



ARPALData: an R package for retrieving and analyzing air quality and weather data from ARPA Lombardia (Italy)

Paolo Maranzano^{1,2} · Andrea Algieri³

Received: 23 September 2023 / Accepted: 14 January 2024 / Published online: 1 March 2024
© The Author(s) 2024

Abstract

We present ARPALData, an R package that can help international users retrieve, handle, and analyze air quality and weather data in the Lombardy region (Northern Italy). The software provides a user-friendly tool that directly inquires into the platform of the regional environmental protection agency and ensures real-time updating of information using standardized syntax. The software provides data in standard statistical formats. Eventually, all measurements, metadata, and subsequent analytical tools are provided to users in English, facilitating accessibility to international and domestic users. Data are collected from the open database of the Regional Agency for Environmental Protection of Lombardy, namely ARPA Lombardia. ARPALData returns measurements at several temporal frequencies (infra-hourly to yearly) collected through air quality and weather ground monitoring networks managed by ARPA Lombardia, as well as estimates of several pollutants at the municipal level. In addition to data download functions, ARPALData provides functions to explore, describe, analyze, and graphically represent air quality and weather data. In particular, users are provided with functions to compute key descriptive statistics and input data maps, temporally aggregate measurements, detect outliers, and study missing-value (gap length) patterns. Herein, we discuss purposes, goals, and functioning of the package, and present three guided examples and case studies in which the software is used to characterize air quality and meteorology in different settings. The examples are designed to provide a step-by-step guide for accomplished analyses using the most relevant tools included in ARPALData.

Keywords Regional Agency for Environmental Protection of Lombardy - ARPA Lombardia · Air quality data · Data retrieving, handling, and analyzing · R language · Weather and meteorological data

Handling Editor: Luiz Duczmal

Extended author information available on the last page of the article

1 Introduction

In recent years, there has been a remarkable increase in the availability and size of georeferenced datasets and data retrieval tools, particularly for environmental studies. In certain research areas, such as air quality monitoring (Vitolo et al. 2016), agriculture and livestock farming (Fassó et al. 2023), climate (Cruz-Alonso et al. 2023), and geology (Ryberg and Vecchia 2012), access to high-quality data is a crucial issue.

For air quality and pollution in particular, access to data is typically linked to finding information from institutional sources, such as public environmental monitoring and protection agencies, government departments, and satellite data repositories. Successful examples of data collection are the `rdefra` (Vitolo et al. 2016) software, which is used to retrieve air pollution data from the UK Department for Environment, Food and Rural Affairs; `SHERPA-city` (Degraeuwe et al. 2021), a web application that assesses the impact of traffic measures on NO₂ in European cities; `saqgetr` (Grange 2019), a package to import AQ monitoring data from the AirBase - The European air quality database; and `openair`, an open source software for air quality data retrieval and analysis from UK national and regional networks (Carslaw and Ropkins 2012; Szulecka et al. 2017). Additional tools have been developed to supplement the aforementioned databases by retrieving satellite data on terrestrial weather conditions and anthropogenic emissions of various pollutant species, e.g., `ecmwf_r` (Hufkens et al. 2019), which downloads data from the European Centre for Medium-Range Weather Forecasts dataset web services (ECMWF) and Copernicus's Climate Data Store (CDS) and on land cover (see `sen2r` by Ranghetti et al. 2020, for automatic discovery and management of satellite data provided by the Sentinel-2 program of the European Space Agency).

In this paper, we present the ARPALData software, an R package (R Core Team 2020) aimed at retrieving, handling, and analyzing air quality and weather data for the Lombardy region in Northern Italy. Lombardy has approximately 10 million inhabitants and the largest population density in Italy, with the highest per capita GDP across the country, sustained by an extensive network of enterprises (mainly services) and industries (Regional Statistical Yearbook 2017). Owing to its high industrial and housing density, coupled with growing land consumption, the region has one of the worst air quality in Europe and worldwide (Carugno et al. 2016; Raffaelli et al. 2020). Therefore, the case study of Lombardy and the Po Valley has received considerable attention from the media and international researchers (Collivignarelli et al. 2020).

Air quality and weather monitoring in Lombardy is entrusted to the Regional Agency for Environmental Protection, namely ARPA Lombardia, which maintains ground monitoring networks and validates and provides data to the public (Maranzano 2022). ARPALData retrieves data directly from the Regione Lombardia Open Database, which provides information on several topics, including the environment, transportation, economy, and social issues in Lombardy.

Specifically, for air quality research, data provided by ARPA Lombardia on pollutants and local meteorology are the main institutional sources for researchers. These data have been used in multiple investigations, including monitoring long-term trends in air concentrations (Bigi and Ghermandi 2014; Arvani et al. 2016), source apportionment (Manigrasso et al. 2012), green areas management in urban settings (Terzaghi et al. 2020), the role of agriculture and livestock in fine particulate emissions (Fassó et al. 2023; Marongiu et al. 2023), the impact on air quality of local traffic policies (Fassó 2013; Gibson and Carnovale 2015; Maranzano et al. 2020), the impact of restrictive policies related to COVID-19 (Camelletti 2020; Bontempi 2020; Lovarelli et al. 2021; Bontempi et al. 2022; Fioravanti et al. 2022; Maranzano and Pelagatti 2023; Maranzano et al. 2023), and epidemiological studies linking pollution, mortality, and diseases (Angelici et al. 2016; Carugno et al. 2017; Zoran et al. 2020a, b).

Despite the wide dissemination of ARPA data for research and local policies, data collection in Lombardy suffers from several critical issues, some of which are directly addressed by ARPALData. In addition, there is currently no national platform that systematically collects, harmonizes, and disseminates data from individual regional agencies. The joint collection of data pertaining to multiple Italian regions requires access to specific databases that are not uniform and have unique peculiarities. Therefore, ARPALData is an attempt to standardize the environmental data discovery process for a specific region and broaden the debate on the need for a unified national system by involving researchers and experts from other agencies and institutions in its development.

The remainder of this paper is organized as follows. Section 2 describes the main issues addressed by the software by explaining how the data collection step is currently possible. Section 3 describes the available data collected using the software and the data quality standards guaranteed by ARPA Lombardia. Section 4 presents the functions/commands provided to the users of ARPALData in detail, specifying their objectives and the main technical details. Section 5 provides three guided examples of how ARPALData can be used to cover the major software capabilities. Throughout the text, the syntax (included in the appropriate code chunks) and results obtained are discussed. Finally, Sect. 6 provides information on software availability, as well as plans for future developments and enhancements.

2 Issues addressed by ARPALData

ARPALData addresses several issues concerning data retrieval and management in the Lombardy region. The first issue regards the potential challenges posed by data access and downloading. Official air quality and weather data for Lombardy can be retrieved using two institutional open data sources. The first source is the online dashboard provided by ARPA Lombardia, accessible at <https://www.arpalombardia.it/temi-ambientali/meteo-e-clima/guida-richiesta-dati/> for weather and <https://www.arpalombardia.it/dati-e-indicatori/aria/> for air quality (both platforms were accessed

on December 7th, 2023). The dashboard allows users to inquire about measurements collected from one or more specific sensors installed on the control units that are currently active on the network. For each query, the user can select a maximum of six sensors simultaneously and data referring to no more than one year (i.e., 365 consecutive days). The output is a spreadsheet sent via email to the requester. Thus, data collection requires considerable time and many repeated queries. The second option is to directly enquire into the official regional open database available at <https://www.datilombardia.it/> (accessed on December 7th, 2023). However, accessing data through a regional open database requires nontrivial knowledge of data management and file management tools, thereby complicating data consultation for non-expert users. To this extent, ARPALData provides a user-friendly tool that does not require access to institutional platforms, is often accessible only through personal credentials, and ensures the real-time updating of information using standardized syntax.

The second issue concerns the original data format and the requirement for remarkable data manipulation and harmonization. ARPA Lombardia provides air quality and weather measurements at a pre-specified sampling frequency, that is, hourly air quality data from the ground monitoring network, daily air quality data at the municipal level, and infra-hourly (10 min) data for meteorological parameters. The concentrations of airborne pollutants or meteorological variables measured through the monitoring network, and daily municipal estimates of airborne pollutant concentrations, can be retrieved by accessing a regional open database. In all cases, users are only allowed to download year-specific spreadsheets containing measurements for the entire network, structured according to a long format. In the case of a multi-year query, the original data should be manually repeated by downloading and reshaping multiple spreadsheets. In addition, each sensor is uniquely identified using a sensor code that is specifically tabulated in an external metadata file. Users are not allowed to prefilter the measurements by date, station, or sensor. The collection procedure requires nontrivial data manipulation (data cleaning and table joining) of the raw measurements and metadata. ARPALData directly collects data from the regional open database source and provides information reshaped according to a two-way panel structure by station (i.e., each entry is univocally identified by the station identification code or name and the time stamp) or a three-way panel format by sensor (i.e., each entry is univocally identified by the sensor identification code and the time stamp). Through user-defined arguments, the functions for downloading data used in the package allow the retrieval of user-specified measurements taken on the entire network or on a subset of the available monitoring stations for any user-specified time period. Furthermore, the user can request data at the original sampling frequency or aggregated according to customized aggregation functions (e.g., monthly averages or weekly maxima). The required observations are directly combined and harmonized into a single data frame object of the class ARPALdf. Eventually, in addition to supplying raw data, ARPALData aims to provide key descriptive statistics and graphic tools to complement official reports and personal analyses, allowing users to learn more about issues related to air quality and meteorology in the region.

The third issue addressed by ARPALData concerns the language. All the original explanatory metadata (i.e., registry tables of monitoring stations or municipalities) and the raw data from ARPA and Regione Lombardia are reported in Italian. This could discourage access by non-Italian-speaking users who are unable to understand the information. Therefore, the dissemination and use of air quality and weather data for Lombardy are greatly reduced and made almost exclusively by domestic affairs. To address this challenge, ARPALData provides all measurements, metadata, and subsequent analytical tools in English, thus facilitating accessibility for both international and domestic users.

Finally, although the data are geographically limited to the Lombardy region, ARPALData allows accessing to environmental measures that are public but not otherwise available through other software or international platforms. For example, as highlighted by Fassó et al. (2023), the European Environmental Agency (EEA) platform provides measures for only a subset of pollutants in Lombardy, from which ammonia or metal concentrations are excluded.

