Environmental Modelling and Software 138 (2021) 104976

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

aiRe - A web-based R application for simple, accessible and repeatable analysis of urban air quality data

Juan José Díaz [a,b], Ivan Mura [c,*], Juan Felipe Franco [d], Raha Akhavan-Tabatabaei [e]

[a] TNG Technology Consulting, Munich, Germany
[b] Department of Industrial Engineering, Universidad de Los Andes, Bogotá, Colombia
[c] Department of Electrical and Computer Engineering, Duke Kunshan University, Kunshan, China
[d] Research and Knowledge Department, Hill Consulting, Bogotá, Colombia
[e] Sabanci Business School, Sabanci University, Istanbul, Turkey

ARTICLE INFO

Keywords:
aiRe
Air quality software
Open source
R
Shiny web application

ABSTRACT

Recent technological advances in collecting data on emission sources, meteorological conditions and concentration of air pollutants in urban areas, offer invaluable opportunities for the better understanding of air quality problems. However, processing large sets of data to extract statistically valid evidence poses many challenges from both the conceptual and technical viewpoints. Air quality data acquisition, cleaning and authentication are necessary and crucial preliminary phases to support descriptive, predictive and prescriptive models and to ensure that aggregated and high-quality information is delivered to the central and local governments, decision makers and citizens. Automated software tools can facilitate drawing conclusions based on the information contained in the data, limiting subjective judgment and providing repeatability. However, the costly state-of-the-art software applications developed by major vendors are inaccessible to many cities and townships in the developing world. Moreover, their usage creates dependency on proprietary solutions, which can hinder the possibility of evolving the data processing and analysis protocols. We present an open-source web application for air quality data analysis and visualization, called aiRe, based on the R statistical framework and Shiny web package. aiRe has been developed in collaboration with the Colombian environmental authorities, and implements best practices validated by experts in air quality. We believe that the process of developing aiRe was extremely valuable with the ultimate purpose of supporting cities in air quality management, while strengthening local capabilities to improve urban air pollution. This open-access tool simplifies and makes air quality data analysis and visualization accessible, with the desirable effect of removing ownership costs, fostering appropriation by non-expert users and ultimately promoting informed decision-making for the general public and the local government authorities. We present the performance of this tool over a series of examples of open data collected by the air quality monitoring network of Bogotá, Colombia.

1. Introduction

1.1. Context

Poor air quality is a growing global concern (Gulia et al., 2015; Baklanov et al., 2016). Recent evidence shows that 8 out of 10 people worldwide are exposed to air pollution levels that are considered harmful to human health (Landrigan et al., 2018). Air quality management in many cities around the world calls for the deployment of surveillance systems to monitor air pollution levels. Such systems are normally composed by a set of stations to measure ambient air pollutant concentrations, as well as meteorological variables. These networks vary widely in reliability and representation, influenced by factors such as topography, demography, meteorology, and socioeconomic development (Air pollution in the world's megacities, 1994).

Due to the many complexities of air quality monitoring, large cities often opt for commercial software packages that help them configure, administrate and maintain air monitoring infrastructures, as well as creating reports for the decision makers in environmental policy. However, most smaller cities and towns in less-developed countries

* Corresponding author.
E-mail addresses: juan.diazbaquero@tngtech.com, jj.diaz1067@uniandes.edu.co (J.J. Díaz), ivan.mura@dukekunshan.edu.cn (I. Mura), jfranco@hill.com.co (J.F. Franco), akhavan@sabanciuniv.edu (R. Akhavan-Tabatabaei).

https://doi.org/10.1016/j.envsoft.2021.104976
Accepted 18 January 2021
Available online 1 February 2021

cannot afford such products. Instead, they generate their reports using very basic tools, which are unsuitable for database management and big-data analysis. Furthermore, there is often lack of expertise and adequate capabilities in local administrations to ensure the reliability of air monitoring networks, data cleaning and the subsequent analyses. With the increasing availability and popularity of cheaper air quality measurement sensors (Gulia et al., 2020), the amount of data to be analyzed will grow and will call more and more for advanced and automated software support.

The impact of poor air quality also has a socioeconomic class facet, as according to the WHO, in low and middle-income countries approximately 98% of city inhabitants are living in environmental conditions that do not meet the guidelines for air pollutant exposure. In high income countries this percentage (yet still very high) reduces to 56% (Osseiran and Chriscaden, 2016). Among the reasons for this difference, are the unavailability of reliable air monitoring infrastructure and/or the lack of analytical capabilities to develop effective public policies. Without the support offered by automation, even guaranteeing compliance with existing air quality standards and regulations for data reporting becomes difficult. For example in Colombia (the subject of our case study), the responsibility of air quality management is assigned to both local and regional environmental authorities, mostly decentralized and independent agencies that have the responsibility to administrate and operate air quality monitoring networks. Such independence may lead to a wide range of interpretations on data treatment and reporting guidelines suggested by the national government, creating the necessity for integrated software solutions and data analysis approaches.

