks, the number of iterations has been empirically determined in the present work. Figure 2.31 represents the variation of some of the considered parameters over the iterations of the procedure of imputation.
While various variables appear to reach convergence after a small number of iterations, some (e.g. temperature variables) are characterized by slow convergence and achieve a stationary trend only after approximately 100 iterations. Then the number of iterations



(a) Slowly-converging variables

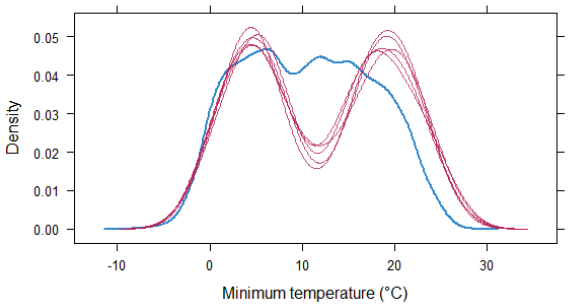


(b) Fastly-converging variables

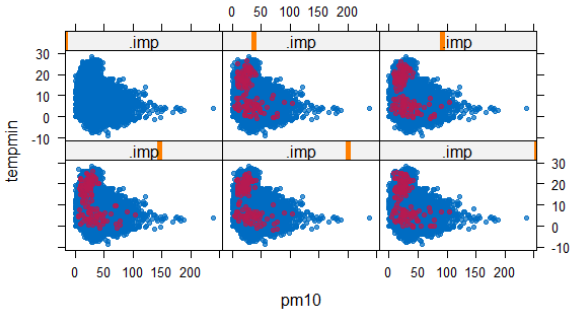
Figure 2.31: Convergence of mean and variance of imputed values vs number of performed iterations.

(i.e. the `maxit` variable in the `mice` function) has been set to that value.

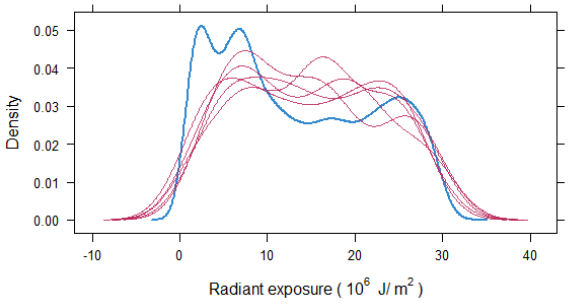
The imputed values have been also checked against the observed values, comparing the original distribution with both the kernel density estimates (KDE) and the distribution of imputed values for the 5 imputed dataset.
Figure 2.32 presents the comparisons performed in the two ways. The distributions of single values for the considered variables, plotted against PM_{10} values, are also shown separately for each of the 5 imputed datasets created in the imputation process.



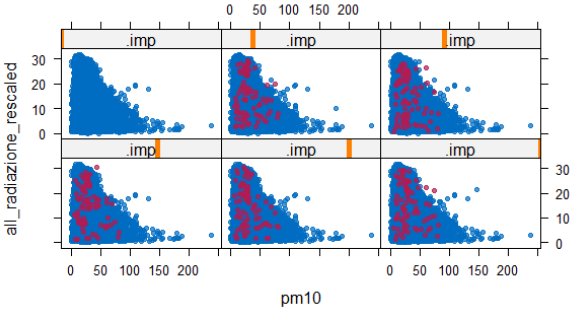
(a) Minimum temperature - Kernel density estimation



(b) Minimum temperature - Scatterplot



(c) Radiant exposure - Kernel density estimation



(d) Radiant exposure - Scatterplot

Figure 2.32: Comparison between observed (blue) and imputed (red) values for some variables.

It can be noticed that different patterns for observed and imputed data are highlighted in the KDE plot, since the imputation algorithm considers the distribution of both observed (for all variables) and imputed (for all variables except the considered variables) data: imputation does not produce a uniform distribution on the interval of the considered data.

Scatterplots allow to understand how missing data have been replaced in the imputed datasets, accounting for the difference among the KDE of those sets.

2.6.2 Linear regression models

Standard multiple linear regression has been performed in R using the `lm()` function. The function takes the following inputs:

- a formula that specifies the relationship to be modelled: the left-hand side contains the response variable, the right-hand side contains the sequence of predictors to be considered;
- the dataset in which the values of the response and the predictors are stored.

Once the model has been trained, the function provides a number of outputs including the linear regression coefficients (specifying both the errors and the significance), the residuals and the parameter that describes the quality of the fit.
The use of the `predict()` allows to apply the trained model to a new dataset of predictors in order to evaluate an estimate of the response variable using the regression coefficients obtained in the training task.

As concerns Ridge and Lasso regression models, R `glmnet` package provides the homonymous `glmnet()` function. In this function a `alpha` argument can be set in order to determine the kind of regularized linear model to consider[10]: the argument `alpha = 0` corresponds to the Ridge regression model, while `alpha = 1` corresponds to the Lasso model.
As said before, standardization of the values of predictors is necessary when regularization is present: in the case of `glmnet` function, this task is automatically performed by default (this setting can be changed using the argument `standardize`).
Once the regression model is chosen, a value (or a sequence of values, when optimization is required) of the tuning parameter λ has to be provided as argument `lambda` to the function call.
`glmnet()` function works similarly to `lm()` and can be trained on a set of data in order to get the regression coefficients.

[10]The name `glmnet` refers to the *elasticnet* regularization, a generalization of the regularized models considered in this work. Practically, the `alpha` argument corresponds to a mixing parameter that determins the reciprocal weight of the Ridge and Lasso penalties in a generalized form of the RSS which contains both terms.

2.6.3 Regression tree models

In order to work with regression trees, the package `tree` is available in R.
As in the previous cases, the main function `tree()` gets as input argument a formula that contains the response variable and the predictors, along with the name of the variable in which they are stored.
Using the function `summary` on a `tree`-class object stored as a variable, the outputs include the variable which have actually been used in the regression, the number of terminal nodes and other informations concerning the error and the residuals.
In the same package, the function `prune.tree()` can be used to perform pruning on a tree object. The function requires an argument `best` that corresponds to the (integer) number of terminal nodes of the pruned tree. So this function can be used iteratively to prune the tree object and train each obtained tree in order to select the best perfoming one.

Concerning bagging and random forest models, the R library `randomForest` provides the function `randomForest()`. This function takes as arguments the usual formula containing the response variable and the predictors, the number `mtry` of predictors considered at each tree split (all the p predictors in the case of bagging; `mtry` $< p$ in the case of random forests) and the number `ntree` of fitted trees in the model.

Boosting models can be obtained using the `gbm` package. The function `gbm()` takes as input the usual formula with response and predictor names and the arguments `distribution` (whose value depends on the kind of modelling task; `"gaussian"` is chosen by default when a regression must be performed), `n.trees` (setting the number of trees to be fitted) and `interaction.depth` (corresponding to the number of terminal nodes of each fitted tree).