3 Available air quality and weather data

The ARPALData software handles the following datasets from the ARPA Lombardia open data:

- hourly or daily air quality measurements from the ground monitoring network (174 stations activated since 1968, thereof 93 currently active) since 1968;
- 10-min weather measurements from the ground monitoring network (331 stations activated since 1989, thereof 267 currently active) since 1989;
- daily air quality measurement for the 1506 municipalities in Lombardy since 2011.

The available data are continuously updated to the current date. Figure 6 presents the maps of either the currently active or historical air quality and ground weather monitoring networks. The two networks are heterotopic, and in only a few cases do they overlap (i.e., few sites monitor both pollution and weather). The air quality ground monitoring network of ARPA Lombardia comprises fixed stations that provide continuous data at regular time intervals by means of automatic analyzers (ARPA Lombardia 2023b), that is, the sensors. Following the EU Air Quality Directive 2008/50/EC (European Parliament 2008), the pollutant species monitored are NO_X , NO_2 , SO_2 , CO , O_3 , PM_{10} , $\text{PM}_{2.5}$, NH_3 , black carbon, arsenic, benzene, benzo(a)pirene, and metals (such as, nickel, cadmium, and lead). The list of pollutants surveyed by the monitoring stations may differ depending on the environmental context in which the unit is being monitored. Therefore, not all stations are equipped with the same analytical toolkit (for instance, ammonia sensors are typically installed at rural sites only, while NO_2 sensors are spread all over the region). The stations are distributed throughout the region according to population density and type of territory

according to the criteria defined by Legislative Decree 155/2010 (Presidenza della Repubblica Italiana 2010). Pollutant data are provided at hourly frequency, except for PM₁₀, PM_{2.5}, and NH₃, for which daily averages are provided.

The meteorological monitoring network (ARPA Lombardia 2023c) comprises stations measuring the following parameters: hydrometric level (cm), snow depth (cm), precipitation of snow and rainfall (mm), temperature (°C), relative Humidity (%), global solar Radiation (W/m²), wind speed and direction (m/s and circular degrees), and speed and direction (m/s and degrees N) of wind gusts. All the weather measurements are provided by the agency with a 10-min frequency.

As for the air quality network, also the surveyed meteorological parameters may be different across stations depending on the reference environmental context (e.g., snow depth is measured only at mountain sites). In both networks, the distribution across the regional territory considerably change among pollutants and among weather parameters. As an illustration, in Figs. 7 and 8 we show the spatial distributions of the installed sensors for some relevant pollutants (i.e., NO₂, PM₁₀, PM_{2.5}, and ammonia) and weather parameters (i.e., temperature, rainfall, snow depth, and wind), respectively. Also, note that there exists a one-to-many relationship between stations and sensors, that is, each sensor is programmed to collect information on a single weather or pollution parameter, while multiple sensors can be installed in each station simultaneously.

Municipal values were obtained from regional-scale simulations performed using a chemical-physical air quality model. Therefore, these are not direct measurements, but estimates that also use data from the ARPA air quality detection network. Available pollutants at the municipal level are the weighted daily averages of PM₁₀, PM_{2.5}, and NO₂; the daily maximum of NO₂ and tropospheric O₃; and the daily maximum of the 8-h moving average of tropospheric O₃. Daily estimates of each of the pollutants mentioned are given for all municipalities.

Data provided by the regional agency for the current and immediately preceding year may contain measurements with a low degree of fidelity that must be checked through an appropriate data validation process. Statistical validation of the data involves an assessment phase for each series collected at each site that is completed by March 30 of the year following the measurement year. Therefore, the data should be considered non-definitive before that date. For municipal data in particular, estimates must be considered provisional until the validation process of measurements is completed, followed by the recalculation of values and the release of consolidated data; consolidated estimates for the previous year are typically available as of June 30 of the current year. Potential causes of the unreliability of certain measurements include malfunctioning or failure of monitoring stations that induce anomalies in the measured values and interruptions in data collection. To overcome these issues, ARPA Lombardia has adopted management procedures and field checks that guarantee that the data collected are as accurate as possible in compliance with European and Global standards. Specifically, the agency implements quality assurance procedures involving a

specialized team in charge of performing specific checks in the context of quality control and quality assurance activities, including preventive maintenance and corrective maintenance interventions. Readers may refer to Maranzano (2022) for an extensive discussion of the monitoring networks maintained by ARPA Lombardia on the type and quality of data collected and provided by the agency and on the agency's role in national and European contexts.

4 Structure and main features of ARPALData

The ARPALData package provides users with 17 functions, which can be grouped into 4 categories: 1) air quality and weather measurements retrieval; 2) data exploration, quality check and graphical representation; 3) class check; and 4) auxiliary functions. A detailed description and summary of ARPALData functions is reported in Table 3. Furthermore, while we discuss in detail the functions to download air quality and weather data, their metadata and exploratory statistics commands, we refer the reader to Supplementary Material 2 for an extended description of other auxiliary functions. In particular, we provide details on class-checking functions, temporal aggregation functions, and commands providing geometries (shapefiles) and maps of Lombardy at different spatial scales. In Supplementary Material 2, we also describe the input arguments of the functions (e.g., type and class, default values and values they can take), the structure of the outputs (e.g., the format of downloaded datasets), and the statistical and computational techniques implemented (e.g., temporal aggregation functions and spatial mapping).

Notably, ARPALData complies with the Comprehensive R Archive Network (CRAN) policy, which requires packages using Internet resources to provide an informative message if the resource is unavailable or has changed. In addition, several functions admit an optional `verbose` argument that prints auxiliary information on the screen that informs the user about which operations the command is performing and some object-specific information.

4.1 Retrieve air quality and weather data using ARPALData

The ARPALData package contains three functions that allow data to be downloaded from the ARPA Lombardia database. Each function is identified by the prefix `get_ARPA_Lombardia`, while the suffix determines the type of data to be downloaded: `AQ_data` returns the observed air quality measurements collected by the ARPA Lombardy ground-based survey system, `W_data` returns the observed meteorological measurements collected by the ground-based monitoring network, and `AQ_municipal_data` returns the air quality estimates for each municipality in Lombardy.

The `get_ARPA_Lombardia` commands share the same set of input arguments, which enable the user to customize the query regarding the variables and period of interest, the sampling frequency, and the grouping statistics. For example, it is possible to concurrently download the average daily concentration of one pollutant and the maximum daily concentration of another pollutant. The period of interest is specified by the user through the start and end dates (optionally stating the exact time); therefore, it is possible to download measurements pertaining to infra-annual sub-periods, but also to periods crossing one or more years. Users can also choose whether to consider the entire monitoring network (or all municipalities) or only a subset by entering the codes of the stations of interest. These codes are provided through the metadata functions presented in the following sub-sections, which enable the filtering and selection of stations based on user-specific criteria. Air quality measurements are originally provided at an hourly frequency (e.g., NO₂ or O₃) or daily frequency (e.g., PM₁₀ and PM_{2.5}), whereas weather measurements are all provided at a 10 min frequency. The user can customize a query such that the output dataset is aggregated to a specific temporal frequency (i.e., hourly, daily, weekly, monthly, or yearly). The temporal sampling rate is reduced by aggregating the measurements at each monitoring site according to a set of aggregation functions available to users, including centrality, variability, and shape measures. The temporal aggregation operation can also be applied to a generic object of the class `ARPALdf` through the `Time_aggregate` command. Examples of aggregation are provided in Sects. 5.1 and 5.2, as well as in Supplementary Material 1.

Download functions return as output an object of the classes `data.frame` and `ARPALdf` fully compatible with the `Tidyverse` (Wickham et al. 2019) and the graphical environment `ggplot2` (Wickham 2016). Recall that each sensor measures one and only one meteorological or pollutant parameter and that a one-to-many relationship exists between stations and sensors. In addition, multiple pollutants or weather parameters can be measured at each station. The default output object is a long-shaped (or two-way panel) `data.frame` with dimensions $(N \times T) \times (K + 3)$, where T is the number of time stamps, N is the number of stations (or municipalities), and K the number of available measurements (airborne pollutants or weather variables). The first three columns are the time stamp (`Date`), the station unique identification number (`IDStation`) and the station name (`NameStation`), respectively. Notice that the two-way panel takes on an exactly rectangular shape as for each station the same number of time stamps is reported. Thus, if a parameter is not measured at a specific station, a sequence of missing values is reported for all the time stamps. These missing values should not be subject to any imputation analysis because they belong to locations not considered in the sampling plan of ARPA Lombardia. The user can also request the dataset in the three-way panel format with respect to the sensors, where each row corresponds to the measurement collected from a given sensor installed at a given station at a given time stamp. Let us denote as S_i the number of sensors installed at a generic station $i = 1, \dots, N$. Then, the number of rows in the three-way panel is $\left(\sum_{i=1}^N S_i\right) \times T$, and the number of columns is

six, that is Date, IDStation, NameStation, IDSensor, Variable (i.e., meteorological parameters or pollutants) and Value.

The user can easily retrieve the metadata (or registry) on the air quality and weather ground monitoring networks owned by ARPA Lombardia and Lombardy municipalities by using the `get_ARPA_Lombardia_*_registry` functions. Similar to the data commands, the `AQ_registry`, `W_registry` and `AQ_municipal_registry`, return explanatory metadata for air quality monitoring stations, the weather monitoring stations and for the municipalities, respectively. Available metadata include the unique identification numbers of sensors and stations, station names, descriptions of the pollutant or weather parameters associated with each sensor, geographical coordinates (latitude, longitude and altitude), and administrative information, such as province, city, and municipality. Metadata may evolve over time as a result of updates performed by the regional agency on the stations. However, the software preserves the entire history of the monitoring networks as it contains information on currently active (as of the date of the request) and disabled sensors. Specifically, for all sensors, the date of initial operation is available, whereas for disabled sensors, the date of deactivation is also given. For the specific case of air quality stations, metadata additionally provide information on the ARPA zone and station type classification. The stations type classification arranges the sites according to the EEA stations type taxonomy (European Environmental Agency 2023), that is the stations are classified according to the urbanisation degree (physical distribution/density of buildings) of the surrounding area (Urban, Suburban, and Rural) and by predominant emission sources (Background, Traffic, and Industrial). On the other hand, the ARPA zone classification (ARPA Lombardia 2023d) has been defined by the Lombardy Region authorities through the Regional Decree 2605/2011 (Giunta regionale della Lombardia 2011), and distinguishes the regional territory into homogenous areas or agglomerations. Zoning reflects the main morphological characteristics of regional territory and is used to evaluate local compliance with national and international environmental objectives and limiting values. The Lombardy region is classified into three large urbanized areas (metropolitan areas of Milano, Bergamo and Brescia), urbanized areas in rural contexts (namely, the urbanized plains), rural areas (namely, the rural plains), mountainous areas, and valley bottoms. The user can retrieve the exact boundaries of the ARPA zones through the `get_ARPA_Lombardia_zoning` function, which returns the geometries (polygonal shape file, `sf` object) and a map of the ARPA zoning for Lombardy. The right panel of Fig. 2 shows a map of the zones that subdivide the Lombardy region.