1.2. Available software packages for air quality analysis

Currently there are several software applications capable of reading structured data sets of air quality measurements and performing statistical or visual analysis. We can distinguish them in two classes: proprietary applications with commercial purposes, and non commercial software, which includes many academic developments.

Commercial software application usually have a friendly user-interface and can perform all the necessary steps for different types of analyses. These tools can load the data sets or even directly interface with the station of a monitoring network to obtain the data, perform pre-configured and user-defined analyses, and create detailed reports of air quality. For instance, the Israeli software suite *Envida/Envista*© (Ltd and Envista air, 2007) provides an integrated solution to retrieve, store and analyze measurements collected by an air quality measurement network. It can retrieve the data from a database and perform analyses such as wind roses, histograms, correlations in different aggregation of time and dynamic GIS visualizations. Another program is the German software *Ambient Air Quality Monitoring* (Kisters), which is intended for public authorities and can connect to different monitoring networks to provide local, regional or national-level data managing and analysis. The main advantage offered by commercial products is the quick access to many types of analyses to support decision making. They automate error-prone data manipulation tasks and reduce subjective judgment in data analysis. A clear disadvantage is of course the cost, since these software packages are usually very expensive to acquire, and many governmental entities, especially in low-income regions may not be able to afford them. Moreover, the usage of these commercial solutions may create dependence on the vendors, which might limit the creation of technical know-how that goes beyond the tool's defined functionality, and hinder the possibility of autonomously adapting to new needs and conditions.

There are several non-commercial solutions that offer nice and intuitive graphical user interface and visualization features. For instance, *Giovanni*, a Web accessible tool developed by NASA Goddard Earth Sciences Data and Information Services Center (Prados et al., 2010), allows loading and analyzing many diverse types of open datasets, including air quality measurements collected by satellites and on-ground stations. Giovanni provides time-series and spatial visualization of data on maps. Moreover, it allows exporting data in various formats, including the compressed Keyhole Markup Language (KML), for reusing them in 3D Google Earth visualizations. The KML data format is also used by the solution described in (Chen, 2019), where the authors propose an application that allows visualizing in near real-time air quality data. They describe in (Chen, 2019) a software residing at a server side, which can be configured to continuously gather air quality open-data and structure it into KML files ready to be consumed by clients using Google Earth. Examples of heat-map visualizations are shown in (Chen, 2019) for China air quality data. Another software application that aims at providing intuitive visualizations of air quality data is proposed in (Lu et al., 2017). A Web interface is provided, which allows displaying pollution at various levels, from country to regions, cities, down to monitoring stations. At each level, the time dimension of the data is summarized by a sunburst like chart, which can be zoomed for changing the granularity level. The software application allows visualizing the AQI index, as well as the concentration of user-selected pollutants.

Some other studies in the literature report about software applications that can load and visualize air quality measurements together with predictions generated with various types of machine learning approaches. For instance (Ofoegbu et al., 2014), proposes to a software application that can connect to a database containing the air quality data, and generate visualizations for pollutant time series and their predictions, as well as AQI predictions, using artificial neuronal networks and decision trees type of models. In (Tomić et al., 2014), the authors present a client-server solution that is capable of acquiring measurement through a TCP/IP network, and display visualizations of the collected data together with prediction obtained with an auto-regressive moving average model. The solution includes a client for devices running the Android operating system, so that the data visualization services can be accessed remotely.

Several solutions for air quality analysis and visualizations based on the statistical software *R* have also been proposed. The open source R-package *openair* is currently the most popular software for air quality data analysis (Carslaw and Ropkins, 2012). The package is developed at King's College in London, by the entity in charge of the city's air monitoring network. It is capable of reading a set of air quality data and performing a variety of possible analyses. In particular, *openair* provides extensive data filtering features, such as sorting by weekdays, seasons, daylight hours, etc., creating wind rose, time series and geo-mapped visualizations. This package is more like a library, and does not have a user interface. Thus, it requires the user to be already familiar with *R* programming language. Nevertheless, the open-access and advanced data-handling functionalities of *R* make of it a very attractive option for air quality research. For instance (Oh and Park, 2015), shows how *R* can be used for exploring, modeling and visualizing a dataset of air quality measurements collected by a monitoring network. In a very recent paper, the authors of (Feenstra et al., 2020) present an open-source R package called *AirSensor* and the web application *DataViewer*, built using *Shiny*. Together, they provide comprehensive data handling and visualization features for exploring data collected by networks of low-cost air quality sensors.

