UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Master Degree in Physics of Data

Final Dissertation

Combination techniques of monitoring data and model estimations of air pollutants concentration in the Veneto region

Thesis supervisor

Prof. Alberto Garfagnini

Thesis co-supervisor

Dr. Alberto Dalla Fontana

Candidate

Lorenzo Buriola

Academic Year 2022/2023

Abstract

Air pollution is a crucial environmental parameter related to an alteration in the chemical composition of air. High levels of air pollutants can be dangerous for health and damage the environment, biological resources and ecosystems. Measuring such concentration is an important task, especially in regions where, for morphological reasons, pollutants tend to persist for more time. The Po Valley is a highly industrialized and densely populated region surrounded by mountains that do not favour enough ventilation. Here the air pollutants concentrations can reach the legal limits set by the European Union. In Veneto, ARPAV (Regional Environmental Protection Agency Veneto) is the agency responsible for air quality control. They monitor it using a network of stations distributed over the Regional territory which measure the concentration of the principal pollutants (PM_{10}, $PM_{2.5}$, NO_x, O_3). A deterministic eulerian Chemical Transport Model (CTM) is also used to have better estimations of the whole territory (even far from stations) and predictions of the concentrations for the upcoming days. The goal of this work is to apply and evaluate some "data fusion" methods, statistical-based, commonly used to merge model estimations and station measurements; these techniques allow to improve the estimate of pollutants concentration in the model domain and are thus of paramount importance for ARPAV.

After an introduction to the context where the problem arises, the work will focus on the data and statistics used for the analysis. Firstly, an overview of the air pollutant concentrations data (PM_{10}, NO_x and O_3) for the Veneto region is presented, describing how the data are gathered and organized. Afterwards, the deterministic model is also briefly described, with a particular interest in the verification with respect to the measurements. The main part of the work is an in-depth analysis of the interpolation methods used for spatial prediction to correct the model predictions using station measurements. Finally, the results of the analysis are discussed.

Contents

Chapter 1

Introduction

This work started following my training period with Dr. Alberto Dalla Fontana of ARPAV-UQA (Agenzia Regionale per la Protezione e Prevenzione Ambientale del Veneto - Unità Operativa Qualità dell'Aria) in the spring of 2022. The tasks of the unit relevant to this work are ([1]):

- the implementation and maintenance of air quality diagnostic and prognostic modelling tools;
- usage of dispersion models at different spatial scales;
- management the regional inventory of atmospheric emissions;
- management of the air quality stations network.

Air quality is a crucial environmental theme for the European Union; It is highly regulated from the legal point of view with specific fixed limits, assessment methods and goals. Multiple techniques were developed to measure the concentration of air pollutants and deterministic models are also exploited to forecast air quality for upcoming days.

Air pollution is the contamination of air due to an alteration in the composition. Some substances can modify the quality of the air, resulting in a potential threat to humans and ecosystems. The legal directive for what concerns air quality in Italy is the Decreto Legislativo n. 155/2010 [2] that implements the European Directive 2008/50/CE [3]. The Ambient Air Quality Directives by the European Union set air quality standards for the most important pollutants and gives also a series of indications for the assessment of air quality. They define common methods to monitor the concentration of such pollutants and rules to inform the public. They also require each Country to prepare an air quality plan to address the action to ensure compliance with AQ (air quality) Limits. Air quality is an important issue in the Po Valley. The presence of the Alps and the Appennini causes the pollutants to be trapped inside the region instead of dispersing. This phenomenon periodically leads to the exceedances of AQ limits (in particular PM_{10}) and can force local authorities to take measures to reduce emissions. The restrictions target traffic, forbidding the circulation of the most polluting vehicles, domestic heating and agricultural practices. Therefore, reliable measurements and predictions of pollutant concentration are of paramount importance.

ARPAV (Agenzia Regionale Protezione e Prevenzione Ambientale del Veneto) is the Veneto regional agency, established in 1997, whose competencies are the protection of the environment and prevention of environmental pollution. Like the other ARPAs, it operates through controlling and monitoring activities aimed to ensure compliance with environmental limits. Objects of surveillance are :

- Water quality
- Physical agents (radiation, noise pollution)
- Air quality
- Soil

- Weather (weather forecasts and ideological risk)
- Waste
- Sky (light pollution)

Air is routinely monitored by a network of air quality stations distributed in the Regional territory that measure the concentration of the main pollutants. Aside from the stationary stations, measurement campaigns with mobile labs are conducted. Furthermore, many ARPAs have implemented a chemical transport model to improve pollutant estimations in the whole territory (even far from stations) and predict the concentration for the upcoming days. Another task, that ARPAs are in charge of is the emissions inventory: from direct measurements of emissions and through emissions estimation techniques a detailed inventory is created and periodically updated. In winter an air pollution bulletin ("Bollettino PM10"), based on model predictions and measurements, is also regularly issued to the public and local authorities [4].

1.1 A brief introduction to air pollutants

Many pollutants can be present in the air and only a few of them are regulated by law. They can be divided into two categories: particulate matter and gaseous pollutants. Here we will present an introduction to the major pollutants that are considered in this work.

Particulate Matter

Particulate matter (PM) is composed of heterogeneous substances, a mixture of solid particles and droplets. There exist two origins of particulate matter: some particles are emitted directly from polluting sources such as biomass combustion for home heating, traffic and agricultural activities, and others are produced in the atmosphere as a result of chemical reactions. Given these sources of emissions and the frequent stagnant meteorological conditions, winter is the period when the concentration of PM is higher.

Particles with a diameter smaller than 10 μm are classified as PM_{10} and can be transported by winds and remain in the atmosphere for some days and move for hundreds of kilometres. Particles with a diameter smaller than 2.5 μm are classified as $PM_{2.5}$.

For what concerns the health risk of particulate matter, the problem is due to the inhalable particles that settle inside the respiratory organs causing inflammations and neoplasms. ([1])

The legal limits for PM_{10} are set for both annual and daily averages. The yearly average cannot exceed 40 $\mu g/m^3$ and the daily concentration must not exceed 50 $\mu g/m^3$ for more than 35 days in a year. For $PM_{2.5}$, a target value of 25 $\mu g/m^3$ is set for the yearly mean. ([2], [5], [6]) In this work, we will concentrate only on PM_{10} measurements.

Nitric Dioxide

Nitric oxides are common gaseous pollutants in the atmosphere and can be found in two major components: the monoxide (NO) and the dioxide (NO_2). Emissions of nitric oxides are composed mainly of monoxide but, in the atmosphere, most of NO is oxidized into dioxide. For this reason in this work, only the dioxide will be taken into consideration. This pollutant is concentrated mainly near streets and other areas of traffic. But it is produced also as a result of home heating, thermoelectric power stations and other combustion processes.

Like PM_{10}, nitric dioxide reaches high levels during winter as well, but it has an important role also during summer when, due to intense UV radiation, NO_2 takes part in the formation of ozone. Nitric dioxide is a toxic gas that can produce inflammation and lung issues. It is particularly dangerous for children and people with respiratory problems.

The legal limits to the concentration of NO_2 are established for annual and hourly averages. The yearly average should not overcome 40 $\mu g/m^3$ while the hourly average must not exceed 200 $\mu g/m^3$ for more than 18 times in a year. ([2], [7], [6])

Ozone

Ozone is a gaseous pollutant that is not directly emitted by human activities. Its concentration is correlated to the presence of heat and sun radiation, with peaks during the summer days. It forms in the atmosphere through chemical reactions that involve other pollutants such as NO_x and hydrocarbons. The ozone can, as well, damage the respiratory organs causing inflammations.

Also for Ozone, the law sets specific thresholds to be respected for hourly averages. There exists an information threshold set to $180\mu g/m^3$ and an alarm threshold of $240\mu g/m^3$

Pollutant	**Limits**	**Source**	**Critic Period**
PM_{10}	Yearly average: 40 $\mu g/m^3$ Daily concentration: 50 $\mu g/m^3$ for more than 35 days in a year	Biomass combustion Domestic heating Traffic Agriculture	Winter
NO_2	Yearly average: $40\mu g/m^3$ Hourly concentration: $200\mu g/m^3$ for more than 18 times in a year	Combustion Domestic heating Traffic Thermoelectric centrals	Winter
O_3	Hourly average: $240\mu g/m^3$	Photochemical reactions in atmosphere	Summer

Table 1.1: Summarising table of pollutants

1.2 Motivation

The goal of this work is to improve the estimations of pollutant concentration in the sites where measurements are not available. Air quality measurements rely on the network of stations that will be described more in detail in Sec. 2.1. Those stations provide accurate results for their specific location, but unfortunately, they are limited in number and, although they are located in strategic locations, near cities and well distributed, they cannot cover the entire territory. The deterministic Eulerian Chemical Transport Model (CTM) (Sec. 2.2) used by ARPAV can provide reliable estimations even far from stations, however, due to intrinsic limitations in the model formulation and inaccuracies in its inputs, the computed concentrations are inevitably different from the measurements. Implementing and testing a method that improves model output using measurements is a valuable procedure to get more reliable estimations in the whole territory. As mentioned before, the Po Valley is a critical area concerning air pollution, therefore any improvement in pollutant concentration estimation is very important not only from the scientific point of view but also because it has a large impact on people's quality of life and health.

Chapter 2

Data Collection

The analysis that will be presented in this work uses data collected by ARPAV. More specifically, we will consider data retrieved during the most polluted period, depending on the specific pollutant. For what concerns particulate matter (PM_{10}) and nitric dioxide, the time interval goes from December 2021 to February 2022, during the winter months. On the contrary, for the ozone, the summer months are considered, from June 2021 to August 2021. As mentioned above two kinds of data are used for the analysis: measurements and model predictions.

2.1 Stations

The network of stations in the territory of Veneto has been built and adapted accordingly to the directives of the Italian Decreto Legislativo 155/2010 ([2]). At present, it counts 43 stations. In this work, we are going to consider only the background stations, which are defined in the Decreto as "stations located in such a way that the level of pollution is not influenced by a specific source (such as traffic, industries or home heating), but by an integrated combination of all sources near the station". There are 22 background stations and their features and locations are shown in Tab. 2.1 and Fig. 2.1. Note that the stations in Bassano (VI), Asiago (VI) and S.Giustina (PD) do not have an automatic instrument for PM_{10}; on the other hand, from the station in Monselice (PD) only data relative to particulate matter are retrieved.

ISTAT	x [km]	y [km]	Stazione	Cod_EOI	h [msl]
025006	748.507	5114.854	BL_belluno	IT1594A	378
025021	724.939	5101.629	BL_feltre	IT1619A	356
025038	759.503	5117.598	BL_pieve	IT1790A	690
028031	707.222	5018.483	PD_p_colli	IT1870A	12
028080	726.438	5053.879	PD_sgiust	IT2071A	24
028037	709.304	5011.618	PD_este	IT1871A	10
029004	700.881	4997.683	RO_badia	IT2072A	0
029001	741.051	4992.600	RO_adria	IT1213A	0
029041	719.758	4991.044	RO_borsea	IT1214A	0
026021	756.582	5087.116	TV_conegliano	IT1328A	61
026037	772.611	5081.918	TV_mansue	IT1596A	8
026086	752.186	5062.675	TV_lancieri	IT1590A	15
027033	779.883	5059.112	VE_san_dona	IT1222A	0
027042	754.790	5043.629	VE_bissuola	IT0963A	0
024009	699.424	5080.452	VI_asiago	IT1791A	1366
024012	712.758	5070.942	VI_bassano	IT1065A	114
024100	684.276	5064.949	VI_schio	IT0663A	190
024116	698.127	5048.248	VI_qitalia	IT1177A	36
023011	658.958	5050.610	VR_bcnuova	IT1848A	814
023044	681.524	5005.835	VR_legnago	IT1535A	15
023091	658.792	5033.059	VR_giarol	IT1343A	48
028060	722.451	5028.089	PD_mandria	IT1453A	9
028055	715.810	5013.200	PD_Monselice	99910	10

Table 2.1: Stations data: for each station, we report the unique ISTAT code and unique European EOI code, coordinates and altitude. Note that coordinates are given following the UTM system (zone 32N, EPSG:32632)

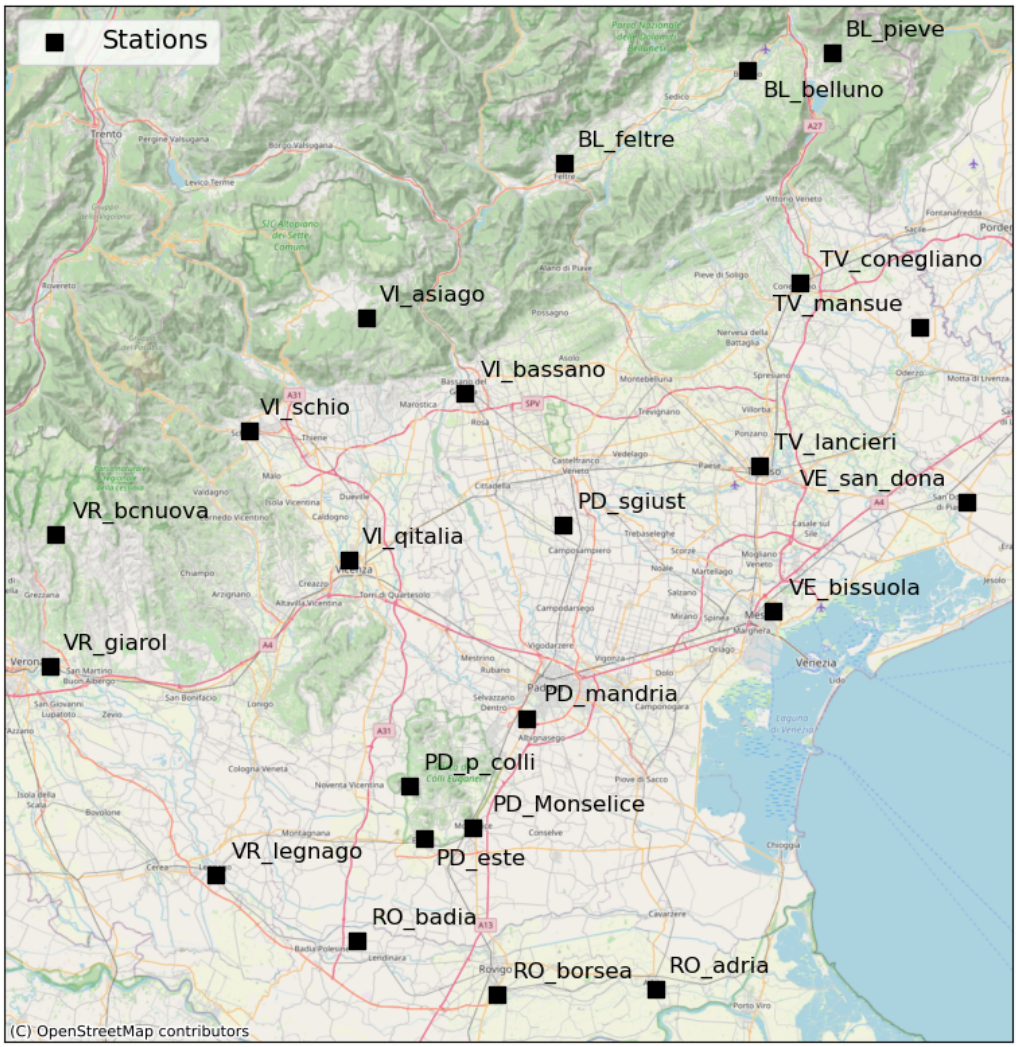


Figure 2.1: Background stations in the Veneto region territory

Measurement methods

For each pollutant, there are rules describing the methods to measure the concentration in the atmosphere ([8], [9], [10]). The measurement instruments that are located inside the stations are certified following these rules. In the analysis, only measurements from automatic instruments will be considered. These instruments perform autonomously the measures without the necessity of chemical analysis in a laboratory.

At the station, each pollutant is measured using a different instrument whose working principle makes use of the chemical and physical properties of the target substance. A simple overview of the operating principles of the instruments for each relevant pollutant is reported below ([8], [9], [10], [1]). Note each instrument measures on an hourly basis.

Particulate Matter

The principle behind the measurement of particulate matter concentration is the absorption of β-rays by samples of particulate (Beta Rays Attenuation). The instruments are provided by an air pump whose geometry allows the filtering of specific particulate matter by size(such as PM_{10}). The measurement process is carried on in two steps:

- At first, a known volume of air is sampled by the pump. Particulate contained in the air settles upon a tape that is used as a filter.

- After the sampling process the tape is moved and the sample of particulate matter deposited on it is exposed to a radioactive source of β-rays. A Geiger counter set on the other side of the tape measures the intensity of radiation coming from the source with the presence of PM collected on the tape along the path. This measure is compared to the baseline beta count. The difference is proportional to the mass of the particulate matter on the sample.

Nitric Dioxide

Measurement of the nitric dioxide concentration is based on a chemiluminescence reaction between O_3 and NO:

$$O_3 + NO \rightarrow NO_2^* + O_2$$
$$NO_2^* \rightarrow NO_2 + h\nu \ (\approx 700nm)$$

A known volume of air is sampled and, in the measurement chamber, is mixed with a separate ozone flux. The ozone and the NO react producing an excited state of NO_2 that quickly returns to the fundamental state emitting UV radiation. From the intensity of the radiation, the mass of NO is retrieved. To obtain the mass of NO_2, after the first measure, the same air is passed over a NO_x-coverter that reduces all the NO_x to NO. Finally, the reduced air reacts again with ozone to find the total concentration of NO_x after reduction. NO_2 concentration is retrieved by a simple subtraction between the total NO_x get in the second measurement and the value of NO found in the first one.

Ozone

This measure uses specific absorption of UV radiation at $\lambda = 254\ nm$ by the ozone. Inside the measurement chamber, a known volume of sampled air is exposed to a UV lamp that emits at the appropriate frequency. Ozone in the air causes part of the radiation to be absorbed, resulting in an intensity reduction. The result of the measurement is compared to a similar one where the air sampled is previously filtered removing ozone. Using the Lambert-Beer law [11] and the difference between the two measures, the concentration of ozone in the air is retrieved.

2.2 Deterministic model

The deterministic model implemented by ARPAV for Veneto can provide an estimate of pollutant concentrations far from the stations and, coupled with a meteorological prognostic model produces forecasts for the upcoming days. Specifically, the system provides forecasts of up to three days. The forecast system is based on the CAMx eulerian photochemical model (Comprehensive Air Quality Model with Extensions) [1] [12]. The domain of the model is a grid of 64×59 cells and 11 vertical levels, from 20 to 6000 meters above ground level (Note that the cells are terrain-following, meaning that the first vertical layer is at the altitude of the terrain. For example, the first cell above the Adriatic Sea is at $20msl$ while cells above dolomites can be even at $2000msl$). The cell size is 4 km, a trade-off between the need to have a good resolution and to avoid computation complexity that arises with high resolution. Geographically speaking the area

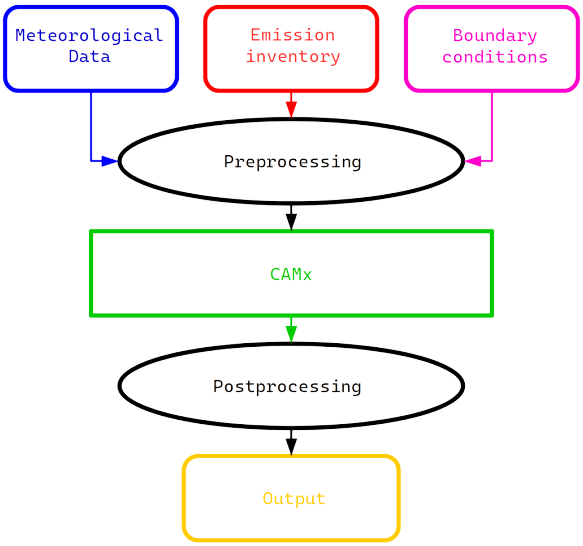


Figure 2.2: Scheme of the deterministic model used by ARPAV

[1] ARPAV uses CAMx version 6.5

covered is the North-East of Italy (Fig. 2.3) including the whole Veneto and part of the nearby regions: Lombardy, Emilia-Romagna, Trentino-Alto Adige and Friuli Venezia Giulia ([13]).

Figure 2.3: CAMx model domain

CAMx requires input information about meteorological conditions, emissions, and boundary conditions. A peculiarity of CAMx is the fact that its computation takes into consideration not only the diffusive processes of pollutants but also the chemical reactions that happen in the atmosphere. As mentioned in Sec. 1.1, some pollutants are produced partially or entirely through reactions and cannot be accounted for in the emission inventory. In Fig. 2.2 the specific structure of the modelling system used by ARPAV is illustrated.

Model Inputs

Emissions are retrieved from the regional inventory INEMAR (Air EMission INventory, [14]). This is an inventory used by all the regions of Northern Italy. It computes the emission of different air pollutants on a municipal basis. In order to match the discrete structure of CAMx domain emission data are also gridded and then assigned to the model levels.

Meteorological data are given by the COSMO-ITA model [15]. It is a non-hydrostatic, limited-area atmospheric model. COSMO outputs, up to three days, are recovered by the national meteorological service.

Boundary conditions data are retrieved by PREV'AIR [16], the French national platform for air quality. A conversion table allows casting PREV'AIR species to CAMx species.

2.3 Datasets

Data from stations and the deterministic model are collected in datasets. For each station, the dataset reports, for each day, the value of concentration measured and also the value of the model in the specific cell where the station lies (see Tab. 2.2 for an example). Note that, as seen above (Sec. 2.1) the raw data gathered in the stations contain hourly measurements, however for this work, we are interested in daily values. In particular, we are interested in the daily average for PM_{10} and NO_2 and in the daily maximum value for the O_3.

day	**month**	**year**	**measurement** $[\mu g/m^3]$	**model** $[\mu g/m^3]$
01	12	2021	39	60.0785
02	12	2021	31	43.2696
03	12	2021	16	38.4391
04	12	2021	26	42.2938
05	12	2021	26	37.3036
06	12	2021	26	36.6818
07	12	2021	43	48.8833
08	12	2021	24	23.3208
09	12	2021	16	24.7045
10	12	2021	29	40.2529
11	12	2021	38	40.4639
12	12	2021	63	36.0657
13	12	2021	61	63.2390
14	12	2021	77	73.4263
15	12	2021	60	62.0124

Table 2.2: Example of daily averages of PM_{10} relative to the station in Mandria(PD), for each row the date, concentration measurements and model estimations are reported

Concerning model estimations, hourly data from CAMx are post-processed, averaging on a daily basis for PM_{10} and NO_2 or taking the maximum value for O_3. The final results are collections of data, one for each cell of the model. For this analysis, only the lowest altitude layer, 10 m above the ground, is taken into consideration.

Note on uncertainties

Concerning precision, the instruments used by ARPAV follow the rules specified in the legal directive ([3]) that set the error target for each pollutant. For PM_{10} the relative uncertainty (expressed at a 95% confidence level) must not be greater than 25%, while for NO_2 and O_3, it cannot exceed 15%. These are relative uncertainties computed at the limit values, which means that values near the legal threshold are affected by this error, while for small values of concentrations, the error is not well quantified (but, of course, it is less important since we are well under the threshold).

On the other hand, for the deterministic model output, the error value is not estimated. When CAMx outputs are reported (2.2) we are going to report all the digits.

Chapter 3

Methods

3.1 Introduction to geostatistics and spatial data

Geostatistics is the branch of statistics that deals with spatiotemporal data structures. The "geo" prefix refers to the fact that geostatistics originally meant studying models and data relative to the earth, with the first application in the geological context. However, the methods and approaches used are universal and designed to tackle any statistical problem relative to processes with continuous spatial index [17]. In many scientific fields, collected data are spatial or time-dependent; geostatistical datasets are collections of samples of some variables $(z_1, z_2, ..., z_N)$ and for each sample, the spatiotemporal information is also given (x_1, x_2, t). Note that in this only the spatial structure of datasets is directly exploited for spatial prediction.

The next definitions and approaches are taken from [18]. Let us now consider only one variable z. This variable is sampled at n locations inside a region $(\mathcal{D})$:

$$z(s_\alpha), \qquad \textbf{with} \quad \alpha = 1, ..., n \tag{3.1}$$

Note that s encodes the coordinates information.

Spatial data are collections of values at specific locations of a variable $z(s)$, which is a function over the whole spatial continuum domain and is called *regionalized variable*.

$$z(s) \quad \textbf{for all} \quad s \in \mathcal{D} \tag{3.2}$$

To take into account randomness and statistical uncertainties sampled values $z(s_\alpha)$ can be considered random draws from *random variables* $Z(s_\alpha)$, which can differ depending on the specific location S_α. To the same extent the *regionalized variable* $\{z(s), s \in \mathcal{D}\}$ can be viewed as one draw from an infinite set of random variables, called *random function*:

$$\{Z(s), s \in \mathcal{D}\} \tag{3.3}$$

To apply the methods in this section to the random function, some assumptions have to be made([17]). Firstly, concerning the first moment:

$$\mathbb{E}(Z(s)) = \mu \qquad for\ all\ s \in \mathcal{D} \tag{3.4}$$

Another important assumption to use the more advanced methods of spatial interpolation is:

$$Cov(Z(s_1) - Z(s_2)) = C(s_1 - s_2) \tag{3.5}$$

where the function C is called covariogram. A random function that satisfies these two assumptions is called second-order stationary.

This approach results to be very useful for solving a wide variety of problems with geostatistical methods as the object of study and modelling is the random function itself [17].

3.2 Interpolation methods

Spatial prediction is one of the most common problems in geostatistics, related originally to the prediction of ores concentration for mining. Many different methods were developed and used. Concerning Air Quality, the European Union gives some guidelines for how to make spatial predictions using both data and models distinguishing eight degrees which vary from spatial prediction using raw data to the simple usage of the unvalidated model ([19], [20]).

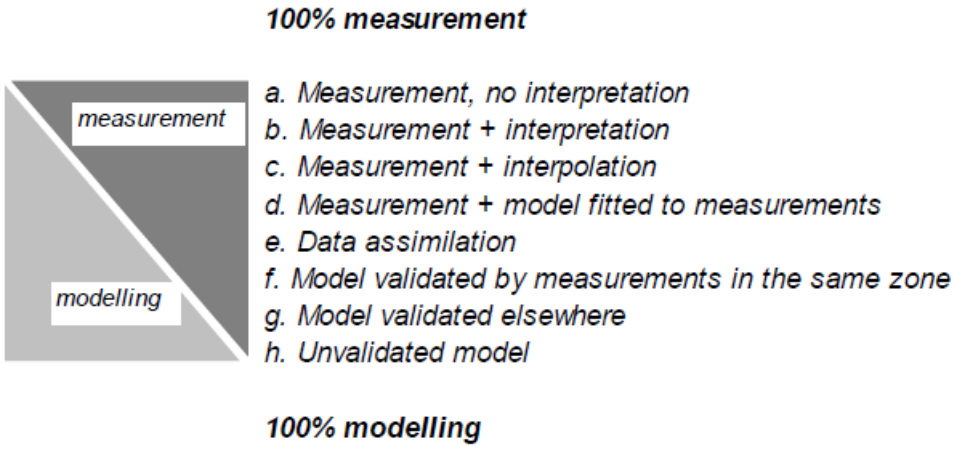


Figure 3.1: Degrees of spatial prediction methods from [20]

For this work, we are going to exploit the so-called interpolation methods using both data and model as explained in the previous chapters ([21]). The approach we are going to use is the interpolation of residual fields. Since at station locations, both the measurements and model estimations are known, one can compute the residuals and interpolate this difference field in the whole domain of the model.

$$R(s_i) = M(s_i) - S(s_i) \tag{3.6}$$

where we refer with R to the residuals, with S to station measurements, with M to model values while s_i is a specific station location (x_i, y_i).

A useful trick that is used ([19,20],) is to calibrate the model values before interpolation. The simplest way to do so is to use linear regression between measures and model estimations:

$$S(s) = a + b \cdot M(s) \tag{3.7}$$

where a and b are the regression parameters. Fitting the linear regression model and finding the best parameters $(\hat{a}, \hat{b})$, the calibrated model is computed

$$M'(s) = \hat{a} + \hat{b} \cdot M(s)$$

The calibrated model is then used to compute the residuals like in Eq.3.6.