4.1.1 Computational issues and parallel download

Regional air quality and weather data are retrieved from the API Socrata Open Data (SODA), which provides systematic access to the Lombardy Region database,

including the capability to filter, query, and aggregate data. ARPALData accesses the Socrata language using the package `RSocrata` (Devlin et al. 2023). Using appropriate queries, the API allows filtering of the stations of interest and the sub-period of dates to be downloaded.

By default, the software retrieves the measurements serially as follows: given a specific date range as the input, to speed up data access, the interval is split into 12¹ non-overlapping contiguous sub-intervals. For each sub-interval, an appropriate query is constructed to sequentially examine the database. Finally, the resulting data blocks are stacked into a single `data.frame` object. However, all the `get_ARPA_Lombardia_*_data` functions can adopt a parallel computation strategy to improve download speed depending on the computational capacity. Parallel computations are performed through the `Futureverse` (Bengtsson 2021). When parallel mode is active, the user can declare the number of parallel workers to be activated. By default the number of active workers is equal to half that of the available local cores. In addition, the user can declare the parallel strategy to be used according to the `Futureverse` syntax, which admits the `multisession` option (i.e., background sessions on the local machine) or a `multicore` setting (i.e., forked R processes on the local machine; not supported by Windows and RStudio).

We provide the results of a MonteCarlo simulation experiment designed to assess the computational benefits derived from parallel computation compared to the serial strategy. In particular, we compared the computing time required to download the air quality data for several combinations of the number of stations and months to be retrieved. For each combination, we replicated the download $N = 50$ times using parallel and serial settings. Randomness is included in the experiment by randomly selecting the stations and the period to be downloaded. At each iteration, the computing times using serial and parallel strategies were calculated from the same set of data to maintain full comparability of the procedures. The code used to reproduce the experiments is presented in Supplementary Material 1. Table 4 reports the average (and standard deviation) computation times under the serial strategy, whereas Table 5 shows the timing under the parallel strategy. By comparing the two tables we can conclude: 1) parallel computation drastically reduces the download time when the number of stations considered is large (average times are reduced by 50% to 65% for 50 stations or for the whole network); 2) regardless the number of stations, it is convenient to use the parallel strategy for periods of at least one year (particularly for a large number of years); 3) when the number of stations of interest and the period of interest are small, the serial strategy is preferable; 4) the computation time appears more affected by the length of the period than by the number of stations (for example, the average times for a single serial month tripled when going from 10 to all stations, while the average times for the number of stations grew substantially); and 5) the variability of the estimated times increases as the number of stations and number of months increases (this could be attributed to the need for a longer connection to the databases via the Socrata API, which is dependent on the accesses and the number of users connected at the time of download).

¹ The number of blocks (twelve) was determined by the authors so that given a one-year interval, each block corresponds to approximately one month of observations.

4.2 Data exploration and graphical analysis using ARPALData

The ARPALData package offers several functions to explore, analyze and represent the air quality and weather measurements previously accessed and processed. The rationale is that the user should be able to verify the statistical properties and the quality of the data of interest immediately after downloading. Indeed, environmental data are often subject to statistical issues, such as outliers or missing data, which compromise their quality and require ad-hoc corrections (Junninen et al. 2004; Gómez-Carracedo et al. 2014; Junger and Ponce de Leon 2015). There are four categories of exploratory tools offered by ARPALData: (1) descriptive statistics and graphs to summarize and depict the measured values globally (over the whole sample) and locally (site-specific or year-specific statistics and graphs); (2) analytical tools and graphs to detect outliers or out-of-scale values (outlier detection analysis); (3) tools describing patterns of missing data to distinguish between cases of missing at random or structural missing caused, for example, by failures in the monitoring network; and (4) correlation analysis between variables at global level (over the whole sample) and local level (correlation for individual monitoring sites).

The above-mentioned tools are available to users through the command `ARPALdf_Summary`, which can be applied to any object of class `ARPALdf`. This command provides the user with descriptive statistics summarizing the dataset and plots. This function pursues the goal to give an overview of the statistical quality of the data at hand. The function can be customized by the user by properly setting the arguments, each of which provides information on the specific statistical tools to be employed. Moreover, the functions admit a `verbose` setting, that screen prints auxiliary information useful for understanding the structure of the `data.frame`, such as the total number of observations, number of monitoring sites included, number of time stamps and time frequency. By default, `ARPALdf_Summary` returns for each variable a set of central tendency measures (e.g., mean, median, and quantiles), variability indicators (e.g., standard deviation and range) and descriptive statistics about missing values. However, the user can automatically compute the essential descriptive statistics (indicators of central tendency, variability, shape of distribution, and missing values) for all variables in the dataset by station and year. Furthermore, when more than one numerical variable is present in the dataset, the user can compute Pearson's linear correlation index among all possible pairs of variables by station and for the overall sample.

When the outlier detection analysis is activated, the function returns an overview of the (potential) outlier values detected using the Hampel Filter (Pearson et al. 2016) and the box-plot rule as follows. Let us denote a sequence of T measurements by X_t with $t = 1, \dots, T$. The Hampel Filter (or identifier) classifies all observations that fall outside the range $\pm 3 \times MADM$ as outliers, where $MADM$ is the median absolute deviation from the median (i.e., $1.4826 \times \text{Median}(|X_t - \text{Median}(X_t)|)$). Notice that the correction factor 1.4826 adjusts the distance such that, on average, the $MADM$ estimate is equal to the standard deviation for Gaussian data. Conversely, the box-plot rule classifies all the observations above the *typical upper value* $X_{0.75} + 1.5 \times IQR$ or below the *typical lower value* $X_{0.25} - 1.5 \times IQR$ as outliers, where $IQR = X_{0.75} - X_{0.25}$ is the interquartile range. Both identifiers use the median

as *typical* value because it is robust to the presence of outliers and the positive skewness that environmental data are most likely to show (Mudelsee and Alkio 2007). In practice, the box-plot rule is often more moderate than the Hampel Filter in identifying outliers. Therefore, their comparison and potential disagreements can serve as a robustness check for users in uncertain situations. Practically, the user is provided with threshold values and the percentage of observations exceeding these values for each variable.

The user can also activate exploratory missing data analysis based on gap length statistics. In particular, the function `ARPALdf_Summary` returns information regarding the temporal distribution of missing values (i.e., sequences of missing values) for each variable by station. Because we are dealing with time series data, we are interested in analyzing the gap length, that is, the temporal distance between one observation and the following. In particular, for each variable and station, the function summarizes the pattern of missing values by reporting summary statistics (e.g., average, variability, and quantiles of gap lengths) in the same temporal unit data (measured in hours, days, weeks or months depending on the frequency of the input data). For example, if the input dataset has a daily frequency, the gap statistics will be expressed in terms of number of days. Missing patterns can be detected by analyzing gap statistics. For instance, if the mean value of the gap length is greater than the minimum value at a certain monitoring site, it means that at least one observation in the time series is missing. When other statistics, such as the median, maximum, and percentiles also differ, then the number of missing observations increases. When all statistics are equal to 1, the variable of interest does not present missing values.

The data exploration phase can be enhanced and supplemented using graphical analysis by station and for the overall sample. The `ARPALdf_Summary` function allows the production of distributional box-plots, histograms, and kernel density plots computed for the full sample. Also, the user can employ the command `ARPALdf_Summary_map` to represent, on a geo-referenced map, any of the descriptive statistics computed by the `ARPALdf_Summary` for the available monitoring sites (or for municipalities if municipal data are used). Extensive examples of the usage of `ARPALdf_Summary` and `ARPALdf_Summary_map` in several contexts are provided in Sect. 5.

5 Case studies and practical examples

We now present three guided examples in which ARPALData is used to retrieve and analyze air quality and weather data across Lombardy.² For brevity, we comment only on the relevant chunks of code that utilize package-specific commands or routines. The full vignette of the three examples is either available in Supplementary Material 1 or at the GitHub webpage <https://github.com/PaoloMaranzano/ARPALData>.

² Data for the three examples have been collected on December, 7th, 2023.

5.1 Case study 1: characterizing air quality according to the local context

In the first example, we download and analyze the weekly PM₁₀ concentrations observed across the entire air quality ground monitoring network. We are interested in describing the average and maximum weekly concentrations from 2019 to 2022 according to the local context in which the stations are installed. Therefore, we characterized PM₁₀ concentrations according to station type and regional zoning, as described in Sect. 4.1.

We begin by downloading the registry (or metadata) concerning the air quality ground monitoring network. We are interested in retrieving information regarding the currently active sites that monitor PM₁₀, and were installed before January 1st, 2019. The full list of stations is filtered based on these criteria.

```
### Download registry table of air quality monitoring sites
reg <- get_ARPA_Lombardia_AQ_registry()

### Filter sensors:
#     Stations activated before 2019
#     Currently on service / active
#     Measuring NO2
reg <- reg %>%
  filter(DateStart <= "2019-01-01",
         is.na(DateStop),
         Pollutant %in% c("PM10"))
```

By computing the frequency table of the available monitoring sites, we notice that the number of stations is not evenly distributed among station types. In fact, Suburban-Traffic (ST), Suburban-Industrial (SI), Urban-Industrial (UI), and Rural-Industrial (RI) counts are less than three stations each. Therefore, to maintain consistency when aggregating observations, we filter the station list by removing the aforementioned classes of sites. The final number of stations is 59. For stations that comply with the original filters and those concerning station typologies, we collect weekly observations of PM₁₀ from 2019 to 2022. The measurements are downloaded using a parallel setup.