Finally, we mention *SISAIRE*, the Web application used by IDEAM (the Meteorological Agency of Colombia) to support the pipeline of air quality data capture and analysis (IDEAM - Instituto de Hidrología, 2018). *SISAIRE* allows 1) collecting information on the meteorological and air quality variables generated by the regional environmental authorities, and 2) facilitating data access and consultation to citizens and institutions in charge of research on environmental issues.

1.3. Our objective

The objective of this work is to introduce aiRe, a web-based, open-source software tool for air quality data cleaning, analysis, visualization,

and reporting. aiRe incorporates best practices validated by experts in the field of air quality management, and offers a repeatable, rigorous, high-quality data analysis approach in air pollution, and a friendly user-interface for the visualization of results and reports, without requiring the intervention of skilled data scientists.

By proposing *aiRe*, we authors intend to provide an open solution for air quality data analysis, which is compliant with the requirements of Colombian protocols for air quality analysis and reporting. Using *aiRe* allows speeding up the execution of tedious and error-prone tasks, and generating a wealth of intuitive data visualizations. *aiRe* can be easily modified and extended: Users killed with *R* can flexibly reuse the existing features and adapt them to new requirements. The software can be freely downloaded from its Gitlab repository hosted at Duke University (see the appendix for details).

The rest of this paper entails the methodology in developing the aiRe package in section 2, its visualization capabilities in section 3, our validation method in section 4 and finally section 5 concludes the work and lights the way forward for research on this topic.

2. Methodology

In this section we present our approach to the development of aiRe. We followed a software engineering incremental and iterative development process (Larman and Basili, 2003). According to this process, the functionalities to be included in aiRe are first defined in terms of user requirements, gathered from the intended users and with the aid of technical experts. The software development was organized according to a combined incremental/spiral life-cycle, whereby subsequent intermediate releases were produced, each one based on a set of requirements. For each requirement set, proof-of-concept prototypes were developed and validated with users and technical experts in Colombia.

Technical requirements included compliance with the process of data cleaning according to the rules for the analysis of air quality data, defined by the Colombian environmental authorities (Ministerio de Ambiente y Desarrollo Sostenible, 2008). This protocol is maintained by the Institute of Hydrology, Meteorology and Environmental Studies (IDEAM), the central and technical agency of the Colombian Ministry of Environment.

We strongly emphasized on the development of an easy-to-use and reproducible data cleaning process, with the objective of allowing any user to execute it and re-obtain the same validated data sets. It is worthwhile mentioning that air quality data collected by monitoring networks in Colombia is openly accessible, and the information contained therein is available to the general public. However, without simple and statistically well-grounded tools to extract such information from the time-series of the variables stored in the data sets, their usefulness is diminished. Therefore, our work was also driven by the imperative need of creating intuitive though flexible visualizations of the validated data sets. Thus, our development process also included methodologies for the design and analysis of information through visual-analytics techniques (Munzner, 2014).

In the next four subsections, we describe the main elements that guided the development of aiRe. First, we present the required input data, and second we discuss the intended users of our tool. Third, we introduce the required functionality of aiRe, and fourth we finally detail the software architecture that is devised in order to facilitate its appropriation, reuse and future improvements.

2.1. Input data

The format in which data is made available by an air quality monitoring network depends on its specific data collection operations, as well as on the technological infrastructure of the network. Any software tool including aiRe, is however bound to specific formats of the input data, which we describe in this section.

Data collected by a monitoring network is usually organized in tables, where each row provides measured values at a given moment in time. A single monitoring station is usually a collection of measurement devices that gather data on several variables of interest, including contaminants, and possibly other environmental variables such as the meteorological factors. Hence, each single measurement can be thought of as a tuple $(time, variable, location, measured\_value)$.

Data is commonly stored and made available in textual file format, and one of the elements of the tuple is implicitly used to identify the file content. If the element is the location, e.g. a monitoring station, then the data would be logically structured as shown on the left of Table 1, with each row providing the values of all variables measured by a station. If the pollutant variable, e.g. carbon monoxide, is used for determining the file content, then the data has a similar format to what is shown in the right-hand-side of Table 1, where each row provides the values of the variable measured by all available monitoring stations.

In Colombia, environmental agencies can choose between the two formats shown in Table 1 to store and interchange their data. We found format a) to be more commonly adopted by environmental agencies that manage smaller monitoring networks and manually collect data, which is the most frequent case in Colombia. The data stored in format a) may not have a constant period of reporting. The format b) is commonly adopted by environmental agencies that own automated monitoring networks, usually located in large cities such as Bogotá and Medellín. The time-series stored in this format have a constant reporting period (hourly values).

We choose format b) in aiRe for storage and calculations. The reason behind this decision is that this format facilitates data processing and the creation of tidy data sets, at the expense of storing missing values. Moreover, most analyses only use a single pollutant and compare its concentration among different locations. However, aiRe can also import data stored in format a), and internally convert it to format b). For this conversion, the software assumes a single measurement is available for each day. All analyses are then based on daily measurements.