Note that the calibration procedure can be applied using other variables are predictors. For the specific case of this work, since we are dealing with a territory with a complex orography, also the altitude ($h(s)$) is used ([22]).

$$S(s) = a + b \cdot M(s) + c \cdot h(s) \tag{3.8}$$

One of the assumptions of the linear regression model is the fact that the variance of the target variable is constant and residuals are normally distributed. Measurement data do not always satisfy this condition. The usage of Box-Cox transformation [23] on the data before the linear regression is, then tested. This method is a power transformation that is commonly used to tackle problems such as nonnormality and nonconstant variance of data. It consists of a non-linear transformation applied to the data, intending to make the data normal distribution-like:

$$y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \lambda \neq 0 \\ log(y) & \lambda = 0 \end{cases} \tag{3.9}$$

In this case, where the linear model used is $y = \beta X$ we want the conditional probability $p(y^{(\lambda)}|\beta X)$ to be a Gaussian distribution. BoxCox parameter λ to use depends on the data and is usually chosen through a Maximum Likelihood approach [23]. For the linear model $y^{(\lambda)} = \beta X$, the normal likelihood is:

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} exp\left\{ -\frac{(y^{(\lambda)} - \beta X)^T (y^{(\lambda)} - \beta X)}{2\sigma^2} \right\} \prod_{i=1}^{n} \left| \frac{dy^{(\lambda)}}{dy} \right| \tag{3.10}$$

Taking the log-likelihood the function to be maximized is:

$$L(\lambda) = (\lambda - 1) \sum_i \log(y_i) - \frac{n}{2} \log\left(\frac{y^{(\lambda)T} (\mathbb{I} - X(X^T X)^{-1} X^T) y^{(\lambda)}}{n} \right) \tag{3.11}$$

After applying linear regression and spatial interpolation methods, the predicted target variable is transformed back, using the inverse transformation.

Although the Box-Cox transformation is used in literature for air quality data [22, 24], its application is somewhat controversial. In fact, after the transformation variables are less interpretable and while the λ parameters found is the one that maximizes the Likelihood, there is no guarantee that the final distribution is Normal. For this work, Box-Cox transformation on measurement data before regression is tested and its results are compared to the basic linear model outputs.

3.2.1 Geometric Methods

These are interpolation methods that take into consideration only the distance among points to compute the value on the unknown locations. In this work, we are going to neglect the simplest possible approaches such as Thiessen Polygon (also called Nearest Neighbour) or Bilinear interpolation (also called Triangulation) [20].

Inverse distance weighting

The simplest method tackled is Inverse Distance Weighting (IDW)([19, 20]), which is the method currently implemented by ARPAV to correct model data with station measurements. The value of the residual field at the location needed s_0 is computed as a linear combination of values measured from all points. Each value is weighted using considering the distance between the unknown point location and the others, see Eq. 3.12.

$$R(s_0) = \frac{\sum_{i=1}^{n} \frac{R(s_i)}{d_{0i}^{\beta}}}{\sum_{i=1}^{n} \frac{1}{d_{0i}^{\beta}}} \tag{3.12}$$

Usually, the β exponent is equal to 2, as we are in a 2D framework.

Radial Basis Function

Radial basis functions were historically introduced for exact function interpolation ([25]). The target smooth function to find is expressed as a linear combination of radial basis functions centred in the data points (Eq. 3.13)([20, 25, 26]). As the IDW method seen above is an exact method: the target value at the data points is the one measured. Radial basis functions do not depend on the specific points but only on the distances among them so no specific information about the spatial structure of data is required.

$$R(s_0) = \sum_{i=1}^{n} w_i \phi(d_{0i}) \tag{3.13}$$

where d_{0i} is the euclidean distance between the test point s_0 and the point s_i.

Given the values of $R(s_i)$, it is easy to compute the unknown weight w_i. Weights are such that the next equation holds:

$$\begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} R(s_1) \\ R(s_2) \\ \vdots \\ R(s_n) \end{bmatrix}$$

where $\phi_{i,j} = \phi(d_{i,j})$. So, they can be computed simply by inverting the matrix:

$$\mathbf{w} = \Phi^{-1}\mathbf{R}$$

Learned weights can now be used to compute the interpolated values at a new location s_0.

Common Radial Basis Functions used are:

- Gaussian: $\phi(r) = e^{-(\varepsilon r)^2}$
- Inverse multiquadric: $\phi(r) = \frac{1}{\sqrt{1+(\varepsilon r)^2}}$
- Inverse quadratic: $\phi(r) = \frac{1}{1+(\varepsilon r)^2}$

Where the ε parameter is a shape parameter that scales the input of RBF. For the case of spatial interpolation, a good guess for ε is the order of magnitude of the mean distance among all the points.

Depending on the RBF used Eq, 3.13 can be modified including a polynomial term ([27]):

$$R(s_0) = \sum_{i=1}^{n} w_i \phi(d_{0i}) + \sum_{i=1}^{m} \gamma_i p_i(s_0) \tag{3.14}$$

where γ_i are parameters and p_i are monomials that span the polynomial (which is computed in s_0). The value m is the order of the polynomial. This can be done to ensure the system is uniquely solvable and improve accuracy.

This variation is the method implemented in SciPy ([28]), which is the Python library used for RBF interpolation.

3.2.2 Kriging

These interpolation methods do use the correlation structure of data to make predictions. Kriging is the most important and diffuse geostatistical method for interpolation ([17,18,29]). It was formalized by the French statistician Matheron and named after Danie G. Krige who applies the method to ore mining in South Africa.

There exist many variants of kriging; here, Ordinary Kriging (OK) is explained as it is the simplest variant and the one used in this work.

Ordinary Kriging is a spatial prediction model that lives under two assumptions:

- The regionalized random function studied has an unknown mean and only the noise term depends on the spatial location.
$$Z(s) = \mu + \delta(s) \tag{3.15}$$
- The random variable at an unknown location depends linearly on the random variables at points s_i.
$$Z(s_0) = \sum_i \lambda_i(s_0) Z(s_i) \qquad \sum_i \lambda_i = 1 \tag{3.16}$$

So, the model idea is similar to the IDW model seen previously: the predicted value of the target function at a given point is given by a linear combination of the known values of the function. However, in this case, the coefficients $\lambda_i(s_0)$ are computed using the covariance structure of data.

To solve the problem and find the λ_i the goal is to minimize Eq. 3.17 with respect to $\lambda_1, \lambda_2, ..., \lambda_n, m$, where m is a Lagrange multiplier for the condition of the sum of λ_i seen in Eq. 3.16

$$\mathbb{E}\left(Z(s_0) - \sum_i \lambda_i Z(s_i)\right)^2 - 2m\left(\sum_i \lambda_i - 1\right) \tag{3.17}$$

For the first term, using the fact that $\sum_i \lambda_i = 1$:

$$\begin{aligned}\left(Z(s_0) - \sum_i \lambda_i Z(s_i)\right)^2 = &- \sum_i \sum_j \lambda_i \lambda_j (Z(s_i) - Z(s_j))^2/2 \\ &+ 2\sum_i \lambda_i (Z(s_0) - Z(s_i))^2/2\end{aligned} \tag{3.18}$$

At this point, one can introduce the definition of variogram 2γ as:

$$2\gamma(d) = Var(Z(s+d) - Z(s)) \tag{3.19}$$

The variogram is a function that evaluates the degree of spatial dependence and is computed using the variance of the random field at two different points.

Thus, using Eq. 3.18 and Eq. 3.19, Eq. 3.17 can be written as:

$$-\sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_j) + 2\sum_i \lambda_i \gamma(s_0 - s_i) - 2m\left(\sum_i \lambda_i - 1\right) \tag{3.20}$$

The solution to the minimization can be found by differentiating Eq. 3.20 with respect to the parameters λ_i and m and equating to 0. The final result [17] obtained is:

$$\boldsymbol{\lambda} = \Gamma^{-1}\boldsymbol{\gamma} \tag{3.21}$$

where

$$\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_n, m)^T \tag{3.22}$$

$$\boldsymbol{\gamma} = (\gamma(s_0 - s_1), ..., \gamma(s_0 - s_n), 1)^T \tag{3.23}$$

$$\Gamma = \begin{pmatrix} \gamma(s_1, s_1) & \cdots & \gamma(s_1, s_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n, s_1) & \cdots & \gamma(s_n, s_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \tag{3.24}$$

In this work, we use kriging as implemented in the Python package `pykrige` ([30])

Regression Kriging

Ordinary Kriging (OK) assume that the random function has constant mean μ, this is a suitable guess for the residuals. Nevertheless, in the previous chapter, we mentioned the possibility of using linear regression before the residual interpolation. In the context of Kriging this procedure is called Regression Kriging (RK) ([29]). Regression Kriging extends the capability of Ordinary Kriging interpolation taking into account the fact that the random function can have a mean that depends on the specific coordinates $\mathbf{s}$:

$$Z(s) = m(s) + \delta(s) \tag{3.25}$$

The two components of Eq. 3.25 are modelled separately and combined, obtaining the final formula for prediction (Eq. 3.26):

$$Z(s_0) = \sum_k \beta_k q_k(s_0) + \sum_i \lambda_i R(s_i) \tag{3.26}$$

where the first part is the fitted deterministic part, while the second is the interpolated residual. The q_k are some independent values used to fit the model and find the β_k parameters. On the other hand, the λ_i are computed using Ordinary Kriging as in Eq 3.21.

Practically, Regression Kriging is a generalization of Ordinary Kriging, where the kriging step is preceded by a linear regression step of the random function as predicted by the independent variables q. Using a linear model before doing the residual interpolation is what was already described in [19,20], as mentioned at the beginning of the section, for geometric methods. In this sense, Regression Kriging is the generalization of this concept for the Kriging approach.

It is worth mentioning that the geostatistical literature uses different terms to refer to this variant of Kriging where the mean value of the random function is not constant. Regression Kriging(RK), Universal Kriging(UK) and Kriging with external drift (KED) refer basically to the same technique and are mathematically equivalent [29].

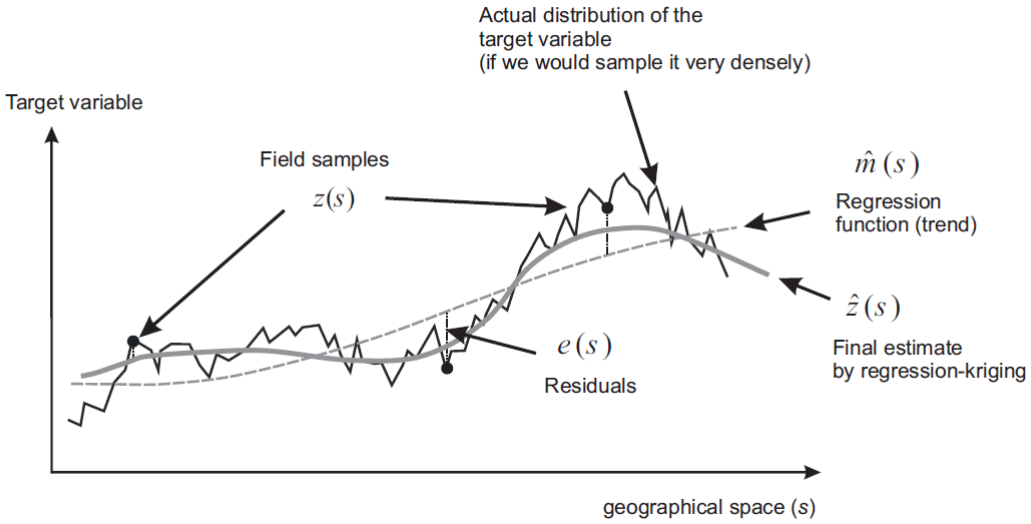
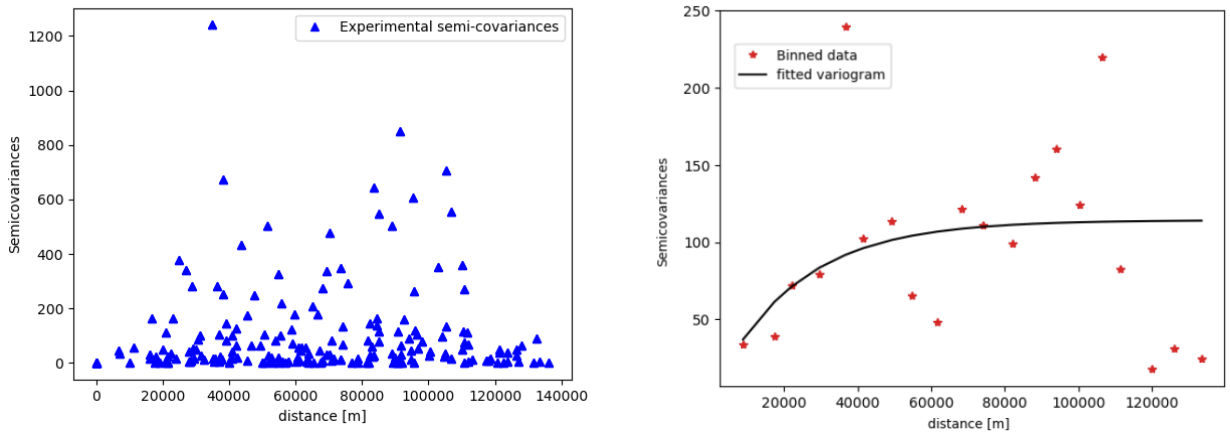


Figure 3.2: Example of Regression Kriging from [29]



(a) Variogram cloud

(b) Fitted variogram

Figure 3.3: Example of cloud variogram plot and fitted variogram with 20 bins

Note on Variongrams

In the Kriging method, the choice of the variogram is fundamental. The theoretical definition of the variogram is given in the Eq. 3.19 but in the algorithm the variogram used is the so-called "Empirical variogram" because it is obtained directly from the data.

The empirical variogram is computed from the empirical covariance among each of the couples of points.

$$\widehat{C}(s_i, s_j) = (z(s_i) - z(s_j))^2 \tag{3.27}$$

One can plot all the empirical semicovariances $\frac{\widehat{C}_{ij}}{2}$ versus the distance $s_i - s_j$ to produce a "variogram cloud" (Fig. 3.3a) that can be interpolated. Usually [17] the points of the "cloud" are grouped into bins $N(d \pm \delta)$ containing points whose distance is $d + \delta$, where $N(d \pm \delta) = (i, j) : s_i - s_j = d \pm \delta$ and for each bin the mean is computed:

$$\hat{\gamma}(d \pm \delta) = \frac{1}{|N(d \pm \delta)|} \sum_{(i,j) \in N(d \pm \delta)} (z(s_i) - z(s_j))^2 \tag{3.28}$$

The experimental variogram is smaller for short distances and tends to reach a constant value for large distances. This happens because the target values are similar (read correlated) for points near each other, while they are more and more independent on each other for long distances.

The fit function is chosen among different families. The choice of the shape of the variogram is a kind of hyperparameter and potentially leads to very different results. The most important kind of variograms used are [17, 31]:

- Exponential

$$\gamma(d) = (\mathcal{S} - \mathcal{N}) \left[1 - exp\left(-\frac{d}{\mathcal{R}/3} \right) \right] + \mathcal{N} \tag{3.29}$$

- Gaussian

$$\gamma(d) = (\mathcal{S} - \mathcal{N}) \left[1 - exp\left(-\frac{d^2}{(\frac{4}{7}\mathcal{R})^2} \right) \right] + \mathcal{N} \tag{3.30}$$

- Spherical

$$\gamma(d) = \begin{cases} (\mathcal{S} - \mathcal{N})\left(\frac{3d}{2\mathcal{R}} - \frac{d^3}{2\mathcal{R}^3}\right) + \mathcal{N} & d \leq \mathcal{R} \\ \mathcal{S} & h > \mathcal{R} \end{cases} \tag{3.31}$$

where $\mathcal{S}$ is called *sill* and represents the asymptotic maximum spatial variance at the longest distances. The *range*, $\mathcal{R}$ is the distance at which the spatial variance has reached $\sim 95\%$ of the *sill* value. The $\mathcal{N}$ is the *nugget* and represents the random deviations from the smooth data trend.

3.2.3 Gaussian Process

Gaussian Process [25, 32, 33] is an important tool widely used in statistics, information theory and machine learning. It is a non-parametric model that has a straightforward Bayesian interpretation. The idea behind this method is to define prior probability distributions over functions and not over parameters (like it is usually done in parametric Bayesian models). Gaussian Process application is usually divided, following the machine learning fashion, into Gaussian Process Regression (GPR) and Gaussian Process Classification, depending on the target value domain.

For obvious reasons, for the specific problem of spatial interpolation, GPR is the method used. Note that in many textbooks ([25, 32]) Kriging method is described as a specific variation of GPR, used for geostatistics context. However, despite Kriging and GPR being essentially the same method, the theory behind them is quite different and also the fitting procedure of the Kernel/Variogram is carried on in distinct ways.

The general definition of a Gaussian Process is [32]:

Definition. A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Gaussian processes inherit some of their properties from Gaussian distributions, in particular, a Gaussian Process $f(z)$ is completely specified by a mean function $m(s)$ and a covariance function also called *kernel* $k(s, s')$:

$$f(s) \sim \mathcal{GP}(m(s), k(s, s')) \tag{3.32}$$

where:

$$m(s) = \mathbb{E}\big[f(s)\big] \tag{3.33}$$
$$k(s, s') = \mathbb{E}\big[(f(s) - m(s))(f(s') - m(s'))\big] \tag{3.34}$$

The Gaussian Process Regression framework is a typical supervised learning problem. The goal is to find $f(s)$ knowing the value of the function at some training points. In the most general case the training points are noisy:

$$y_i = f(s_i) + \varepsilon_i \tag{3.35}$$

In this case, ε is a noise term that does not allow us to know precisely the function values at target points. Behind the assumption of additive independent identically distributed Gaussian noise with a variance σ_ε^2, the covariance (kernel) for the observations is:

$$Cov(y_i, y_j) = k(s_i, s_j) + \sigma_\varepsilon^2 \delta_{ij} \tag{3.36}$$

The Bayesian approach to the problem consists in selecting a GP prior to the $f(s)$ specifying mean and kernel function. The knowledge of the training data 3.35 will modify the prior into a posterior distribution for $f(s)$. Specifically, we are interested in the value of the function f_0 at location s_0. The prior selected is the Gaussian Process at Eq. 3.32. We can write the joint distribution of the observed target values and function values at test locations under GP prior as:

$$\begin{bmatrix} y \\ f_0 \end{bmatrix} \sim \mathcal{GP}\left(0, \begin{bmatrix} k(S, S + \sigma_\varepsilon^2 \mathbb{I}) & k(S, s_0) \\ k(s_0, S) & k(s_0, s_0) \end{bmatrix}\right) \tag{3.37}$$

where $S = (s_1, s_2, \ldots, s_n)$.

The probability distribution for f_0 can be obtained by conditioning the joint Gaussian prior distribution on the observations:

$$f_0 | S, y, s_0 \sim \mathcal{GP}\big(\bar{f}_0, Cov(f_0)\big) \tag{3.38}$$
$$\bar{f}_0 = \mathbb{E}[f_0 | S, y, s_0] = k(s_0, S)[k(S, S) + \sigma_\varepsilon^2 \mathbb{I}]^{-1} y \tag{3.39}$$
$$Cov(f_0) = k(s_0, s_0) - k(S, s_0)[k(S, S) + \sigma_\varepsilon^2 \mathbb{I}]^{-1} k(s_0, S) \tag{3.40}$$

Note that Eq. 3.39 states that the mean prediction value for the function f at point s_0 is a linear combination of observation y. These are the same results of the Kriging method, but in this framework, it was derived and not a starting assumption.

Similarly to what was mentioned for the Kriging method, in Gaussian Process regression the choice of kernel $k(x, x')$ is crucial. In this work two different kernels are used:

- RBF kernel: $k(s_i, s_j) = A \cdot exp\left(-\frac{d(s_i, s_j)^2}{2l^2}\right) + B \cdot \delta(s_i, s_j)$
- Matern kernel: $k(s_i, s_j) = A \cdot \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}}{l} d(s_i, s_j)\right)^\nu K_\nu \left(\frac{\sqrt{2\nu}}{l} d(s_i, s_j)\right) + B \cdot \delta(s_i, s_j)$

where is $d(.,.)$ is the Euclidean distance, K_ν is a modified Bessel function and is $\Gamma(.)$ is the gamma function. For both kernels the parameters to be optimized are the amplitude A, the length scale l and B. Note that a WhiteNoise term with amplitude parameter B was added to both kernels to take into account the presence of independently and identically normal-distributed noise for target variable. The ν parameter in Mater kernel is set to 2.5 which is a standard choice GPR ([32]).

Gaussian Process Regression is tested using the package `Scikit-Learn` [34], in particular, the class `GaussianProcessRegressor`.

GPR and Kriging

GPR and Kriging are, basically, the same method. They were, historically, introduced to solve different problems but it is possible to assume that kriging is, practically, Gaussian Process Regression applied to geostatistics. There is, nonetheless, one little difference in how the Kernel/Variogram is optimized in the two methods implemented for this work. The semivariogram in Ordinary Kriging is fitted, using the empirical variogram extracted from the data and then used for the interpolation process. On the other hand, the hyperparameters of the kernel in GPR are not fitted using the data, instead, they are optimized through multiple runs for finding the values that maximize the log-marginal likelihood.

Since the two methods are linked, we can formulate a connection between Kriging variogram and Gaussian process Kernel, based on the definition of these two fundamental objects.

$$\begin{aligned} k(s,s') &= Cov(z(s), z(s')) = \mathbb{E}[(z(s) - \mu(s))(z(s') - \mu(s'))] \\ &= \mathbb{E}[z(s)z(s')] + (\mu(s) - \mu(s'))^2 \end{aligned} \tag{3.41}$$

Where we used the fact that $\mu(s) = \mathbb{E}[z(s)]$. Now, given the definition of variogram:

$$\begin{aligned} 2\gamma(s,s') &= Var(z(s) - z(s')) = \mathbb{E}\{[(z(s) - \mu(s))(z(s') - \mu(s'))]^2\} \\ &= k(s,s) + k(s',s') - 2k(s,s') \\ &= 2(\mathcal{S} - k(s,s')) \end{aligned} \tag{3.42}$$

In Eq.3.42 taking the limit for $|x - x'| \to \infty$, we can recognize the definition of sill ($\mathcal{S}$) (Sec. 3.2.2).

$$\lim_{|x-x'|\to\infty} \gamma(x,x') = \mathcal{S} \tag{3.43}$$

3.3 Method comparison

3.3.1 Cross-Validation

The difficulty of a proper comparison among the previously seen methods arise from the fact that there are no other points apart from stations that can be used as test points. Moreover, as seen in section 2.1, the number of stations is quite limited and devoting a part of them to validate the model can reduce the capability of our interpolation. The strategy we decided to adopt to perform the comparison is a Cross-Validation strategy. Every time a method is tested, the interpolation procedure is repeated a number of times equal to the number of stations and every time a different station is left out and used for validation (Left-One-Out strategy)(Fig. 3.4).

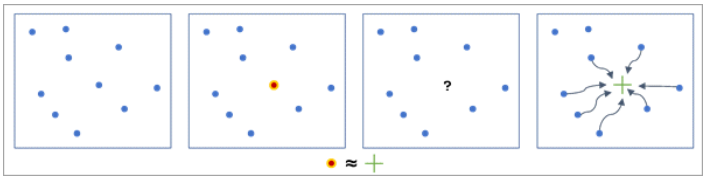


Figure 3.4: Diagram of cross-validation, from [35]

This procedure is then repeated for each day of the analyzed period. Note that this procedure can become computationally demanding when using more complex algorithms (*e.g.* GPR). The final result

is a predicted value of the concentration of air pollutants for each station, for each day. This final estimation depends on the measured values retrieved from all the other stations. Those values are then analyzed using the statistical estimator described in the next section.

3.3.2 Statistical Indicators

The statistical indicators used to evaluate the previously seen methods are listed in this section. Note that these indicators contemplate the knowledge of both station measurements and model values (or post-processed model values, after the application of spatial interpolation). Given a certain interval of time, which corresponds to a certain amount of data, they are computed for each different station, measuring how well model estimations are close to the observations.

- Correlation Coefficient:

$$r = \frac{\sum_i (S_i - \bar{S})(M_i - \bar{M})}{\sqrt{\sum_i (S_i - \bar{S})^2 (M_i - \bar{M})^2}}$$

- Normalized Mean Bias:

$$NMB = \frac{\sum_i (M_i - S_i)}{\sum_i S_i} \cdot 100$$

- Normalized Mean Error:

$$NME = \frac{\sum_i |M_i - S_i|}{\sum_i S_i} \cdot 100$$

While the NME is the more suitable estimator for the overall performance of methods we want to check also NMB that can spot the presence of positive or negative biases. Correlation is also important to detect if some time patterns in air pollutant concentrations are well reproduced by the model.

Following [13] and [36], one can establish some benchmarks for method evaluation. In this work, the goals adopted can be seen in Tab. 3.1. Note that there is a minimum threshold and a goal.

Estimator	PM_{10}		NO_2		O_3	
	Criteria	Goal	Criteria	Goal	Criteria	Goal
NME	$< 50\%$	$< 35\%$	$< 50\%$	$< 35\%$	$< 25\%$	$< 15\%$
NMB	$< \pm 30\%$	$< \pm 10\%$	$< \pm 30\%$	$< \pm 10\%$	$< \pm 15\%$	$< \pm 5\%$
r	> 0.4	> 0.7	> 0.4	> 0.7	> 0.5	> 0.75

Table 3.1: Benchmarks for method evaluation

Chapter 4

Results

4.1 Particulate matter

Raw model performance

Before evaluating the methods seen in the previous chapter, an analysis of the model introduced at Sec. 2.2 is performed. A complete verification of the model performance can be found at [13]. Here we report the analysis for the period from December 2021 to February 2022. For each station, the statistical indicators described in detail in Sec.3.3.2 are computed (see Tab. A.1, A.2,A.3, first column) and visualized in Fig. 4.1. The evaluation is done using the station measurements S_i and the model estimations M_i. The goal is to highlight which are the locations where the estimations differ more from the measured values. In this chapter, each postprocessing method applied to the model output is going to be tested using the same indicators used for the raw model. For each station, S_i are the observations, while M_i are the results of Cross-Validation where the testing station was left out.

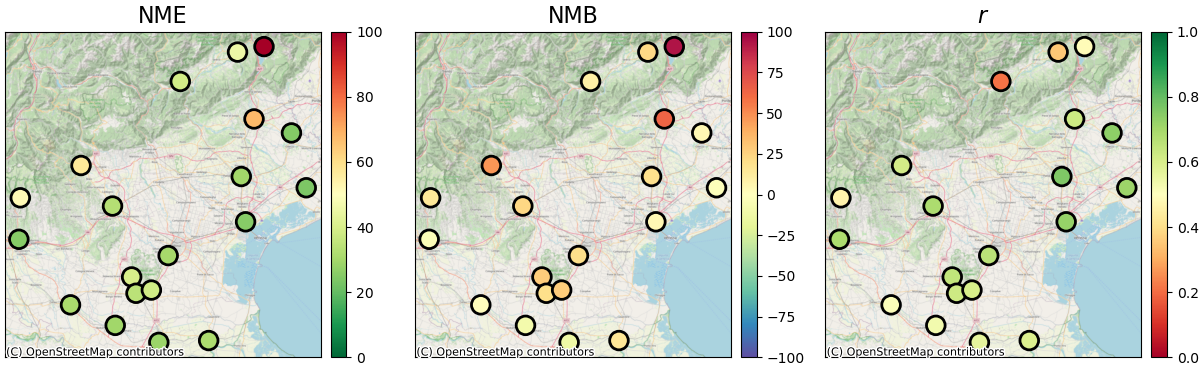


Figure 4.1: PM_{10}: NME, NMB and correlation (r) at each station location for model estimations

The model has good performance for stations that are located in the plains. Note that this is where the exceeding of the legal threshold is more expected to happen. On the other hand, the NME plot, and correlation plot underline some errors for stations on hills and mountains.