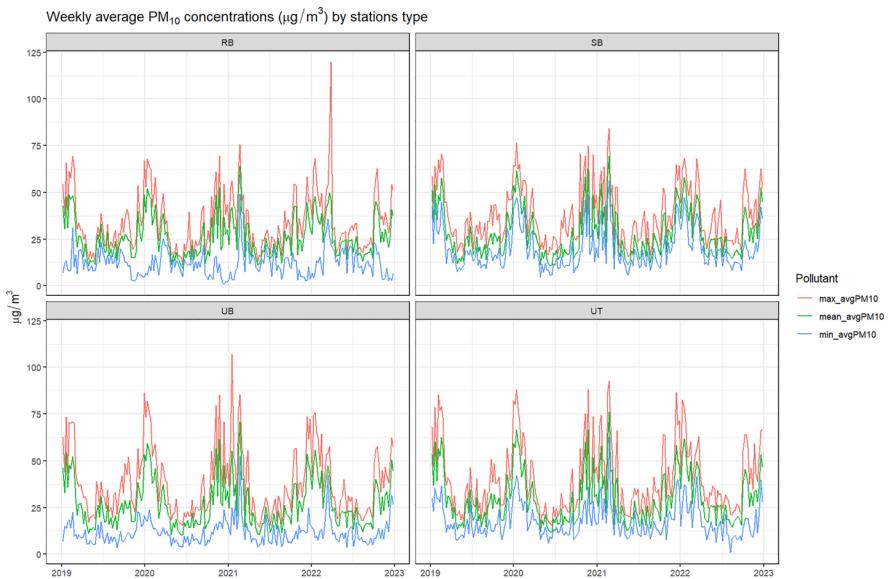


Fig. 1 Weekly PM₁₀ minimum (blue), average (green), and maximum (red) concentrations observed from January 2018 to December 2021 by station type, that is Rural-Background (RB), Suburban-Background (SB), Urban-Background (UB), and Urban-Traffic (UT)

```
### Download weekly average and weekly maximum of PM10 for all the available
      stations from 2019 to 2022. Parallel computation activated
Data_AQ <- get_ARPA_Lombardia_AQ_data(ID_station = reg$IDStation,
                                         Date_begin = "2019-01-01",
                                         Date_end = "2022-12-31",
                                         Frequency = "weekly",
                                         Var_vec = c("PM10", "PM10"),
                                         Fns_vec = c("mean", "max"),
                                         parallel = TRUE)
```

To exploit information about the local contexts in which the stations are located, we aggregate the weekly average and maximum values by station type and zoning, respectively. We start by aggregating the weekly average PM₁₀ concentrations and computing the lower and upper bounds (minimum and maximum), as well the central value (mean), by station type. The resulting time series are then plotted (Fig. 1) into four sub-panels, each corresponding to a different station type. In each panel, the maximum weekly, average, and minimum concentrations are depicted in red, green, and blue, respectively.

The charts show that the average weekly concentrations exhibit different characteristics depending on the type of station. For instance, the weekly average PM₁₀ concentrations observed at Urban-Traffic (UT) stations frequently exceed 75 µg/m³

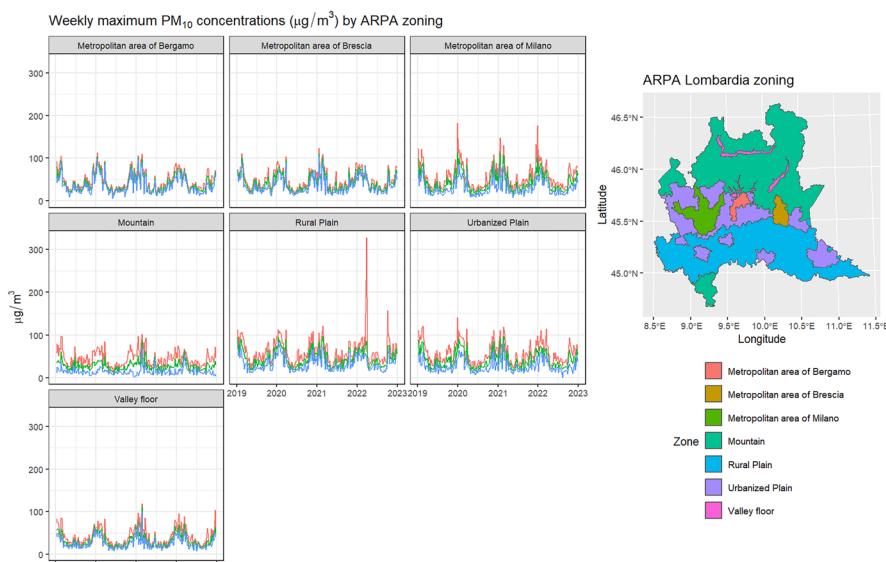


Fig. 2 Weekly PM_{10} minimum (blue), average (green), and maximum (red) concentrations observed from January 2018 to December 2021 by ARPA zone

and are compact during summer but more volatile in winter. The vertical distance between the maximum (red series) and the minimum values (blue series) is approximately $20\mu\text{g}/\text{m}^3$ in summer and $40\mu\text{g}/\text{m}^3$ in winter. Similar considerations can be made for Suburban-Background (SB) stations. In contrast, in Rural-Background (RB) stations, average concentrations rarely exceed $75\mu\text{g}/\text{m}^3$, but vary greatly in the winter months (e.g., the vertical distance between the minimum and maximum of the average in January 2020, 2021, and 2022, is approximately $50\mu\text{g}/\text{m}^3$).

Similar to the process for station types, we aggregate the weekly PM_{10} maxima by zones (obtained through the command `get_ARPA_Lombardia_zoning`). The maximal concentrations are aggregated by computing their ranges (from maximum and minimum values) and the average quantities. The aggregated time series are shown in Fig. 2 where the plot presents seven subpanels in which the weekly minimum (blue), weekly average (green), and weekly maximum concentrations (red) are depicted for each zone.

The temporal plots suggest that the maximum PM_{10} values are particularly high in the Milan metropolitan area (the socio-economic center of the region with a high degree of urbanization and vehicles) and in the plains (both urbanized and rural). In the latter case, particulate matter could be attributed to the numerous agricultural activities and livestock farms, which are responsible for the emission of large amounts of secondary particulate matter and ammonia through the spreading of slurry and animal manure (Fassó et al. 2023; Marongiu et al. 2023). The Alps to the north and the valleys had significantly lower maximum values. The other two urban agglomerations (Bergamo and Brescia) exhibit an intermediate situation, where the weekly PM_{10} maxima are steady and less volatile. Stations

located in northern Alpine or pre-Alpine areas register significantly lower maximum values.

5.2 Case study 2: air quality conditions during COVID-19 pandemic at municipal level in Lombardy

In the second example, we use ARPALData to investigate how air quality improved across the region due to the shutdown imposed during the COVID-19 pandemic in spring 2020. Our goal is to compare the average values observed across the 1516 municipalities of Lombardy during the lockdown period (March 8th to May 18th) with the concentrations observed in the same subperiods in 2018, 2019, and 2021.

As a first step, we download the daily estimates of NO₂ concentrations at the municipal level from January 2018 to December 2021.

```
### Download daily NO2 concentrations at municipal level from 2018 to 2020
Data_mun <- get_ARPA_Lombardia_AQ_municipal_data(
  Date_begin = "2018-01-01",
  Date_end = "2021-12-31",
  Frequency = "daily",
  Var_vec = c("NO2_mean"),
  Fns_vec = c("mean")
)
```

Then, the full dataset is filtered to retain the values recorded at each municipality during the lockdown subperiod of 2018, 2019, 2020, and 2021 (the output dataset is labelled as Data_spring). The average concentrations during this period are then computed by aggregating the daily values using the function Time_aggregate with argument Frequency = "yearly".

```
### Filter observations with respect to the date: only measures from 8th March to
### 18th May of each year
Data_spring <- Data_mun %>%
  filter(Date >= "2021-03-08" & Date <= "2021-05-18" |
         Date >= "2020-03-08" & Date <= "2020-05-18" |
         Date >= "2019-03-08" & Date <= "2019-05-18" |
         Date >= "2018-03-08" & Date <= "2018-05-18")

### Computing period averages (8th March - 18th May) of NO2 concentrations
Data_y <- Time_aggregate(
  Dataset = Data_spring,
  Frequency = "yearly"
)
```

The average NO₂ concentrations for the available municipalities and years are graphically represented on a map by employing the `ARPALdf_Summary_map` command. Each map adopts a color scale ranging from green (low concentration), to yellow (medium concentration), to red (high concentration). To make the color scale uniform throughout the maps, we set a common central value by imposing a mid point equal to the regional average concentration over the period of interest in 2018 (i.e., `val_midpoint = mid_conc_2018`).

```
# Mapping NO2 concentrations for 2021 by municipality
map_21 <- ARPALdf_Summary_map(
  Data = Data_y %>% filter(lubridate::year(Date) == 2021),
  Title_main = TeX("March 8^{th} to May 18^{th}, 2021"),
  Variable = "NO2_mean",
  val_midpoint = mid_conc_2018
)
```

Figure 3 shows the combined maps of NO₂ concentrations. During spring 2020 the entire region shifted from a yellow-orange color (concentrations above the 2018 average) to greenish colors (concentrations below the 2018 average). In particular, the alpine zone in the north and the Apennines in the south-west experience remarkable improvements. The situation in the highly industrialized and urbanized central belt remains critical. Notably, the four main cities in the region (Milan, Monza, Bergamo, and Brescia) are connected by an orange stripe, perfectly overlapping with the route of the main highway in Northern Italy, that is the A4 Turin-Trieste highway, which is approximately 220 km long and is travelled by about 153 thousand vehicles daily in the Lombardy segment (AISCAT 2017). Among others, supportive findings were found by Fioravanti et al. (2022), who revealed a statistically significant decrease in NO₂ concentrations across the whole Po Valley during March and April 2020 (corresponding to the hard lockdown phase) compared to 2019.