2.6.4 Cross-validation, model selection and comparison

In order to perform cross-validation on the considered models and select the best performer, the first performed task has been the identification of the training set and the test set for each separate city, both for LWD dataset[11] and for the MI datasets. As a first step, the `createDataPartition()` function from the `caret` package has been exploited: it performs a stratified sampling by splitting the selected dataset, sectioning the interval of the values of the response variable and maintaining the proportions of samples for each identified subintervals in the training and test sets.

[11] These steps have been performed once with the largest ensemble of predictors; then the appropriate sub-ensemble has been selected appropriately for each modelling task, depending on the set of predictors of interest.

The arguments of this function include the name of the variable corresponding to the dataset to be split, the proportion `p` of samples of the dataset to be assigned to the training set and the number `groups` of the percentiles to be used in identifying the subinterval of values of the response variable. In this work the proportion `p` has always been set to 0.8 (so that 80% of the samples were assigned to the training set), while the argument `groups` has been maintained to its default value of 5.

Once the test sets have been identified, the training sets have still to be subfolded in order to perform the 10-fold cross-validation. To perform this task, the `caret` package `createFolds()` function has been used. It takes as input arguments the variable containing the response values of the training set and the value `k` of folds to be created. The manual specifies that the number of groups identified by the function in order to stratify the sampling procedure is "set dynamically based on the sample size and `k`" [34].

Then the cross-validation has been applied on all the models for each city separately, in order to assess the performance in terms of MSE and R^2. For the models that contains one or more hyperparameters (see Table 2.18), a grid search has been performed in order to find the best combination of values.

At the end of the cross-validation the values of MSE and R^2 are compared for all the models and the best performer in terms of R^2 for each city is selected.
Finally, the values for MSE and R^2 are evaluated for each chosen model applied on the test set previously defined.

In order to perform the aforementioned Wilcoxon and Mann-Whitney-Wilcoxon tests for paired and non-paired samples respectively, so to compare the perfomance of two models, R provides a unique function, `wilcox.test()`.
When it is used to perform Wilcoxon test, the two samples containing the R^2 values (obtained by validating the considered models) have to be given as input together with the argument `paired=TRUE`; in order to perform Mann-Whitney-Wilcoxon test, on the contrary, the same argument must be set to `FALSE`.

2.6.5 Classification task

In order to evaluate the performance of regressors with respect to the two classes of response values identified in section 2.5.6, the predicted values of the response variable and the corresponding binarized measured ones have been provided to the function `prediction()` of the package `ROCR`: a vector of `labels` and one of `predictions` must be provided, together with a (optional) argument `label.ordering` that helps the function in recognizing the correct negative and positive labels in the first vector. If the provided `predictions` vector contains more than two values, the function automatically assumes

those values as a scoring parameter (i.e. a continuous variable).

The `prediction` object created as output by the previous function can be feeded into the `performance()` function in order to obtain a `performance` object that allows to evaluate the quality of the matching between predicted and true values.
Providing the function with the additional arguments `measure="tpr"` and `x.measure = "fpr"` allows to print the standard version of the ROC curve by simply using the `plot()` function with the `performance` object as argument.
Furthermore, the value of AUC can be obtained by calling the `performance()` function with the additional argument `measure="auc"`.

Chapter 3

Results

In this section the results of the modelling tasks are reported.
In section 3.1 the performance of the models trained on the LWD dataset (considering only the *basic* set of predictors) is shown for each city: for each kind of model the best performance (in terms of R^2) is reported, providing the values of the hyperparameters of that model and the corresponding values of *MSE* and R^2. In section 3.2 a city-wise comparison is made between best models trained on the LWD dataset and on the MI datasets (the method followed to obtain these results is described in paragraph 2.5.7): in this case, the comparison is limited to the best performing models for each city, the one trained with the LDW dataset and the one trained with MI datasets. In section 3.3 the performance of the models trained with the datasets integrated with non-meteorological predictors (as explained in paragraph 2.5.2) is presented: also in this case, the comparative presentation is limited to the best performing models. Finally, in section 3.4 the results of the classification task performed by the best models chosen in previous stages are compared.

3.1 Performance on LWD-*basic* datasets

As concerns the regression models trained with the LWD dataset considering only the *basic* set of predictors, Tables 3.1 and 3.2 presents the best performing models in the cross-validation step with respect to each city, respectively for standard linear, Ridge and Lasso regression and for tree-based models.
It can be noticed that standard and regularized linear model never exceed $R^2 = 0.481\,(0.005)$ (observed in Ferrara); in general linear regression don't show particular improvements with the addition of regularization. On the other hand, both random forest and boosting models are able to exceed that value, reaching up to $R^2 = 0.599\,(0.003)$ (for Ravenna).

In order to test the statistical significance of the differences between the best models for

Model	Linear	Ridge		Lasso	
City	Validation error	Hypar.	Validation error	Hypar.	Validation error
Piacenza	MSE = 186 (3) $R^2 = 0.367\,(0.004)$	$\lambda = 0.339$	MSE = 185 (3) $R^2 = 0.369\,(0.003)$	$\lambda = 0.0531$	MSE = 185 (3) $R^2 = 0.369\,(0.004)$
Parma	MSE = 239 (6) $R^2 = 0.369\,(0.004)$	$\lambda = 0.680$	MSE = 239 (6) $R^2 = 0.369\,(0.004)$	$\lambda = 0.0531$	MSE = 239 (6) $R^2 = 0.369\,(0.004)$
Reggio Emilia	MSE = 198 (4) $R^2 = 0.371\,(0.004)$	$\lambda = 1.080$	MSE = 198 (4) $R^2 = 0.371\,(0.004)$	$\lambda = 0.339$	MSE = 197 (4) $R^2 = 0.376\,(0.005)$
Modena	MSE = 190 (5) $R^2 = 0.454\,(0.005)$	$\lambda = 0.339$	MSE = 190 (5) $R^2 = 0.455\,(0.005)$	$\lambda = 0.0265$	MSE = 190 (5) $R^2 = 0.455\,(0.005)$
Bologna	MSE = 172 (9) $R^2 = 0.381\,(0.007)$	$\lambda = 1.362$	MSE = 172 (9) $R^2 = 0.382\,(0.007)$	$\lambda = 0.134$	MSE = 172 (9) $R^2 = 0.383\,(0.008)$
Ferrara	MSE = 204 (5) $R^2 = 0.480\,(0.005)$	$\lambda = 0.427$	MSE = 203 (5) $R^2 = 0.481\,(0.005)$	$\lambda = 0.0844$	MSE = 203 (5) $R^2 = 0.481\,(0.005)$
Ravenna	MSE = 186 (2) $R^2 = 0.4758\,(0.0019)$	$\lambda = 10^{-5}$	MSE = 149 (2) $R^2 = 0.4758\,(0.0019)$	$\lambda = 0.00164$	MSE = 149 (2) $R^2 = 0.4758\,(0.0019)$
Forlì	MSE = 137 (4) $R^2 = 0.460\,(0.005)$	$\lambda = 1.080$	MSE = 136 (4) $R^2 = 0.462\,(0.005)$	$\lambda = 0.0844$	MSE = 137 (4) $R^2 = 0.461\,(0.005)$
Cesena	MSE = 128 (4) $R^2 = 0.411\,(0.008)$	$\lambda = 0.539$	MSE = 128 (4) $R^2 = 0.412\,(0.008)$	$\lambda = 0.0421$	MSE = 128 (4) $R^2 = 0.412\,(0.008)$
Rimini	MSE = 165 (3) $R^2 = 0.449\,(0.006)$	$\lambda = 0.0105$	MSE = 165 (3) $R^2 = 0.449\,(0.006)$	$\lambda = 0.134$	MSE = 165 (3) $R^2 = 0.449\,(0.006)$