Moreover, from the NMB plot, it is clear that the model overestimates the pollutant concentration for many station locations. This can be seen also in Fig. 4.2 where the points lie asymmetrically with respect to the bisector. The effect is particularly accentuated in the station of Pieve d'Alpago and Conegliano (Fig. 4.3), where the model struggles to give good estimations. The overestimation is the main reason to apply a linear regression before any spatial interpolation of residuals. The fact that the biggest errors are located in stations with higher altitude suggest the possibility of using h_{msl} as a predictor

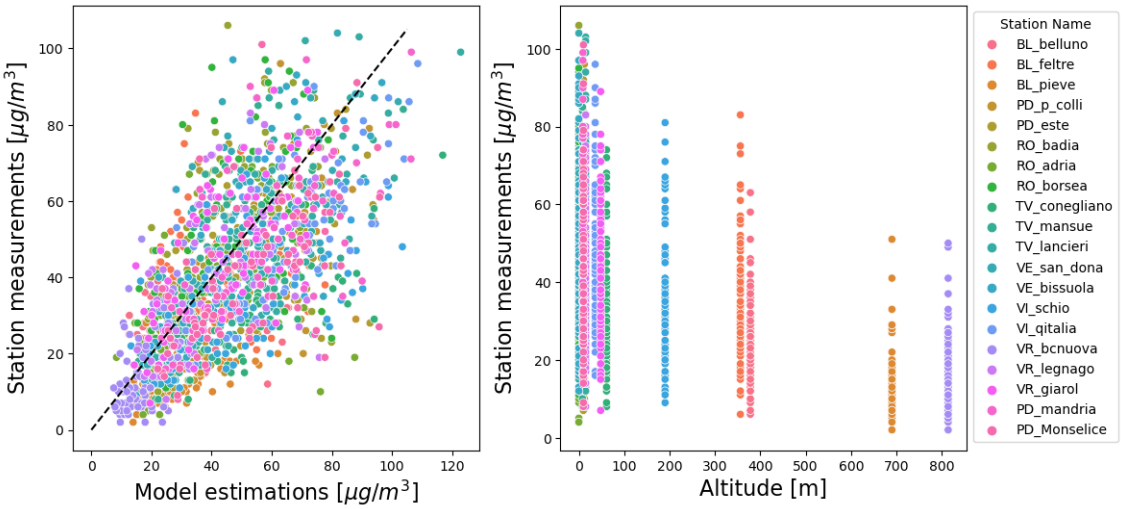


Figure 4.2: PM_{10}: Scatterplots of measurements and estimations and measurement and altitudes at station locations

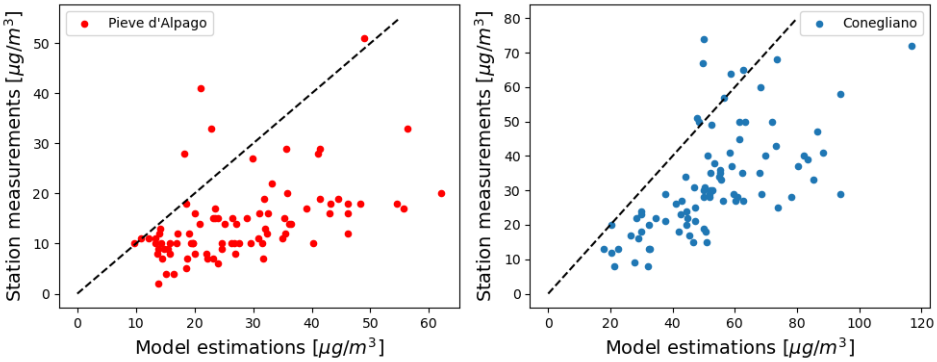


Figure 4.3: PM_{10}: Scatterplot of measurements and estimations at the station in Pieve d'Alpago and Conegliano

Regression

Four different linear regression methods are compared (see Sec. 3.2):

- Model estimations $M(s)$ as predictor;
- Model estimations as a predictor and Box-Cox on measurement data $S(s)$;
- Model estimations and altitude as predictors;
- Model estimations and altitude as predictors and Box-Cox on measurement data.

For Box-Cox transformation the parameter λ needs to be found by studying the log-Likelihood. In Fig. 4.4 we report the results of maximization.

- For the linear model $S(s) = a + b \cdot M(s)$, $\lambda = 0.53 \pm 0.02$.
- For the linear model $S(s) = a + b \cdot M(s) + c \cdot h(s)$, $\lambda = 0.43 \pm 0.02$

The results of the comparison of regression models are shown in Tab. A.1, A.2, A.3 and Fig. 4.5. Note that linear regression solves the overestimation shown in the previous paragraph with the exception of the two critical stations mentioned above. Nevertheless, the results are better than the raw model for every station. The best regression model is the one that uses two predictors. Box-Cox does not improve the model performances so it will not be used for PM_{10} data. Because of the success of the linear regression model, in the next analysis regression is applied to data before the spatial interpolation of residuals.

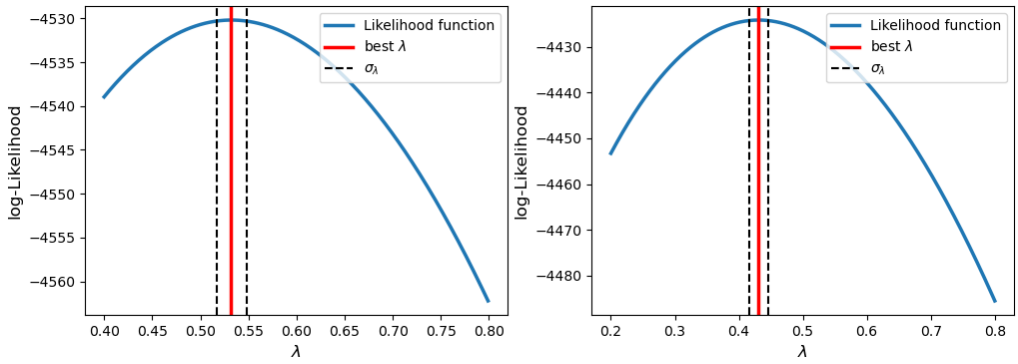


Figure 4.4: PM_{10}: Box-Cox likelihood function and best λ estimation for the two different linear regression models

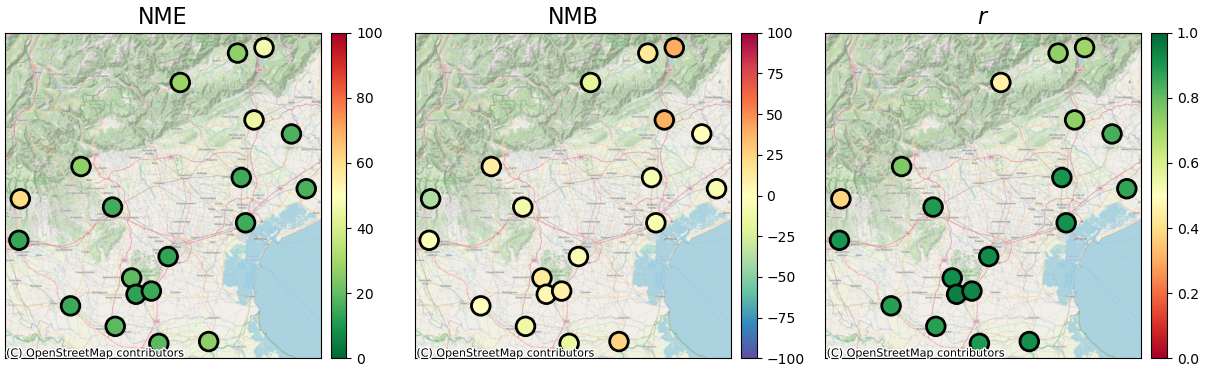


Figure 4.5: PM_{10}: NME, NMB and correlation (r) at each station location for linear regression results

Inverse Distance Weighting

The simplest interpolation method, which is the one already in use by ARPAV, is to apply inverse distance weighting as seen in Sec. 3.2.1. Results are visualized in Fig. 4.6. Comparing these plots with the one in Fig. 4.1 one can see that overall the NME is better using IDW but still the corrected model seems to overestimate the concentration at some station locations. In order to solve this issue, linear regression can be performed before spatial interpolation (Fig. 4.7).

Using linear regression (Regression IDW) improves the final results of IDW. In particular, it lowers NME and correlations in mountain stations like Bosco Chiesanuova(VR) and Schio(VI). It also reduces the overall positive bias.

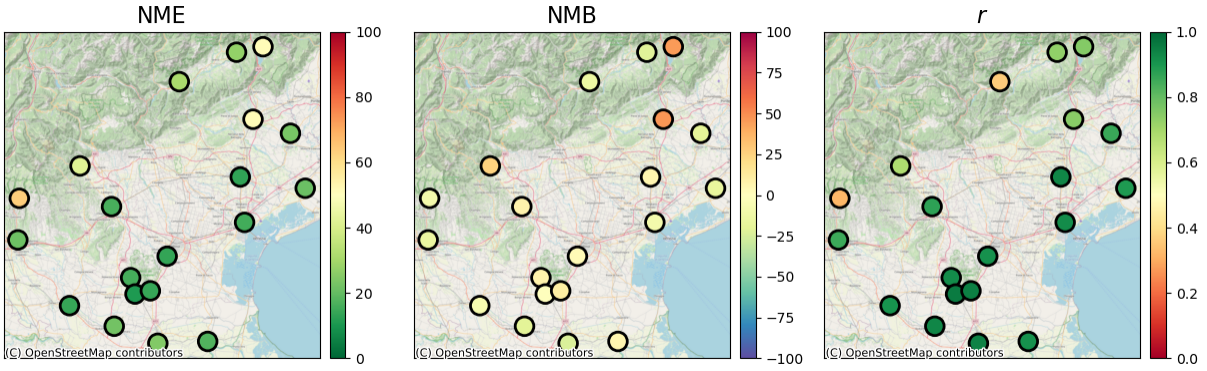


Figure 4.6: PM_{10}: NME, NMB and correlation (r) at each station location for IDW results

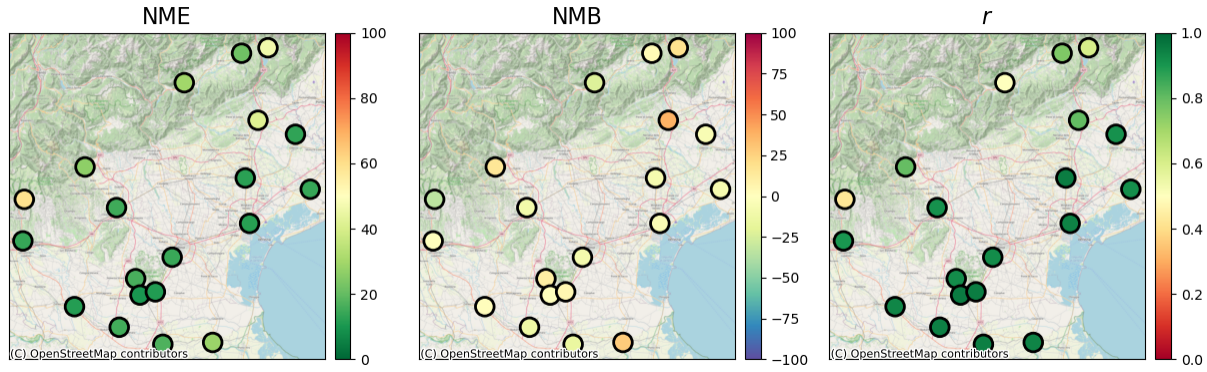


Figure 4.7: PM_{10}: NME, NMB and correlation (r) at each station location for Regression IDW results

Radial Basis Function interpolation

For the RBF interpolation, we will use a procedure similar to the previous method: at first, a linear model is fitted and the residual field is computed using RBF interpolation on top of it. Different Radial Basis Functions were tried and here (Fig. 4.8) we report the results for inverse multiquadric radial basis function: $\phi(r) = \frac{1}{\sqrt{1+(\varepsilon r)^2}}$ which performs better than the others (Tab. A.7). The best ε for PM_{10} data is equal to $10^{-4}\ m^{-1}$, the same order of magnitude that the inverse of mean distance among points which is $6.2 \cdot 10^4\ m$.

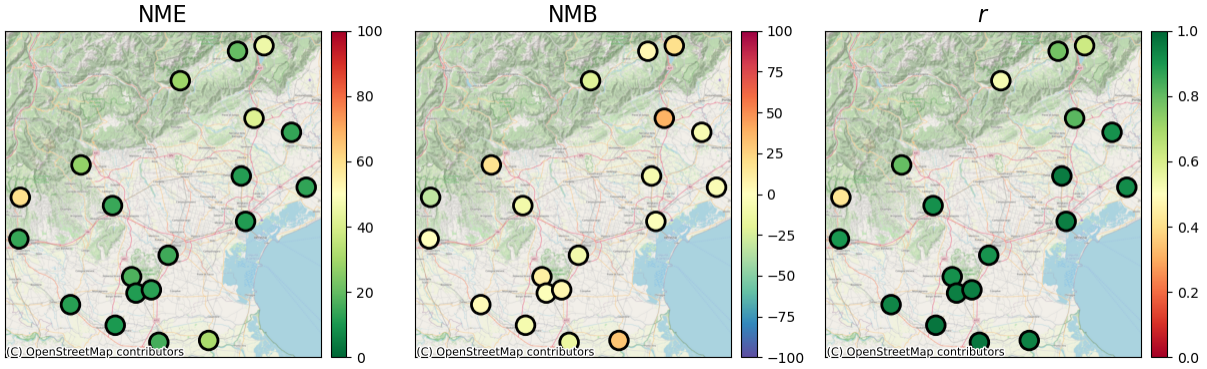


Figure 4.8: PM_{10}: NME, NMB and correlation (r) at each station location for Regression and RBF interpolation results

Regression Kriging and Gaussian Process Regression

For the Regression Kriging different variogram models were studied (see Sec. 3.2.2), obtaining similar results. In Fig. 4.10 we report the plots for the Exponential variogram. Regression Kriging performs well, improving the raw model estimation by reducing NME, NMB and correlation.

Also in the case of Gaussian Process Regression, the spatial interpolation of residual is preceded by a linear regression. The actual GPR is performed using the `GaussianProcessRegression` algorithm of `Scikit-learn` ([34]). For the fitting procedure, target data (measurements) are standardized and the hyperparameters of the kernels are optimized through, multiple runs of the model. After the comparison (see Tab A.14),

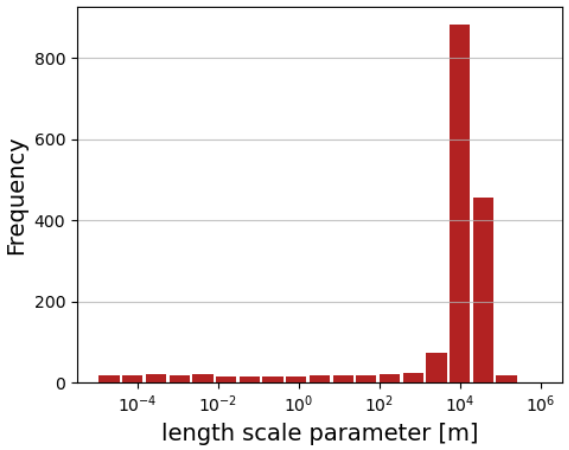


Figure 4.9: PM_{10}: Distribution of length scale parameter in Cross-Validation

Matern kernel is the best of the two tested. The results are shown in Fig. 4.11 and are very similar to the Regression Kriging ones. Among the Gaussian Process Regression results, it is worth analyzing the distribution of the kernel parameter l, called length scale, fitted during Cross-Validation. In the case of PM_{10} the length scale distribution is peaked at 10^4 4.9. This gives an idea of the scale which regulates the speed of decay of correlation between points and confirms the result obtained in RBF interpolation of a characteristic length of the system in the order of magnitude of 10 km.

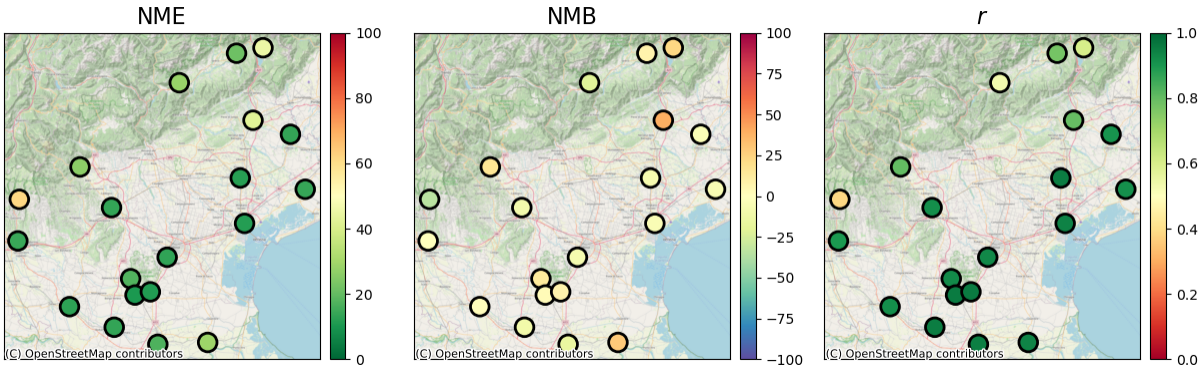


Figure 4.10: PM_{10}: NME, NMB and correlation (r) at each station location for RK results

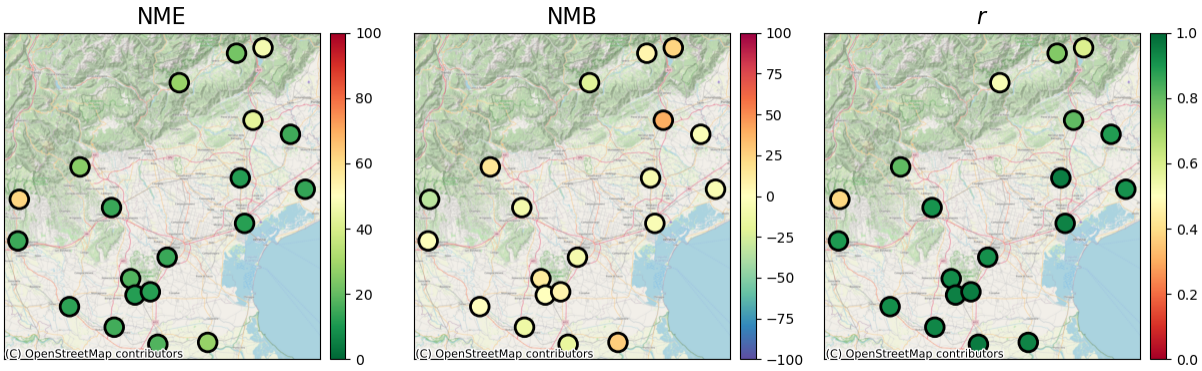


Figure 4.11: PM_{10}:NME, NMB and correlation (r) at each station location for Regression and GPR results

Method comparison

The results of the best variant for each method are collected in Tab. 4.1, 4.2, 4.3.

From the NME data, it is easy to see that most of the time the spatial interpolation methods improve the performance of the model. Linear regression is an important tool to use before interpolating the residuals to improve the estimation and, more importantly, remove the positive bias of the raw model. The final computed indicators meet the benchmarks set in Tab. 3.1 with the only exceptions of the stations in Pieve d'Alpago(BL) and Conegliano(TV) whose indicators respect the "Criteria" threshold but not the "Goal" one, and Bosco Chiesanuova(VR). These are all mountain stations where the raw model has poor performances that are difficult to correct. Note that results for the station in Feltre(BL) have some issues: the corrected model underestimates concentrations (also the correlation is not good). This is due to the morphological shape of the territory near the city where pollutants tend to reach high concentrations even if at 300 msl. In general spatial interpolation models work well for stations on plains, especially for stations that are geographically close to other measurement sites (for example in the area of Padua there is a cluster of stations). This is because they share similar altitudes and similar errors in model estimations. The complex landscape of mountain territory and the minor number of mountain stations are limitations to the performance of spatial interpolation. An example to mention regards the station of Belluno and Pieve di Alpago. These two stations are $\sim 10\ km$ apart but they sample very different areas in terms of pollutants concentration: while Belluno

is a city, Pieve d'Alpago is a small town with less anthropic activities and emissions. Despite the levels of PM_10 being much smaller in Pieve D'Alpago(BL), the raw model estimates similar concentrations at the two locations and geospatial interpolation does not manage to correct it because the station in Pieve d'Alpago highly depends on the value of Belluno, since they are close. The problem is probably due to the difficulty of the meteorological prediction (input of CAMx) to model sufficiently well this mountain area, operating at a coarse resolution of 5 km.

The overall performance of the geospatial interpolation methods is similar. Despite being much more complex methods, Kriging and Gaussian Process regression do not outperform simpler geometric methods such as IDW and RBF. This is probably due to the limited number of stations used. Only 20 stations do not seem to be sufficient to train complex models, too few points are used to retrieve the covariance structure of the data to fit the variogram in Kriging and kernel in GPR. All be considered, a safe choice to be implemented in ARPAV is the combination of linear regression and inverse distance weighting. It is a very simple method and the IDW part is already used by ARPAV. The introduction of the regression can avoid bias in model output and boost the performance.

The results are visualized in Fig. 4.13, where we show the time series of different methods vs measurements for ten stations. The series of data reported are the raw model, the classic inverse distance weighting and the best method among those compared in this paragraph (in this case Regression IDW).

Finally, Fig 4.12 shows the steps to correct the raw model for a specific day, using all the stations. In particular, the four images show the initial model estimation over the whole domain, the results after the application of linear regression with two predictors, the residual "field" computed by IDW and finally the sum of the regression and the residuals. The fourth panel contains the final corrected estimations of the concentrations of PM_{10}. The residual plot in the third panel presents the characteristic shape of IDW residuals, with high absolute values near station locations. It is worthwhile to mention that the final results are not reliable for the whole model domain. In fact, far from stations, the effectiveness of data fusion techniques is doubtful. Strictly speaking, geospatial interpolation methods can be successfully applied only in the areas near stations.

Station	**model**	**idw**	**reg+idw**	**RBF**	**RK**	**GPR**
BL_belluno	45.27	26.40	21.71	20.29	21.26	22.42
BL_feltre	38.79	28.49	29.42	28.56	28.04	27.74
BL_pieve	100.26	47.07	46.35	45.38	45.18	46.88
PD_p_colli	39.60	19.21	16.17	16.80	17.25	17.13
PD_este	34.01	12.55	9.98	11.00	10.43	11.56
RO_badia	29.51	18.92	15.40	10.31	13.61	15.39
RO_adria	31.42	26.20	28.49	31.04	28.53	27.34
RO_borsea	28.50	18.91	16.99	15.84	16.94	17.17
TV_conegliano	67.83	44.85	41.71	41.55	42.53	42.93
TV_mansue	24.72	17.02	13.23	13.44	13.52	14.80
TV_lancieri	29.55	14.90	12.19	11.39	11.84	11.72
VE_san_dona	23.65	16.74	13.90	12.98	13.70	13.57
VE_bissuola	25.16	14.97	11.93	10.99	11.66	12.00
VI_schio	57.28	26.52	26.44	26.97	25.97	25.66
VI_qitalia	33.17	15.57	14.87	14.43	13.53	14.32
VR_bcnuova	51.80	60.90	59.30	59.59	61.56	61.63
VR_legnago	30.88	14.60	11.43	11.98	12.85	12.79
VR_giarol	25.11	14.03	13.99	13.93	14.23	14.82
PD_mandria	30.43	13.53	14.28	15.15	13.25	14.05
PD_Monselice	38.65	14.47	11.15	12.45	11.26	11.92

Table 4.1: PM_{10}: NME results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	21.93	-21.34	3.13	4.47	7.63	6.13
BL_feltre	9.48	-14.19	-22.91	-22.26	-21.68	-21.53
BL_pieve	93.40	45.98	18.56	19.08	22.69	24.03
PD_p_colli	27.94	7.51	10.41	10.31	12.97	12.14
PD_este	20.21	-0.96	-1.23	-3.07	2.46	-0.99
RO_badia	-9.54	-20.93	-12.62	-8.11	-10.88	-12.44
RO_adria	14.89	6.03	28.00	31.04	28.14	26.90
RO_borsea	-12.52	-23.74	-15.74	-15.05	-15.71	-16.53
TV_conegliano	63.95	47.42	37.84	38.04	39.18	39.81
TV_mansue	4.40	-20.00	-6.10	-6.99	-4.20	-4.46
TV_lancieri	19.41	5.93	-8.09	-7.49	-7.42	-7.34
VE_san_dona	-2.77	-17.41	-7.20	-5.09	-6.21	-5.59
VE_bissuola	3.95	-9.78	-4.99	-2.77	-4.78	-5.08
VI_schio	47.21	25.77	15.47	17.70	16.19	15.39
VI_qitalia	22.96	7.41	-10.81	-10.82	-9.66	-9.94
VR_bcnuova	14.26	-9.48	-33.92	-33.19	-34.53	-33.67
VR_legnago	1.55	-7.58	0.86	2.84	1.11	0.09
VR_giarol	-3.29	-14.86	-0.83	-0.23	-1.30	-2.02
PD_mandria	19.54	3.40	-9.31	-9.97	-8.04	-9.16
PD_Monselice	27.51	10.21	5.16	6.02	7.30	6.54

Table 4.2: PM_{10}: NMB results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	0.35	0.73	0.76	0.78	0.78	0.75
BL_feltre	0.21	0.36	0.50	0.53	0.53	0.52
BL_pieve	0.49	0.75	0.60	0.62	0.61	0.60
PD_p_colli	0.65	0.92	0.92	0.91	0.93	0.92
PD_este	0.62	0.96	0.96	0.95	0.96	0.95
RO_badia	0.53	0.93	0.95	0.97	0.96	0.94
RO_adria	0.59	0.91	0.94	0.94	0.94	0.94
RO_borsea	0.56	0.94	0.95	0.97	0.95	0.96
TV_conegliano	0.62	0.74	0.80	0.81	0.80	0.81
TV_mansue	0.74	0.86	0.91	0.91	0.91	0.89
TV_lancieri	0.76	0.94	0.95	0.96	0.95	0.96
VE_san_dona	0.71	0.89	0.92	0.92	0.92	0.91
VE_bissuola	0.73	0.92	0.95	0.95	0.94	0.94
VI_schio	0.61	0.68	0.80	0.80	0.80	0.80
VI_qitalia	0.68	0.87	0.91	0.92	0.92	0.91
VR_bcnuova	0.46	0.32	0.42	0.42	0.38	0.38
VR_legnago	0.51	0.91	0.94	0.93	0.92	0.92
VR_giarol	0.69	0.85	0.90	0.90	0.90	0.89
PD_mandria	0.66	0.91	0.92	0.91	0.93	0.92
PD_Monselice	0.61	0.95	0.96	0.95	0.96	0.95

Table 4.3: PM_{10}: Correlation results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

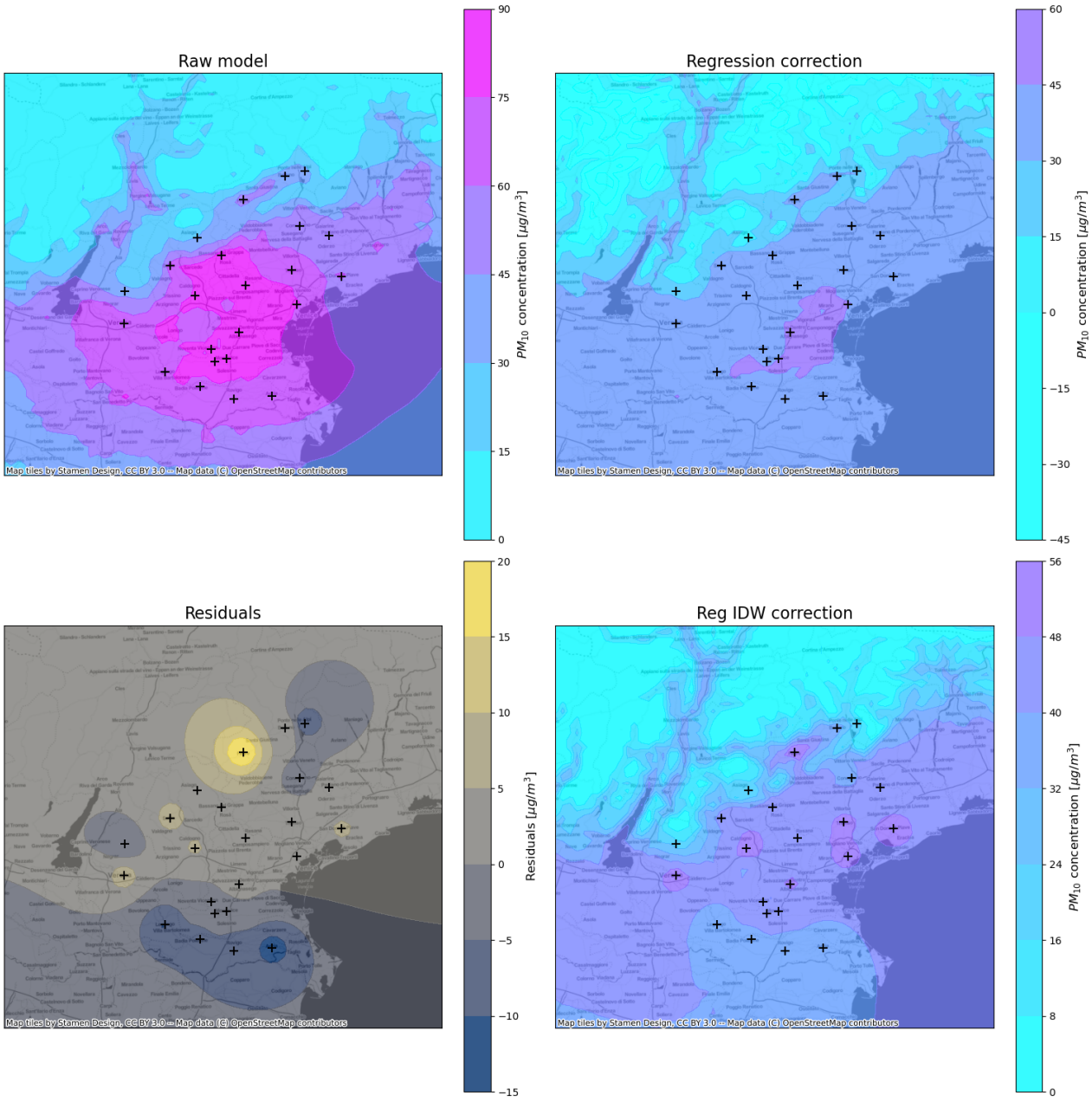


Figure 4.12: PM_{10}: model correction steps for 03/01/2022. The first panel shows raw model data, the second the results of linear regression, the third shows the residuals after IDW application and the last shows the final concentration estimations after correction.