5.3 Example 3: characterization of meteorological phenomena across Lombardy

In the third example, we use ARPALData to retrieve weather data for Lombardy to depict the main characteristics of local meteorology across Lombardy and study the heterogeneity due to the morphology of the territory. In addition, we show the use of data quality check functions provided in ARPALData to assess the actual statistical properties of the available data. In particular, as described in Sect. 4.2, ARPALData contains functions that allow the statistical exploration of a given input dataset, including an analytical tool for outlier detection and identification of recurrent patterns as regards missing values.

We download weather data for the entire ground monitoring network for 2021. Our interest lies in characterizing the monthly measurements of cumulative rainfall, average temperature, and average wind speed and direction.

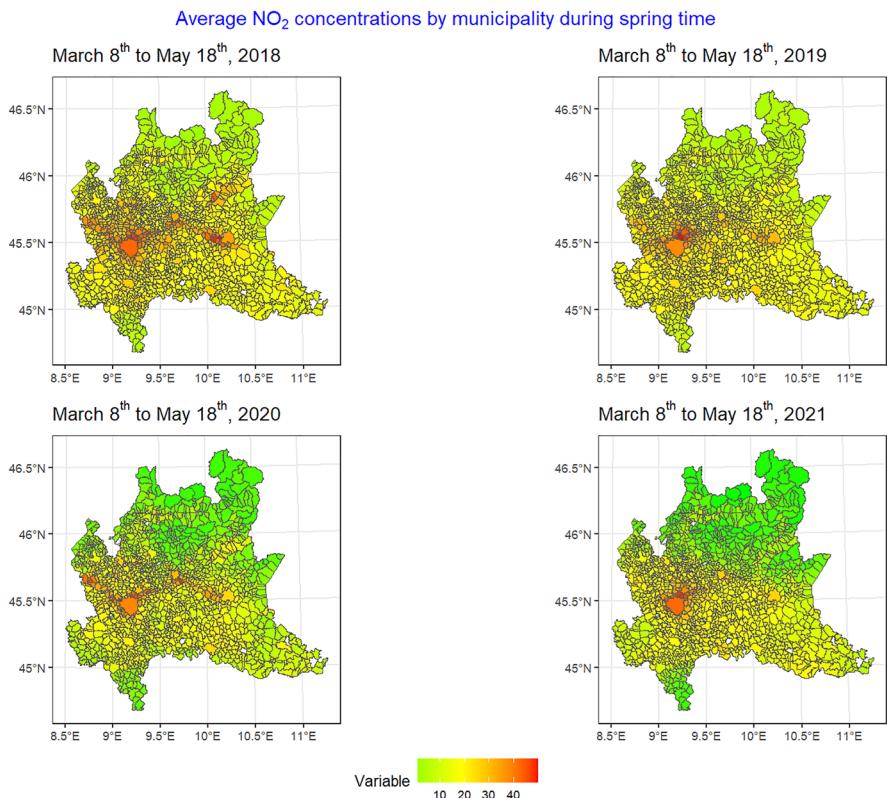


Fig. 3 Average NO₂ concentrations observed in 2018, 2019, 2020 and 2021 during the subperiod 8th March to 18th May. Data regarding eleven small municipalities are not available (grey unnoticeable areas)

```
### Downloading weather measurements for 2021 at monthly frequency
### Cumulated rainfall, average temperature and average wind speed and direction
Data_W <- get_ARPA_Lombardia_W_data(
  ID_station = NULL,
  Date_begin = "2021-01-01", Date_end = "2021-12-31",
  Frequency = "monthly",
  Var_vec = c("Rainfall", "Temperature", "Wind_speed", "Wind_direction"),
  Fns_vec = c("sum", "mean", "mean", "mean"),
  parallel = TRUE
)
```

After downloading the dataset of interest, we employ the `ARPALdf_Summary` function to explore the quality of the downloaded data.

Table 1 Outlier analysis for each variable (Hampel Filter)

Variable	Lower_bound	Obs_below_low_count	Obs_below_low_perc	Upper_bound	Obs_above_upp_count	Obs_above_upp_perc
Wind_speed	0.00	0.00	0.00	1.42	84.00	3.15
Wind_direction	0.00	0.00	0.00	360.00	0.00	0.00
Temperature	-18.17	3.00	0.11	27.68	0.00	0.00
Rainfall	0.00	0.00	0.00	269.89	51.00	1.91

```
### Summary statistics and plots for the dataset:
# Descriptive statistics for the whole sample
# Descriptive statistics for each station and for each year
# Gap length (missing data) analysis
# Outlier detection
# Correlation analysis for each station
# Density plots for each variable
summ_data <- ARPALdf_Summary(
  Data = Data_W,
  by_IDStat = TRUE,
  by_Year = TRUE,
  gap_length = TRUE,
  outlier = TRUE,
  correlation = TRUE,
  density = TRUE,
  histogram = FALSE,
  verbose = FALSE
)
```

The Hampel Filter output (Table 1) shows that at the aggregate level, all the considered meteorological parameters are mainly affected by slight positive skewness and that potential outliers are abundant. For example, considering wind speed, almost 3% of the observations (84 values) are above the upper bound (column Obs_above_upp_count), whereas there are no values below the lower bound. Similar considerations can be applied to rainfall. However, we must consider that the upper bound of the wind speed is just above 1.42 m/s , which is moderate. The table also suggests that three potential negative outlier measurements (column Obs_below_low_count) could be detected for the average temperature. In other words, there were three months in 2021 at specific locations where the recorded temperatures were particularly lower than those of the rest of the sample.

Moving to the output of the gap length (missing values) analysis by station (summ_data\$Gap_length), no station present missing values, as all gap

statistics are always equal to 1 month. In Table 2, we report the gap statistics for five stations taken as an example.

As described in Sect. 4.2, descriptive statistics at the station or municipal levels, as well as time-specific measurements, can be graphically represented through maps using the ARPALdf_Summary_map function. In Example 2, we used such command, that depicts the average observed NO₂ values at the municipal level for several years. Similarly, the average temperatures observed at each monitoring station are mapped. This can be easily done by using as argument of ARPALdf_Summary_map, the site-specific descriptive statistics computed through ARPALdf_Summary. To show how the temperature is strongly affected by seasonality, geography, and local conditions, we represent the average temperature across Lombardy in two non-overlapping sub-periods, that is October to March 2021 (winter) and April to September 2021 (summer). Starting from the full dataset, we split the sample into the winter and summer subperiods and compute summary descriptive statistics by station for both subsets (namely, the object summ_data_spring_summer for the summer subperiod). Eventually, we plot the average seasonal temperature for each station using the ARPALdf_Summary_map function by setting as Data argument the Descr_by_IDStat\$Mean_by_stat table contained in the output of ARPALdf_Summary. To allow for the maps comparability, we set up a common reference scale for the colors by imposing an average color equal to the yearly grand mean across the region. The two maps are shown in Fig. 4.

```
### Summary statistics for spring/summer subperiod
summ_data_spring_summer <- ARPALdf_Summary(Data = Data_spring_summer,
                                              gap_length = FALSE,
                                              outlier = FALSE,
                                              correlation = TRUE,
                                              density = FALSE)

### Plot the map of average temperature by station during spring/summer period
map_T_summer <- ARPALdf_Summary_map(
  Data = summ_data_spring_summer$Descr_by_IDStat$Mean_by_stat,
  Variable = "Temperature",
  Title_main = "Average temperature in Spring/Summer 2021",
  col_scale = c("#FFFF00", "#FF9933", "#FF0000"),
  val_midpoint = mean_temp)
```

From the left map, the average summer temperature recorded in the Alps is significantly lower than that in the Po Valley in the south, whereas the thermal excursion may depend on the yearly subperiod. In fact, the summer range is approximately 30°, while the winter range is from -10° to 5°.

Gap length and outlier statistics by station can be summarized well either in tabular form or on a map. However, linear correlations are suitable for graphical representation and mapping, particularly when the number of stations is large. In particular, we propose to represent this on a map of average site-specific linear correlations between the observed average temperature and rainfall. This plot can be obtained

Table 2 Gap length (sequences of missing data) analysis for wind speed measurements at five stations

IDStation	Name station	Min_gap	q25_gap	Mean_gap	Median_gap	q75_gap	Max_gap	SD_gap
58	Carona Lago Fregabolgia	1.00	1.00	1.00	1.00	1.00	1.00	0.00
60	Villa Di Chiavenna	1.00	1.00	1.00	1.00	1.00	1.00	0.00
100	Milano Lambrate	1.00	1.00	1.00	1.00	1.00	1.00	0.00
102	San Colombano Al Lambro	1.00	1.00	1.00	1.00	1.00	1.00	0.00
106	Varzi Nivione	1.00	1.00	1.00	1.00	1.00	1.00	0.00

Since observations have monthly frequency, gap lengths statistics are measured as months

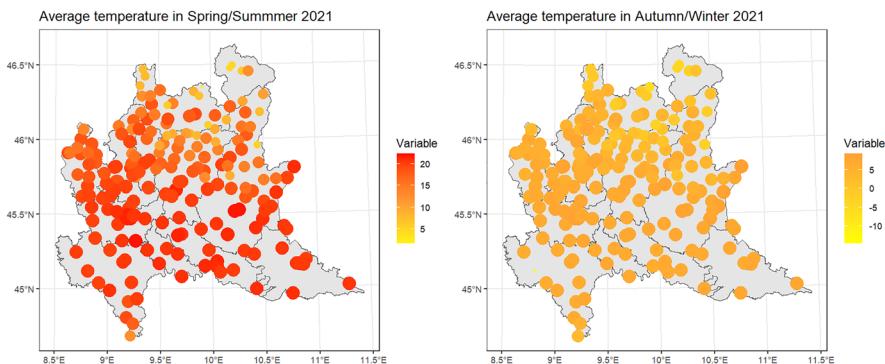


Fig. 4 Average temperature observed in the subperiod April - September 2021 for each station (left panel) and average temperature observed in the subperiod October - March 2021 for each station (right panel). The central value of the colors scales is the regional grand mean temperature recorded in 2021

by employing the ARPALdf_Summary_map filled with the argument Data = summ_data_spring_summer\$Cor_matrix, where Cor_matrix is the linear correlation matrix computed using the ARPALdf_Summary function on the subsample. By repeating this operation on the winter and summer subsamples, we compare the dynamics of Pearson's linear correlation index across the climatic seasons. The correlation maps for the two sub-samples are shown in Fig. 5.

```
### Plot map of linear correlation among maximum temperature and rainfall by
# station during spring/summer period
map_corr1 <- ARPALdf_Summary_map(
  Data = summ_data_spring_summer$Cor_matrix,
  Variable = "Temperature_max_Rainfall",
  Title_main = "Correlation among max temperature and rainfall in
Spring/Summer 2021",
  val_midpoint = 0)
```