Table 3.1: Best LWD-*basic*-trained standard, Ridge- and Lasso-regularized linear regression models for each city. MSE values are provided in unit of $(\mu g/m^3)^2$.

Model	Regression tree		Random forest		Boosting	
City	Hypar.	Validation error	Hypar.	Validation error	Hypar.	Validation error
Piacenza	size = 7	MSE = 196 (4) $R^2 = 0.334\,(0.006)$	$N_{trees} = 500$ $p = 3$	MSE = 166 (4) $R^2 = 0.437\,(0.005)$	$N_{trees} = 5000$ $d = 4$, $\lambda = 10^{-2.75}$	MSE = 168 (3) $R^2 = 0.427\,(0.005)$
Parma	size = 10	MSE = 254 (8) $R^2 = 0.331\,(0.006)$	$N_{trees} = 225$ $p = 3$	MSE = 200 (4) $R^2 = 0.463\,(0.007)$	$N_{trees} = 3500$ $d = 4$, $\lambda = 10^{-2.5}$	MSE = 212 (5) $R^2 = 0.433\,(0.008)$
Reggio Emilia	size = 11	MSE = 212 (5) $R^2 = 0.328\,(0.007)$	$N_{trees} = 300$ $p = 3$	MSE = 165 (4) $R^2 = 0.478\,(0.007)$	$N_{trees} = 500$ $d = 4$, $\lambda = 10^{-1.75}$	MSE = 173 (5) $R^2 = 0.452\,(0.008)$
Modena	size = 9	MSE = 213 (5) $R^2 = 0.384\,(0.006)$	$N_{trees} = 50$ $p = 6$	MSE = 157 (4) $R^2 = 0.548\,(0.007)$	$N_{trees} = 2500$ $d = 4$, $\lambda = 10^{-2.5}$	MSE = 164 (4) $R^2 = 0.522\,(0.008)$
Bologna	size = 9	MSE = 177 (9) $R^2 = 0.359\,(0.007)$	$N_{trees} = 125$ $p = 3$	MSE = 147 (7) $R^2 = 0.467\,(0.006)$	$N_{trees} = 2000$ $d = 4$, $\lambda = 10^{-2.5}$	MSE = 149 (7) $R^2 = 0.455\,(0.006)$
Ferrara	size = 13	MSE = 231 (4) $R^2 = 0.406\,(0.004)$	$N_{trees} = 375$ $p = 3$	MSE = 160 (4) $R^2 = 0.592\,(0.007)$	$N_{trees} = 4000$ $d = 4$, $\lambda = 10^{-2.5}$	MSE = 164 (5) $R^2 = 0.580\,(0.008)$
Ravenna	size = 11	MSE = 164 (2) $R^2 = 0.420\,(0.006)$	$N_{trees} = 100$ $p = 5$	MSE = 113.3 (1.3) $R^2 = 0.599\,(0.003)$	$N_{trees} = 4000$ $d = 4$, $\lambda = 10^{-2.5}$	MSE = 117.5 (1.3) $R^2 = 0.583\,(0.004)$
Forlì	size = 12	MSE = 160 (4) $R^2 = 0.356\,(0.004)$	$N_{trees} = 150$ $p = 3$	MSE = 108 (3) $R^2 = 0.572\,(0.003)$	$N_{trees} = 3500$ $d = 4$, $\lambda = 10^{-2.5}$	MSE = 108 (3) $R^2 = 0.571\,(0.005)$
Cesena	size = 11	MSE = 139 (4) $R^2 = 0.352\,(0.007)$	$N_{trees} = 375$ $p = 3$	MSE = 97 (3) $R^2 = 0.549\,(0.007)$	$N_{trees} = 3500$ $d = 4$, $\lambda = 10^{-2.25}$	MSE = 100 (3) $R^2 = 0.533\,(0.007)$
Rimini	size = 4	MSE = 213 (3) $R^2 = 0.285\,(0.009)$	$N_{trees} = 225$ $p = 3$	MSE = 128 (2) $R^2 = 0.573\,(0.006)$	$N_{trees} = 500$ $d = 4$, $\lambda = 10^{-1.75}$	MSE = 130 (2) $R^2 = 0.564\,(0.007)$

Table 3.2: Best LWD-*basic*-trained regression tree, random forest and boosting models for each city. MSE values are provided in unit of $(\mu g/m^3)^2$.

each city, 2-sample Wilcoxon statistical tests for paired samples have been performed between all the possible pairs of best models reported in Tables 3.1 and 3.2, for each city separately. Results are reported in Table 3.3 and have to be compared with a significance level $\alpha = 0.05$.

PC	RegR	RegL	RT	RF	B
L	0.375	0.322	0.131	0.004	0.0019
RegR		0.922	0.432	0.010	0.010
RegL			0.432	0.010	0.010
RT				0.002	0.002
RF					0.695

PR	RegR	RegL	RT	RF	B
L	0.193	0.193	0.064	0.004	0.037
RegR		0.432	0.922	0.020	0.064
RegL			0.922	0.020	0.064
RT				0.002	0.002
RF					0.049

RE	RegR	RegL	RT	RF	B
L	0.846	0.922	0.027	0.002	0.020
RegR		0.375	0.557	0.020	0.084
RegL			0.492	0.020	0.084
RT				0.002	0.004
RF					0.193

MO	RegR	RegL	RT	RF	B
L	0.846	0.846	0.002	0.002	0.002
RegR		0.922	0.020	0.064	0.037
RegL			0.020	0.064	0.037
RT				0.002	0.002
RF					0.375

BO	RegR	RegL	RT	RF	B
L	0.695	0.695	0.232	0.002	0.002
RegR		0.846	0.846	0.020	0.020
RegL			0.846	0.020	0.020
RT				0.004	0.006
RF					0.375