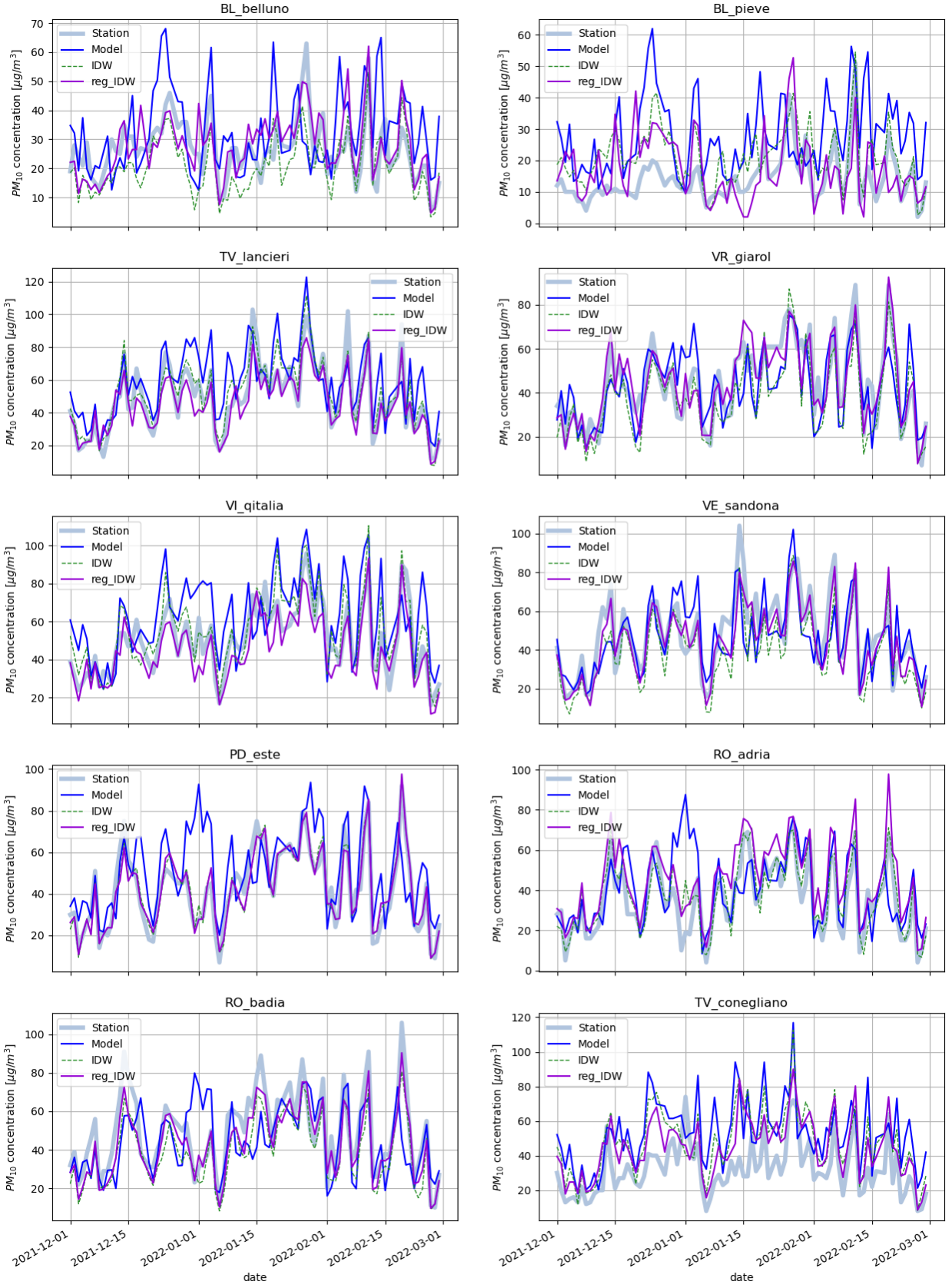


Figure 4.13: PM_{10}: Time series of station measures, model estimations and model estimations corrected by IDW and Regression IDW.

4.2 Nitric Dioxide

Raw model performance

Also in the NO_2 case, let us first analyze the raw model (Fig. 4.14). We can easily see that NO_2 estimations are not optimal, especially near the mountains. From the NMB plot, it is clear that the model strongly overestimates the concentrations. The worse performances of CAMx for Nitric Dioxide are probably caused by the fact that anthropogenic diffusive emissions, which are used as inputs to the model, are all allocated to the first vertical layer ([37]) and not distributed in the higher ones. This affects specifically nitric dioxide more than particulate matter and ozone because NO_2 forms very quickly in the atmosphere through oxidation of NO emitted from the sources.

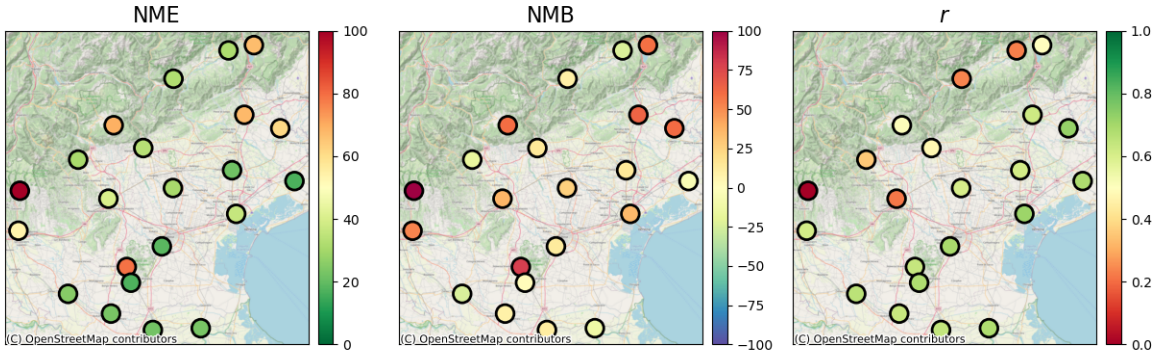


Figure 4.14: NO_2: NME, NMB and correlation (r) at each station location for model estimations

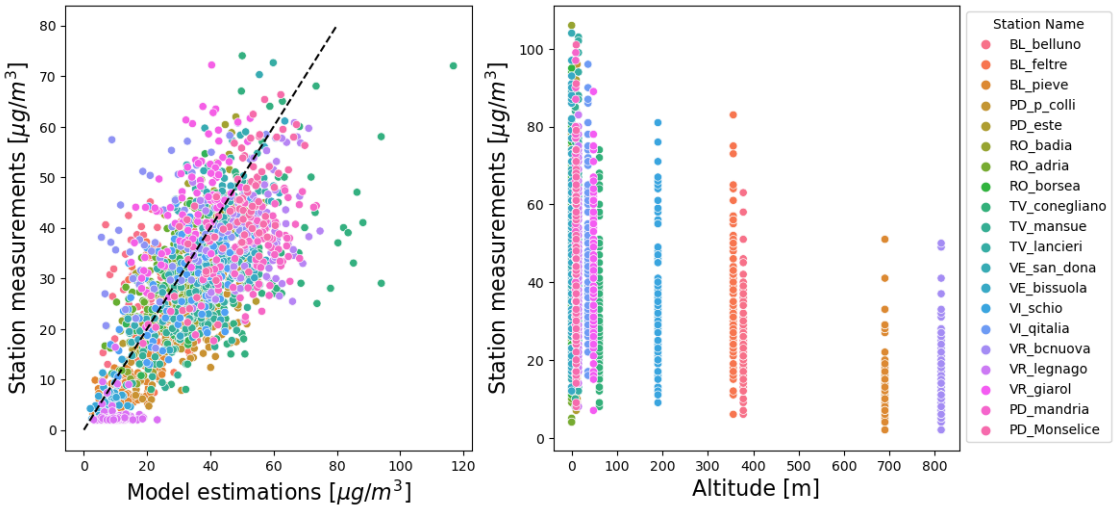


Figure 4.15: NO_2: Scatterplots of measures and estimations at station locations and measures and altitudes

Note that Bosco Chiesanuova(VR) has very poor performance. The scatterplot of the dataset (Fig. 4.15) and the time series plot (Fig. 4.16) reveal something strange about that station. There is, in fact, a big discrepancy between measurement and model, with the model that strongly overestimates the measurements. This kind of problem cannot be solved by any kind of data fusion method, and since this difference can affect the interpolation procedures, we are going to ignore the station for the next analysis.

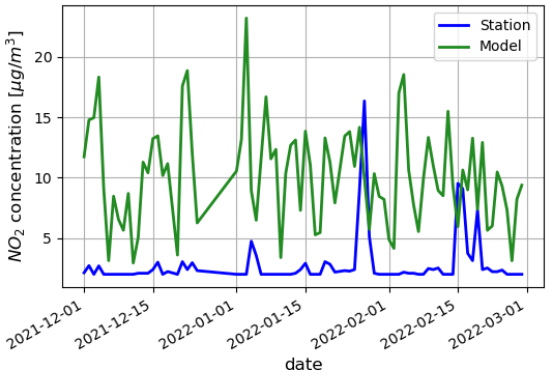


Figure 4.16: NO_2: Time series of measurements and estimations at Bosco Chiesanuova (VR)

Regression

The same linear regression methods applied for PM_{10} are used also in this case and, similarly, the linear regression with altitude as the second predictor has overall the best performance. The Box-Cox transformation is also implemented with λ parameters shown below:

- For the linear model $S(s) = a + b \cdot M(s)$, $\lambda = 0.79 \pm 0.02$.
- For the linear model $S(s) = a + b \cdot M(s) + c \cdot h(s)$, $\lambda = 0.45 \pm 0.02$

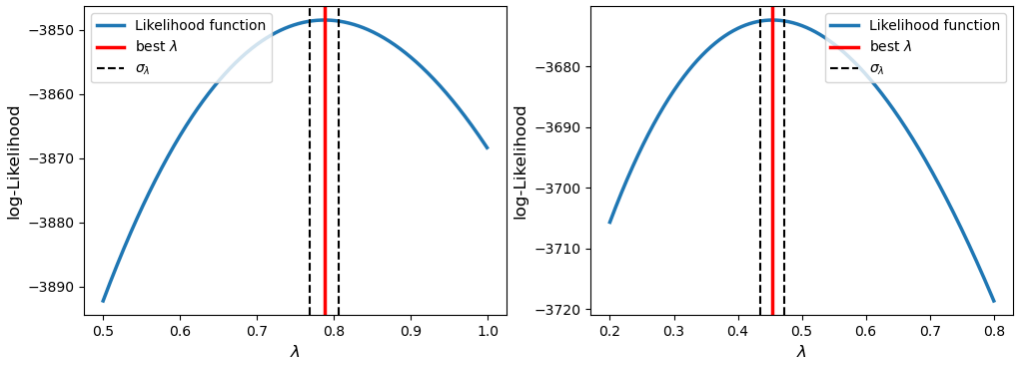


Figure 4.17: NO_2: Box-Cox likelihood function and λ best estimation for the two different linear regression models

In this case, using the Box-Cox transformation slightly improves the performance of the linear regression with two predictors. Note that from Tab. A.17, all the regression models perform very similarly. Our choice of linear regression method to be used for future analysis is based on the fact that the Box-Cox, two-predictor regression improves results for a larger number of stations. The results are plotted in Fig. 4.18.

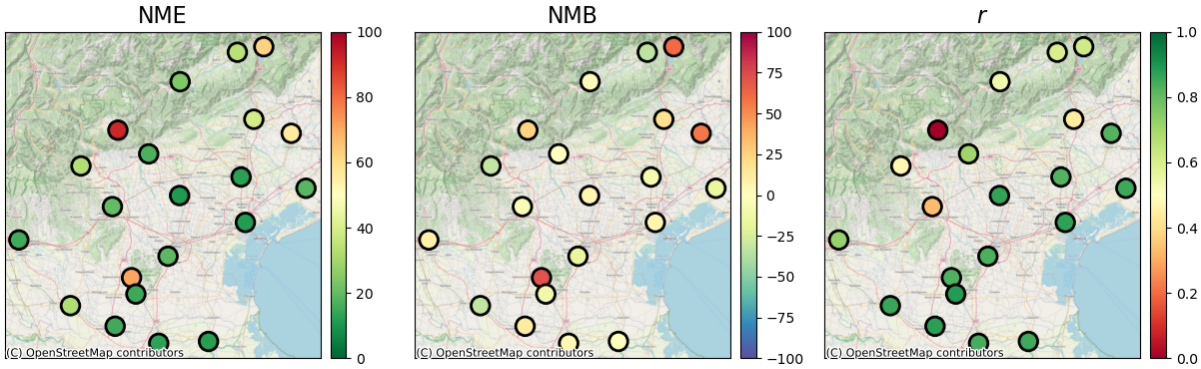


Figure 4.18: NO_2:NME, NMB and correlation (r) at each station location for Linear Regression results

Linear regression partially removes the overestimation, reducing NME and NMB for most stations.

Inverse Distance Weighting

In Fig. 4.19 the result of simple IDW interpolation is shown. This method works for almost every station, increasing the performance of stations at low altitudes. For Regression IDW, the same method applied to the PM_{10} data is followed. Regression with two predictors improves the simple IDW performances for many stations but there are some (Pieve d'Alpago, Asiago, Padova colli (PD)) where pollutant concentration is still highly overestimated (see Tab. A.20,A.21,A.22).

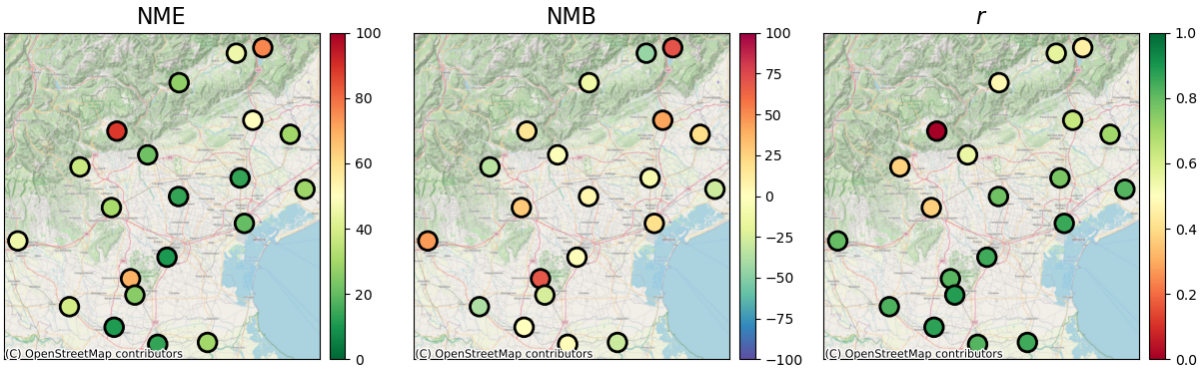


Figure 4.19: NO_2:NME, NMB and correlation (r) at each station location for IDW results

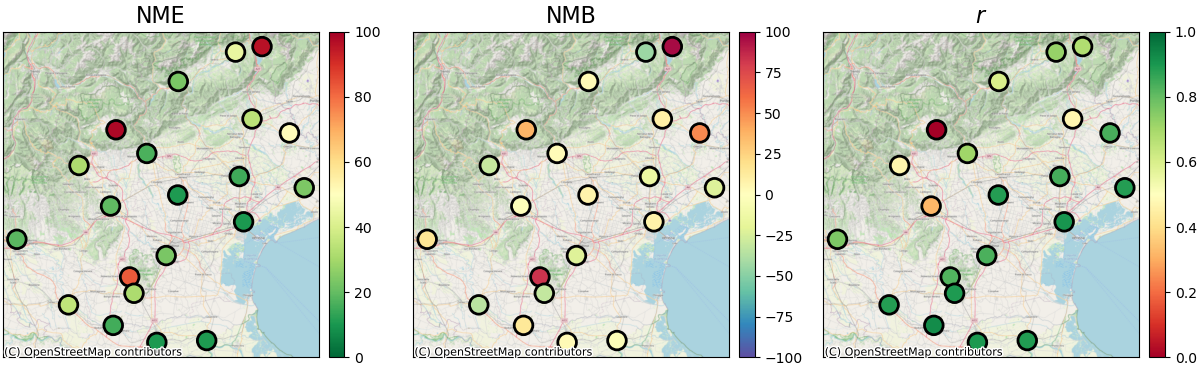


Figure 4.20: NO_2:NME, NMB and correlation (r) at each station location for Regression IDW

Radial Basis Function interpolation

RBF method is also applied to NO_2 data. Among the radial basis function tested, the best performing one is Gaussian RBF with a scale parameter $\varepsilon = 10^{-3}$. The usage of a smaller scale parameter is probably due to the fact that nitric dioxide tends to be highly concentrated near the sources of emission like streets and cities and the scale of variation of its concentration is smaller. RBF interpolation has slightly better performance than the previous method.

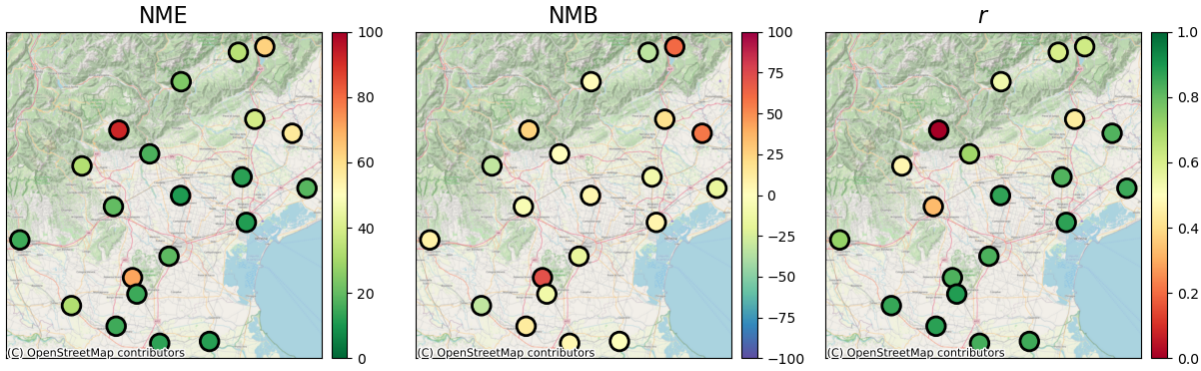


Figure 4.21: NO_2:NME, NMB and correlation (r) at each station location for Regression and RBF interpolation results

Regression Kriging and Gaussian Process Regression

For nitric dioxide, regression Kriging (Fig. 4.23) and Gaussian Process regression results (Fig. 4.24) are very similar between each other and similar with RBF interpolation. However, they are slightly better than the IDW methods. The best regression kriging variogram among those tested is the Gaussian variogram.

Concerning GPR, the Matern kernel is the final choice, slightly outperforming the RBF kernel. Also in this case one can plot the distribution of the length scale parameter fitted in Cross-Validation; but for NO_2, the distribution (4.22) is not clearly peaked at some value. This is an index that the correlation structure of the data is very weak and the l parameter is going to zero for many sets of points in Cross-Validation. This is not a surprise, for example, stations like Asiago(VI), or Colli Euganei are not correlated with the station nearby and are difficult to be incorporated into spatial interpolation methods.

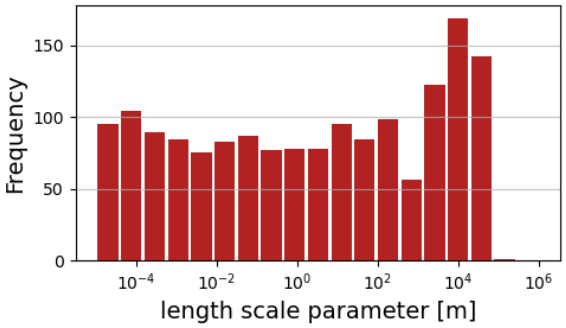


Figure 4.22: NO_2: Distribution of length scale parameter in Cross-Validation

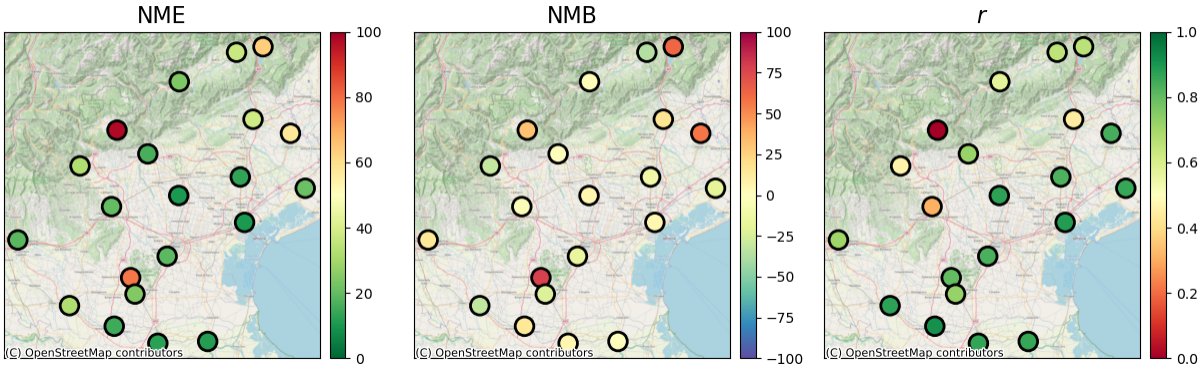


Figure 4.23: NO_2: NME, NMB and correlation (r) at each station location for Regression Kriging results

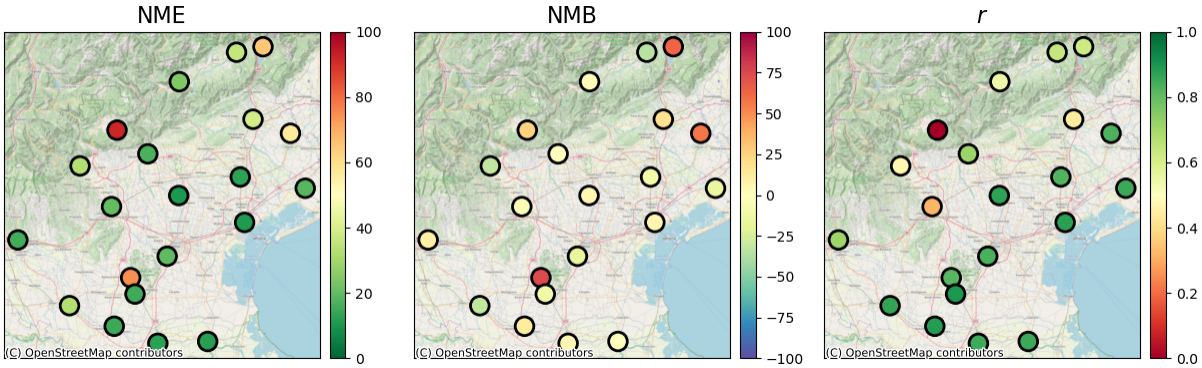


Figure 4.24: NME, NMB and correlation (r) at each station location for Regression and GPR results

Method comparison

Nitric Dioxide results, reported in Tab 4.4, 4.5, 4.6, are not as good as particulate matter ones. Nevertheless, the geostatistical interpolation methods meet the estimator criteria for stations in the plains, which are the most important because it is where the value of the concentration can reach legal limits. On the other hand, there are some stations where the error is very big and very difficult to correct. For some stations (*e.g.* Asiago) the spatial interpolation methods degrade the model performances. The station on Colli Euganei (PD) has bad indicators, with the raw model that overestimates the concentration of nitric dioxide. On the contrary, the model has good performance for nearby stations in Este(PD) and Padova. This discrepancy among close stations affects negatively the spatial interpolation. Concerning the comparison among applied methods, in the NO_2 case, RBF, Regression Kriging and GPR are slightly better than IDW method. Nevertheless, considering the small differences in results may not justify the implementation of a new method bt ARPAV.

Also in this case in Fig. 4.26 we report the time series of different methods vs measurements. Note that in the plots, the corrected model via Regression Kriging (the most performant spatial interpolation model) is usually the most similar to the Station series. Finally, we report the plot showing the final correction result on a single day in Fig.4.25. It is important to mention that, in the second and third panels, quantities are transformed using Box-Cox, so they are not directly comparable with the concentrations. From this plot, we can see that Regression Kriging returns small corrections, positive for the mountain area and negative for the southern area of the region. Nevertheless, for the final results, the major contribution is given by the regression.