2.2. Intended users

Although developed based on our interaction with the Colombian air quality managers, aiRe provides a general framework for the analysis of open-access air quality data. Many countries and cities in Latin America share similar concerns and needs for air quality data analysis and reporting (Franco et al., 2019). In particular, medium and small-sized cities and towns could complement their air quality management practices by applying aiRe to feed and support the decision-making process of environmental authorities. Moreover, aiRe could be of interest to scholars working in air quality, as it offers a first cut approximation tool for exploring the dynamics of air pollution. Finally, since the software is endowed with a friendly and intuitive web-based interface, we believe it could improve the access to air quality descriptive analysis, generating interest in a larger audience, including non-specialists.

The agencies in charge of environmental monitoring in Colombia can be split into two groups according to the type of monitoring network infrastructure:

- agencies with automatic stations, typically found in larger cities, endowed with specialized software applications that support the data analysis and reporting;
- agencies managing manual stations, commonly found in smaller cities and towns around the country, using general-purpose software to perform data analysis, often without the assistance of specialized personnel.

A common problem of all agencies is the lack of resources devoted to knowledge management and research in these topics. As a result, the personnel in such agencies is usually hired on short contract terms, and they may change quickly in a relatively short time period. This situation hinders the build-up and development of technical know-how in the

Table 1
Data formats used by environmental agencies in Colombia: data coming from manual collection in format a), on the left, and from an automated network in format b), on the right.

Time&Hour	Pollutant 1	Pollutant 2	Pollutant 3	Time&Hour	Station 1	Station 2	Station 3
January 1, 2012	30	0.2	93	January 1, 2012 11:00	35	40	No Data
January 3, 2012	35	0.4	94	January 1, 2012 12:00	37	42	31
January 4, 2012	40	0.3	No Data	January 1, 2012 13:00		No Data	No Data
January 6, 2012	38	0.2	97	January 1, 2012 13:00	31	34	21
	Format a)				Format b)		

agencies, and may generate inconsistencies in the handling and analysis of data over time.

In such circumstances, the adoption of aiRe, a tool endowed with a Web interface, which facilitates data treatment according to agreed policies, can be of help in providing consistent air quality analysis results. Moreover, the open-source nature of aiRe makes it possible to customize and adapt accordingly to the evolving needs and regulations in air quality management and analysis.

2.3. *Functional requirements of* aiRe

We collected the data analytics needs of the environmental agencies with the help of academic experts on air quality. The requirements were gathered with the continuous feedback of the main national authorities of air quality management. Although many different types of functionalities could be automated, the requirements that we chose to implement in aiRe are those considered to provide the most value, and remove the necessity of tedious and error-prone handling of data.

The following requirements of data treatment, cleaning, authentication and analysis have been expressed by the Colombian environmental agencies:

- **Data availability:** Determine where (for which measuring locations), when (for which time periods), and how much valid data is available. This allows an understanding of whether there is sufficient data to perform the desired analyses, and if so, how representative results will be. The norms may prescribe lower bounds on the amount of available data necessary for conducting certain types of statistical analyses. For instance, according to Colombian regulations, at least 75% out of the 24 hourly measurements collected in a day must be valid for determining whether that day exceeded or not the threshold set by national standards.
- **Days that exceed national standards:** Determine the number of days that the contamination has exceeded the norms of the daily maximum contaminant concentration allowed, an analysis task required by the national standards and a key indicator for all environmental agencies. Verify that the concentration values are expressed at reference conditions (25 °C and 760 mm Hg).
- **Comparative analysis:** Identify the locations with the highest concentration of contaminants, compare different moments of the day, such as traffic peak hours with valley hours, compare different days such as holidays or weekdays with weekends. These requirements call for a variety of different analyses and therefore require a flexible mechanism to allow environmental agencies compare the data in all meaningful ways.
- **Time series:** Visualize the historical data of the concentration of contaminants in the form of time-series, for a given measurement location. This is a common step in air quality analysis, which allows determining trends and visually identifying seasonality in pollutant concentration.
- **Analysis of indexes:** Use the data to calculate and visualize aggregate indices defined by the national regulations. For example, in Colombia national guidelines prescribe the use of a color-scale index to facilitate the communication of air quality levels.
- **Pollutant time series:** Estimate the pattern of pollutant concentrations through the day, for the whole set of historical data or restricted to a specific time range and/or location.

These distinct classes of requirements represent the basis for the development of aiRe and are reflected into the software architecture of the tool, as detailed in the next subsection.

2.4. *Software architecture of* aiRe

We designed a modular architecture for aiRe, which will facilitate the future evolution and addition of software features with limited implementation efforts. Fig. 1 shows a diagram of the tool architecture, where each module is a different R script. Most scripts are implemented using a Shiny (Chang et al.) module that defines both the user interface and the server logic.