FE	RegR	RegL	RT	RF	B
L	0.695	0.695	0.002	0.004	0.004
RegR		0.695	0.002	0.002	0.004
RegL			0.002	0.002	0.004
RT				0.002	0.002
RF					0.492

RA	RegR	RegL	RT	RF	B
L	0.846	0.846	0.020	0.002	0.002
RegR		0.922	0.020	0.002	0.002
RegL			0.020	0.002	0.002
RT				0.002	0.002
RF					0.131

FO	RegR	RegL	RT	RF	B
L	0.846	0.770	0.002	0.002	0.002
RegR		0.557	0.002	0.010	0.006
RegL			0.002	0.010	0.006
RT				0.002	0.002
RF					0.695

CE	RegR	RegL	RT	RF	B
L	0.193	0.193	0.006	0.002	0.002
RegR		1.000	0.193	0.004	0.002
RegL			0.193	0.004	0.002
RT				0.002	0.002
RF					0.232

RM	RegR	RegL	RT	RF	B
L	0.432	0.322	0.002	0.002	0.002
RegR		0.375	0.002	0.002	0.002
RegL			0.002	0.002	0.004
RT				0.002	0.002
RF					0.492

Table 3.3: Resulting p-values from 2-sample Wilcoxon statistic test for paired samples performed on R^2 values obtained for best performing LWD-*basic*-trained models. The test has been performed on all the pairs of models. The models considered are those reported on the left-hand side of Table 3.4. The distributions of R^2 values are the ones obtained in the cross-validation process.

The null hypothesis, i.e. that the distributions of R^2 values calculated for the pairs of models come from the same distribution, can be rejected on a number of cases. Noticeably, while random forest models always outperform the other models in terms of R^2, the statistical test shows that the corresponding R^2 are not significantly different from the ones obtained with boosting models in all cases with the exception of Parma (p-value $= 0.049$); furthermore, in the case of Modena they are not significantly different also from the ones obtained from the regularized regressions (Ridge and Lasso).

For the purpose of model selection, the random forest model is however chosen as the reference one for all the considered cities. Testing these models on the whole dataset (split into training and test data) has led to the results shown on the left hand side of Table 3.4: the performance in terms of R^2 has improved with respect to the validation step in 8 out of 10 cities, while it has worsened in the cases of Parma and Ferrara.

3.2 Comparison of performances on LWD-*basic* and MI-*basic* datasets

In order to evaluate the performance of the aforementioned models on the MI datasets, the cross-validation procedure has been applied also for those datasets.
Just like in the case of LWD-trained models, the best performing algorithm has been the random forest one for all the 10 cities. Right hand side of Table 3.4 shows the validation error obtained by cross-validating these models and the corresponding test errors.

City	LWD dataset - *basic* set of predictors				MI dataset - *basic* set of predictors			
	Model	Hypar.	Validation error	Test error	Model	Hypar.	Validation error	Test error
Piacenza	RF	$N_{trees} = 500$ $p = 3$	$MSE = 166\,(4)$ $R^2 = 0.437\,(0.005)$	$MSE = 174$ $R^2 = 0.449$	RF	$N_{trees} = 250$ $p = 3$	$MSE = 169\,(3)$ $R^2 = 0.430\,(0.007)$	$MSE = 139$ $R^2 = 0.456$
Parma	RF	$N_{trees} = 225$ $p = 3$	$MSE = 200\,(4)$ $R^2 = 0.463\,(0.007)$	$MSE = 238$ $R^2 = 0.439$	RF	$N_{trees} = 200$ $p = 3$	$MSE = 211\,(5)$ $R^2 = 0.453\,(0.006)$	$MSE = 171$ $R^2 = 0.426$
Reggio Emilia	RF	$N_{trees} = 300$ $p = 3$	$MSE = 165\,(4)$ $R^2 = 0.478\,(0.007)$	$MSE = 134$ $R^2 = 0.486$	RF	$N_{trees} = 400$ $p = 3$	$MSE = 160\,(4)$ $R^2 = 0.460\,(0.007)$	$MSE = 172$ $R^2 = 0.446$
Modena	RF	$N_{trees} = 50$ $p = 6$	$MSE = 157\,(4)$ $R^2 = 0.548\,(0.007)$	$MSE = 152$ $R^2 = 0.571$	RF	$N_{trees} = 450$ $p = 5$	$MSE = 160\,(4)$ $R^2 = 0.535\,(0.005)$	$MSE = 161$ $R^2 = 0.567$
Bologna	RF	$N_{trees} = 125$ $p = 3$	$MSE = 147\,(7)$ $R^2 = 0.467\,(0.006)$	$MSE = 115$ $R^2 = 0.563$	RF	$N_{trees} = 400$ $p = 4$	$MSE = 137\,(8)$ $R^2 = 0.478\,(0.010)$	$MSE = 154$ $R^2 = 0.409$
Ferrara	RF	$N_{trees} = 375$ $p = 3$	$MSE = 160\,(4)$ $R^2 = 0.592\,(0.007)$	$MSE = 167$ $R^2 = 0.574$	RF	$N_{trees} = 500$ $p = 3$	$MSE = 168\,(3)$ $R^2 = 0.568\,(0.004)$	$MSE = 162$ $R^2 = 0.569$
Ravenna	RF	$N_{trees} = 100$ $p = 5$	$MSE = 113.3\,(1.3)$ $R^2 = 0.599\,(0.003)$	$MSE = 125$ $R^2 = 0.621$	RF	$N_{trees} = 475$ $p = 3$	$MSE = 169\,(3)$ $R^2 = 0.430\,(0.007)$	$MSE = 112$ $R^2 = 0.591$
Forlì	RF	$N_{trees} = 150$ $p = 3$	$MSE = 108\,(3)$ $R^2 = 0.572\,(0.003)$	$MSE = 88$ $R^2 = 0.638$	RF	$N_{trees} = 350$ $p = 5$	$MSE = 107\,(3)$ $R^2 = 0.572\,(0.004)$	$MSE = 92$ $R^2 = 0.558$
Cesena	RF	$N_{trees} = 375$ $p = 3$	$MSE = 97\,(3)$ $R^2 = 0.549\,(0.007)$	$MSE = 97$ $R^2 = 0.562$	RF	$N_{trees} = 75$ $p = 3$	$MSE = 100.6\,(1.7)$ $R^2 = 0.506\,(0.004)$	$MSE = 96$ $R^2 = 0.567$
Rimini	RF	$N_{trees} = 225$ $p = 3$	$MSE = 128\,(2)$ $R^2 = 0.573\,(0.006)$	$MSE = 120$ $R^2 = 0.616$	RF	$N_{trees} = 275$ $p = 3$	$MSE = 129\,(3)$ $R^2 = 0.561\,(0.007)$	$MSE = 135$ $R^2 = 0.566$

Table 3.4: Performances of best LWD- and MI-trained models (basic set of predictors). MSE values are provided in unit of $(\mu g/m^3)^2$.