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	31.64	45.97	44.39	33.98	36.84	35.80
BL_feltre	32.31	26.92	23.29	24.21	23.62	24.20
BL_pieve	67.06	76.14	96.12	62.63	63.37	64.64
PD_p_colli	79.05	68.89	82.89	71.44	78.26	74.90
PD_sgiust	30.61	13.47	10.78	10.56	10.35	10.36
PD_este	17.07	25.76	31.62	14.64	24.40	14.77
RO_badia	22.96	10.62	15.33	14.84	15.10	14.78
RO_adria	22.90	29.32	10.81	11.92	11.71	11.92
RO_borsea	22.26	13.30	10.25	12.63	12.15	12.42
TV_conegliano	67.83	49.74	34.07	38.74	38.16	38.73
TV_mansue	61.03	29.32	50.90	56.77	57.26	57.01
TV_lancieri	21.66	13.67	14.86	12.30	13.02	12.62
VE_san_dona	16.62	28.61	23.58	18.00	20.71	18.29
VE_bissuola	35.98	20.19	10.34	11.01	10.26	10.86
VI_asiago	69.49	88.64	98.34	91.84	97.66	91.95
VI_bassano	33.52	21.74	16.64	16.48	16.28	16.46
VI_schio	30.62	37.00	31.51	32.58	32.12	32.55
VI_qitalia	39.86	30.56	19.25	19.59	19.70	19.69
VR_legnago	25.37	38.87	34.85	32.22	32.21	32.43
VR_giarol	53.66	45.93	18.70	15.20	18.38	15.50
PD_mandria	18.46	10.28	23.09	18.92	18.86	19.17

Table 4.4: NO_2: NME results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	-23.74	-45.81	-44.39	-33.91	-36.80	-35.74
BL_feltre	10.04	-12.22	3.47	1.76	1.68	1.77
BL_pieve	59.25	70.97	96.12	61.40	62.61	63.40
PD_p_colli	78.95	68.89	82.89	71.44	78.26	74.90
PD_sgiust	25.98	6.10	7.49	6.16	6.16	6.38
PD_este	2.14	-25.70	-31.62	-11.55	-22.82	-12.66
RO_badia	9.26	-0.39	14.30	12.12	14.16	12.37
RO_adria	-12.91	-29.05	-3.68	-1.82	-1.66	-1.79
RO_borsea	11.22	2.62	3.67	5.40	6.21	5.42
TV_conegliano	63.95	42.80	9.94	18.41	17.15	18.20
TV_mansue	60.92	19.67	50.90	56.77	57.26	57.01
TV_lancieri	14.14	-7.65	-14.20	-9.90	-11.31	-10.30
VE_san_dona	-5.44	-28.31	-23.35	-16.51	-20.00	-16.97
VE_bissuola	35.51	19.44	7.88	6.28	5.50	6.35
VI_asiago	59.87	16.08	36.84	24.42	31.26	25.23
VI_bassano	14.33	-3.90	2.61	0.62	1.06	0.76
VI_schio	-16.96	-35.12	-30.89	-32.13	-31.64	-32.11
VI_qitalia	37.34	28.23	-2.01	-5.50	-4.46	-5.22
VR_legnago	-24.85	-38.87	-34.85	-32.22	-32.21	-32.43
VR_giarol	52.91	45.88	16.32	9.34	14.55	9.82
PD_mandria	13.13	-2.66	-22.42	-17.41	-17.48	-17.69

Table 4.5: NO_2: NMB results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	0.23	0.57	0.73	0.60	0.65	0.64
BL_feltre	0.24	0.47	0.60	0.54	0.57	0.54
BL_pieve	0.50	0.45	0.68	0.62	0.66	0.62
PD_p_colli	0.64	0.82	0.83	0.83	0.80	0.82
PD_sgiust	0.60	0.78	0.89	0.87	0.87	0.88
PD_este	0.68	0.89	0.89	0.89	0.73	0.90
RO_badia	0.64	0.87	0.92	0.88	0.91	0.89
RO_adria	0.68	0.85	0.89	0.85	0.86	0.85
RO_borsea	0.64	0.82	0.89	0.84	0.87	0.85
TV_conegliano	0.62	0.63	0.47	0.44	0.44	0.44
TV_mansue	0.73	0.70	0.84	0.83	0.84	0.83
TV_lancieri	0.61	0.77	0.84	0.83	0.83	0.83
VE_san_dona	0.69	0.82	0.89	0.85	0.86	0.86
VE_bissuola	0.71	0.86	0.91	0.87	0.89	0.88
VI_asiago	0.51	-0.21	-0.20	-0.21	-0.22	-0.21
VI_bassano	0.48	0.55	0.71	0.71	0.72	0.71
VI_schio	0.35	0.37	0.47	0.47	0.46	0.47
VI_qitalia	0.22	0.37	0.32	0.33	0.31	0.32
VR_legnago	0.66	0.83	0.88	0.86	0.87	0.87
VR_giarol	0.61	0.80	0.76	0.73	0.71	0.71
PD_mandria	0.69	0.85	0.83	0.83	0.83	0.83

Table 4.6: NO_2: Correlation results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

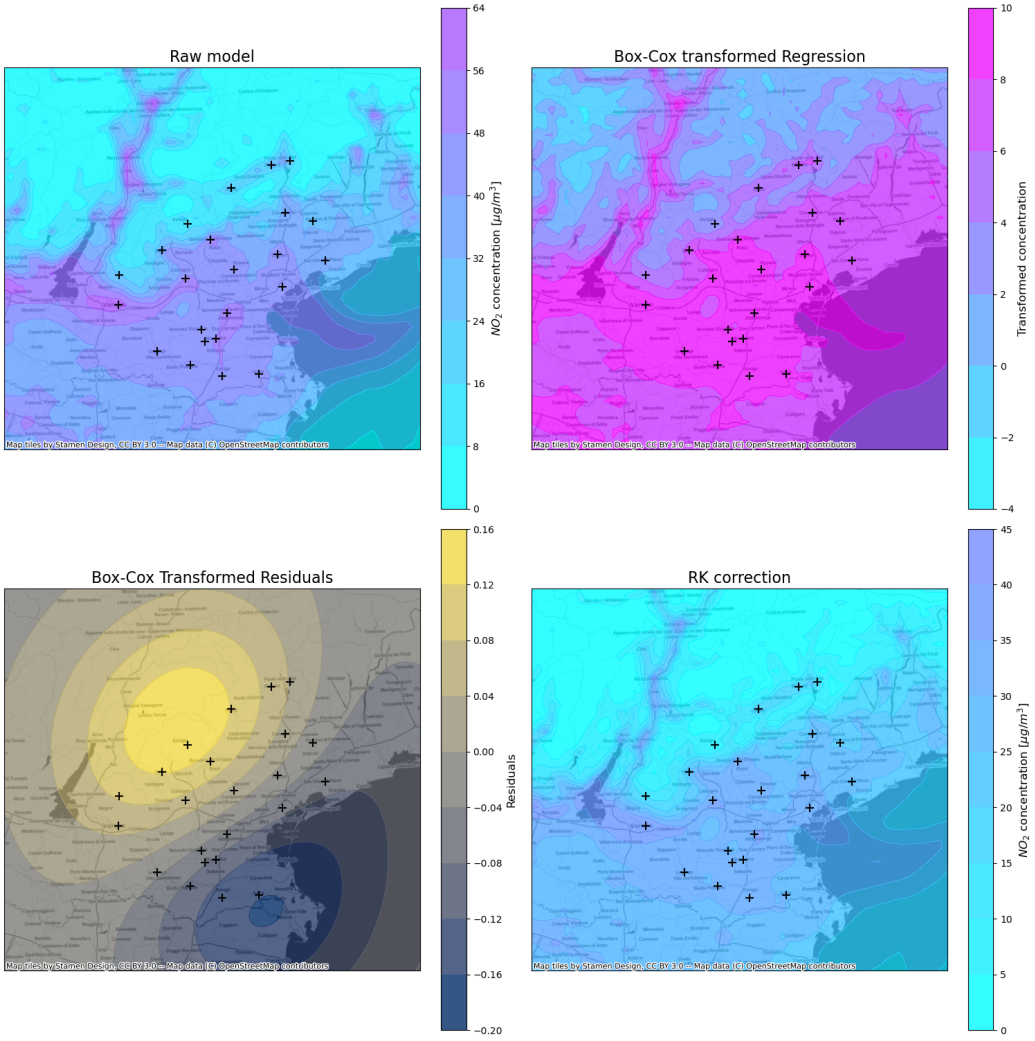


Figure 4.25: NO_2: model correction steps for 03/01/2022. The first panel shows raw model data, the second the results of linear regression, the third shows the residuals after RK application and the last shows the final concentration estimations after correction.

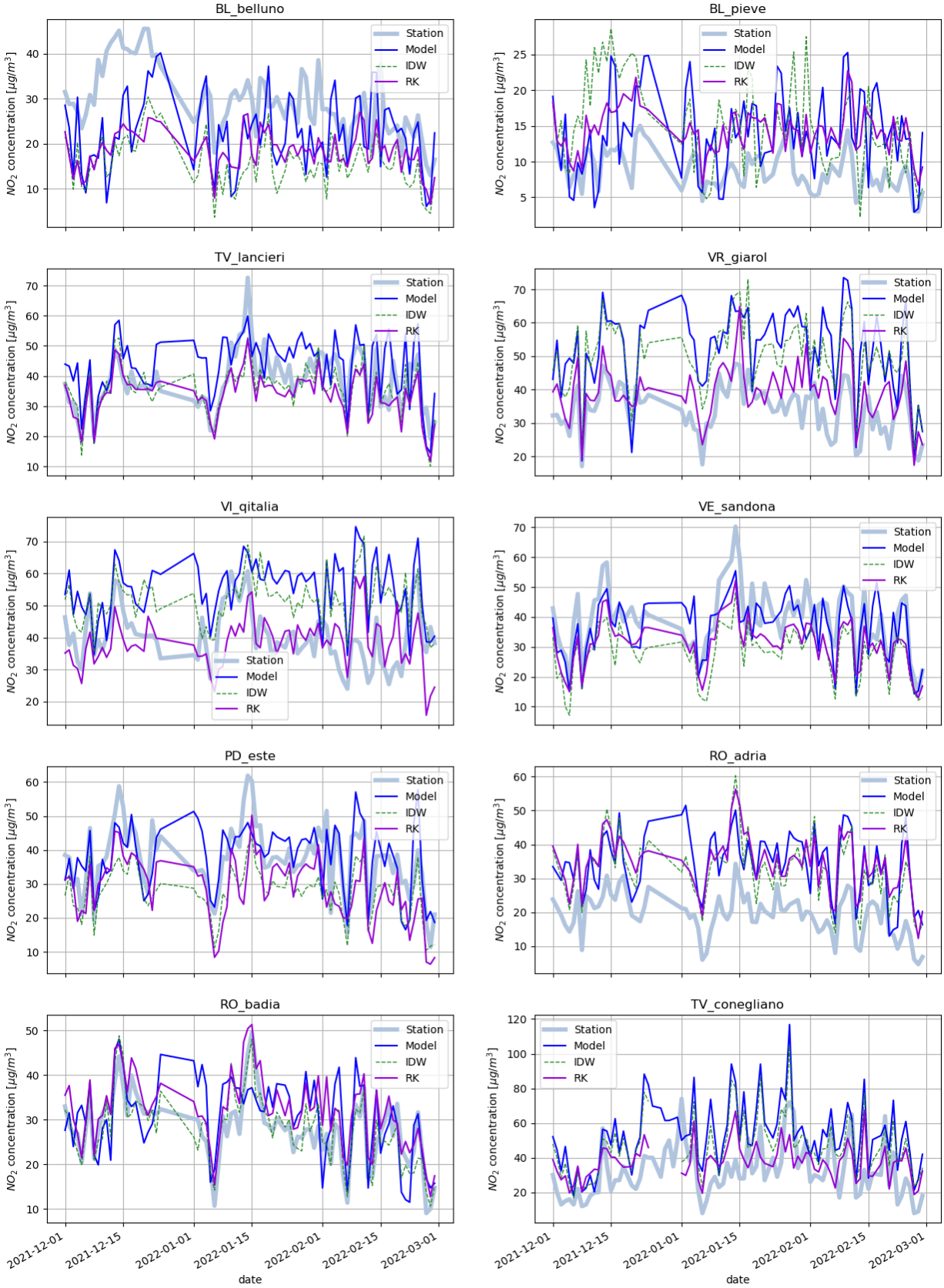


Figure 4.26: NO_2: Time series of station measures, model estimations and model estimations corrected by IDW and Regression Kriging.

4.3 Ozone

Raw model performance

Differently from the other two pollutants, the model estimation of Ozone is quite satisfactory without any interpolation method. Nevertheless, also in this case the model seems to overestimate the values (NMB plot in Fig. 4.27).

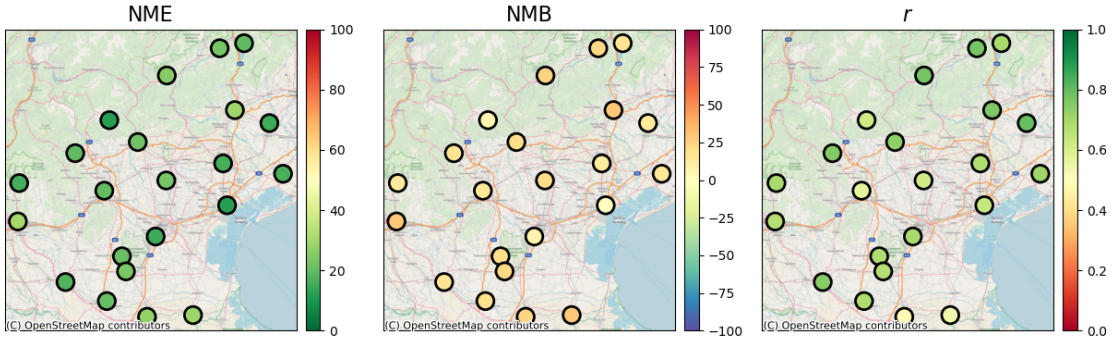


Figure 4.27: O_3: NME, NMB and correlation at each station location for model estimations

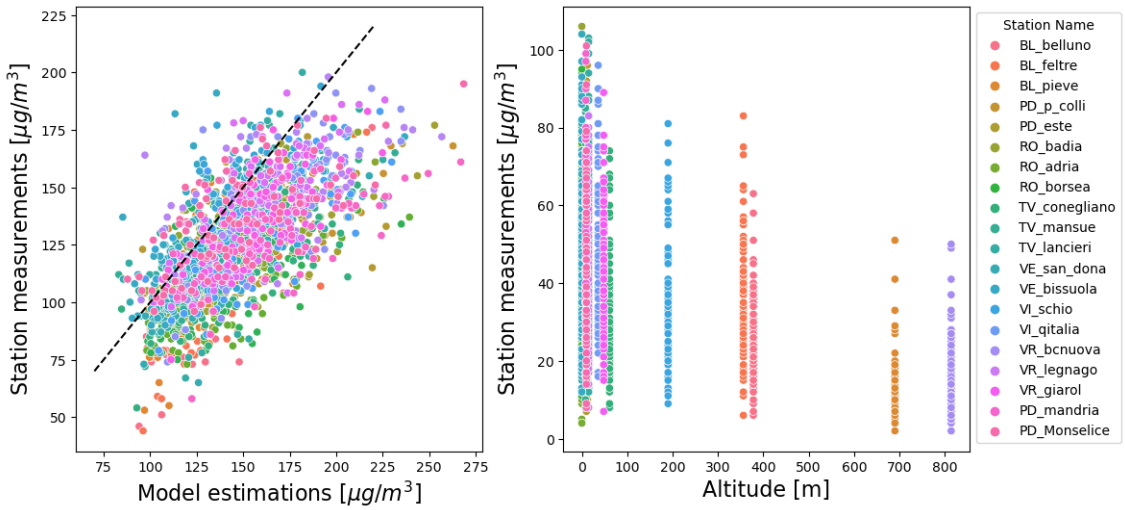


Figure 4.28: O_3: Scatterplots of measures and estimations at station locations and measurements and altitudes

Regression

In this case, the Box-Cox transformation has no impact on the data. In fact, estimating the λ via maximum likelihood the results are:

- For the linear model $S(s) = a + b \cdot M(s)$, $\lambda = 1.04 \pm 0.05$.
- For the linear model $S(s) = a + b \cdot M(s) + c \cdot h(s)$, $\lambda = 1.08 \pm 0.05$

When λ parameters are close to one the Box-Cox transformation becomes an identity. For this reason, for O_3 data, no power transformation is used. Among the two linear models, in this case, the one with better performance is a regression with only the model as a single predictor (Tab. A.32). The results are plotted in Fig. 4.30

Inverse Distance Weighting

IDW method and IDW with regression perform similarly (Tab. A.35, A.36, A.37). Note that in this case, the overestimation of the model seems to be corrected by the IDW method, even without the regression. Nonetheless, linear regression has an impact on the correlations that are generally better.

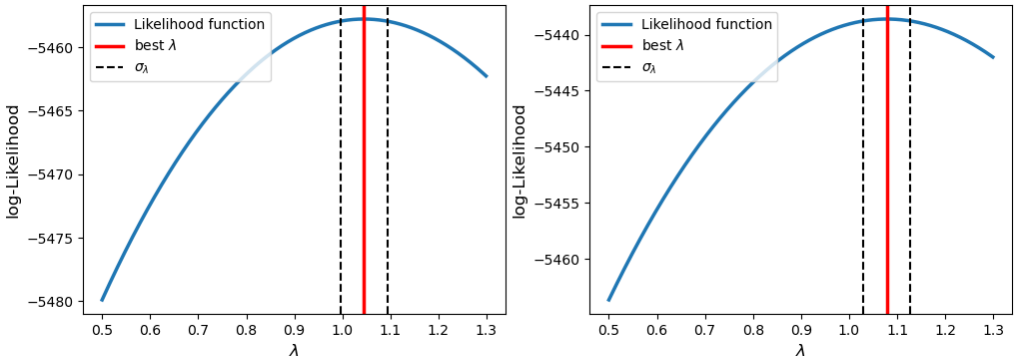


Figure 4.29: O_3: Box-Cox likelihood function and λ best estimation for the two different linear regression models

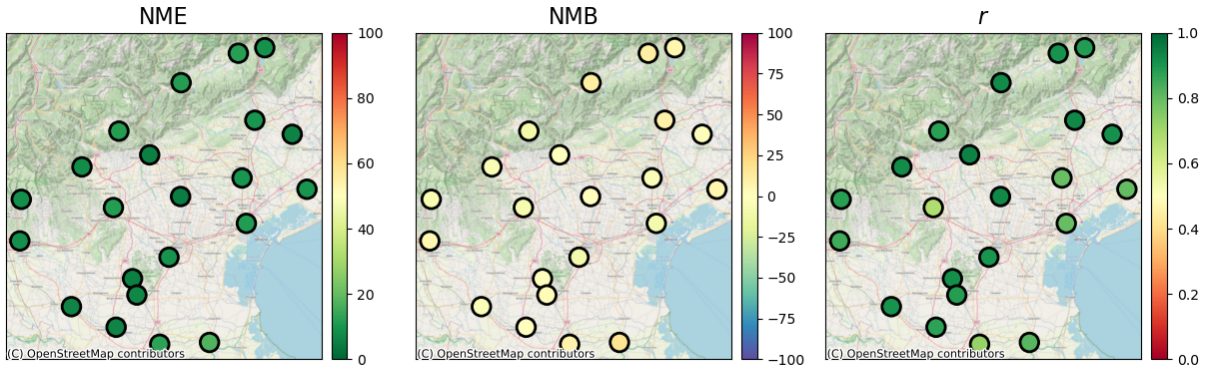


Figure 4.30: O_3: NME, NMB and correlation at each station location for Linear Regression results

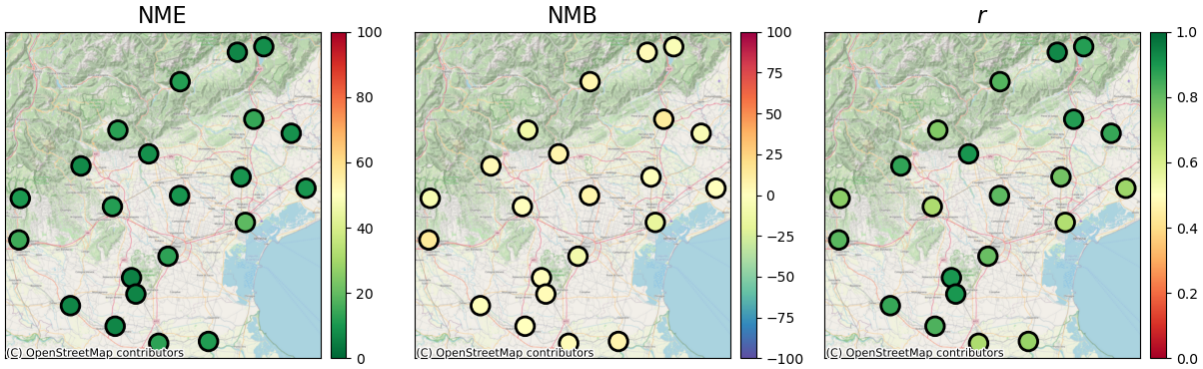


Figure 4.31: O_3: NME, NMB and correlation at each station location for IDW results

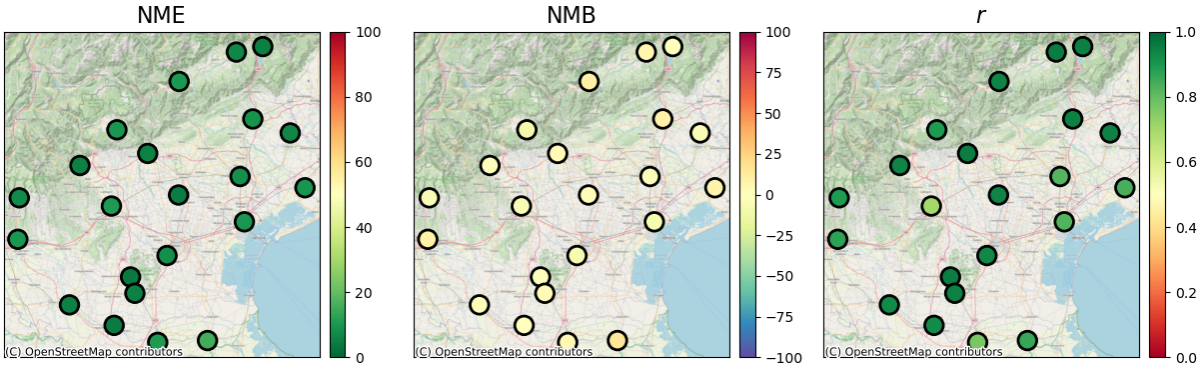


Figure 4.32: O_3:NME, NMB and correlation at each station location for Regression IDW results

Radial Basis Function interpolation

RBF method, with inverse multiquadric radial basis function, is used in Fig. 4.33. Other RBF functions (the inverse quadratic and Gaussian) perform similarly and the results are compatible with IDW model ones. Like the PM_{10} case the best scale parameter ε that maximize the results is 10^{-4} which has the same order of magnitude of the mean distance among stations. This guess of the characteristic length scale of the problem is confirmed, in the next paragraph, by the distribution of fitted length scale parameters in Gaussian Process Regression.

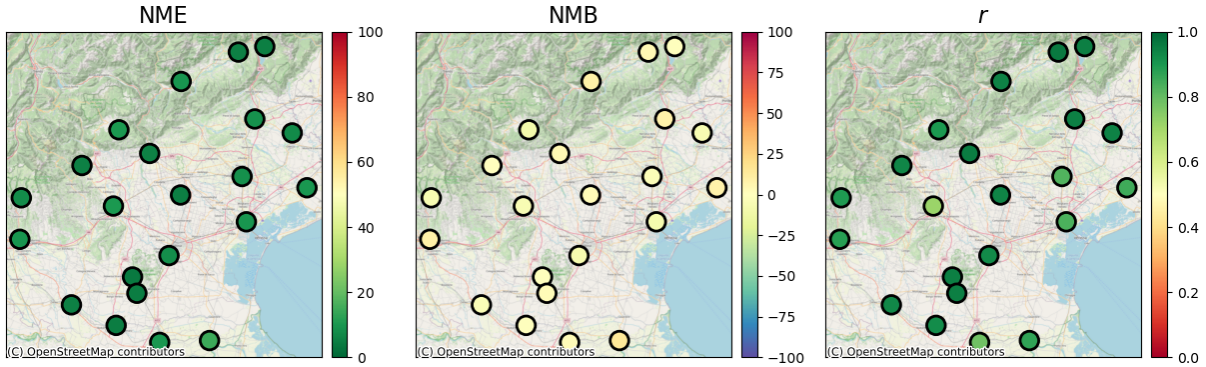


Figure 4.33: O_3: NME, NMB and correlation at each station location for Regression and RBF interpolation results

Regression Kriging and Gaussian Process Regression

Regression Kriging has good results independently on the variogram shape chosen. Here (Fig. 4.35) we report the results for a spherical variogram which is slightly better (Tab.A.41) for a large number of stations.

Gaussian Process Regression performances resemble the Kriging ones, also in this case the Mater kernel is the better choice over the RBF kernel. NME, NMB and correlation plot are shown in Fig. 4.36 and results are reported in Tab. A.44 A.45, A.46. Similarly to the Particulate matter analysis, the length scale parameter distribution, shown in Fig. 4.34, which quantifies the characteristic correlation length of the system, is peaked for $l = 10^4$.

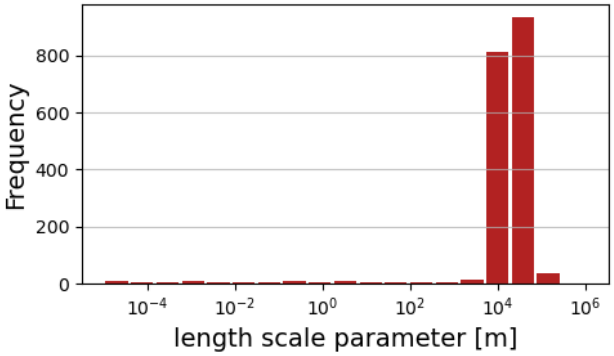


Figure 4.34: O_3: Distribution of length scale parameter in Cross-Validation

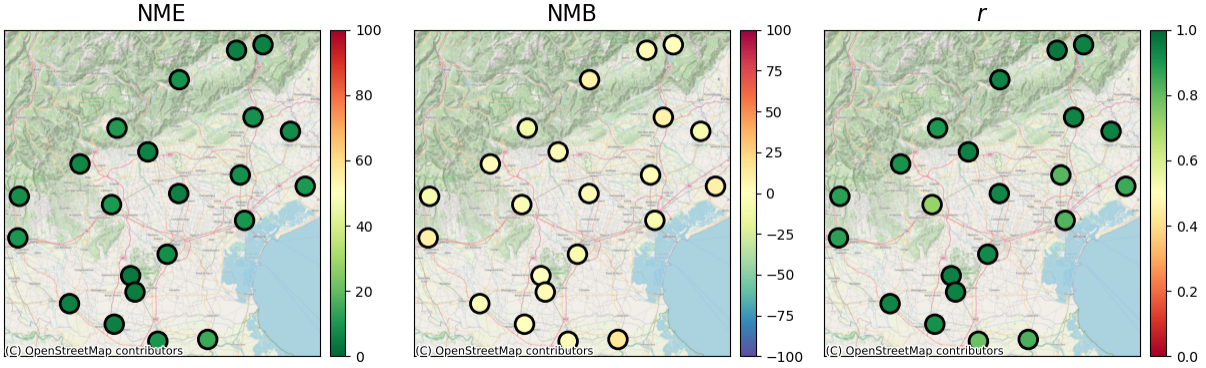


Figure 4.35: O_3: NME, NMB and correlation at each station location for Regression Kriging results

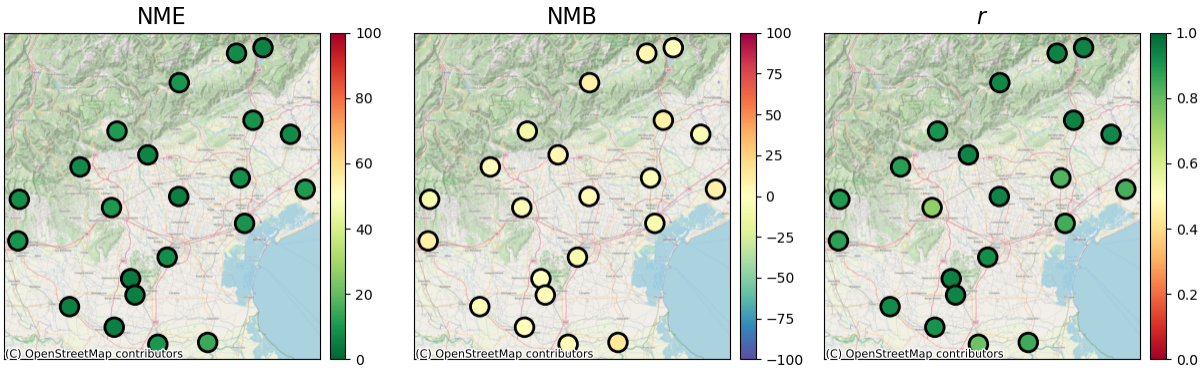


Figure 4.36: O_3:NME, NMB and correlation at each station location for regression and GPR results

Method comparison

Looking at Tab. 4.7, 4.8, 4.9, we can see the impact of geostatistical interpolation methods on predictions. The usage of simple IDW increases model performances and decreases the overestimation. However, results can be improved again, by applying the linear regression. Different spatial interpolation method results are similar, as happens for the other two pollutants. RBF seems to be the worst with two values of NME over 15 for the station of Adria(RO) and San Donà(Ve). Also in this case complex methods do not perform better than simpler methods, probably due to the low number of stations. Similarly to the PM_{10} case, the method suggested for the implementation is regression IDW. Despite the simplicity, its results are satisfying. In Fig. 4.26 the time series of Regression IDW is compared to raw model estimations and simple IDW.

We show also the effect of Regression IDW method correction to the model domain for a single day in Fig. 4.37. The residuals show the characteristic structure of IDW results, with high absolute values near the stations.