The maps show how the correlation between rainfall and temperature strongly depends on the morphology and season. In particular, the mountainous areas (Alps) in the north exhibit strong positive correlations in both summer and winter (red or orange dots), while the Po Valley in the south shows strong negative correlations (green points). In the former case, the positive correlation might indicate that, at higher altitudes, lower temperatures are accompanied by low rainfall during summer. In contrast, plains are characterized by a scarcity of rainfall and high temperatures, particularly during summer. The correlations shown in these two maps agree with the negative monthly rainfall anomalies that characterized the year 2022 in the entire Lombardy region, as well as the trends in the maximum cumulative rainfall values across the 24-hours day. An extensive discussion of rainfall and temperature anomalies registered across

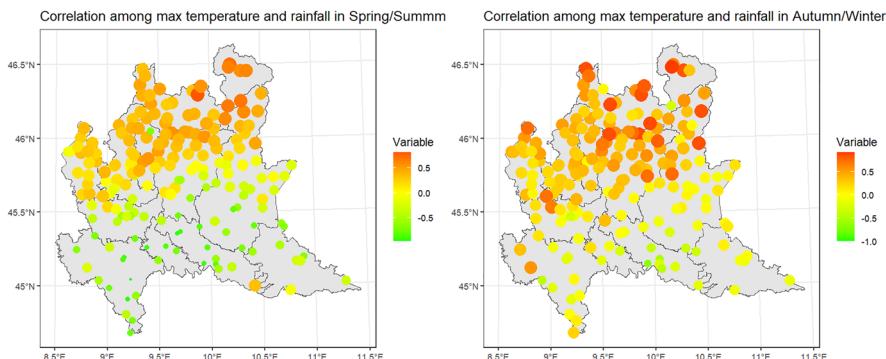


Fig. 5 Linear correlation among maximum temperature and rainfall observed in the subperiod April – September 2021 for each station (left panel) and linear correlation among average temperature and rainfall observed in the subperiod October – March 2021 for each station (right panel). Green points are associated with negative values of linear correlations, while orange points are associated with positive values. Yellow points mean uncorrelated data

Lombardy during the considered period can be found in the report "Climate, natural risks and availability water supply in Lombardy in 2022" (ARPA Lombardia 2023a). In particular, the report states that the climate in 2022 was classified as *hot-dry* with negative rainfall anomalies distributed for most of the year in the plain areas, whereas the Alpine and pre-Alpine areas experienced positive anomalies only in the first autumn months.

6 Software availability and future developments

The ARPALData software is freely available on the CRAN of R at <https://cran.r-project.org/web/packages/ARPALData/index.html> since July 2021. The currently available release is 1.5.0 (September 2023). ARPALData runs on Microsoft Windows, Linux and Apple Mac machines. No special hardware other than a standard desktop computer is required to run it. Future software releases will potentially include station or sensor selection using coordinates (i.e., quadrants, single coordinates, and NUTS levels); interactive and dynamic maps; and spatial interpolation (e.g., IDW) for an entire region or a user-defined area. To extend the discussion on the need to develop a unified platform at the national level, since 2023, part of the ARPALData team has been involved in the development of complementary software to retrieve air quality data at the European level, namely EEAaq (Tassan Mazzocco et al. 2023), from the open database of the European Environment Agency.

Available functions in ARPALData

See Table 3

Table 3 List of functions/commands available for the users of ARPALData (release 1.5.0, September 2023)

Command	Category	Description
get_ARPA_Lombardia_AQ_data	Download	Retrieves air quality data from the ground monitoring network
get_ARPA_Lombardia_W_municipal_data	Download	Retrieves meteorological data from the ground monitoring network
get_ARPA_Lombardia_AQ_municipal_data	Download	Retrieves air quality data at the municipal level
get_ARPA_Lombardia_AQ_registry	Download	Retrieves the metadata/registry for ground air quality network
get_ARPA_Lombardia_W_registry	Download	Retrieves the metadata/registry for ground weather network
get_ARPA_Lombardia_AQ_municipal_registry	Download	Retrieves the metadata/registry for air quality data at the municipal level
ARPALdf_Summary	Analysis	Computes summary descriptive statistics by station and by year, and produces descriptive plots of the sample
ARPALdf_Summary_map	Analysis and Mapping	Maps (reference) summary descriptive statistics (see ARPALdf_Summary) by station, by year or by municipality
map_Lombardia_stations	Mapping	Maps (georeference) the coordinates of the stations in a ARPALdf object
get_Lombardia_geospatial	Mapping	Returns the geometries (spatial polygon) of Lombardy at the municipal (LAU), provincial (NUTS3), and regional (NUTS2) levels.
get_ARPA_Lombardia_zoning	Mapping	Returns the geometries (spatial polygon) of the ARPA Lombardia's zoning of the region (see Sect. 4.1)
Time_aggregate	Auxiliary	Aggregates an ARPALdf object to a lower time frequency (e.g., from daily to monthly) using several aggregation functions
registry_KNN_dist	Auxiliary	Maps (reference) summary descriptive statistics (see ARPALdf_Summary) by station, by year or by municipality
is_ARPALdf	Class check	Checks if a R object is of class ARPALdf
is_ARPALdf_AQ	Class check	Checks if a R object is of class ARPALdf_AQ
is_ARPALdf_W	Class check	Checks if a R object is of class ARPALdf_W
is_ARPALdf_AQ_mun	Class check	Checks if a R object is of class ARPALdf_AQ_mun

Extended results of the MonteCarlo experiment for computing time under serial and parallel computation

See Tables 4, 5.

Simulations were computed between September 13th and 15th 2023 using a Windows virtual machine with Intel(R) Xeon(R) Gold 5120 CPU @ 2.20GHz and 16 virtual processors. The virtual machine accessed to an effective internet connection speed (tested on September 14th, 2023 using Ookla speed tester) of 1632 Mbps for download and 915 Mbps for upload.

We recall that, for some large data sets or complex analyses, the download step may require a remarkable computation time. In particular, when using the `get_ARPA_Lombardia_W_data` function to download meteorological data (which are provided at a 10 min frequency), the download time can be high even in parallel mode. It is therefore advisable to break long sequences of dates into smaller sequences (e.g., one year at a time) and repeat the download in a loop.

Table 4 Serial computation: average and standard deviation (under parentheses) computing time (in seconds) across the $N = 50$ MonteCarlo replications for several combinations of number of stations and number of months to be retrieved

Serial	Number of months				
	1	3	6	12	36
Number of stations					
10	13.06 (1.73)	20.44 (2.52)	31.56 (3.89)	53.90 (6.00)	146.66 (14.4)
20	15.22 (1.47)	33.86 (4.34)	55.94 (7.25)	104.82 (13.2)	273.92 (40.7)
50	30.14 (3.50)	63.06 (6.81)	140.48 (22.2)	210.00 (35.7)	548.30 (81.4)
All	48.72 (6.04)	130.72 (18.3)	253.18 (37.4)	538.62 (103)	1400.32 (222)

Table 5 Parallel computation (`parworkers` = 8 and `parfuturetype` = "multisession" options): average and standard deviation (under parentheses) computing time (in seconds) across the $N = 50$ MonteCarlo replications for several combinations of number of stations and number of months to be retrieved

Parallel	Number of months				
	1	3	6	12	36
Number of stations					
10	27.50 (1.40)	28.72 (0.858)	31.86 (1.43)	36.34 (2.98)	61.34 (4.62)
20	27.62 (0.725)	31.40 (1.68)	37.70 (2.51)	49.52 (5.37)	100.90 (11.4)
50	31.22 (1.20)	39.30 (2.82)	59.84 (7.27)	83.70 (11.9)	188.74 (14.2)
All	37.60 (2.29)	64.30 (6.35)	106.12 (14.4)	202.52 (37.1)	503.20 (75.5)

Historical and current air quality and weather ground monitoring network of ARPA Lombardia

See Fig 6

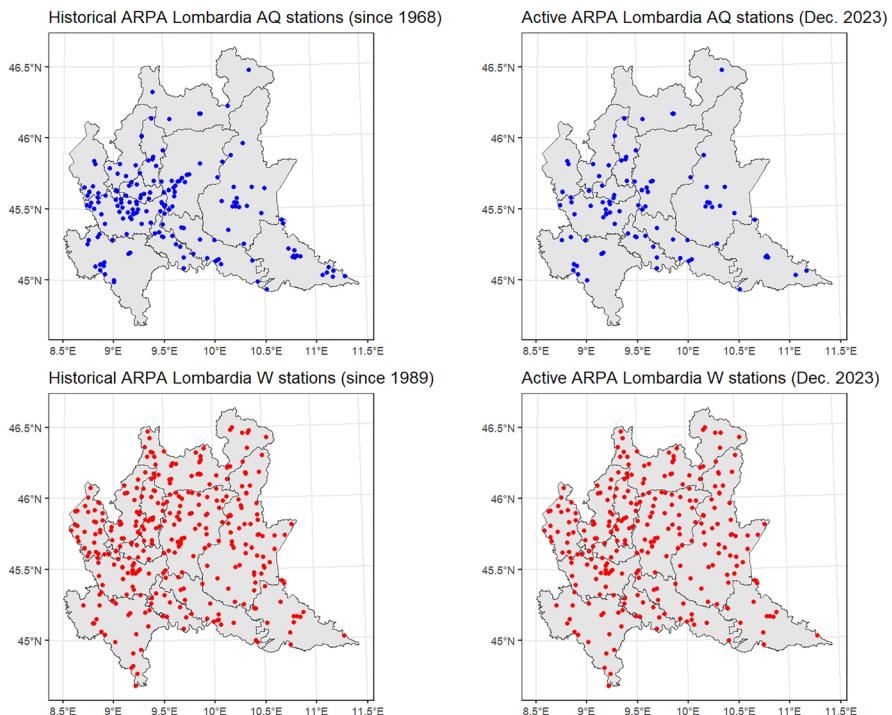


Fig. 6 Historical (left panesl) and currently active on December 2023 (right panels) air quality (upper panels) and weather (lower panels) ground monitoring network of ARPA Lombardia