Concerning the statistical significance of the difference between the models shown on the right hand side of Table 3.4 and the others evaluated for each city separately, paired 2-sample Wilcoxon tests have been performed as in the previous paragraph. Keeping $\alpha = 0.05$ as significance level, the following results have been obtained (table not present):

- in the cases of Piacenza, Modena, Bologna, Ferrara, Ravenna, Forlì, Cesena and Rimini the chosen (random forest) model's performance is not significantly different from the one recorded for the best performing boosting model (p-value > 0.05);
- in the case of Parma and Reggio Emilia the chosen (random forest) model's performance is significantly different from all the others R^2 values obtained for models of different kind (p-value < 0.05 for all the tests between the random forest model's and other models' performances).

As before, random forest models whose performances have not been found significantly different from the ones of other models have been kept as "best" models for the following tasks.

Comparing the results obtained on the MI-*basic* datasets with the ones obtained for the LWD-*basic* dataset in the test step, it can be seen that the performance in terms of R^2 improves slightly on MI datasets only in one case (Cesena), while it worsens by more than 0.01 in 5 cases; in the other cases a smaller reduction of R^2 is observed.

In order to evaluate the significance of these differences, in this case a non paired 2-sample Mann-Whitney-Wilcoxon test has been performed on the pair of considered models (again on the R^2 values obtained in cross-validation). The results are reported in Table 3.5.

	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
p-values	1	0.436	0.481	0.481	0.796	0.218	0.579	0.912	0.075	0.684

Table 3.5: Resulting p-values from 2-sample Mann-Whitney-Wilcoxon statistic test for non-paired samples of R^2 values obtained for best performing models trained on the LWD-*basic* and the MI-*basic* datasets, for each city separately. The models considered are those reported in Table 3.4. The distributions of R^2 values are the ones obtained in the cross-validation process.

Taking again a significance level $\alpha = 0.05$, none of the considered pairs shows a significant difference in the performances.

3.3 Performance of models with non-meteorological predictors

Following the definition of *nonmet* and (*nonmet*+*lag*[PM_{10}]) set of predictors, cross-validation has been applied to the corresponding LWD datasets in order to evaluate the contribution of the added predictors to the overall performance.
Table 3.6 reports, for each city, the best performing models for the two considered datasets.

Comparing the test performances of the best models trained on the *basic* and those trained on the *nonmet* dataset, only in the case of Bologna an improvement can be observed; in 8 cities out of 10 the addition of time-related predictors has worsened the performance in terms of R^2 by more than 0.01.
Concerning the models trained on the (*nonmet*+*lag*[PM_{10}]) datasets, they seem to perform better in the test step than both *basic*- and *nonmet*-trained models with improve-

City	LWD dataset - *nonmet* set of predictors				LWD dataset - (*nonmet*+*lag*[PM_{10}]) set of predictors			
	Model	Hypar.	Validation error	Test error	Model	Hypar.	Validation error	Test error
Piacenza	B	$N_{trees} = 1500$ $d = 3$, $\lambda = 10^{-2}$	$MSE = 164\,(3)$ $R^2 = 0.442\,(0.003)$	$MSE = 183$ $R^2 = 0.422$	B	$N_{trees} = 2000$ $d = 4$, $\lambda = 10^{-2}$	$MSE = 87.5\,(0.8)$ $R^2 = 0.699\,(0.006)$	$MSE = 97.3$ $R^2 = 0.688$
Parma	RF	$N_{trees} = 200$ $p = 4$	$MSE = 198\,(5)$ $R^2 = 0.471\,(0.008)$	$MSE = 250$ $R^2 = 0.412$	B	$N_{trees} = 500$ $d = 4$, $\lambda = 10^{-1.5}$	$MSE = 122\,(2)$ $R^2 = 0.673\,(0.006)$	$MSE = 122$ $R^2 = 0.722$
Reggio Emilia	RF	$N_{trees} = 100$ $p = 4$	$MSE = 162\,(4)$ $R^2 = 0.488\,(0.006)$	$MSE = 145$ $R^2 = 0.443$	B	$N_{trees} = 2000$ $d = 3$, $\lambda = 10^{-1.75}$	$MSE = 86\,(2)$ $R^2 = 0.712\,(0.006)$	$MSE = 97$ $R^2 = 0.648$
Modena	RF	$N_{trees} = 225$ $p = 4$	$MSE = 157\,(4)$ $R^2 = 0.547\,(0.007)$	$MSE = 164$ $R^2 = 0.536$	RF	$N_{trees} = 225$ $p = 6$	$MSE = 90\,(3)$ $R^2 = 0.746\,(0.003)$	$MSE = 103$ $R^2 = 0.703$
Bologna	RF	$N_{trees} = 50$ $p = 8$	$MSE = 148\,(8)$ $R^2 = 0.468\,(0.008)$	$MSE = 106$ $R^2 = 0.596$	RF	$N_{trees} = 75$ $p = 5$	$MSE = 102\,(6)$ $R^2 = 0.641\,(0.006)$	$MSE = 83$ $R^2 = 0.612$
Ferrara	RF	$N_{trees} = 150$ $p = 5$	$MSE = 158\,(5)$ $R^2 = 0.597\,(0.007)$	$MSE = 188$ $R^2 = 0.520$	RF	$N_{trees} = 225$ $p = 6$	$MSE = 94.1\,(1.4)$ $R^2 = 0.745\,(0.005)$	$MSE = 109.7$ $R^2 = 0.757$
Ravenna	RF	$N_{trees} = 250$ $p = 6$	$MSE = 113.3\,(1.2)$ $R^2 = 0.600\,(0.003)$	$MSE = 126$ $R^2 = 0.615$	B	$N_{trees} = 1500$ $d = 4$, $\lambda = 10^{-2.25}$	$MSE = 71.7\,(1.1)$ $R^2 = 0.749\,(0.004)$	$MSE = 68.7$ $R^2 = 0.759$
Forlì	B	$N_{trees} = 2000$ $d = 4$, $\lambda = 10^{-2.25}$	$MSE = 105\,(3)$ $R^2 = 0.579\,(0.004)$	$MSE = 91$ $R^2 = 0.627$	RF	$N_{trees} = 225$ $p = 6$	$MSE = 71\,(2)$ $R^2 = 0.721\,(0.006)$	$MSE = 77$ $R^2 = 0.671$
Cesena	RF	$N_{trees} = 75$ $p = 5$	$MSE = 97\,(3)$ $R^2 = 0.551\,(0.006)$	$MSE = 111$ $R^2 = 0.497$	RF	$N_{trees} = 150$ $p = 6$	$MSE = 63.2\,(1.3)$ $R^2 = 0.690\,(0.004)$	$MSE = 82.4$ $R^2 = 0.687$
Rimini	RF	$N_{trees} = 100$ $p = 3$	$MSE = 126\,(2)$ $R^2 = 0.578\,(0.005)$	$MSE = 142$ $R^2 = 0.546$	B	$N_{trees} = 2000$ $d = 4$, $\lambda = 10^{-2.25}$	$MSE = 83\,(2)$ $R^2 = 0.726\,(0.006)$	$MSE = 89$ $R^2 = 0.687$

Table 3.6: Performances of best LWD-*nonmet*- and LWD-(*nonmet*+*lag*[PM_{10}])-trained models. MSE values are provided in unit of $(\mu g/m^3)^2$.

ments in the performance larger than 0.01 in terms of R^2 for all the considered cities.