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	23.02	11.78	6.36	5.68	5.45	6.94
BL_feltre	25.55	11.36	10.41	10.88	9.20	10.02
BL_pieve	19.26	8.75	4.74	6.08	5.36	5.66
PD_p_colli	20.90	4.95	3.87	4.55	3.67	3.71
PD_sgiust	22.75	4.42	4.48	9.66	5.26	5.03
PD_este	23.40	6.39	5.10	4.39	4.65	4.64
RO_badia	20.23	5.99	4.28	5.49	4.87	4.85
RO_adria	28.82	17.29	15.32	17.95	14.67	14.67
RO_borsea	27.07	12.56	10.57	12.30	9.73	10.05
TV_conegliano	29.29	9.74	9.07	11.01	8.91	9.24
TV_mansue	15.00	5.67	6.41	9.07	7.49	7.19
TV_lancieri	16.38	9.51	8.54	11.32	8.60	8.54
VE_san_dona	16.73	9.38	9.75	18.76	10.53	10.76
VE_bissuola	12.03	11.02	9.93	9.25	9.67	9.40
VI_asiago	12.07	11.41	9.96	11.56	10.67	10.77
VI_bassano	21.92	4.82	5.22	9.00	5.70	6.19
VI_schio	19.46	7.53	5.27	13.04	6.37	7.23
VI_qitalia	19.65	11.62	10.57	13.04	10.30	10.22
VR_bcnuova	15.93	7.90	7.18	13.51	8.46	8.17
VR_legnago	17.22	6.04	5.25	9.33	5.36	5.39
VR_giarol	29.87	8.35	9.81	14.85	9.80	9.53
PD_mandria	15.08	9.02	8.17	8.62	7.96	8.21

Table 4.7: O_3: NME results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	22.21	10.42	5.52	-3.56	3.78	5.31
BL_feltre	24.73	10.77	9.88	8.89	8.32	9.20
BL_pieve	16.78	5.60	-1.06	2.97	-0.38	0.82
PD_p_colli	20.10	-1.61	-1.94	0.02	-0.27	-0.29
PD_sgiust	21.64	0.04	2.31	7.59	3.16	3.31
PD_este	22.48	2.62	3.80	1.05	3.18	2.92
RO_badia	19.29	0.24	-0.31	-0.71	-0.94	-1.42
RO_adria	28.36	16.78	15.13	8.06	14.37	14.26
RO_borsea	24.01	7.38	5.21	-2.84	2.36	2.70
TV_conegliano	29.04	9.20	8.77	8.57	8.21	8.91
TV_mansue	13.16	-2.92	-5.93	-8.48	-7.18	-6.60
TV_lancieri	12.30	-4.02	-3.43	-7.75	-3.72	-3.45
VE_san_dona	14.81	6.17	7.26	18.17	9.01	9.07
VE_bissuola	-1.36	-8.89	-7.90	-0.22	-7.89	-7.62
VI_asiago	5.77	-10.84	-9.30	-7.93	-10.17	-10.41
VI_bassano	20.83	-0.35	3.43	2.15	3.77	4.08
VI_schio	17.60	-5.28	-1.12	9.87	1.67	1.53
VI_qitalia	14.32	-8.53	-6.09	-6.01	-5.64	-5.80
VR_bcnuova	13.26	-6.50	-5.85	-12.84	-7.54	-7.22
VR_legnago	16.36	-3.96	-3.58	-5.08	-4.09	-3.69
VR_giarol	29.66	5.41	9.09	14.34	9.02	8.58
PD_mandria	8.04	-8.68	-7.85	-7.17	-7.65	-7.68

Table 4.8: O_3: NMB results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

Station	model	idw	reg+idw	RBF	RK	GPR
BL_belluno	0.77	0.91	0.96	0.96	0.96	0.94
BL_feltre	0.78	0.93	0.94	0.88	0.94	0.93
BL_pieve	0.69	0.89	0.95	0.94	0.94	0.94
PD_p_colli	0.68	0.92	0.95	0.92	0.95	0.95
PD_sgiust	0.61	0.93	0.94	0.88	0.93	0.94
PD_este	0.67	0.88	0.95	0.94	0.94	0.94
RO_badia	0.68	0.87	0.92	0.89	0.91	0.91
RO_adria	0.54	0.82	0.86	0.71	0.84	0.85
RO_borsea	0.48	0.72	0.77	0.66	0.79	0.79
TV_conegliano	0.76	0.93	0.95	0.89	0.94	0.95
TV_mansue	0.79	0.91	0.94	0.87	0.94	0.93
TV_lancieri	0.68	0.80	0.82	0.81	0.82	0.83
VE_san_dona	0.71	0.80	0.84	0.81	0.85	0.84
VE_bissuola	0.64	0.80	0.82	0.79	0.82	0.84
VI_asiago	0.62	0.88	0.89	0.86	0.90	0.90
VI_bassano	0.74	0.94	0.95	0.77	0.94	0.94
VI_schio	0.75	0.92	0.94	0.75	0.91	0.88
VI_qitalia	0.57	0.68	0.70	0.72	0.72	0.74
VR_bcnuova	0.69	0.88	0.89	0.83	0.88	0.89
VR_legnago	0.75	0.91	0.93	0.79	0.93	0.92
VR_giarol	0.67	0.85	0.87	0.74	0.88	0.87
PD_mandria	0.69	0.91	0.93	0.89	0.93	0.92

Table 4.9: O_3: Correlation results for raw model data, IDW, Regression IDW, RBF, Regression Kriging and GPR

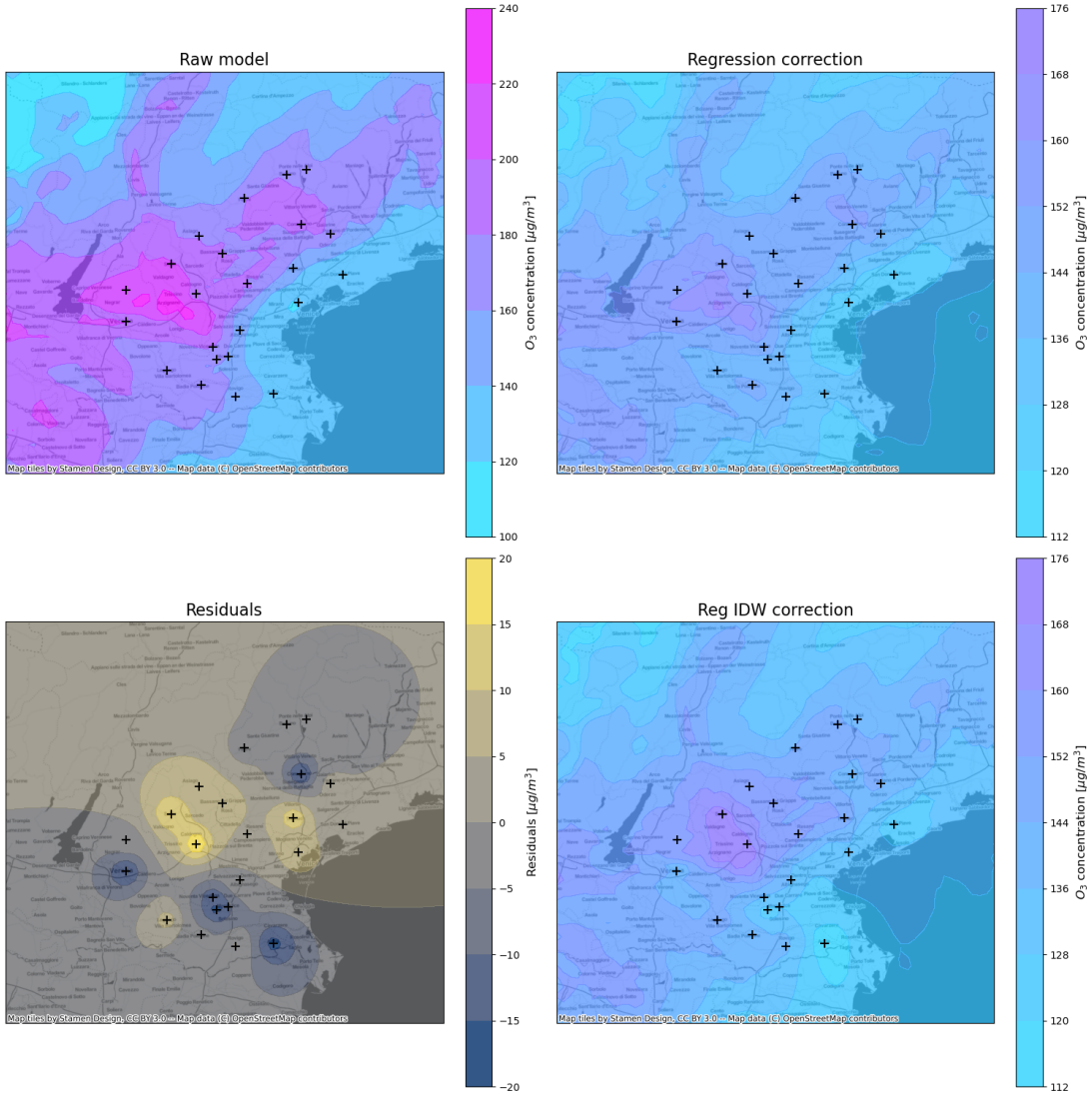


Figure 4.37: O_3: model correction steps for 12/07/2021. The first panel shows raw model data, the second the results of linear regression, the third shows the residuals after IDW application and the last shows the final concentration estimations after correction.

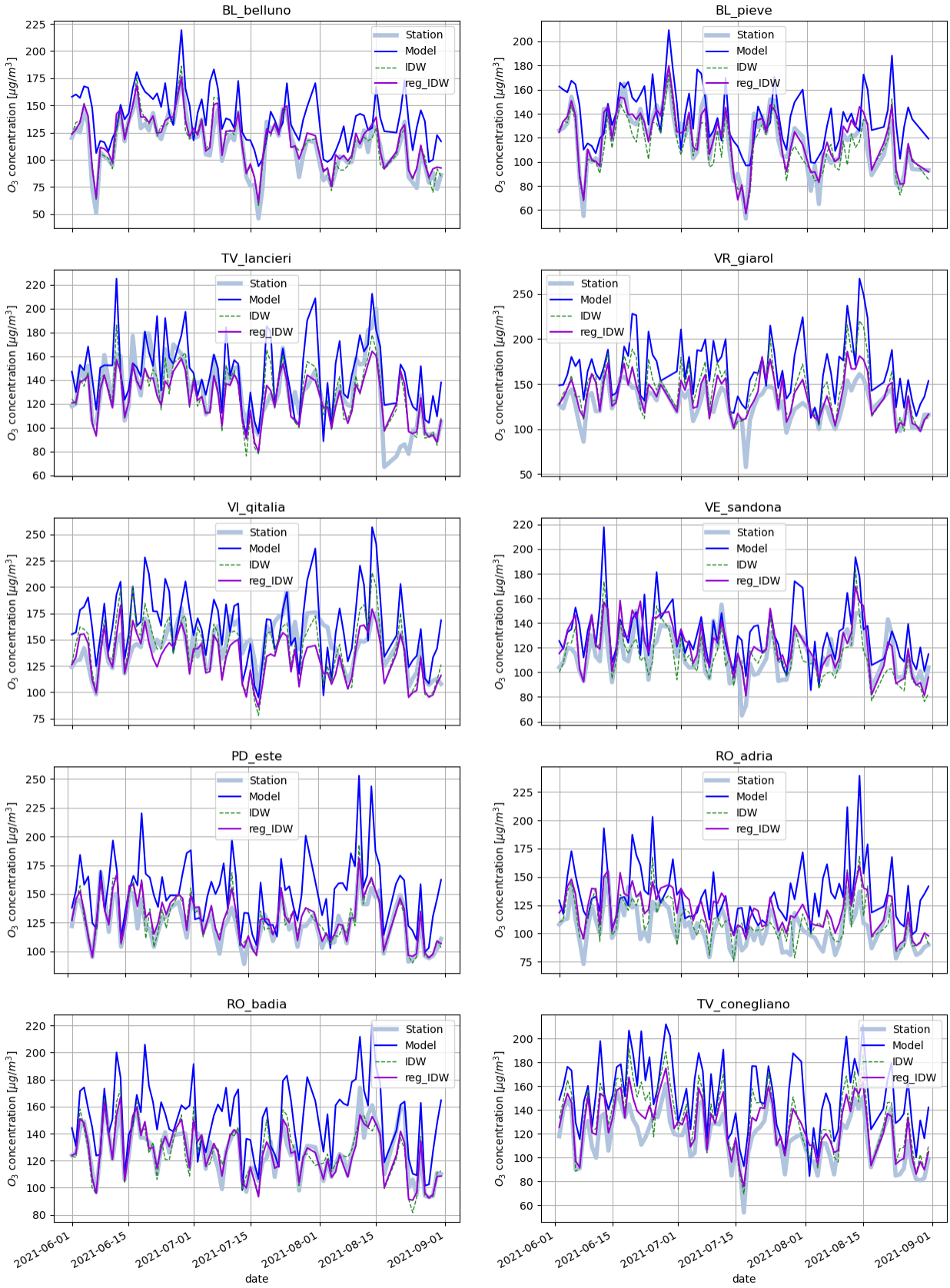


Figure 4.38: O_3: Time series of station measures, model estimations and model estimations corrected by IDW and Regression IDW.

Chapter 5

Conclusion

In this work different methods of "data fusion", to combine model predictions and measurements, were implemented, tested and compared. The analysis has been done for the specific case of air pollutants concentration data in the Veneto region, to increase the reliability of ARPAV estimations. The tested techniques vary from simple geometrical methods, (the simplest among them, IDW, is already implemented at ARPAV,) to more complex Geostatistic methods.

For each pollutant, the different interpolation models are compared using a Cross-Validation technique, with a Left-One-Out strategy. A common feature of all datasets is the fact that the model data tend to overestimate the concentrations. To solve this problem, the usage of linear regression before the residual interpolation was exploited with success. A summary of the findings for each pollutant follows:

- PM_{10}. Using "data fusion" methods increase significantly the performance of the model. Linear regression was applied to the raw data with good results, especially when using both the model estimation and the altitude as predictors. Different spatial interpolation methods perform similarly, despite the increasing complexity of the techniques. For this reason, the best method that we suggest is the simplest one, which is Linear Regression with two predictors and IDW on top of it.

- NO_2. The performance of the raw model is not optimal, especially for some stations in complex terrain so it is hard to improve the results with any postprocessing. However, a partial improvement was obtained using Regression Kriging and Gaussian Process.

- O_3. The raw estimations of Ozone concentration are already performing well, still, there is a systematic positive bias for most of the stations. As for PM_{10}, the application of linear regression and a simple interpolation of the residual as IDW improve the results. Also in this case there is no need for a more complex method since all spatial interpolation methods give very similar results.

We want to point out the fact that the more complex interpolation techniques (Kriging and GPR) do not always improve the results and their performances are very similar to the simpler ones, such as IDW method (or RBF). This result is probably a consequence of the small number of stations that are used. Further analysis can be done in order to explore different hyperparameters and different techniques, but the small amount of data limits the effectiveness of these methods.

Concerning possible future research work, there are two main ways to continue this project:

- The first one consists in increasing the model domain. The logical step is to apply the same methodologies to the whole Po Valley [38]. Of course, in order to do so, one has to have access to a model that runs over a much bigger domain and to the measurements of all the stations scattered in the region, which are of competence of the different regional ARPAs. Nevertheless, increasing the model domain should improves the performance of data fusion methods because the number of stations (data used for the interpolation) considered would be much higher (more

than 150 stations). Another advantage is the fact that the whole Po Valley is a closed and homogeneous area surrounded by mountains, on which the boundary conditions have less effect on the model estimation, in comparison with what happens in the case of the actual domain. Finally, CMT models such as CAMx work better for larger domains.

- On the other hand, some data assimilation (DA) [39] techniques could be successfully applied to the problem. In this work, all the proposed methods, are postprocessing tools applied to the raw model output and cannot, by any means affect how the model computation. Data Assimilation takes the interpolation of measurements with model estimations to the next logical level. The observational data are used to affect the model output in a way that accounts for the uncertainties in both, in order to estimate the possible state of the system. Data Assimilation is commonly used in Meteorological models but can be applied also to Air Quality models. The most important Data Assimilation techniques are Optimal Interpolation, Variational methods and Kalman filter methods ([20]). The application of data assimilation can, in principle, improve the results but the algorithms and models needed are much more complex to be implemented and to be studied.

Finally, the methodologies explored in this work can be the starting point to devise a technique to correct the forecast of the model for the upcoming days on the basis of Model Output Statistics (MOS) ([40]), widely used in weather forecasts.

Appendix A

Tables

In this Appendix, we report the tables of the statistical estimators in Sec. 3.3.2 computed for each implemented method. Tables are organized depending on the used technique. There are some acronyms used in this section for table readability that are not used in the rest of the thesis:

- **reg1**: Linear Regression with only model estimations as a predictor
- **reg2**: Linear Regression with model estimations and altitudes as predictors
- **BC**: Box-Cox transformation
- **gau**: gaussian, it can be a variogram or a radial basis function
- **invm**: inverse multiquadric radial basis function
- **invq**: inverse quadratic radial basis function
- **exp**: exponential variogram
- **sph**: spherical variogram

Remember that, even if it is not written RBF and GPR methods are always preceded by some kind of linear regression.

Station	model	reg1	reg1+BC	reg2	reg2+BC
BL_belluno	45.27	32.37	29.98	26.40	23.70
BL_feltre	38.79	27.83	27.87	28.49	31.27
BL_pieve	100.26	122.50	113.09	47.07	41.69
PD_p_colli	39.60	18.36	18.08	19.21	18.69
PD_este	34.01	14.57	15.62	12.55	12.59
RO_badia	29.51	24.79	26.45	18.92	19.06
RO_adria	31.42	18.26	18.64	26.20	26.00
RO_borsea	28.50	25.77	27.62	18.91	19.17
TV_conegliano	67.83	46.03	45.82	44.85	43.20
TV_mansue	24.72	19.35	20.97	17.02	16.77
TV_lancieri	29.55	14.57	15.36	14.90	15.31
VE_san_dona	23.65	21.17	22.56	16.74	16.87
VE_bissuola	25.16	17.59	18.60	14.97	15.11
VI_schio	57.28	30.71	30.73	26.52	24.72
VI_qitalia	33.17	15.45	16.67	15.57	16.18
VR_bcnuova	51.80	79.71	75.21	60.90	45.44
VR_legnago	30.88	18.19	19.88	14.60	14.82
VR_giarol	25.11	18.32	20.64	14.03	14.81
PD_mandria	30.43	13.62	14.57	13.53	14.12
PD_Monselice	38.65	16.30	17.24	14.47	14.19

Table A.1: PM_{10}: NME values for different linear regression models

Station	model	reg1	reg1+BC	reg2	reg2+BC
BL_belluno	21.93	21.88	15.54	14.21	4.86
BL_feltre	9.48	-3.99	-7.30	-18.36	-24.72
BL_pieve	93.40	122.50	113.09	40.25	36.17
PD_p_colli	27.94	9.30	7.31	14.05	12.87
PD_este	20.21	1.99	0.63	5.11	3.97
RO_badia	-9.54	-21.06	-23.29	-12.59	-13.08
RO_adria	14.89	6.75	2.84	24.56	23.79
RO_borsea	-12.52	-23.20	-25.50	-14.55	-15.13
TV_conegliano	63.95	38.44	37.41	39.05	37.05
TV_mansue	4.40	-9.03	-11.27	1.08	0.61
TV_lancieri	19.41	-4.44	-3.51	-7.26	-7.99
VE_san_dona	-2.77	-16.01	-18.12	-6.83	-7.21
VE_bissuola	3.95	-11.30	-12.61	-6.75	-7.40
VI_schio	47.21	19.23	17.07	10.80	6.11
VI_qitalia	22.96	-5.25	-4.12	-10.28	-10.81
VR_bcnuova	14.26	72.28	67.82	-39.04	-22.98
VR_legnago	1.55	-10.49	-13.23	-1.36	-2.36
VR_giarol	-3.29	-14.78	-18.24	-4.54	-6.14
PD_mandria	19.54	-4.30	-3.53	-6.99	-7.58
PD_Monselice	27.51	7.78	7.17	8.89	7.85

Table A.2: PM_{10}:NMB values for different linear regression models

Station	model	reg1	reg1+BC	reg2	reg2+BC
BL_belluno	0.35	0.69	0.69	0.73	0.74
BL_feltre	0.21	0.41	0.40	0.46	0.46
BL_pieve	0.49	0.81	0.83	0.71	0.77
PD_p_colli	0.65	0.90	0.90	0.92	0.92
PD_este	0.62	0.92	0.90	0.95	0.95
RO_badia	0.53	0.85	0.84	0.88	0.88
RO_adria	0.59	0.89	0.87	0.91	0.91
RO_borsea	0.56	0.87	0.86	0.89	0.89
TV_conegliano	0.62	0.66	0.63	0.73	0.74
TV_mansue	0.74	0.80	0.77	0.84	0.84
TV_lancieri	0.76	0.90	0.89	0.91	0.90
VE_san_dona	0.71	0.86	0.85	0.87	0.87
VE_bissuola	0.73	0.90	0.89	0.91	0.91
VI_schio	0.61	0.77	0.76	0.76	0.78
VI_qitalia	0.68	0.86	0.84	0.89	0.89
VR_bcnuova	0.46	0.65	0.68	0.38	0.57
VR_legnago	0.51	0.85	0.84	0.88	0.88
VR_giarol	0.69	0.90	0.89	0.90	0.90
PD_mandria	0.66	0.91	0.89	0.92	0.92
PD_Monselice	0.61	0.90	0.88	0.94	0.93

Table A.3: PM_{10}: Correlation values for different linear regression models

Station	model	reg2	idw	reg2+idw
BL_belluno	45.27	26.40	27.21	21.71
BL_feltre	38.79	28.49	31.10	29.42
BL_pieve	100.26	47.07	51.24	46.35
PD_p_colli	39.60	19.21	15.80	16.17
PD_este	34.01	12.55	10.71	9.98
RO_badia	29.51	18.92	21.89	15.40
RO_adria	31.42	26.20	17.55	28.49
RO_borsea	28.50	18.91	24.40	16.99
TV_conegliano	67.83	44.85	51.21	41.71
TV_mansue	24.72	17.02	22.90	13.23
TV_lancieri	29.55	14.90	12.92	12.19
VE_san_dona	23.65	16.74	20.94	13.90
VE_bissuola	25.16	14.97	15.48	11.93
VI_schio	57.28	26.52	42.11	26.44
VI_qitalia	33.17	15.57	16.37	14.87
VR_bcnuova	51.80	60.90	63.30	59.30
VR_legnago	30.88	14.60	13.91	11.43
VR_giarol	25.11	14.03	21.12	13.99
PD_mandria	30.43	13.53	13.31	14.28
PD_Monselice	38.65	14.47	13.93	11.15

Table A.4: PM_{10}: NME values for IDW and Regression IDW

Station	model	reg2	idw	reg+idw
BL_belluno	21.93	14.21	-21.34	3.13
BL_feltre	9.48	-18.36	-14.19	-22.91
BL_pieve	93.40	40.25	45.98	18.56
PD_p_colli	27.94	14.05	7.51	10.41
PD_este	20.21	5.11	-0.96	-1.23
RO_badia	-9.54	-12.59	-20.93	-12.62
RO_adria	14.89	24.56	6.03	28.00
RO_borsea	-12.52	-14.55	-23.74	-15.74
TV_conegliano	63.95	39.05	47.42	37.84
TV_mansue	4.40	1.08	-20.00	-6.10
TV_lancieri	19.41	-7.26	5.93	-8.09
VE_san_dona	-2.77	-6.83	-17.41	-7.20
VE_bissuola	3.95	-6.75	-9.78	-4.99
VI_schio	47.21	10.80	25.77	15.47
VI_qitalia	22.96	-10.28	7.41	-10.81
VR_bcnuova	14.26	-39.04	-9.48	-33.92
VR_legnago	1.55	-1.36	-7.58	0.86
VR_giarol	-3.29	-4.54	-14.86	-0.83
PD_mandria	19.54	-6.99	3.40	-9.31
PD_Monselice	27.51	8.89	10.21	5.16

Table A.5: PM_{10}: NMB values for IDW and Regression IDW

Station	model	reg2	idw	reg+idw
BL_belluno	0.35	0.73	0.76	0.76
BL_feltre	0.21	0.46	0.50	0.50
BL_pieve	0.49	0.71	0.60	0.60
PD_p_colli	0.65	0.92	0.92	0.92
PD_este	0.62	0.95	0.96	0.96
RO_badia	0.53	0.88	0.95	0.95
RO_adria	0.59	0.91	0.94	0.94
RO_borsea	0.56	0.89	0.95	0.95
TV_conegliano	0.62	0.73	0.80	0.80
TV_mansue	0.74	0.84	0.91	0.91
TV_lancieri	0.76	0.91	0.95	0.95
VE_san_dona	0.71	0.87	0.92	0.92
VE_bissuola	0.73	0.91	0.95	0.95
VI_schio	0.61	0.76	0.80	0.80
VI_qitalia	0.68	0.89	0.91	0.91
VR_bcnuova	0.46	0.38	0.42	0.42
VR_legnago	0.51	0.88	0.94	0.94
VR_giarol	0.69	0.90	0.90	0.90
PD_mandria	0.66	0.92	0.92	0.92
PD_Monselice	0.61	0.94	0.96	0.96

Table A.6: PM_{10}: Correlation values for IDW and Regression IDW

Station	model	reg2	RBF gau	RBF invq	RBF invm
BL_belluno	45.27	24.01	24.91	21.98	20.29
BL_feltre	38.79	30.60	28.13	28.12	28.56
BL_pieve	100.26	42.37	44.86	44.39	45.38
PD_p_colli	39.60	18.78	16.92	16.81	16.80
PD_este	34.01	12.58	11.18	10.73	11.00
RO_badia	29.51	19.05	17.57	13.47	10.31
RO_adria	31.42	25.97	27.38	28.75	31.04
RO_borsea	28.50	19.13	18.41	16.40	15.84
TV_conegliano	67.83	43.52	45.22	42.92	41.55
TV_mansue	24.72	16.81	16.21	13.74	13.44
TV_lancieri	29.55	15.21	14.31	12.40	11.39
VE_san_dona	23.65	16.86	16.21	14.27	12.98
VE_bissuola	25.16	15.09	14.32	12.36	10.99
VI_schio	57.28	25.09	26.76	26.71	26.97
VI_qitalia	33.17	16.02	15.26	14.76	14.43
VR_bcnuova	51.80	47.76	59.98	59.34	59.59
VR_legnago	30.88	14.78	14.77	12.84	11.98
VR_giarol	25.11	14.63	13.82	13.65	13.93
PD_mandria	30.43	13.98	13.73	14.03	15.15
PD_Monselice	38.65	14.23	12.79	12.19	12.45

Table A.7: PM_{10}: NME values for RBF interpolation with different radial basis functions (gaussian, inverse quadratic and inverse multiquadric)

Station	model	reg2	RBF gau	RBF invq	RBF invm
BL_belluno	21.93	6.75	10.48	7.24	4.47
BL_feltre	9.48	-23.42	-17.55	-19.70	-22.26
BL_pieve	93.40	37.12	33.58	26.25	19.08
PD_p_colli	27.94	13.08	10.57	10.74	10.31
PD_este	20.21	4.18	-3.66	-2.67	-3.07
RO_badia	-9.54	-13.01	-11.17	-9.93	-8.11
RO_adria	14.89	23.91	25.92	28.39	31.04
RO_borsea	-12.52	-15.04	-13.86	-14.53	-15.05
TV_conegliano	63.95	37.44	39.96	38.81	38.04
TV_mansue	4.40	0.66	0.32	-3.59	-6.99
TV_lancieri	19.41	-7.87	-6.47	-7.12	-7.49
VE_san_dona	-2.77	-7.18	-6.11	-5.79	-5.09
VE_bissuola	3.95	-7.30	-5.87	-4.38	-2.77
VI_schio	47.21	7.11	12.01	14.65	17.70
VI_qitalia	22.96	-10.75	-9.71	-10.29	-10.82
VR_bcnuova	14.26	-25.55	-36.68	-34.83	-33.19
VR_legnago	1.55	-2.18	-0.43	1.27	2.84
VR_giarol	-3.29	-5.83	-3.39	-1.77	-0.23
PD_mandria	19.54	-7.49	-7.19	-8.57	-9.97
PD_Monselice	27.51	8.04	6.16	6.22	6.02

Table A.8: PM_{10}: NMB values for RBF interpolation with different radial basis functions (gaussian, inverse quadratic and inverse multiquadric)

Station	model	reg2	RBF gau	RBF invq	RBF invm
BL_belluno	0.35	0.74	0.73	0.76	0.78
BL_feltre	0.21	0.46	0.46	0.49	0.53
BL_pieve	0.49	0.76	0.68	0.65	0.62
PD_p_colli	0.65	0.92	0.92	0.92	0.91
PD_este	0.62	0.95	0.95	0.95	0.95
RO_badia	0.53	0.88	0.89	0.95	0.97
RO_adria	0.59	0.91	0.91	0.93	0.94
RO_borsea	0.56	0.89	0.90	0.94	0.97
TV_conegliano	0.62	0.74	0.75	0.79	0.81
TV_mansue	0.74	0.84	0.85	0.89	0.91
TV_lancieri	0.76	0.90	0.91	0.95	0.96
VE_san_dona	0.71	0.87	0.87	0.90	0.92
VE_bissuola	0.73	0.91	0.92	0.94	0.95
VI_schio	0.61	0.78	0.76	0.78	0.80
VI_qitalia	0.68	0.89	0.89	0.91	0.92
VR_bcnuova	0.46	0.53	0.38	0.40	0.42
VR_legnago	0.51	0.88	0.88	0.91	0.93
VR_giarol	0.69	0.90	0.90	0.91	0.90
PD_mandria	0.66	0.92	0.92	0.92	0.91
PD_Monselice	0.61	0.93	0.95	0.95	0.95

Table A.9: PM_{10}: Correlation values for RBF interpolation with different radial basis functions (gaussian, inverse quadratic and inverse multiquadric)

Station	model	reg2	RK gau	RK exp	RK sph
BL_belluno	45.27	26.40	23.68	21.26	20.95
BL_feltre	38.79	28.49	29.33	28.04	28.56
BL_pieve	100.26	47.07	48.45	45.18	45.55
PD_p_colli	39.60	19.21	18.15	17.25	17.44
PD_este	34.01	12.55	11.81	10.43	9.89
RO_badia	29.51	18.92	13.73	13.61	13.37
RO_adria	31.42	26.20	28.57	28.53	29.08
RO_borsea	28.50	18.91	17.18	16.94	16.87
TV_conegliano	67.83	44.85	42.63	42.53	42.48
TV_mansue	24.72	17.02	14.21	13.52	13.93
TV_lancieri	29.55	14.90	11.83	11.84	12.12
VE_san_dona	23.65	16.74	13.64	13.70	13.53
VE_bissuola	25.16	14.97	11.53	11.66	11.35
VI_schio	57.28	26.52	28.10	25.97	25.91
VI_qitalia	33.17	15.57	14.61	13.53	13.67
VR_bcnuova	51.80	60.90	61.66	61.56	61.26
VR_legnago	30.88	14.60	13.18	12.85	13.23
VR_giarol	25.11	14.03	14.72	14.23	14.48
PD_mandria	30.43	13.53	14.93	13.25	13.59
PD_Monselice	38.65	14.47	12.64	11.26	11.79