The main advantage of this architecture is the use of a replication structure that allows to easily plug in new modules. In particular, the modules that implement the analysis functionalities have low or no coupling at all. If a new feature is implemented, then the developers would only need to write the feature code in a single R script and then integrate it with the data analysis module. This facilitates the addition of further analysis functionality in aiRe.

The main disadvantage of this architecture is that modules have to be independent of each other, which may require the duplication of functionalities to some extent. However, we decided to trade code optimization and computational performance with the ease in extension and understanding.

3. Visualizations

In this section we describe the visualization functionality of aiRe to address each of the requirements mentioned in the previous section. Throughout this section, we present examples based on the data from the Bogotá air quality monitoring network (RMCAB hereafter, by its initials in Spanish *Red de Monitoreo de Calidad del Aire de Bogotá*). This network has been operating since 1998, collecting hourly measurements of several meteorological variables and air pollutants. After being collected, authenticated and analyzed, the data is stored in an open-data repository, from where it can be retrieved in the form of textual files.

We have processed the data of 12 out of the 14 stations of the network, excluding the mobile monitoring station that has been dedicated to the evaluation of air pollution at several locations along heavy traffic roads and a station that only reports meteorological data. We focus on the analysis of $PM_{2.5}$ and PM_{10} pollutants, which have been on the watch for their frequently high concentrations within the urban area of Bogotá (Mura et al., 2020).

3.1. *Analysis of data availability*

To address the requirement of data cleaning and authentication, we develop two visualization tools in the form of a stacked bar chart and a heat-map. Fig. 2 depicts an example of the stacked bar-chart, providing an overall description of the data availability for each monitoring station in our case study of Bogotá. Each bar in Fig. 2 represents the

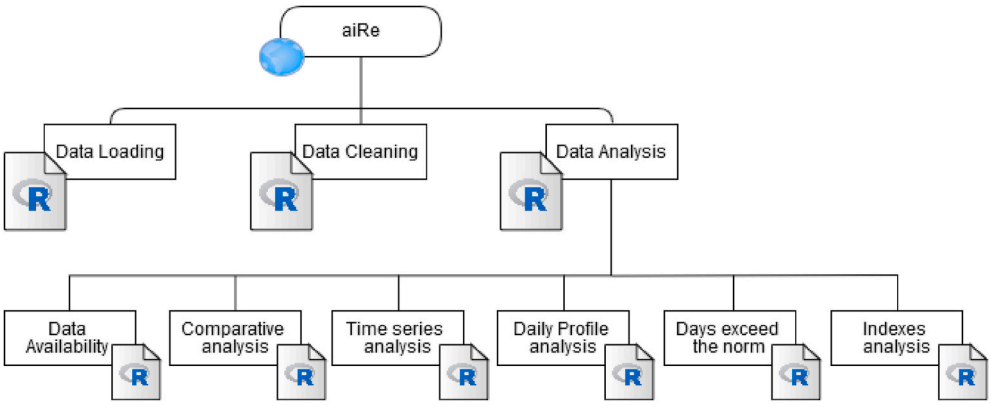

Fig. 1. Architectural diagram of aiRe.

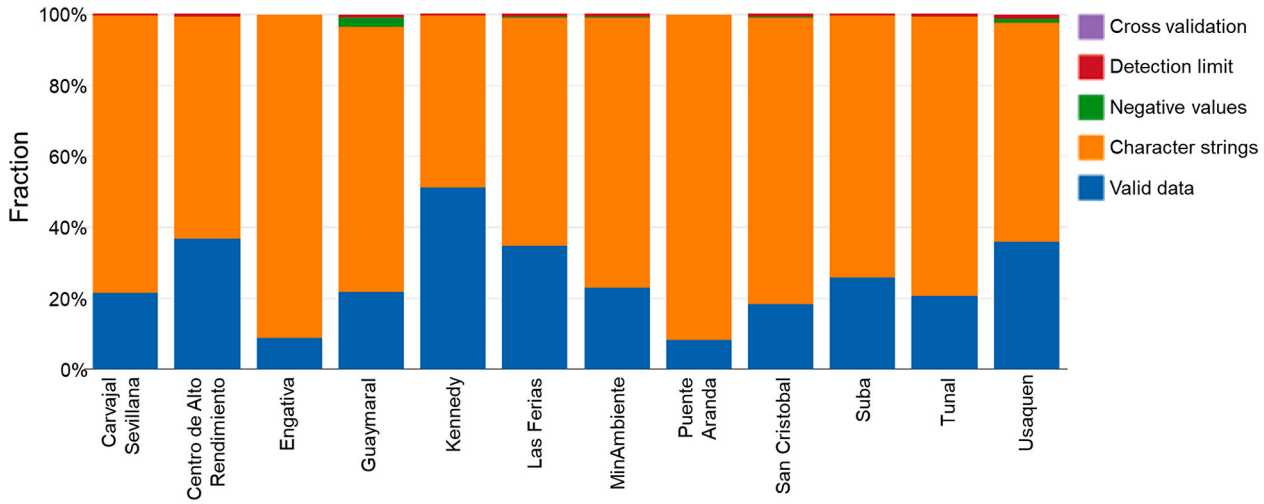

Fig. 2. $PM_{2.5}$ availability of authenticated data per station from 1998 to 2018, Bogotá monitoring network.