In order to assess the significance of these variations in the models' performances, 2-sample Wilcoxon statistic tests have been performed on the three pairs of selected model for each city (respectively cross-validated on the LWD-*basic*, LWD-*nonmet* and LWD-(*nonmet*+*lag*[PM_{10}]) datasets). The results are reported in Table 3.7.
The values reported testify that the null hypothesis can be rejected (p-value $< \alpha = 0.05$) for at least two cases per city: in particular, it can be seen that the (*nonmet*+*lag*[PM_{10}]) set of predictor gives always significantly different results from the *basic* one and the *nonmet* one; on the other hand, the *nonmet* dataset provides significantly different performances from the *basic* one in the cases of Modena, Bologna, Ferrara and Cesena.

A useful comparison can be made between the (*nonmet*+*lag*[PM_{10}])-trained models and a simple *persistence* model that takes into account only the information on the persistency of PM_{10} in the atmosphere. Such a model can be obtained by predicting a value of the response variable $y(t)$ equal to the previous value in the time series, i.e. in this case choosing the measured value of PM_{10} on the previous day as the prediction of the same variable for the day of interest; being t the variable that identifies the days in the time series, the relationship becomes:

$$\hat{y}(t) = y(t-1) \tag{3.1}$$

Following its definition, the best estimate for the performance of the model in terms of R^2 corresponds to the squared value of the temporal autocorrelation value (calculated using

	Piacenza	Parma	Reggio Emilia	Modena	Bologna
basic vs *nonmet*	0.922	0.375	0.695	0.002	0.002
basic vs (*nonmet*+lag)	0.004	0.004	0.002	0.002	0.002
nonmet vs (*nonmet*+lag)	0.002	0.004	0.002	0.049	0.037
	Ferrara	Ravenna	Forlì	Cesena	Rimini
basic vs *nonmet*	0.004	0.492	0.275	0.004	0.695
basic vs (*nonmet*+lag)	0.004	0.002	0.004	0.002	0.010
nonmet vs (*nonmet*+lag)	0.009	0.002	0.004	0.049	0.014

Table 3.7: Resulting p-values from 2-sample Wilcoxon statistic test for paired samples of R^2 values obtained for best performing models trained on the LWD-*basic*, the LWD-*nonmet* and the LWD-(*nonmet*+*lag*[PM_{10}]) datasets, for each city separately. The models considered are those reported in Tables 3.4 (left-hand side) and 3.6. The distributions of R^2 values are the ones obtained in the cross-validation process.

Equation 2.7) on the series of considered values setting the lag $k = 1$, i.e. $R^2 = (r_1)^2$. The values of the autocorrelation are reported in the first row of Table 2.6, while Table 3.8 shows the aforementioned squared values for the 10 cities.

Variable	Persistence model									
	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
R^2	0.586	0.563	0.544	0.641	0.572	0.643	0.582	0.604	0.580	0.600

Table 3.8: Performance of persistence models.

Comparing them with the test R^2 values obtained for the (*nonmet*+*lag*[PM_{10}])-trained models shows that the latter ones have always a better performance. A statistical test is considered unnecessary in this case.

3.4 Comparing models for classification tasks

As outlined in section 2.5.6, the outcomes of the test tasks on the four groups of city-level trained models make it possible to evaluate the performance of the same models in the aforementioned classification task of predicting a response value that falls into the same class of the true value.
Evaluating the AUC of ROC values for the considered models has led to the values reported in Table 3.9.

It can be seen that the best models in the regression task (the ones trained on the LWD-(*nonmet*+*lag*[PM_{10}]) dataset) are also the best ones in the classification task. It is interesting to notice that, depending on the city, the second best performer does not

	Piacenza	Parma	Reggio Emilia	Modena	Bologna	Ferrara	Ravenna	Forlì	Cesena	Rimini
LWD-*basic*	0.902	0.861	0.851	0.893	0.863	0.909	0.923	0.957	0.930	0.953
MI-*basic*	0.861	0.859	0.849	0.917	0.888	0.944	0.935	0.906	0.941	0.913
LWD-*nonmet*	0.880	0.859	0.837	0.899	0.893	0.899	0.925	0.950	0.922	0.952
LWD-(*nonmet*+lag)	0.957	0.960	0.930	0.955	0.943	0.980	0.970	0.935	0.965	0.954

Table 3.9: Resulting AUROC values obtained for best performing models on the LWD-*basic*, the MI-*basic*, the LWD-*nonmet* and the LWD-(*nonmet*+*lag*[PM_{10}]) datasets, for each city separately. The models considered are those reported in Tables 3.4 and 3.6.

necessarily belong to a specific group of models. However, a good prediction quality is observed in all the considered cases.

Chapter 4

Conclusions

In this work an analysis of the relationship between meteorology-related variables and PM_{10} concentration levels in the capitals of the provinces of Emilia-Romagna has been performed in order to understand how the meteorological conditions affect PM_{10} concentration. The considered meteorological variables have been subsequently input as predictors to statistical regression models based on machine learning in order to obtain predictions for the daily mean value of PM_{10} concentration.
A dataset containing time series of daily values of 10 meteorological variables and those of PM_{10} urban background concentration for the 10 cities, spanning a time interval of 2008 days, has been created. It has been subjected to an exploratory data analysis that has allowed to point out the main features of each variable and its relationship with PM_{10} concentration, evaluated on a daily basis. This analysis has showed that ranked correlations between meteorological variables (considered separately) and PM_{10} concentration can not be considered negligible in most cases; in one case (that of precipitation), correlation between previous-day intensity and PM_{10} concentration is stronger than that obtained with same-day intensity.

After being preprocessed, data have been used to train regression models with the aim of predicting PM_{10} daily mean concentration values starting from the same-day values of the meteorological variables. All the considered models, which include standard and regularized linear regressions and regression tree-based ones, have been trained separately with the data of each city, in order to reproduce specifically the patterns observed at a local level.

The cross-validation and testing tasks have shown that random forest and boosting models provide better performances than the other models that have been considered. Almost always, however, the difference in the performance of this two models is not significant.

The use of multiple imputation in order to address the problem of missing values in the

original dataset has not produced significant differences in the performance, although a slight worsening has been observed for the test error in some cases.