Table A.10: heading

Table A.11: PM_{10}: NME values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2	RK gau	RK exp	RK sph
BL_belluno	21.93	14.21	9.35	7.63	7.13
BL_feltre	9.48	-18.36	-23.00	-21.68	-22.94
BL_pieve	93.40	40.25	20.54	22.69	21.20
PD_p_colli	27.94	14.05	14.12	12.97	13.17
PD_este	20.21	5.11	0.81	2.46	2.62
RO_badia	-9.54	-12.59	-9.22	-10.88	-10.82
RO_adria	14.89	24.56	28.15	28.14	28.68
RO_borsea	-12.52	-14.55	-16.17	-15.71	-15.78
TV_conegliano	63.95	39.05	39.23	39.18	39.14
TV_mansue	4.40	1.08	-4.92	-4.20	-5.03
TV_lancieri	19.41	-7.26	-7.74	-7.42	-7.67
VE_san_dona	-2.77	-6.83	-6.32	-6.21	-5.94
VE_bissuola	3.95	-6.75	-5.07	-4.78	-4.80
VI_schio	47.21	10.80	18.22	16.19	16.48
VI_qitalia	22.96	-10.28	-10.16	-9.66	-9.71
VR_bcnuova	14.26	-39.04	-34.33	-34.53	-33.45
VR_legnago	1.55	-1.36	-0.55	1.11	1.45
VR_giarol	-3.29	-4.54	-0.71	-1.30	-1.04
PD_mandria	19.54	-6.99	-8.68	-8.04	-8.18
PD_Monselice	27.51	8.89	8.12	7.30	7.54

Table A.12: PM_{10}: NMB values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2	RK gau	RK exp	RK sph
BL_belluno	0.35	0.73	0.66	0.78	0.78
BL_feltre	0.21	0.46	0.54	0.53	0.55
BL_pieve	0.49	0.71	0.55	0.61	0.60
PD_p_colli	0.65	0.92	0.92	0.93	0.93
PD_este	0.62	0.95	0.94	0.96	0.96
RO_badia	0.53	0.88	0.94	0.96	0.96
RO_adria	0.59	0.91	0.94	0.94	0.94
RO_borsea	0.56	0.89	0.96	0.95	0.96
TV_conegliano	0.62	0.73	0.81	0.80	0.80
TV_mansue	0.74	0.84	0.90	0.91	0.91
TV_lancieri	0.76	0.91	0.95	0.95	0.95
VE_san_dona	0.71	0.87	0.92	0.92	0.91
VE_bissuola	0.73	0.91	0.95	0.94	0.95
VI_schio	0.61	0.76	0.78	0.80	0.81
VI_qitalia	0.68	0.89	0.89	0.92	0.92
VR_bcnuova	0.46	0.38	0.39	0.38	0.39
VR_legnago	0.51	0.88	0.91	0.92	0.92
VR_giarol	0.69	0.90	0.89	0.90	0.89
PD_mandria	0.66	0.92	0.90	0.93	0.92
PD_Monselice	0.61	0.94	0.95	0.96	0.96

Table A.13: PM_{10}: Correlation values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2	GPR gaussian	GPR matern
BL_belluno	45.27	26.40	22.47	22.04
BL_feltre	38.79	28.49	27.86	27.73
BL_pieve	100.26	47.07	47.50	46.76
PD_p_colli	39.60	19.21	17.63	17.02
PD_este	34.01	12.55	12.36	11.47
RO_badia	29.51	18.92	15.20	15.42
RO_adria	31.42	26.20	27.07	27.39
RO_borsea	28.50	18.91	17.74	17.16
TV_conegliano	67.83	44.85	43.17	42.94
TV_mansue	24.72	17.02	15.07	14.60
TV_lancieri	29.55	14.90	11.69	11.75
VE_san_dona	23.65	16.74	13.64	13.59
VE_bissuola	25.16	14.97	11.91	11.97
VI_schio	57.28	26.52	25.72	25.58
VI_qitalia	33.17	15.57	14.29	14.30
VR_bcnuova	51.80	60.90	61.81	61.63
VR_legnago	30.88	14.60	12.99	12.82
VR_giarol	25.11	14.03	14.98	14.83
PD_mandria	30.43	13.53	14.62	14.12
PD_Monselice	38.65	14.47	11.56	12.01

Table A.14: PM_{10}: NME values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg2	GPR gaussian	GPR matern
BL_belluno	21.93	14.21	6.94	6.43
BL_feltre	9.48	-18.36	-21.72	-21.52
BL_pieve	93.40	40.25	25.08	24.31
PD_p_colli	27.94	14.05	12.44	11.97
PD_este	20.21	5.11	-0.34	-0.34
RO_badia	-9.54	-12.59	-12.35	-12.52
RO_adria	14.89	24.56	26.67	26.94
RO_borsea	-12.52	-14.55	-17.00	-16.53
TV_conegliano	63.95	39.05	40.11	39.82
TV_mansue	4.40	1.08	-4.82	-4.91
TV_lancieri	19.41	-7.26	-7.49	-7.40
VE_san_dona	-2.77	-6.83	-5.40	-5.75
VE_bissuola	3.95	-6.75	-4.92	-5.09
VI_schio	47.21	10.80	15.57	15.54
VI_qitalia	22.96	-10.28	-9.61	-9.92
VR_bcnuova	14.26	-39.04	-34.01	-33.67
VR_legnago	1.55	-1.36	-0.21	-0.02
VR_giarol	-3.29	-4.54	-2.23	-2.06
PD_mandria	19.54	-6.99	-9.80	-9.16
PD_Monselice	27.51	8.89	6.15	6.28

Table A.15: PM_{10}: NMB values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg2	GPR gaussian	GPR matern
BL_belluno	0.35	0.73	0.76	0.76
BL_feltre	0.21	0.46	0.52	0.52
BL_pieve	0.49	0.71	0.59	0.60
PD_p_colli	0.65	0.92	0.93	0.92
PD_este	0.62	0.95	0.94	0.95
RO_badia	0.53	0.88	0.94	0.94
RO_adria	0.59	0.91	0.94	0.94
RO_borsea	0.56	0.89	0.95	0.96
TV_conegliano	0.62	0.73	0.81	0.81
TV_mansue	0.74	0.84	0.88	0.89
TV_lancieri	0.76	0.91	0.96	0.95
VE_san_dona	0.71	0.87	0.91	0.91
VE_bissuola	0.73	0.91	0.94	0.94
VI_schio	0.61	0.76	0.80	0.80
VI_qitalia	0.68	0.89	0.91	0.91
VR_bcnuova	0.46	0.38	0.39	0.38
VR_legnago	0.51	0.88	0.92	0.92
VR_giarol	0.69	0.90	0.89	0.89
PD_mandria	0.66	0.92	0.91	0.92
PD_Monselice	0.61	0.94	0.96	0.95

Table A.16: PM_{10}: Correlation values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg1	reg1+BC	reg2	reg2+BC
BL_belluno	31.64	26.95	29.22	26.88	33.98
BL_feltre	32.31	27.59	26.12	27.88	24.21
BL_pieve	67.06	123.95	116.29	91.92	62.63
PD_p_colli	79.05	58.67	55.72	71.96	71.44
PD_sgiust	30.61	10.41	10.27	11.20	10.56
PD_este	17.07	17.41	18.29	14.49	14.64
RO_badia	22.96	10.87	11.44	14.52	14.84
RO_adria	22.90	19.63	21.64	12.09	11.92
RO_borsea	22.26	11.96	12.56	12.21	12.63
TV_conegliano	67.83	41.90	43.33	39.04	38.74
TV_mansue	61.03	42.59	39.64	56.85	56.77
TV_lancieri	21.66	12.05	12.42	11.60	12.30
VE_san_dona	16.62	21.53	22.71	17.68	18.00
VE_bissuola	35.98	12.18	12.44	11.35	11.01
VI_asiago	69.49	291.17	290.09	109.31	91.84
VI_bassano	33.52	18.10	18.21	16.86	16.48
VI_schio	30.62	30.41	31.77	29.05	32.58
VI_qitalia	39.86	18.46	19.39	18.54	19.59
VR_legnago	25.37	36.39	37.63	31.85	32.22
VR_giarol	53.66	21.70	23.48	16.26	15.20
PD_mandria	18.46	15.47	14.86	18.49	18.92

Table A.17: NO_2: NME values for different lienar regression models

Station	model	reg1	reg1+BC	reg2	reg2+BC
BL_belluno	-23.74	-25.19	-27.96	-25.88	-33.91
BL_feltre	10.04	12.38	8.05	13.94	1.76
BL_pieve	59.25	123.95	116.29	91.79	61.40
PD_p_colli	78.95	58.67	55.72	71.96	71.44
PD_sgiust	25.98	4.68	3.39	7.68	6.16
PD_este	2.14	-14.67	-15.91	-10.75	-11.55
RO_badia	9.26	-0.85	-3.41	12.01	12.12
RO_adria	-12.91	-17.77	-20.33	-2.80	-1.82
RO_borsea	11.22	-2.99	-4.90	5.55	5.40
TV_conegliano	63.95	27.77	30.22	19.45	18.41
TV_mansue	60.92	42.59	39.63	56.85	56.77
TV_lancieri	14.14	-8.83	-9.28	-8.83	-9.90
VE_san_dona	-5.44	-20.83	-22.11	-16.21	-16.51
VE_bissuola	35.51	7.95	8.07	6.84	6.28
VI_asiago	59.87	291.17	290.09	26.68	24.42
VI_bassano	14.33	-0.31	-2.59	4.47	0.62
VI_schio	-16.96	-28.60	-30.16	-27.89	-32.13
VI_qitalia	37.34	4.61	6.81	-4.39	-5.50
VR_legnago	-24.85	-36.39	-37.63	-31.85	-32.22
VR_giarol	52.91	20.08	21.85	12.06	9.34
PD_mandria	13.13	-13.54	-12.91	-16.82	-17.41

Table A.18: NO_2: NMB values for different linear regression models

Station	model	reg1	reg1+BC	reg2	reg2+BC
BL_belluno	0.23	0.53	0.52	0.59	0.60
BL_feltre	0.24	0.51	0.49	0.54	0.54
BL_pieve	0.50	0.49	0.51	0.59	0.62
PD_p_colli	0.64	0.80	0.80	0.82	0.83
PD_sgiust	0.60	0.85	0.84	0.87	0.87
PD_este	0.68	0.84	0.84	0.88	0.89
RO_badia	0.64	0.83	0.83	0.88	0.88
RO_adria	0.68	0.83	0.82	0.85	0.85
RO_borsea	0.64	0.80	0.79	0.85	0.84
TV_conegliano	0.62	0.49	0.49	0.45	0.44
TV_mansue	0.73	0.81	0.80	0.83	0.83
TV_lancieri	0.61	0.82	0.82	0.83	0.83
VE_san_dona	0.69	0.85	0.84	0.86	0.85
VE_bissuola	0.71	0.85	0.84	0.87	0.87
VI_asiago	0.51	-0.14	-0.13	-0.26	-0.21
VI_bassano	0.48	0.67	0.65	0.71	0.71
VI_schio	0.35	0.44	0.43	0.47	0.47
VI_qitalia	0.22	0.41	0.40	0.38	0.33
VR_legnago	0.66	0.81	0.81	0.86	0.86
VR_giarol	0.61	0.80	0.79	0.75	0.73
PD_mandria	0.69	0.84	0.84	0.83	0.83

Table A.19: NO_2: Correlation values for different linear regression models

Station	model	reg2+BC	idw	reg2+idw	reg2+BC+idw
BL_belluno	31.64	33.98	45.97	39.36	44.39
BL_feltre	32.31	24.21	26.92	26.58	23.29
BL_pieve	67.06	62.63	76.14	119.83	96.12
PD_p_colli	79.05	71.44	68.89	83.80	82.89
PD_sgiust	30.61	10.56	13.47	11.37	10.78
PD_este	17.07	14.64	25.76	29.05	31.62
RO_badia	22.96	14.84	10.62	16.56	15.33
RO_adria	22.90	11.92	29.32	11.04	10.81
RO_borsea	22.26	12.63	13.30	10.53	10.25
TV_conegliano	67.83	38.74	49.74	34.59	34.07
TV_mansue	61.03	56.77	29.32	50.99	50.90
TV_lancieri	21.66	12.30	13.67	13.88	14.86
VE_san_dona	16.62	18.00	28.61	22.79	23.58
VE_bissuola	35.98	11.01	20.19	10.91	10.34
VI_asiago	69.49	91.84	88.64	122.75	98.34
VI_bassano	33.52	16.48	21.74	17.24	16.64
VI_schio	30.62	32.58	37.00	28.12	31.51
VI_qitalia	39.86	19.59	30.56	18.18	19.25
VR_legnago	25.37	32.22	38.87	33.99	34.85
VR_giarol	53.66	15.20	45.93	20.26	18.70
PD_mandria	18.46	18.92	10.28	21.47	23.09

Table A.20: NO_2 NME values for IDW and regression IDW

Station	model	reg2+BC	idw	reg2+idw	reg2+BC+idw
BL_belluno	-23.74	-33.91	-45.81	-39.36	-44.39
BL_feltre	10.04	1.76	-12.22	13.73	3.47
BL_pieve	59.25	61.40	70.97	119.83	96.12
PD_p_colli	78.95	71.44	68.89	83.80	82.89
PD_sgiust	25.98	6.16	6.10	8.88	7.49
PD_este	2.14	-11.55	-25.70	-29.05	-31.62
RO_badia	9.26	12.12	-0.39	15.73	14.30
RO_adria	-12.91	-1.82	-29.05	-3.98	-3.68
RO_borsea	11.22	5.40	2.62	4.86	3.67
TV_conegliano	63.95	18.41	42.80	11.64	9.94
TV_mansue	60.92	56.77	19.67	50.93	50.90
TV_lancieri	14.14	-9.90	-7.65	-12.94	-14.20
VE_san_dona	-5.44	-16.51	-28.31	-22.57	-23.35
VE_bissuola	35.51	6.28	19.44	8.61	7.88
VI_asiago	59.87	24.42	16.08	41.63	36.84
VI_bassano	14.33	0.62	-3.90	6.60	2.61
VI_schio	-16.96	-32.13	-35.12	-26.36	-30.89
VI_qitalia	37.34	-5.50	28.23	-1.40	-2.01
VR_legnago	-24.85	-32.22	-38.87	-33.99	-34.85
VR_giarol	52.91	9.34	45.88	18.83	16.32
PD_mandria	13.13	-17.41	-2.66	-20.69	-22.42

Table A.21: NO_2: NMB values for IDW and regression IDW

Station	model	reg2+BC	idw	reg2+idw	reg2+BC+idw
BL_belluno	0.23	0.60	0.57	0.69	0.73
BL_feltre	0.24	0.54	0.47	0.61	0.60
BL_pieve	0.50	0.62	0.45	0.62	0.68
PD_p_colli	0.64	0.83	0.82	0.82	0.83
PD_sgiust	0.60	0.87	0.78	0.89	0.89
PD_este	0.68	0.89	0.89	0.90	0.89
RO_badia	0.64	0.88	0.87	0.92	0.92
RO_adria	0.68	0.85	0.85	0.88	0.89
RO_borsea	0.64	0.84	0.82	0.90	0.89
TV_conegliano	0.62	0.44	0.63	0.47	0.47
TV_mansue	0.73	0.83	0.70	0.84	0.84
TV_lancieri	0.61	0.83	0.77	0.84	0.84
VE_san_dona	0.69	0.85	0.82	0.89	0.89
VE_bissuola	0.71	0.87	0.86	0.91	0.91
VI_asiago	0.51	-0.21	-0.21	-0.24	-0.20
VI_bassano	0.48	0.71	0.55	0.71	0.71
VI_schio	0.35	0.47	0.37	0.45	0.47
VI_qitalia	0.22	0.33	0.37	0.38	0.32
VR_legnago	0.66	0.86	0.83	0.89	0.88
VR_giarol	0.61	0.73	0.80	0.79	0.76
PD_mandria	0.69	0.83	0.85	0.84	0.83

Table A.22: NO_2: Correlation values for IDW and regression IDW

Station	model	reg2+BoxCox	RBF gau	RBF invq	RBF invm
BL_belluno	31.64	26.88	33.98	34.09	35.59
BL_feltre	32.31	27.88	24.21	24.22	24.12
BL_pieve	67.06	91.92	62.63	63.08	65.79
PD_p_colli	79.05	71.96	71.44	71.78	74.52
PD_sgiust	30.61	11.20	10.56	10.58	10.67
PD_este	17.07	14.49	14.64	15.02	16.92
RO_badia	22.96	14.52	14.84	14.87	15.13
RO_adria	22.90	12.09	11.92	11.90	11.65
RO_borsea	22.26	12.21	12.63	12.61	12.22
TV_conegliano	67.83	39.04	38.74	38.70	38.03
TV_mansue	61.03	56.85	56.77	56.76	56.32
TV_lancieri	21.66	11.60	12.30	12.30	12.53
VE_san_dona	16.62	17.68	18.00	18.01	18.50
VE_bissuola	35.98	11.35	11.01	11.01	10.81
VI_asiago	69.49	109.31	91.84	91.94	93.23
VI_bassano	33.52	16.86	16.48	16.48	16.43
VI_schio	30.62	29.05	32.58	32.55	32.25
VI_qitalia	39.86	18.54	19.59	19.58	19.38
VR_legnago	25.37	31.85	32.22	32.23	32.48
VR_giarol	53.66	16.26	15.20	15.23	15.71
PD_mandria	18.46	18.49	18.92	18.95	19.34

Table A.23: NO_2: NME values for RBF interpolation using different radial basis functions (gaussian, inverse quadratic, inverse multiquadric)

Station	model	reg2+BoxCox	RBF gau	RBF invq	RBF invm
BL_belluno	-23.74	-25.88	-33.91	-34.04	-35.59
BL_feltre	10.04	13.94	1.76	1.82	2.21
BL_pieve	59.25	91.79	61.40	61.89	64.98
PD_p_colli	78.95	71.96	71.44	71.78	74.52
PD_sgiust	25.98	7.68	6.16	6.21	6.64
PD_este	2.14	-10.75	-11.55	-12.24	-15.45
RO_badia	9.26	12.01	12.12	12.19	13.04
RO_adria	-12.91	-2.80	-1.82	-1.80	-1.78
RO_borsea	11.22	5.55	5.40	5.42	5.51
TV_conegliano	63.95	19.45	18.41	18.34	17.26
TV_mansue	60.92	56.85	56.77	56.76	56.32
TV_lancieri	14.14	-8.83	-9.90	-9.93	-10.63
VE_san_dona	-5.44	-16.21	-16.51	-16.54	-17.43
VE_bissuola	35.51	6.84	6.28	6.33	6.62
VI_asiago	59.87	26.68	24.42	24.62	27.09
VI_bassano	14.33	4.47	0.62	0.68	1.31
VI_schio	-16.96	-27.89	-32.13	-32.10	-31.77
VI_qitalia	37.34	-4.39	-5.50	-5.42	-4.59
VR_legnago	-24.85	-31.85	-32.22	-32.23	-32.48
VR_giarol	52.91	12.06	9.34	9.42	10.57
PD_mandria	13.13	-16.82	-17.41	-17.46	-18.03

Table A.24: NO_2: NMB values for RBF interpolation using different radial basis functions (gaussian, inverse quadratic, inverse multiquadric)

Station	model	reg2+BoxCox	RBF gau	RBF invq	RBF invm
BL_belluno	0.23	0.59	0.60	0.60	0.63
BL_feltre	0.24	0.54	0.54	0.54	0.55
BL_pieve	0.50	0.59	0.62	0.62	0.65
PD_p_colli	0.64	0.82	0.83	0.83	0.83
PD_sgiust	0.60	0.87	0.87	0.87	0.87
PD_este	0.68	0.88	0.89	0.89	0.90
RO_badia	0.64	0.88	0.88	0.88	0.89
RO_adria	0.68	0.85	0.85	0.85	0.86
RO_borsea	0.64	0.85	0.84	0.84	0.86
TV_conegliano	0.62	0.45	0.44	0.44	0.45
TV_mansue	0.73	0.83	0.83	0.83	0.84
TV_lancieri	0.61	0.83	0.83	0.83	0.83
VE_san_dona	0.69	0.86	0.85	0.85	0.86
VE_bissuola	0.71	0.87	0.87	0.87	0.88
VI_asiago	0.51	-0.26	-0.21	-0.21	-0.21
VI_bassano	0.48	0.71	0.71	0.71	0.71
VI_schio	0.35	0.47	0.47	0.47	0.47
VI_qitalia	0.22	0.38	0.33	0.33	0.33
VR_legnago	0.66	0.86	0.86	0.86	0.87
VR_giarol	0.61	0.75	0.73	0.73	0.73
PD_mandria	0.69	0.83	0.83	0.83	0.84

Table A.25: NO_2: Correlation values for RBF interpolation using different radial basis functions (gaussian, inverse quadratic, inverse multiquadric)

Station	model	reg2+BC	RK gau	RK exp	RK sph
BL_belluno	31.64	26.88	36.84	35.80	36.31
BL_feltre	32.31	27.88	23.62	23.71	23.65
BL_pieve	67.06	91.92	63.37	65.99	64.85
PD_p_colli	79.05	71.96	78.26	79.51	78.89
PD_sgiust	30.61	11.20	10.35	10.50	10.45
PD_este	17.07	14.49	24.40	18.24	22.62
RO_badia	22.96	14.52	15.10	14.90	15.07
RO_adria	22.90	12.09	11.71	11.72	11.75
RO_borsea	22.26	12.21	12.15	12.27	12.18
TV_conegliano	67.83	39.04	38.16	38.45	38.18
TV_mansue	61.03	56.85	57.26	57.05	57.30
TV_lancieri	21.66	11.60	13.02	12.84	13.01
VE_san_dona	16.62	17.68	20.71	19.44	20.18
VE_bissuola	35.98	11.35	10.26	10.73	10.47
VI_asiago	69.49	109.31	97.66	95.15	96.28
VI_bassano	33.52	16.86	16.28	16.43	16.40
VI_schio	30.62	29.05	32.12	32.42	32.25
VI_qitalia	39.86	18.54	19.70	19.91	19.83
VR_legnago	25.37	31.85	32.21	32.18	32.22
VR_giarol	53.66	16.26	18.38	16.87	17.69
PD_mandria	18.46	18.49	18.86	19.00	19.01

Table A.26: NO_2: NME values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2+BC	RK gau	RK exp	RK sph
BL_belluno	-23.74	-25.88	-36.80	-35.74	-36.27
BL_feltre	10.04	13.94	1.68	1.81	1.76
BL_pieve	59.25	91.79	62.61	64.75	63.89
PD_p_colli	78.95	71.96	78.26	79.51	78.89
PD_sgiust	25.98	7.68	6.16	6.11	6.16
PD_este	2.14	-10.75	-22.82	-16.65	-21.07
RO_badia	9.26	12.01	14.16	13.71	14.13
RO_adria	-12.91	-2.80	-1.66	-1.55	-1.60
RO_borsea	11.22	5.55	6.21	5.82	6.08
TV_conegliano	63.95	19.45	17.15	17.61	17.26
TV_mansue	60.92	56.85	57.26	57.05	57.30
TV_lancieri	14.14	-8.83	-11.31	-10.93	-11.14
VE_san_dona	-5.44	-16.21	-20.00	-18.42	-19.39
VE_bissuola	35.51	6.84	5.50	6.11	5.89
VI_asiago	59.87	26.68	31.26	28.23	29.97
VI_bassano	14.33	4.47	1.06	0.79	0.81
VI_schio	-16.96	-27.89	-31.64	-31.98	-31.79
VI_qitalia	37.34	-4.39	-4.46	-4.90	-4.63
VR_legnago	-24.85	-31.85	-32.21	-32.18	-32.22
VR_giarol	52.91	12.06	14.55	12.05	13.56
PD_mandria	13.13	-16.82	-17.48	-17.61	-17.60

Table A.27: NO_2: NMB values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2+BC	RK gau	RK exp	RK sph
BL_belluno	0.23	0.59	0.65	0.64	0.65
BL_feltre	0.24	0.54	0.57	0.57	0.57
BL_pieve	0.50	0.59	0.66	0.64	0.66
PD_p_colli	0.64	0.82	0.80	0.80	0.80
PD_sgiust	0.60	0.87	0.87	0.87	0.87
PD_este	0.68	0.88	0.73	0.86	0.78
RO_badia	0.64	0.88	0.91	0.91	0.91
RO_adria	0.68	0.85	0.86	0.86	0.86
RO_borsea	0.64	0.85	0.87	0.86	0.86
TV_conegliano	0.62	0.45	0.44	0.44	0.44
TV_mansue	0.73	0.83	0.84	0.83	0.84
TV_lancieri	0.61	0.83	0.83	0.83	0.83
VE_san_dona	0.69	0.86	0.86	0.86	0.86
VE_bissuola	0.71	0.87	0.89	0.88	0.89
VI_asiago	0.51	-0.26	-0.22	-0.22	-0.22
VI_bassano	0.48	0.71	0.72	0.71	0.72
VI_schio	0.35	0.47	0.46	0.47	0.46
VI_qitalia	0.22	0.38	0.31	0.30	0.30
VR_legnago	0.66	0.86	0.87	0.86	0.87
VR_giarol	0.61	0.75	0.71	0.71	0.71
PD_mandria	0.69	0.83	0.83	0.83	0.83

Table A.28: NO_2: Correlation values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2	GPR gaussian	GPR matern
BL_belluno	31.64	26.88	36.08	35.80
BL_feltre	32.31	27.88	24.19	24.20
BL_pieve	67.06	91.92	65.74	65.07
PD_p_colli	79.05	71.96	74.84	74.87
PD_sgiust	30.61	11.20	10.31	10.36
PD_este	17.07	14.49	15.02	14.96
RO_badia	22.96	14.52	14.77	14.79
RO_adria	22.90	12.09	11.84	11.87
RO_borsea	22.26	12.21	12.36	12.42
TV_conegliano	67.83	39.04	38.74	38.73
TV_mansue	61.03	56.85	57.04	57.00
TV_lancieri	21.66	11.60	12.68	12.62
VE_san_dona	16.62	17.68	18.35	18.29
VE_bissuola	35.98	11.35	11.03	10.86
VI_asiago	69.49	109.31	92.04	91.95
VI_bassano	33.52	16.86	16.39	16.41
VI_schio	30.62	29.05	32.54	32.55
VI_qitalia	39.86	18.54	19.69	19.69
VR_legnago	25.37	31.85	32.48	32.43
VR_giarol	53.66	16.26	15.50	15.50
PD_mandria	18.46	18.49	19.19	19.12

Table A.29: NO_2: NME values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg2	GPR gaussian	GPR matern
BL_belluno	-23.74	-25.88	-36.02	-35.74
BL_feltre	10.04	13.94	1.73	1.77
BL_pieve	59.25	91.79	64.50	63.84
PD_p_colli	78.95	71.96	74.84	74.87
PD_sgiust	25.98	7.68	6.41	6.38
PD_este	2.14	-10.75	-12.24	-12.85
RO_badia	9.26	12.01	12.44	12.38
RO_adria	-12.91	-2.80	-1.84	-1.83
RO_borsea	11.22	5.55	5.44	5.42
TV_conegliano	63.95	19.45	18.20	18.20
TV_mansue	60.92	56.85	57.04	57.00
TV_lancieri	14.14	-8.83	-10.36	-10.30
VE_san_dona	-5.44	-16.21	-17.07	-16.97
VE_bissuola	35.51	6.84	6.18	6.35
VI_asiago	59.87	26.68	25.32	25.23
VI_bassano	14.33	4.47	0.82	0.80
VI_schio	-16.96	-27.89	-32.10	-32.11
VI_qitalia	37.34	-4.39	-5.19	-5.22
VR_legnago	-24.85	-31.85	-32.48	-32.43
VR_giarol	52.91	12.06	9.90	9.82
PD_mandria	13.13	-16.82	-17.69	-17.63

Table A.30: NO_2: NMB values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg2+Box-Cox	GPR gaussian	GPR matern
BL_belluno	0.23	0.59	0.64	0.64
BL_feltre	0.24	0.54	0.54	0.54
BL_pieve	0.50	0.59	0.63	0.63
PD_p_colli	0.64	0.82	0.82	0.82
PD_sgiust	0.60	0.87	0.88	0.88
PD_este	0.68	0.88	0.88	0.89
RO_badia	0.64	0.88	0.89	0.89
RO_adria	0.68	0.85	0.86	0.85
RO_borsea	0.64	0.85	0.85	0.85
TV_conegliano	0.62	0.45	0.44	0.44
TV_mansue	0.73	0.83	0.83	0.83
TV_lancieri	0.61	0.83	0.83	0.83
VE_san_dona	0.69	0.86	0.86	0.86
VE_bissuola	0.71	0.87	0.87	0.88
VI_asiago	0.51	-0.26	-0.21	-0.21
VI_bassano	0.48	0.71	0.72	0.72
VI_schio	0.35	0.47	0.47	0.47
VI_qitalia	0.22	0.38	0.32	0.32
VR_legnago	0.66	0.86	0.87	0.87
VR_giarol	0.61	0.75	0.71	0.71
PD_mandria	0.69	0.83	0.83	0.83