Location of air quality and meteorology sensors for some relevant pollutants and parameters

See Figs. 7, 8

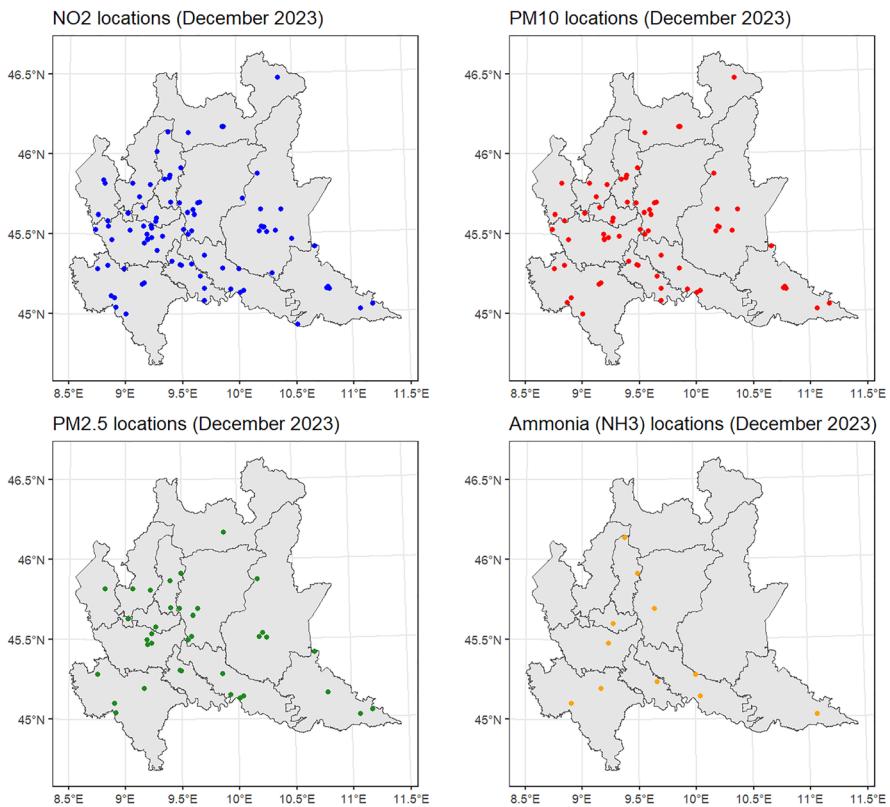


Fig. 7 Location of NO₂, PM₁₀, PM_{2.5}, and ammonia (NH₃) sensors managed by ARPA Lombardia (sensors active on December 2023)

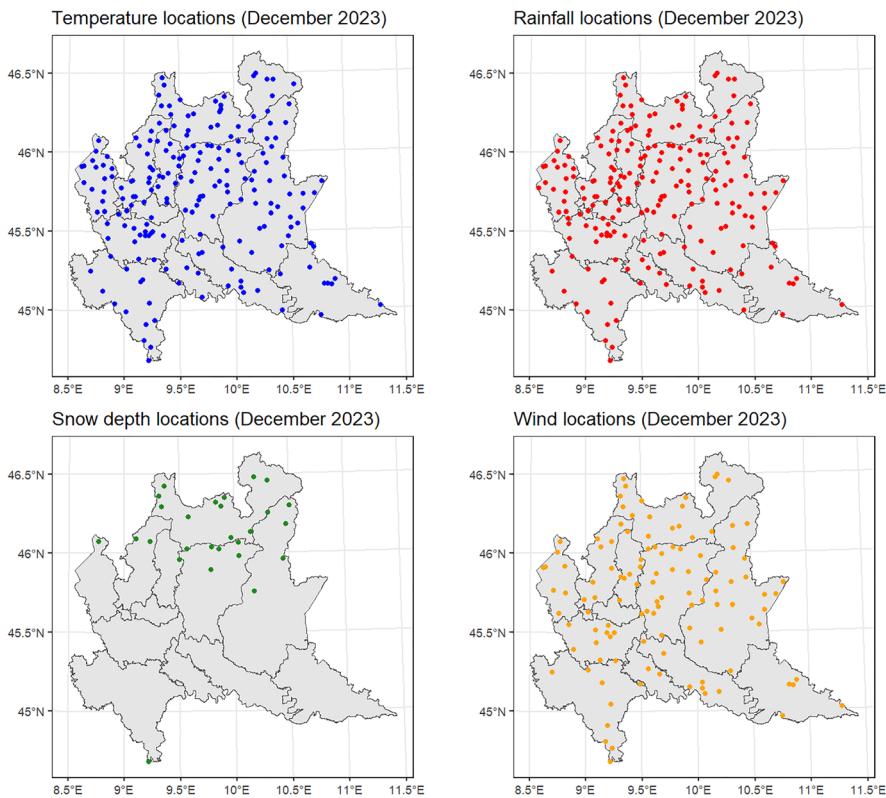


Fig. 8 Location of temperature, rainfall, snow depth, and wind speed/direction sensors managed by ARPA Lombardia (sensors active on December 2023)

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10651-024-00599-6>.

Acknowledgements The authors thank their colleagues from the "AgrImOnIA - The Impact of Agriculture on Italian Air Quality" research group, in particular Prof. Alessandro Fassó, for the support they have provided during the past few years while developing the ARPALData package; their colleagues from ARPA Lombardia, in particular, the Air Quality and INEMAR operational units; and the members of the SIS-GRASPA group for the discussion and "food-for-thought" provided during the 2023 annual conference held in Palermo.

Author contributions PM: Sect. 1 "Introduction" (design, original draft, literature review, software comparison, and revision); Sect. 2 "Issues addressed by ARPALData" (design, original draft, and revision); Sect. 3 "Available air quality and weather data" (design, original draft and revision); Sect. 4 "Structure and main features of ARPALData" (design of the contents, original draft, design and implementation of the R code, design and implementation of the Monte Carlo experiment, revision); Sect. 5 "Case studies and practical examples" (design of the contents, original draft, design and implementation of the code for the examples, revision of the text); Sect. 6 (original draft and revision). PM is the main developer and designer of ARPALData and is the maintainer on the CRAN. AA: Sect. 1 "Introduction" (revision); Sect. 2 "Issues addressed by ARPALData" (design and revision); Sect. 3 "Available air quality and weather data" (check and revision); Sect. 4 "Structure and main features of ARPALData" (design and revision of the code, design and revision of the Monte Carlo experiment, revision); Sect. 5 "Case studies

and practical examples" (design of the contents and revision of the code for the examples, revision of the text); Sect. 6 (revision). AA is co-developer of ARPALData.

Funding Open access funding provided by Università degli Studi di Milano - Bicocca within the CRUI-CARE Agreement.

Code availability The paper is accompanied by two Supplementary Material files.

Declarations

Conflict of interest None.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- AISCAT AISCAeT (2017) Traffico autostradale, veicoli teorici medi giornalieri e veicoli-km-autostrada (2015). Report, AISCAT, Associazione Italiana Società Concessionarie Autostrade e Trafori, https://www.asr-lombardia.it/asrlomb/it/14032traffico-autostradale-veicoli-teorici-medi-giornalieri-e-veicoli-km-autostrada?t=Tabella_&restrictBy=CCAUTOSTRADE_E_1831249587=Milano-Brescia%7CBrescia-Milano%7CTorino-Milano,CCANNO_63889777=2015
- Angelici L, Piola M, Cavalleri T et al (2016) Effects of particulate matter exposure on multiple sclerosis hospital admission in Lombardy region, Italy. Environ Res 145:68–73. <https://doi.org/10.1016/j.envres.2015.11.017>
- ARPA Lombardia (2023a) Clima, rischi naturali e disponibilità idrica in lombardia nel 2022. Report, ARPA Lombardia, https://www.arpalombardia.it/media/fd2je4v5/report_arpa_riscus_lombardia_2023.pdf
- ARPA Lombardia (2023b) Eelenco e posizione delle stazioni di monitoraggio qualità dell'aria e dei sensori. <https://www.dati.lombardia.it/Ambiente/Stazioni-qualit-dell-aria/ib47-atvt>. Accessed 22 Sep 2023
- ARPA Lombardia (2023c) Elenco e posizione delle stazioni meteorologiche e dei sensori. <https://www.dati.lombardia.it/Ambiente/Stazioni-Meteorologiche/nf78-nj6b>. Accessed 22 Sep 2023
- ARPA Lombardia (2023d) Mappa della zonizzazione della lombardia. <https://www.arpalombardia.it/temi-ambientali/aria/rete-di-rilevamento/zonizzazione/>. Accessed 22 Sep 2023
- Arvani B, Pierce RB, Lyapustin AI (2016) Seasonal monitoring and estimation of regional aerosol distribution over Po Valley, Northern Italy, using a high-resolution Maiac product. Atmos Environ 141:106–121
- Bengtsson H (2021) A unifying framework for parallel and distributed processing in r using futures. R J 13(2):208–227. <https://doi.org/10.32614/RJ-2021-048>
- Bigi A, Ghermandi G (2014) Long-term trend and variability of atmospheric pm₁₀ concentration in the po valley. Atmos Chem Phys 14(10):4895–4907. <https://doi.org/10.5194/acp-14-4895-2014>
- Bontempi E (2020) First data analysis about possible covid-19 virus airborne diffusion due to air particulate matter (pm): The case of Lombardy (Italy). Environ Res 186:109636. <https://doi.org/10.1016/j.envres.2020.109639>