Concerning the addition on non-meteorological variables, it has been proved that time-related descriptors are able to significantly improve the performance for less than half the considered cities.
On the other hand, adding the previous-day PM_{10} mean concentration as a predictor gives always a significantly better performance with respect to the baseline set of predictors and the one integrated with time-related descriptors.
Given this larger set of predictors, the best models also outperforms a simple persistence model based only on the previous-day PM_{10} mean concentration.

Finally, an assessment on the ability of assigning each predicted response to the correct side of the threshold of $50\,\mu g/m^3$ has been performed on the best models for each of the previously identified training groups.
The results from this task confirmed the conclusions on the better modelling capacity of the models that have better performed in regression, even if good results have been obtained for all the considered models.

4.1 Further developments

At the end of the present work, a number of proposals can be made in order to improve the quality of the modelling process and get better results in the task of value prediction.

The first aspect concerns the **temporal window** in which the predictors are picked. The inclusion of the previous-day PM_{10} concentration has provided a significant improvements and it is possible that the inclusion of other meteorological variables measured on the previous day can improve as well the performances. Also a widening of the temporal windows to more than 1 day before the date of interest could further enhance the quality of the prediction.
Still concerning the temporal dimension, it could be interesting to group samples by the year in which they have been taken and to evaluate the contribution that each group gives to the model, e.g. by training the model with different combinations of these groups. If set appropriately, this analysis can offer an understanding of how a variation of the distribution of meteorological variables on a long timescale (e.g. because of climate change) influence the measured levels of PM_{10} pollution. A different analysis could similarly assess (in a more indirect way) how the changes in the anthropogenic emissions over time affect PM_{10} concentration levels.

As regards the **tipologies of models**, this work has been necessarily limited to a number of regression algorithms. The choice of the models (some of which are characterized by a low complexity) was related to the purpose of building a relatively simple model that can be eventually considered for concrete applications (such as the tree model used by ARPAE for the classification task, as said in section 1): nonetheless, the best performances have been attained by the most complex models, i.e. random forest and boosting ones. Further analysis can consider other models such as neural networks, that could be usefully applied also for tasks involving the temporal dimension as said above.

Finally, since this work only considered each city separately, it could be useful to reproduce this analysis considering all the cities at the same time in order to understand if a **regional-level model** would be able to achieve the same quality of prediction.

Bibliography

[1] Agenzia Regionale per la Protezione dell'Ambiente e l'Energia dell'Emilia-Romagna, 2018, *La qualità dell'aria in Emilia-Romagna. Edizione 2018* (https://www.arpae.it/cms3/documenti/aria/rapporto_finale_inventario_emissioni_2015.pdf

[2] Agenzia Regionale per la Protezione dell'Ambiente e l'Energia dell'Emilia-Romagna, *Rete di monitoraggio della qualità dell´aria* (https://www.arpae.it/dettaglio_generale.asp?id=2892&idlivello=846; last visited: 06/07/2019)

[3] Agenzia Regionale per la Protezione dell'Ambiente e l'Energia dell'Emilia-Romagna, Centro tematico regionale Qualità dell'aria, 2019, *Aggiornamento dell'inventario regionale delle emissioni in atmosfera dell'Emilia-Romagna relativo all'anno 2015 (INEMAR-ER 2015)* (https://www.arpae.it/cms3/documenti/aria/rapporto_finale_inventario_emissioni_2015.pdf

[4] Giovanni Bonafè, ARPA Emilia-Romagna, Servizio IdroMeteoClima, 2011, *Indicatori meteo a supporto delle valutazioni di qualità dell'aria* (https://www.arpae.it/cms3/documenti/simc/meteoambiente/scheda-indicatori-meteo.pdf)

[5] L. Bai, J. Wang, X. Ma, H. Lu, 2018, *Air Pollution Forecasts: An Overview*, International journal of environmental research and public health, 15, 4, 780 (https://doi.org/10.3390/ijerph15040780; last visited: 20/11/2019).

[6] W.R. Burrows, M. Benjamin, S. Beauchamp, E.R. Lord, D. McCollor, and B. Thomson, 1995, *CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada*, J. Appl. Meteor., 34, 1848–1862 (https://doi.org/10.1175/1520-0450(1995)034<1848:CDTSAA>2.0.CO;2; last visited: 20/11/2019).

[7] A. Chaloulakou, D. Assimacopoulos, T. Lekkas, 1999, *Forecasting Daily Maximum Ozone Concentrations in the Athens Basin*, Environmental Monitoring and Assessment, 56, 1, 97-112 (https://doi.org/10.1023/A:1005943201063; last visited: 20/11/2019).

[8] A. Chaloulakou, P. Kassomenos, N. Spyrellis, P. Demokritou, P. Koutrakis, 2003, *Measurements of PM10 and PM2.5 particle concentrations in Athens, Greece*, Atmospheric Environment, 37, 5, 649-660, (https://doi.org/10.1016/S1352-2310(02)00898-1; last visited: 20/11/2019).

[9] A. Chaloulakou and G. Grivas and N. Spyrellis, 2003, *Neural Network and Multiple Regression Models for PM10 Prediction in Athens: A Comparative Assessment*, Journal of the Air & Waste Management Association, 53, 10, 1183-1190 (https://doi.org/10.1080/10473289.2003.10466276; last visited: 20/11/2019).

[10] A. Chaloulakou, M. Saisana, N. Spyrellis, 2003, *Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens*, Science of The Total Environment, 313, 1, 1-13 (https://doi.org/10.1016/S0048-9697(03)00335-8; last visited 20/11/2019).

[11] A. C. Comrie, 1997, *Comparing Neural Networks and Regression Models for Ozone Forecasting*, Journal of the Air & Waste Management Association, 47, 6, 653-663 (https://doi.org/10.1080/10473289.1997.10463925; last visited: 20/11/2019)

[12] Frank De Leeuw, 2012, *Limit values & Particulate Matter*, PM workshop, Brussels, 18-19 June 2012 (https://circabc.europa.eu/webdav/CircaBC/env/ambient/Library/extension_notifications/workshop_18-19_2012/presentations_day/deLeeuw_limitvalues.pdf)

[13] J. Demšar, 2006, *Statistical Comparisons of Classifiers over Multiple Data Sets*, The Journal of Machine Learning Research, 7, 1-30 (https://dl.acm.org/doi/10.5555/1248547.1248548; last visited: 01/02/2020)

[14] European Commission, 2017, *Special Eurobarometer 468: Attitudes of European citizens towards the environment* (http://data.europa.eu/euodp/en/data/dataset/S2156_88_1_468_ENG)

[15] European Commission, 2018, *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions — 'A Europe that protects: Clean air for all'* (COM(2018) 330 final) (http://ec.europa.eu/environment/air/pdf/clean_air_for_all.pdf)