Table A.31: NO_2: Correlation values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg1	reg2
BL_belluno	23.02	11.78	11.88
BL_feltre	25.55	11.36	11.63
BL_pieve	19.26	8.75	10.84
PD_p_colli	20.90	4.95	5.17
PD_sgiust	22.75	4.42	4.54
PD_este	23.40	6.39	6.09
RO_badia	20.23	5.99	5.54
RO_adria	28.82	17.29	16.60
RO_borsea	27.07	12.56	12.04
TV_conegliano	29.29	9.74	9.31
TV_mansue	15.00	5.67	6.04
TV_lancieri	16.38	9.51	9.95
VE_san_dona	16.73	9.38	9.21
VE_bissuola	12.03	11.02	11.69
VI_asiago	12.07	11.41	13.80
VI_bassano	21.92	4.82	5.22
VI_schio	19.46	7.53	8.07
VI_qitalia	19.65	11.62	11.97
VR_bcnuova	15.93	7.90	7.09
VR_legnago	17.22	6.04	6.54
VR_giarol	29.87	8.35	8.49
PD_mandria	15.08	9.02	9.43

Table A.32: O_3: NME values for different linear regression models

Station	model	reg1	reg2
BL_belluno	22.21	10.42	11.01
BL_feltre	24.73	10.77	11.16
BL_pieve	16.78	5.60	9.00
PD_p_colli	20.10	-1.61	-2.36
PD_sgiust	21.64	0.04	-0.79
PD_este	22.48	2.62	1.94
RO_badia	19.29	0.24	-0.58
RO_adria	28.36	16.78	16.09
RO_borsea	24.01	7.38	6.66
TV_conegliano	29.04	9.20	8.43
TV_mansue	13.16	-2.92	-3.90
TV_lancieri	12.30	-4.02	-4.82
VE_san_dona	14.81	6.17	5.17
VE_bissuola	-1.36	-8.89	-9.79
VI_asiago	5.77	-10.84	-11.70
VI_bassano	20.83	-0.35	-1.34
VI_schio	17.60	-5.28	-5.86
VI_qitalia	14.32	-8.53	-9.37
VR_bcnuova	13.26	-6.50	-4.10
VR_legnago	16.36	-3.96	-4.87
VR_giarol	29.66	5.41	4.64
PD_mandria	8.04	-8.68	-9.20

Table A.33: O_3: NMB values for different linear regression models

Station	model	reg1	reg2
BL_belluno	0.77	0.91	0.92
BL_feltre	0.78	0.93	0.93
BL_pieve	0.69	0.89	0.88
PD_p_colli	0.68	0.92	0.92
PD_sgiust	0.61	0.93	0.93
PD_este	0.67	0.88	0.89
RO_badia	0.68	0.87	0.88
RO_adria	0.54	0.82	0.82
RO_borsea	0.48	0.72	0.72
TV_conegliano	0.76	0.93	0.93
TV_mansue	0.79	0.91	0.91
TV_lancieri	0.68	0.80	0.79
VE_san_dona	0.71	0.80	0.78
VE_bissuola	0.64	0.80	0.78
VI_asiago	0.62	0.88	0.82
VI_bassano	0.74	0.94	0.94
VI_schio	0.75	0.92	0.92
VI_qitalia	0.57	0.68	0.69
VR_bcnuova	0.69	0.88	0.89
VR_legnago	0.75	0.91	0.91
VR_giarol	0.67	0.85	0.84
PD_mandria	0.69	0.91	0.91

Table A.34: O_3: Correlation values for different linear regression models

Station	model	reg1	idw	reg1+idw	reg2+idw
BL_belluno	23.02	11.78	6.51	6.36	5.92
BL_feltre	25.55	11.36	11.33	10.41	10.30
BL_pieve	19.26	8.75	7.97	4.74	5.42
PD_p_colli	20.90	4.95	5.30	3.87	3.90
PD_sgiust	22.75	4.42	9.38	4.48	4.33
PD_este	23.40	6.39	6.01	5.10	5.00
RO_badia	20.23	5.99	6.58	4.28	4.20
RO_adria	28.82	17.29	10.95	15.32	15.19
RO_borsea	27.07	12.56	12.57	10.57	10.45
TV_conegliano	29.29	9.74	13.87	9.07	8.65
TV_mansue	15.00	5.67	8.73	6.41	6.81
TV_lancieri	16.38	9.51	9.68	8.54	8.72
VE_san_dona	16.73	9.38	9.58	9.75	9.74
VE_bissuola	12.03	11.02	18.91	9.93	10.20
VI_asiago	12.07	11.41	12.50	9.96	13.33
VI_bassano	21.92	4.82	8.30	5.22	5.19
VI_schio	19.46	7.53	7.40	5.27	6.43
VI_qitalia	19.65	11.62	11.59	10.57	10.74
VR_bcnuova	15.93	7.90	10.30	7.18	6.59
VR_legnago	17.22	6.04	6.56	5.25	5.48
VR_giarol	29.87	8.35	14.86	9.81	8.82
PD_mandria	15.08	9.02	12.54	8.17	8.23

Table A.35: O_3 NME values for IDW and regression IDW

Station	model	reg1	idw	reg1+idw	reg2+idw
BL_belluno	22.21	10.42	2.93	5.52	4.62
BL_feltre	24.73	10.77	5.52	9.88	9.58
BL_pieve	16.78	5.60	-3.93	-1.06	1.86
PD_p_colli	20.10	-1.61	0.91	-1.94	-2.10
PD_sgiust	21.64	0.04	6.35	2.31	1.69
PD_este	22.48	2.62	2.90	3.80	3.76
RO_badia	19.29	0.24	-0.29	-0.31	-0.55
RO_adria	28.36	16.78	6.48	15.13	14.94
RO_borsea	24.01	7.38	2.93	5.21	5.08
TV_conegliano	29.04	9.20	12.96	8.77	8.16
TV_mansue	13.16	-2.92	-5.21	-5.93	-6.52
TV_lancieri	12.30	-4.02	-1.85	-3.43	-3.79
VE_san_dona	14.81	6.17	-1.03	7.26	6.75
VE_bissuola	-1.36	-8.89	-17.80	-7.90	-8.26
VI_asiago	5.77	-10.84	-11.60	-9.30	-10.08
VI_bassano	20.83	-0.35	6.45	3.43	1.15
VI_schio	17.60	-5.28	3.28	-1.12	-2.67
VI_qitalia	14.32	-8.53	-1.46	-6.09	-6.99
VR_bcnuova	13.26	-6.50	-6.81	-5.85	-3.36
VR_legnago	16.36	-3.96	-1.97	-3.58	-4.05
VR_giarol	29.66	5.41	13.51	9.09	7.36
PD_mandria	8.04	-8.68	-9.60	-7.85	-7.88

Table A.36: O_3: NMB values for IDW and regression IDW

Station	model	reg1	idw	reg1+idw	reg2+idw
BL_belluno	0.77	0.91	0.93	0.96	0.96
BL_feltre	0.78	0.93	0.82	0.94	0.94
BL_pieve	0.69	0.89	0.88	0.95	0.94
PD_p_colli	0.68	0.92	0.90	0.95	0.96
PD_sgiust	0.61	0.93	0.81	0.94	0.93
PD_este	0.67	0.88	0.90	0.95	0.95
RO_badia	0.68	0.87	0.83	0.92	0.92
RO_adria	0.54	0.82	0.72	0.86	0.86
RO_borsea	0.48	0.72	0.68	0.77	0.77
TV_conegliano	0.76	0.93	0.89	0.95	0.94
TV_mansue	0.79	0.91	0.86	0.94	0.94
TV_lancieri	0.68	0.80	0.78	0.82	0.82
VE_san_dona	0.71	0.80	0.71	0.84	0.83
VE_bissuola	0.64	0.80	0.68	0.82	0.81
VI_asiago	0.62	0.88	0.75	0.89	0.83
VI_bassano	0.74	0.94	0.90	0.95	0.94
VI_schio	0.75	0.92	0.87	0.94	0.92
VI_qitalia	0.57	0.68	0.68	0.70	0.70
VR_bcnuova	0.69	0.88	0.74	0.89	0.89
VR_legnago	0.75	0.91	0.86	0.93	0.93
VR_giarol	0.67	0.85	0.81	0.87	0.85
PD_mandria	0.69	0.91	0.80	0.93	0.93

Table A.37: O_3: Correlation values for IDW and regression IDW

Station	model	reg2+BoxCox	RBF gau	RBF invq	RBF invm
BL_belluno	23.02	11.88	9.59	7.14	5.39
BL_feltre	25.55	11.63	11.50	10.48	9.54
BL_pieve	19.26	10.84	6.52	5.41	4.83
PD_p_colli	20.90	5.17	4.02	3.60	3.62
PD_sgiust	22.75	4.54	4.36	4.29	4.81
PD_este	23.40	6.09	5.08	4.58	4.35
RO_badia	20.23	5.54	5.63	4.51	4.34
RO_adria	28.82	16.60	17.36	15.84	14.48
RO_borsea	27.07	12.04	12.50	10.70	10.14
TV_conegliano	29.29	9.31	9.85	9.03	8.71
TV_mansue	15.00	6.04	5.60	6.03	7.15
TV_lancieri	16.38	9.95	9.30	8.78	8.51
VE_san_dona	16.73	9.21	9.42	9.52	9.93
VE_bissuola	12.03	11.69	10.86	10.22	9.63
VI_asiago	12.07	13.80	11.20	10.60	10.38
VI_bassano	21.92	5.22	4.63	4.85	5.51
VI_schio	19.46	8.07	7.35	5.65	5.26
VI_qitalia	19.65	11.97	11.57	10.83	10.25
VR_bcnuova	15.93	7.09	7.86	7.68	7.71
VR_legnago	17.22	6.54	6.00	5.63	5.38
VR_giarol	29.87	8.49	8.42	8.79	9.66
PD_mandria	15.08	9.43	8.84	8.37	7.95

Table A.38: O_3: NME values for RBF interpolation using different radial basis functions (gaussian, inverse quadratic, inverse multiquadric)

Station	model	reg2+BoxCox	RBF gau	RBF invq	RBF invm
BL_belluno	22.21	11.01	8.70	6.28	4.18
BL_feltre	24.73	11.16	10.94	9.91	8.88
BL_pieve	16.78	9.00	2.96	0.98	-0.85
PD_p_colli	20.10	-2.36	-2.57	-1.53	-1.01
PD_sgiust	21.64	-0.79	0.16	1.72	3.00
PD_este	22.48	1.94	3.34	3.02	2.78
RO_badia	19.29	-0.58	-0.05	-0.82	-1.37
RO_adria	28.36	16.09	16.89	15.56	14.36
RO_borsea	24.01	6.66	7.33	4.81	2.61
TV_conegliano	29.04	8.43	9.48	8.73	8.26
TV_mansue	13.16	-3.90	-3.30	-5.37	-7.00
TV_lancieri	12.30	-4.82	-3.74	-3.67	-3.57
VE_san_dona	14.81	5.17	6.34	6.90	7.82
VE_bissuola	-1.36	-9.79	-8.69	-8.29	-7.88
VI_asiago	5.77	-11.70	-10.64	-10.09	-9.85
VI_bassano	20.83	-1.34	0.54	2.67	4.08
VI_schio	17.60	-5.86	-5.01	-2.15	0.50
VI_qitalia	14.32	-9.37	-8.42	-6.81	-5.49
VR_bcnuova	13.26	-4.10	-6.58	-6.69	-6.69
VR_legnago	16.36	-4.87	-3.87	-4.01	-4.03
VR_giarol	29.66	4.64	5.82	7.29	8.86
PD_mandria	8.04	-9.20	-8.47	-8.07	-7.65

Table A.39: O_3: NMB values for RBF interpolation using different radial basis functions (gaussian, inverse quadratic, inverse multi quadratic) for O_3

Station	model	reg2+BoxCox	RBF gau	RBF invq	RBF invm
BL_belluno	0.77	0.92	0.94	0.96	0.96
BL_feltre	0.78	0.93	0.92	0.94	0.94
BL_pieve	0.69	0.88	0.93	0.94	0.95
PD_p_colli	0.68	0.92	0.96	0.96	0.95
PD_sgiust	0.61	0.93	0.93	0.94	0.94
PD_este	0.67	0.89	0.94	0.95	0.95
RO_badia	0.68	0.88	0.88	0.92	0.92
RO_adria	0.54	0.82	0.82	0.85	0.87
RO_borsea	0.48	0.72	0.72	0.76	0.78
TV_conegliano	0.76	0.93	0.93	0.95	0.95
TV_mansue	0.79	0.91	0.92	0.94	0.94
TV_lancieri	0.68	0.79	0.80	0.82	0.82
VE_san_dona	0.71	0.78	0.80	0.83	0.85
VE_bissuola	0.64	0.78	0.80	0.82	0.83
VI_asiago	0.62	0.82	0.88	0.89	0.90
VI_bassano	0.74	0.94	0.95	0.95	0.95
VI_schio	0.75	0.92	0.93	0.94	0.93
VI_qitalia	0.57	0.69	0.68	0.70	0.72
VR_bcnuova	0.69	0.89	0.88	0.89	0.89
VR_legnago	0.75	0.91	0.91	0.92	0.93
VR_giarol	0.67	0.84	0.85	0.87	0.88
PD_mandria	0.69	0.91	0.91	0.92	0.93

Table A.40: O_3: Correlation values for RBF interpolation using different radial basis function (gaussian, inverse quadratic, inverse multiquadric) for O_3

Station	model	reg2+BC	RK gau	RK exp	RK sph
BL_belluno	23.02	11.88	5.74	5.69	5.45
BL_feltre	25.55	11.63	9.61	9.40	9.20
BL_pieve	19.26	10.84	5.23	5.45	5.36
PD_p_colli	20.90	5.17	3.91	3.48	3.67
PD_sgiust	22.75	4.54	4.94	4.79	5.26
PD_este	23.40	6.09	4.93	4.69	4.65
RO_badia	20.23	5.54	4.75	4.68	4.87
RO_adria	28.82	16.60	14.57	15.33	14.67
RO_borsea	27.07	12.04	9.92	10.00	9.73
TV_conegliano	29.29	9.31	9.21	8.50	8.91
TV_mansue	15.00	6.04	7.50	7.06	7.49
TV_lancieri	16.38	9.95	8.87	8.61	8.60
VE_san_dona	16.73	9.21	11.01	9.83	10.53
VE_bissuola	12.03	11.69	9.46	9.69	9.67
VI_asiago	12.07	13.80	10.70	10.60	10.67
VI_bassano	21.92	5.22	6.40	5.46	5.70
VI_schio	19.46	8.07	7.40	5.94	6.37
VI_qitalia	19.65	11.97	10.75	10.27	10.30
VR_bcnuova	15.93	7.09	8.92	7.95	8.46
VR_legnago	17.22	6.54	6.00	5.37	5.36
VR_giarol	29.87	8.49	10.24	9.71	9.80
PD_mandria	15.08	9.43	8.15	8.16	7.96

Table A.41: O_3: NME values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2+BC	RK gau	RK exp	RK sph
BL_belluno	22.21	11.01	3.43	4.46	3.78
BL_feltre	24.73	11.16	8.69	8.63	8.32
BL_pieve	16.78	9.00	-0.25	-0.32	-0.38
PD_p_colli	20.10	-2.36	-0.02	-0.57	-0.27
PD_sgiust	21.64	-0.79	2.98	2.83	3.16
PD_este	22.48	1.94	2.77	3.18	3.18
RO_badia	19.29	-0.58	-0.57	-0.77	-0.94
RO_adria	28.36	16.09	13.87	15.08	14.37
RO_borsea	24.01	6.66	2.33	3.23	2.36
TV_conegliano	29.04	8.43	8.67	8.21	8.21
TV_mansue	13.16	-3.90	-7.17	-6.70	-7.18
TV_lancieri	12.30	-4.82	-4.26	-3.72	-3.72
VE_san_dona	14.81	5.17	9.42	7.85	9.01
VE_bissuola	-1.36	-9.79	-7.69	-8.19	-7.89
VI_asiago	5.77	-11.70	-10.19	-10.12	-10.17
VI_bassano	20.83	-1.34	4.37	3.44	3.77
VI_schio	17.60	-5.86	2.75	0.03	1.67
VI_qitalia	14.32	-9.37	-6.22	-5.56	-5.64
VR_bcnuova	13.26	-4.10	-8.16	-6.86	-7.54
VR_legnago	16.36	-4.87	-4.45	-4.01	-4.09
VR_giarol	29.66	4.64	9.56	8.88	9.02
PD_mandria	8.04	-9.20	-7.70	-7.86	-7.65
PD_Monselice					

Table A.42: O_3: NMB values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg2+BC	RK gau	RK exp	RK sph
BL_belluno	0.77	0.92	0.96	0.96	0.96
BL_feltre	0.78	0.93	0.93	0.94	0.94
BL_pieve	0.69	0.88	0.95	0.94	0.94
PD_p_colli	0.68	0.92	0.94	0.96	0.95
PD_sgiust	0.61	0.93	0.94	0.94	0.93
PD_este	0.67	0.89	0.93	0.94	0.94
RO_badia	0.68	0.88	0.91	0.92	0.91
RO_adria	0.54	0.82	0.81	0.85	0.84
RO_borsea	0.48	0.72	0.78	0.78	0.79
TV_conegliano	0.76	0.93	0.94	0.95	0.94
TV_mansue	0.79	0.91	0.94	0.94	0.94
TV_lancieri	0.68	0.79	0.82	0.82	0.82
VE_san_dona	0.71	0.78	0.84	0.85	0.85
VE_bissuola	0.64	0.78	0.82	0.83	0.82
VI_asiago	0.62	0.82	0.89	0.90	0.90
VI_bassano	0.74	0.94	0.93	0.95	0.94
VI_schio	0.75	0.92	0.88	0.92	0.91
VI_qitalia	0.57	0.69	0.73	0.72	0.72
VR_bcnuova	0.69	0.89	0.87	0.89	0.88
VR_legnago	0.75	0.91	0.91	0.93	0.93
VR_giarol	0.67	0.84	0.87	0.87	0.88
PD_mandria	0.69	0.91	0.92	0.93	0.93

Table A.43: O_3: Correlation values for Regression Kriging with different types of variogram (gaussian, exponential, spherical)

Station	model	reg1	GPR gaussian	GPR matern
BL_belluno	23.02	11.88	6.93	6.93
BL_feltre	25.55	11.63	9.82	10.02
BL_pieve	19.26	10.84	5.84	5.66
PD_p_colli	20.90	5.17	4.04	3.70
PD_sgiust	22.75	4.54	4.97	5.03
PD_este	23.40	6.09	4.84	4.64
RO_badia	20.23	5.54	4.88	4.85
RO_adria	28.82	16.60	14.69	14.67
RO_borsea	27.07	12.04	10.07	10.05
TV_conegliano	29.29	9.31	9.37	9.23
TV_mansue	15.00	6.04	7.32	7.17
TV_lancieri	16.38	9.95	8.61	8.54
VE_san_dona	16.73	9.21	10.84	10.76
VE_bissuola	12.03	11.69	9.39	9.40
VI_asiago	12.07	13.80	10.76	10.76
VI_bassano	21.92	5.22	6.36	6.19
VI_schio	19.46	8.07	7.37	7.23
VI_qitalia	19.65	11.97	10.16	10.22
VR_bcnuova	15.93	7.09	8.49	8.17
VR_legnago	17.22	6.54	5.37	5.39
VR_giarol	29.87	8.49	9.55	9.53
PD_mandria	15.08	9.43	8.30	8.21

Table A.44: O_3: NME values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg1	GPR gaussian	GPR matern
BL_belluno	22.21	11.01	5.23	5.31
BL_feltre	24.73	11.16	8.99	9.20
BL_pieve	16.78	9.00	0.95	0.82
PD_p_colli	20.10	-2.36	-0.16	-0.29
PD_sgiust	21.64	-0.79	3.25	3.231
PD_este	22.48	1.94	3.05	2.92
RO_badia	19.29	-0.58	-1.51	-1.43
RO_adria	28.36	16.09	14.20	14.26
RO_borsea	24.01	6.66	2.51	2.70
TV_conegliano	29.04	8.43	9.06	8.91
TV_mansue	13.16	-3.90	-6.72	-6.60
TV_lancieri	12.30	-4.82	-3.57	-3.45
VE_san_dona	14.81	5.17	9.08	9.08
VE_bissuola	-1.36	-9.79	-7.47	-7.62
VI_asiago	5.77	-11.70	-10.38	-10.41
VI_bassano	20.83	-1.34	4.17	4.08
VI_schio	17.60	-5.86	1.74	1.52
VI_qitalia	14.32	-9.37	-5.93	-5.80
VR_bcnuova	13.26	-4.10	-7.50	-7.22
VR_legnago	16.36	-4.87	-3.76	-3.69
VR_giarol	29.66	4.64	8.64	8.58
PD_mandria	8.04	-9.20	-7.66	-7.68

Table A.45: O_3: NMB values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Station	model	reg1	GPR gaussian	GPR matern
BL_belluno	0.77	0.92	0.95	0.94
BL_feltre	0.78	0.93	0.94	0.93
BL_pieve	0.69	0.88	0.94	0.94
PD_p_colli	0.68	0.92	0.94	0.95
PD_sgiust	0.61	0.93	0.93	0.94
PD_este	0.67	0.89	0.94	0.94
RO_badia	0.68	0.88	0.91	0.91
RO_adria	0.54	0.82	0.85	0.85
RO_borsea	0.48	0.72	0.78	0.79
TV_conegliano	0.76	0.93	0.94	0.95
TV_mansue	0.79	0.91	0.93	0.93
TV_lancieri	0.68	0.79	0.83	0.83
VE_san_dona	0.71	0.78	0.84	0.84
VE_bissuola	0.64	0.78	0.84	0.84
VI_asiago	0.62	0.82	0.90	0.90
VI_bassano	0.74	0.94	0.93	0.94
VI_schio	0.75	0.92	0.87	0.88
VI_qitalia	0.57	0.69	0.74	0.74
VR_bcnuova	0.69	0.89	0.88	0.89
VR_legnago	0.75	0.91	0.93	0.92
VR_giarol	0.67	0.84	0.87	0.87
PD_mandria	0.69	0.91	0.91	0.92

Table A.46: O_3: Correlation values for Gaussian Process regression with different types of kernel (Gaussian and Matern)

Bibliography

[1] ARPAV, "Aria." https://www.arpa.veneto.it/temi-ambientali/aria, 2020.

[2] D. L. 155/2010, "Attuazione della direttiva 2008/50/ce relativa alla qualità dell'aria ambiente e per un'aria più pulita in europa.," 15 settembre 2010.

[3] "Directive 2008/50/ce of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for europe," 2008.

[4] ARPAV, "Bollettino pm10." https://www.arpa.veneto.it/dati-ambientali/bollettini/aria/bollettino-livelli-di-allerta-pm10, 2023.

[5] L. Rouil and B. Bessagnet, "How to start with pm modelling for air quality assessment and planning relevant to the air quality directive," *ETC/ACM Technical Paper*, vol. 11, no. 2014, p. 5, 2013.

[6] ARPAV, "Relazione regionale dalle qualità dell'aria - anno di riferimento: 2021," tech. rep., ARPAV, 2022.

[7] B. Denby, "Guide on modelling nitrogen dioxide (no2) for air quality assessment and planning relevant to the european air quality directive, etc/acm technical paper 2011/15," *European Topic Centre on Air Pollution and Climate Change Mitigation*, 2011.

[8] U. E. 16450:2017, "Aria ambiente - sistemi di misura automatici per la misurazione della concentrazione del particolato (pm10; pm2,5)," 2017.

[9] U. E. 14211:2012, "Qualità dell'aria ambiente - metodo normalizzato per la misrazione della concentrazione di diossido di azoto mediante chemioluminescenza," 2012.

[10] U. E. 14625:2012, "Qualità dell'aria ambiente - metodo normalizzato per la misurazione della concentrazione di ozono mediante fotometria ultravioletta," 2012.

[11] D. F. Swinehart, "The beer-lambert law," *Journal of Chemical Education*, vol. 39, no. 7, p. 333, 1962.

[12] Ramboll, "Camx, a multi-scale photochemical modeling system for gas and particulate air pollution." https://www.camx.com/, 2022.

[13] A. Dalla Fontana, S. Pillon, and S. Patti, "A performance evaluation of the camx air quality model to forecast ozone and pm10 over the italian region of veneto," *J. Mediterr. Meteorol. Climatol*, vol. 18, pp. 1–13, 2021.

[14] "Inventario emissioni aria." https://www.inemar.eu/.

[15] "Consortium for smal-scale modelling." http://www.cosmo-model.org/.

[16] "Plate-forme nationale de prévision de la qualité de l'air." http://www2.prevair.org/.

[17] N. Cressie, *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, Wiley, 1993.

[18] H. Wackernagel, *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2003.

[19] J. Horálek, B. Denby, P. de Smet, F. de Leeuw, P. Kurfürst, R. Swart, and T. van Noije, "Spatial mapping of air quality for european scale assessment," tech. rep., ETC/ACC, 2006.

[20] B. Denby, J. Horálek, S. E. Walker, K. Eben, and J. Fiala, "Interpolation and assimilation methods for european scale air quality assessment and mapping," *Part I: Review and Recommendations. European Topic Centre on Air and Climate Change Technical Paper*, vol. 7, 2005.

[21] L. Tarrason, A. Semb, A. Hjellbrekke, S. Tsyro, J. Schaug, J. Batnicki, and S. Solberg, *Geographical distribution of sulphur and nitrogen compounds in Europe derived both from modelled and observed concentrations*. Norwegian Met. Inst., 1998.

[22] S. Bande, M. Stortini, R. Amorati, G. Giovannini, G. Bonafé, L. Matavz, E. Angelino, L. Colombo, G. Fossati, and A. Marongiu, "Prepair action d5. air quality assessment."

[23] G. E. P. Box and D. R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 26, no. 2, pp. 211–252, 1964.

[24] R. Ignaccolo, S. Ghigo, and S. Bande, "Functional zoning for air quality," *Environmental and ecological statistics*, vol. 20, pp. 109–127, 2013.

[25] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006.

[26] G. E. Fasshauer, *Meshfree Approximation Methods with Matlab*. WORLD SCIENTIFIC, 2007.

[27] N. Flyer, B. Fornberg, V. Bayona, and G. A. Barnett, "On the role of polynomials in rbf-fd approximations: I. interpolation and accuracy," *Journal of Computational Physics*, vol. 321, pp. 21–38, 2016.

[28] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[29] T. Hengl, *A Practical Guide to Geostatistical Mapping of Environmental Variables*. Office for Official Publications of the European Communities, 2007.

[30] "Pykrige." `https://geostat-framework.readthedocs.io/projects/pykrige/en/stable/`.

[31] J.-P. Chiles and P. Delfiner, *Geostatistics: modeling spatial uncertainty*, vol. 713. John Wiley & Sons, 2012.

[32] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

[33] D. J. MacKay, D. J. Mac Kay, *et al.*, *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] "Using cross validation to assess interpolation results." `https://pro.arcgis.com/en/pro-app/3.0/help/analysis/geostatistical-analyst/performing-cross-validation-and-validation.htm`.

[36] C. Emery, Z. Liu, A. G. Russell, M. T. Odman, G. Yarwood, and N. Kumar, "Recommendations on statistics and benchmarks to assess photochemical model performance," *Journal of the Air & Waste Management Association*, vol. 67, no. 5, pp. 582–598, 2017. PMID: 27960634.

[37] G. Veratti, M. Stortini, R. Amorati, L. Bressan, G. Giovannini, S. Bande, F. Bissardella, S. Ghigo, E. Angelino, L. Colombo, G. Fossati, G. Malvestiti, A. Marongiu, A. Dalla Fontana, B. Intini, and S. Pillon, "Impact of nox and nh3 emission reduction on particulate matter across po valley: A life-ip-prepair study," *Atmosphere*, vol. 14, no. 5, 2023.

[38] S. Bande, M. Stortini, R. Amorati, G. Giovannini, G. Bonafé, L. Matavz, E. Angelino, L. Colombo, G. Fossati, A. I. B. Marongiu, Alessandro Dalla Fontana, and S. Pillon, "Prepair, air quaily assessment 2022."

[39] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, "Data assimilation in the geosciences: An overview of methods, issues, and perspectives," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 9, no. 5, p. e535, 2018.

[40] H. Petetin, D. Bowdalo, P.-A. Bretonnière, M. Guevara, O. Jorba, J. Mateu Armengol, M. Samso Cabre, K. Serradell, A. Soret, and C. Pérez Garcia-Pando, "Model output statistics (mos) applied to copernicus atmospheric monitoring service (cams) o_3 forecasts: trade-offs between continuous and categorical skill scores," *Atmospheric Chemistry and Physics*, vol. 22, no. 17, pp. 11603–11630, 2022.