- Bontempi E, Carnevale C, Cornelio A et al (2022) Analysis of the lockdown effects due to the covid-19 on air pollution in Brescia (Lombardy). Environ Res 212:113193. <https://doi.org/10.1016/j.envres.2022.113193>
- Cameletti M (2020) The effect of corona virus lockdown on air pollution: evidence from the city of Brescia in Lombardia region (Italy). Atmos Environ 239:117794. <https://doi.org/10.1016/j.atmosenv.2020.117794>
- Carslaw DC, Ropkins K (2012) Openair—an R package for air quality data analysis. Environ Model Softw 27(28):52–61. <https://doi.org/10.1016/j.envsoft.2011.09.008>
- Carugno M, Consonni D, Randi G et al (2016) Air pollution exposure, cause-specific deaths and hospitalizations in a highly polluted Italian region. Environ Res 147:415–424
- Carugno M, Consonni D, Bertazzi PA et al (2017) Temporal trends of pm10 and its impact on mortality in Lombardy, Italy. Environ Pollut 227:280–286
- Collivignarelli MC, Abbà A, Bertanza G et al (2020) Lockdown for covid-2019 in Milan: what are the effects on air quality? Sci Total Environ 732:139280. <https://doi.org/10.1016/j.scitotenv.2020.139280>
- Cruz-Alonso V, Pucher C, Ratcliffe S et al (2023) The Easyclimate R package: easy access to high-resolution daily climate data for Europe. Environ Model Softw 8:105627. <https://doi.org/10.1016/j.envsoft.2023.105627>
- Degraeuwe B, Pisoni E, Christidis P et al (2021) Sherpa-city: a web application to assess the impact of traffic measures on no2 pollution in cities. Environ Model Softw 135:104904. <https://doi.org/10.1016/j.envsoft.2020.104904>
- Devlin HDP, Schenk Jr. T, et al (2023) RSocrata: Download or Upload 'Socrata' Data Sets. <https://CRAN.R-project.org/package=RSocrata>, r package version 1.7.15-1
- European Environmental Agency E (2023) Classification of monitoring stations and criteria to include them in eea's assessments products. Report, European Environmental Agency, EEA, <https://www.eea.europa.eu/themes/air/air-quality-concentrations/classification-of-monitoring-stations-and>
- European Parliament (2008) Directive 2008/50/ec of the European parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for Europe. Official Journal of the European Union <https://eur-lex.europa.eu/TodaysOJ/>
- Fassò A, Rodeschini J, Moro AF et al (2023) Agrimonia: a dataset on livestock, meteorology and air quality in the Lombardy region, Italy. Sci Data 10(1):143. <https://doi.org/10.1038/s41597-023-02034-0>
- Fassò A (2013) Statistical assessment of air quality interventions. Stoch Environ Res Risk Assess 27(7):1651–1660. <https://doi.org/10.1007/s00477-013-0702-5>
- Fioravanti G, Cameletti M, Martino S et al (2022) A spatiotemporal analysis of no2 concentrations during the Italian 2020 covid-19 lockdown. Environmetrics 33(4):e2723. <https://doi.org/10.1002/env.2723>
- Gibson M, Carnovale M (2015) The effects of road pricing on driver behavior and air pollution. J Urban Econ 89:62–73. <https://doi.org/10.1016/j.jue.2015.06.005>
- Grange SK (2019) Technical note: saqget R package. <https://drive.google.com/open?id=1IgDODHqBHeWCTKLDAAxRyR7ml8ht6Od>
- Gómez-Carracedo MP, Andrade JM, López-Mahía P et al (2014) A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemom Intell Lab Syst 134:23–33. <https://doi.org/10.1016/j.chemolab.2014.02.007>
- Hufkens K, Stauffer R, Campitelli E (2019) The ecwmfr package: an interface to ECMWF api endpoints. <https://doi.org/10.5281/zenodo.2647531>, <https://bluegreen-labs.github.io/ecmwfr/>
- Junger WL, de Leon AP (2015) Imputation of missing data in time series for air pollutants. Atmos Environ 102:96–104. <https://doi.org/10.1016/j.atmosenv.2014.11.049>
- Junninen H, Niska H, Tuppurainen K et al (2004) Methods for imputation of missing values in air quality data sets. Atmos Environ 38(18):2895–2907. <https://doi.org/10.1016/j.atmosenv.2004.02.026>
- Giunta regionale della Lombardia (2011) Delibera di giunta regionale (dgr) n. 2605 del 30 novembre 2011. <https://www.regione.lombardia.it/wps/portal/istituzionale/HP/DettaglioRedazionale/servizi-e-informazioni/Enti-e-Operatori/ambiente-ed-energia/Inquinamento-atmosferico/zonizzazione-territorio-regionale/zonizzazione-territorio-regionale>
- Lovarelli D, Fugazza D, Costantini M et al (2021) Comparison of ammonia air concentration before and during the spread of covid-19 in lombardy (italy) using ground-based and satellite data. Atmos Environ 259:118534. <https://doi.org/10.1016/j.atmosenv.2021.118534>

- Manigrasso M, Febo A, Guglielmi F et al (2012) Relevance of aerosol size spectrum analysis as support to qualitative source apportionment studies. Environ Pollut 170:43–51. <https://doi.org/10.1016/j.envpol.2012.06.002>
- Maranzano P (2022) Air quality in Lombardy, Italy: a overview of the environmental monitoring system of ARPA Lombardia. Earth 3(1):172–203
- Maranzano P, Pelagatti M (2023) Spatiotemporal event studies for environmental data under cross-sectional dependence: an application to air quality assessment in lombardy. J Agric Biol Environ Stat. <https://doi.org/10.1007/s13253-023-00564-z>, <https://link.springer.com/article/10.1007/s13253-023-00564-z>
- Maranzano P, Fassò A, Pelagatti M et al (2020) Statistical modeling of the early-stage impact of a new traffic policy in Milan, Italy. Int J Environ Res Public Health 17(3):1088. <https://doi.org/10.3390/ijerph17031088>
- Maranzano P, Otto P, Fassò A (2023) Adaptive lasso estimation for functional hidden dynamic geostatistical model. Stoch Environ Res Risk Assess. Accepted for publication on May 1st 2023. <https://doi.org/10.1007/s00477-023-02466-5>, <https://link.springer.com/article/10.1007/s00477-023-02466-5>
- Marongiu A, Collalto AG, Distefano GG, et al (2023) Application of machine learning to estimate ammonia atmospheric emissions. Preprints <https://doi.org/10.20944/preprints202309.0607.v1>,
- Mudelsee M, Alkio M (2007) Quantifying effects in two-sample environmental experiments using bootstrap confidence intervals. Environ Modell Softw 22(1):84–96. <https://doi.org/10.1016/j.envsoft.2005.12.001>
- Pearson RK, Neuvo Y, Astola J et al (2016) Generalized Hampel filters. EURASIP J Adv Signal Process 1:1–18
- R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, <https://www.R-project.org/>
- Raffaelli K, Deserti M, Stortini M et al (2020) Improving air quality in the PO alley, Italy: Some results by the life-IP-Prepair project. Atmosphere 11(4):429
- Ranghetti L, Boschetti M, Nutini F et al (2020) “sen2r”: An r toolbox for automatically downloading and preprocessing sentinel-2 satellite data. Comput Geosci 137:104473. <https://doi.org/10.1016/j.cageo.2020.104473>
- Regional Statistical Yearbook R (2017) Regional statistical yearbook of lombardia in europe 2017/2018. Regional Statistical Yearbook, RSY
- Presidenza della Repubblica Italiana (2010) Decreto legislativo 155/2010, 13 agosto 2010-attuazione della direttiva 2008/50/ce relativa alla qualita' dell'aria ambiente e per un'aria piu' pulita in europa. <https://www.gazzettaufficiale.it/eli/id/2010/09/15/010G0177/sg>
- Ryberg KR, Vecchia AV (2012) waterData—An R package for retrieval, analysis, and anomaly calculation of daily hydrologic time series data, version 1.0. <https://doi.org/10.3133/ofr20121168>, <http://pubs.er.usgs.gov/publication/ofr20121168>
- Szulecka A, Oleniacz R, Rzeszutek M (2017) Functionality of openair package in air pollution assessment and modeling-a case study of Krakow. Ochrona Srodowiska i Zasobów Naturalnych 28(2):22–7
- Tassan Mazzocco A, Maranzano P, Borgoni R (2023) EEAaq: handle air quality data from the European environment agency data portal. <https://CRAN.R-project.org/package=EEAaq>, r package version 0.0.3
- Terzaghi E, de Nicola F, Cerabolini BEL et al (2020) Role of photo- and biodegradation of two pahs on leaves: modelling the impact on air quality ecosystem services provided by urban trees. Sci Total Environ 739:139893. <https://doi.org/10.1016/j.scitotenv.2020.139893>
- Vitolo C, Russell A, Tucker A (2016) rdefra: interact with the UK air pollution database from Defra. J Open Source Softw 1(4):51
- Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, New York
- Wickham H, Averick M, Bryan J et al (2019) Welcome to the Tidyverse. J Open Source Softw 4(43):1686. <https://doi.org/10.21105/joss.01686>
- Zoran MA, Savastru RS, Savastru DM et al (2020) Assessing the relationship between ground levels of ozone (o3) and nitrogen dioxide (no2) with coronavirus (covid-19) in Milan, Italy. Scie Total Environ 740:140005. <https://doi.org/10.1016/j.scitotenv.2020.140005>
- Zoran MA, Savastru RS, Savastru DM (2020) Assessing the relationship between surface levels of pm2.5 and pm10 particulate matter impact on covid-19 in Milan, Italy. Sci Total Environ 738:139825. <https://doi.org/10.1016/j.scitotenv.2020.139825>

Authors and Affiliations

Paolo Maranzano^{1,2} · Andrea Algieri³

 Paolo Maranzano
paolo.maranzano@unimib.it

Andrea Algieri
a.algieri@arpalombardia.it

¹ Department of Economics, Management and Statistics (DEMS), University of Milano-Bicocca (UniMiB), Piazza dell'Ateneo Nuovo, 1, 20126 Milan, Italy

² Fondazione Eni Enrico Mattei (FEEM), Corso Magenta, 63, 20123 Milan, Italy

³ Department of Lodi and Pavia, U.O. Attività Produttive e Controlli, Agenzia Regionale per la Protezione dell'Ambiente Lombardia (ARPA Lombardia), Via Nino Bixio, 13, 27100 Pavia, Italy