[16] European Environment Agency, 2018, *Air quality in Europe — 2018 report*, EEA Report N.12/2018 (https://www.eea.europa.eu/publications/air-quality-in-europe-2018)

[17] *Directive 2004/107/EC of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air* (OJ L 23, 26.1.2005, pp. 3-16) (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2005:023:0003:0016:EN:PDF)

[18] *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe* (OJ L 152, 11.6.2008, p. 1-44) (http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:152:0001:0044:EN:PDF)

[19] T. Fawcett, 2006, *An introduction to ROC analysis*, Pattern Recognition Letters, 27, 8, 861-874 (https://doi.org/10.1016/j.patrec.2005.10.010; last visited: 01/02/2020

[20] A. Feelders, 1999, *Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation?*, Principles of Data Mining and Knowledge Discovery, 329-334, Springer Berlin Heidelberg (https://doi.org/10.1007/978-3-540-48247-5_38; last visited: 01/02/2020)

[21] Y. Feng, W. Zhang, D. Sun, L. Zhang, 2011, *Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification*, Atmospheric Environment, 45, 11, 1979-1985 (https://doi.org/10.1016/j.atmosenv.2011.01.022; last visited: 20/11/2019).

[22] H.J.S. Fernando, M.C. Mammarella, G. Grandoni, P. Fedele, R. Di Marco, R. Dimitrova, P. Hyde, 2012, *Forecasting PM10 in metropolitan areas: Efficacy of neural networks*, Environmental Pollution 163, 62-67 (https://doi.org/10.1016/j.envpol.2011.12.018; last visited: 20/11/2019).

[23] M.W. Gardner, S.R. Dorling, 1999, *Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London*, Atmospheric Environment, 33, 5,709-719 (https://doi.org/10.1016/S1352-2310(98)00230-1; last visited: 20/11/2019).

[24] M.W. Gardner, S.R. Dorling, 2000, *Statistical surface ozone models: an improved methodology to account for non-linear behaviour*, Atmospheric Environment, 34, 1,21-34 (https://doi.org/10.1016/S1352-2310(99)00359-3; last visited: 20/11/2019).

[25] G. Grivas, A. Chaloulakou, 2006, *Artificial neural network models for prediction of PM10 hourly concentrations, in the Greater Area of Athens, Greece*, Atmospheric Environment, 40, 7, 1216-1229 (https://doi.org/10.1016/j.atmosenv.2005.10.036; last visited: 20/11/2019).

[26] IARC, 2013, *Outdoor air pollution a leading environmental cause of cancer deaths*, Press release No 221, International Agency for Research on Cancer (http://www.iarc.fr/en/media-centre/iarcnews/pdf/pr221_E.pdf)

[27] G. James, D. Witten, T. Hastie, R. Tibshirani, 2013, *An Introduction to Statistical Learning Book with Applications in R*, Springer-Verlag New York (https://doi.org/10.1007/978-1-4614-7138-7; last visited: 20/11/2019).

[28] Knorr W. et al., 'Wildfire air pollution hazard during the 21st century', 2017, Atmospheric Chemistry and Physics 17, pp. 9223-9236

[29] A. Kumar and P. Goyal, 2011, *Forecasting of air quality in Delhi using principal component regression technique*, Atmospheric Pollution Research, 2, 4, 436-444 (https://doi.org/10.5094/APR.2011.050; last visited: 20/11/2019).

[30] A. Kurt, A. B. Oktay, 2010, *Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks*, Expert Systems with Applications, 37, 12, 7986-7992 (https://doi.org/10.1016/j.eswa.2010.05.093; last visited: 20/11/2019).

[31] Jos Lelieveld, Klaus Klingmüller, Andrea Pozzer, Ulrich Pöschl, Mohammed Fnais, Andreas Daiber, Thomas Münzel, *Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions*, European Heart Journal, Volume 40, Issue 20, 21 May 2019, Pages 1590–1596

[32] D. Melas, I. Kioutsioukis, I.C. Ziomas, 2000, *Neural network model for predicting peak photochemical pollutant levels*, Journal of the Air & Waste Management Association, 50, 4, 495-501 (https://doi.org/10.1080/10473289.2000.10464039; last visited: 20/11/2019).

[33] Querol X. et al., 2004, *Speciation and origin of PM10 and PM2.5 in selected European cities*, Atmospheric Environment, 38:6547–6555.

[34] *createDataPartition function — R Documentation* on R Documentation website (https://www.rdocumentation.org/packages/caret/versions/6.0-85/topics/createDataPartition; last visited: 01/02/2020).

[35] H. Taheri Shahraiyni, S. Sodoudi, 2016, *Statistical Modeling Approaches for PM10 Prediction in Urban Areas; A Review of 21st-Century Studies*, Atmosphere, 7, 15 (https://doi.org/10.3390/atmos7020015; last visited: 20/11/2019)

[36] S. van Buuren, K. Groothuis-Oudshoorn, 2011, *mice: Multivariate Imputation by Chained Equations in R*, Journal of Statistical Software, 45, 3, 1-67 (https://www.jstatsoft.org/v045/i03; last visited: 30/12/2019)

[37] S. van Buuren, 2018, *Flexible imputation of missing data*, Chapman & Hall/CRC Interdisciplinary Statistics Series (https://stefvanbuuren.name/fimd/; last visited: 30/12/2019)

[38] S. Weichenthal, T. Olaniyan, T. Christidis, E. Lavigne, M. Hatzopoulou, K. Van Ryswyk, M. Tjepkema, R. Burnett, 2020, *Within-city Spatial Variations in Ambient Ultrafine Particle Concentrations and Incident Brain Tumors in Adults*, Epidemiology, 31, 2, 177-183 (https://doi.org/10.1097/EDE.0000000000001137; last visited 27/02/2020)

[39] World Health Organization, Regional Office for Europe, 2000, *Air quality guidelines for Europe*, Copenhagen (http://www.euro.who.int/__data/assets/pdf_file/0005/74732/E71922.pdf)

[40] World Health Organization, Regional Office for Europe, 2006, *Air quality guidelines: global update 2005 — particulate matter, ozone, nitrogen dioxide and sulphur dioxide*, Copenhagen.

[41] Daniel S. Wilks, 2011, *Statistical Methods in the Atmospheric Sciences*, Third Edition, International Geophysics series, Elsevier

[42] J. Yi, V. R. Prybutok, 1996, *A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area*, Environmental Pollution, 92, 3, 349-357 (https://doi.org/10.1016/0269-7491(95)00078-X; last visited: 20/11/2019).

[43] I. C. Ziomas, D. Melas, C. S. Zerefos, A. F. Bais, A. G. Paliatsos, 1995, *Forecasting peak pollutant levels from meteorological variables*, Atmospheric Environment 29, 24, 3703-3711 (https://doi.org/10.1016/1352-2310(95)00131-H; last visited: 20/11/2019).