concentration of $PM_{2.5}$ in one of the 12 stations throughout the city over ten years of data collection period, and the stacked colors indicate the percentage of data that is affected by the user-defined cleaning and authentication rules set in the application. Each rule states one of the properties that the measurement should possess to be considered valid. These rules can be flexibly chosen among a configured list of user-defined requirements for data cleaning and authentication. The blue colored sections show the percentage of valid data that satisfies all the rules, orange for data that did not pass the *numeric-only* criteria (some measurements may be replaced with text), green for the data that did not pass the *positive-number* rule (some concentrations are erroneously reported as negative), red for data that violate the *detection limit* rule (values that are outside of the equipment detection limit) and purple for the data that violates the *cross-validation* rule (some simultaneously measured concentrations of $PM_{2.5}$ are higher than those of PM_{10}). As it can be appreciated in Fig. 2, the station located in *Kennedy* provides the highest percentage of valid data measurements for $PM_{2.5}$ in our case study.

The heat-map in Fig. 3 shows how data availability changes over time. In this map, the x-axis shows the time and the y-axis has a row for

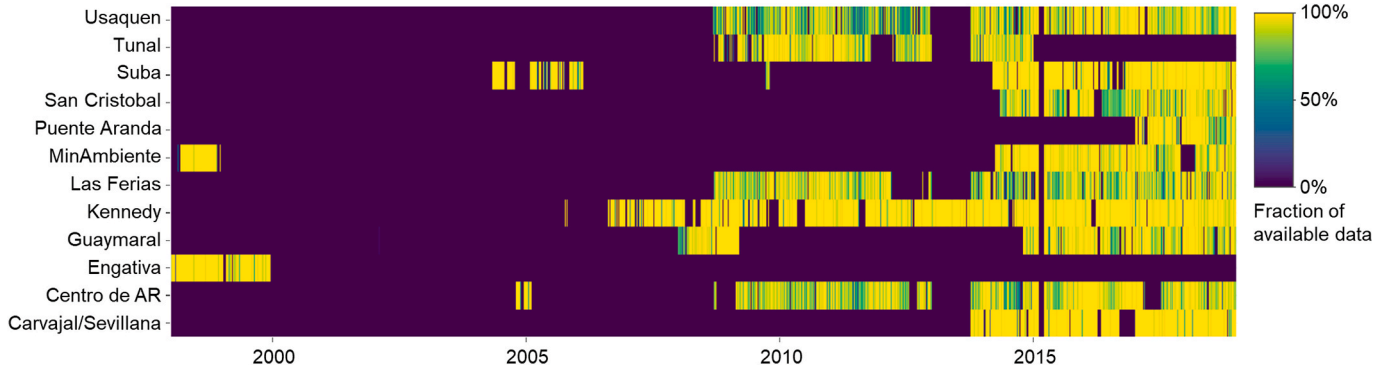

Fig. 3. Heat-map with the $PM_{2.5}$ data availability in Bogotá from 1998 to 2018. Color on a scale from purple to yellow, with yellow indicating high data availability and purple very low data availability, green used for intermediate values.

each one of the monitoring stations. For each monitoring station, the percentage of available data for each single day in the time interval is shown according to a color scheme: yellow is used when availability is 100%, i.e. the number of valid hourly data samples collected during that day is equal to 24, and purple for availability near 0%, i.e. no valid records were taken during that day. To calculate the daily average of the data we used the R-package *openair* (Carslaw and Ropkins, 2012).

The heat-map visualization allows environmental agencies to determine at a glance, which periods of time and which locations have useful data to conduct air quality analyses. In the case of Bogotá, it is easy to see that there are many periods of time in color green, which indicates limited quality of data, as well as periods for which data is not available at all.

3.2. Comparison against national standards

Like many other countries, Colombian laws require all environmental agencies to report the number of days in a year that the concentration of contaminants is higher than the threshold allowed by the national standards. To address this requirement, we developed in aiRe the visualization tool shown in Fig. 4. To perform this analysis, handling of missing data is critical, although not explicitly mentioned by the norms. The chart shows the number of days below the maximum level permitted by national standards, the number of days without enough valid data to reach a conclusion, and the number of days that the air quality violates the thresholds set by the national standards (calculated by the daily averages). Most countries require a minimum proportion of valid data per unit time to declare this measure conclusive. In the case of Colombia, the norm suggests at least 75% of valid data measurements in a day. The bars for the days without enough valid data (in gray) are purposefully located in the middle of each stacked bar chart, to provide a clue of the fact that some of them may belong to the green category and some others to the red one. We have added a slider control for this visualization in the graphical user interface of aiRe, so that the user can change the threshold set on the minimum percentage of valid data and observe the changes in the updated chart.

This visualization tool provides a clear view of the number of days above the limit permitted by the national standards (the red section of the bars). It is easy to see in Fig. 4 that in the case of Bogotá most stations do not exceed the limit set by the law. This plot can be easily interpreted by the general public, as one of the intended audience of our package.

3.3. Comparative analyses

The task of conducting comparative analyses requires a high degree of flexibility and generality by the environmental agencies, and therefore it is among the most challenging for them. This requirement calls for two major pieces of functionality: a flexible sub-setting of the data, so that the user can adequately define the sets of data to be used for the comparative analysis, and a general graphical rendering of the metrics computed on the selected data sets.

To implement a solution for the flexible sub-setting of the data, we have developed slide-controls shown in Fig. 5. With a simple and intuitive process, the user can define multiple comparison ranges and assign a name to each of them. Each comparison range identifies a subset of the data. The data is further segmented by monitoring stations. For instance, if the user were interested in comparing the concentrations of $PM_{2.5}$ measured on Fridays, Saturdays and Sundays in the period 2014–2016, she could define three comparison ranges and name them according to the day, then use the year selection control to choose the period of time for the analysis and check the corresponding box of the specific day to complete the data subsetting. Each comparison range would be associated with a subset of the measured values of the variable (i.e., $PM_{2.5}$ in the example of Bogotá in Fig. 5), which would then be sent to data analysis and rendering.

We visualize the comparisons through grouped box-plots as the general rendering functionality. This visualization functionality allows for a quick comparison of the data subsets within each station and among different stations. For instance, Fig. 6 shows the result of the per-day analysis for Bogotá, as mentioned above. From the chart, we can easily observe that Sundays were cleaner than Fridays at all the stations for the selected time period. Also, it is possible to pinpoint the stations with higher contamination levels, e.g., Carvajal-Sevillana and Kennedy in our case study. Notice that results for 2 out of the 12 stations (Engativá and MinAmbiente) do not appear in Fig. 6, as they have no valid data for this analysis.

This same functionality can be used for a variety of analyses. For instance, one that is periodically conducted by the environmental authorities in Colombia is the comparison of air quality between regular weekdays and the days when special vehicular traffic restrictions are in place. Specifically, once a year Bogotá declares a car-free day where the transit of private vehicles (cars and motorcycles) is banned. This day (called *Día* sin *carro*), usually takes place on the first Thursday of February, as it is one of the most polluted months for the city (Lozano, 2004; Mura et al., 2020). It is clear that such one-day initiatives as *Día* sin *carro* are more apt to raise awareness among the general public rather than meaningfully contributing to an immediate reduction in the air pollution. However, it is interesting to analyze the air pollution data collected by the monitoring network during such special occasions and compare with a set of properly chosen control days. Fig. 7 shows the outcome of a comparative analysis conducted by this functionality in aiRe for the 2018 *Día* sin *carro* edition, which occurred on Thursday, February the 1st. In Fig. 7 we focus on $PM_{2.5}$ measured concentrations, and the control days are the other Thursdays of February of the same

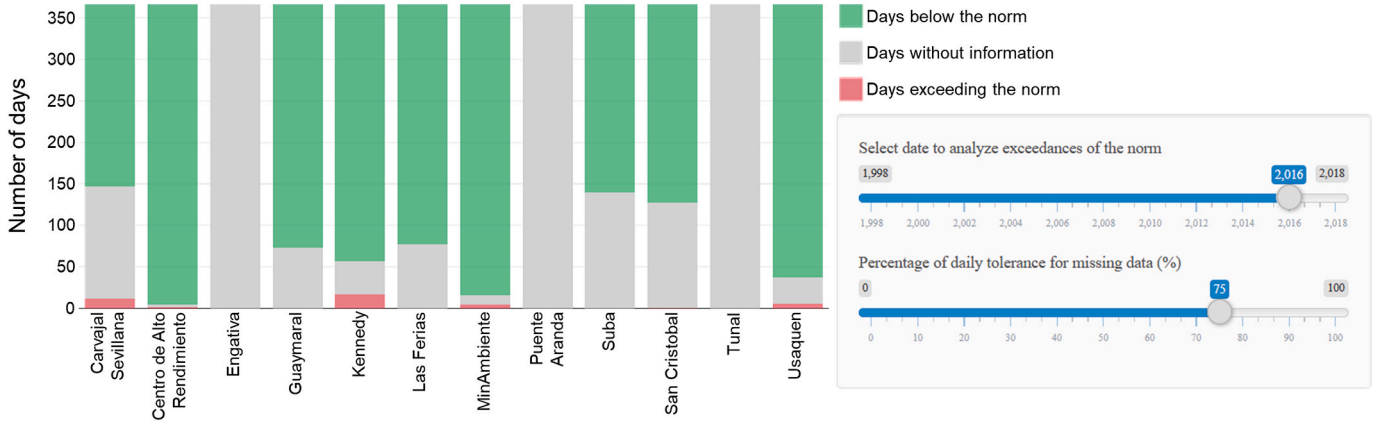

Fig. 4. Number of days exceeding national $PM_{2.5}$ standards per station in 2016. Green: days with concentrations lower than the limit dictated by national standards; Gray: days that do not meet the requirement of data availability: Red: days with concentrations higher than the limits set by the national standards.

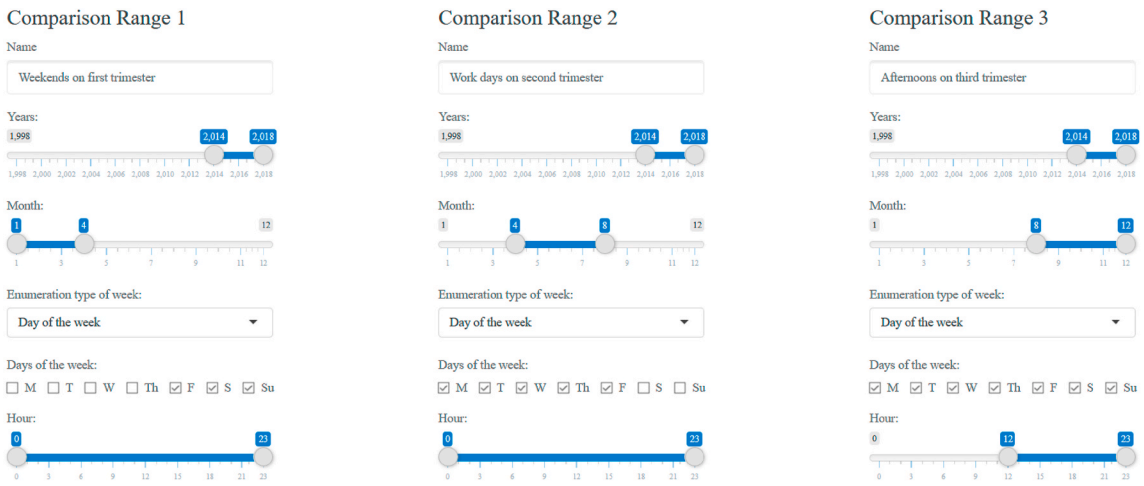

Fig. 5. Form filled with example inputs of the developed interface for the comparative analyses functionality.

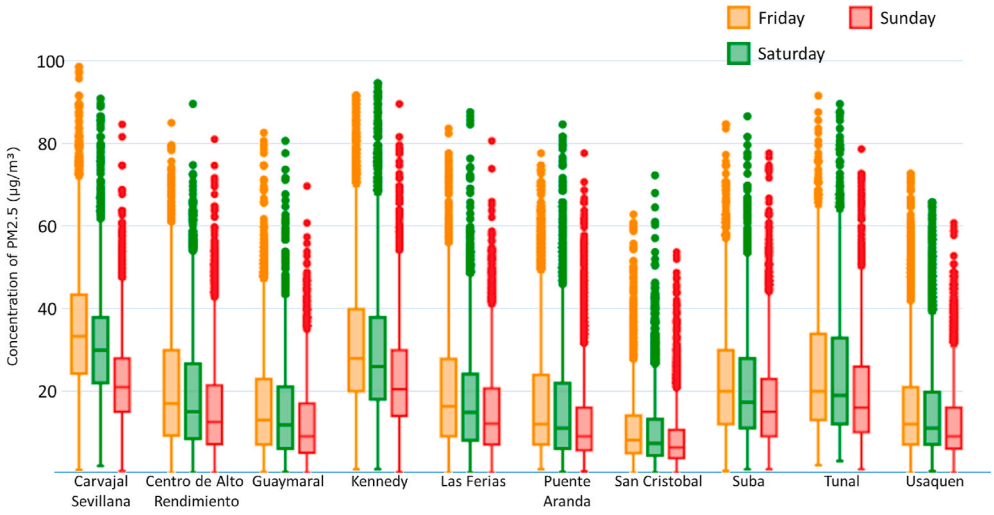

Fig. 6. Comparison of $PM_{2.5}$ concentrations between weekends and Fridays in all stations of Bogotá monitoring network, from 2014 to 2016.

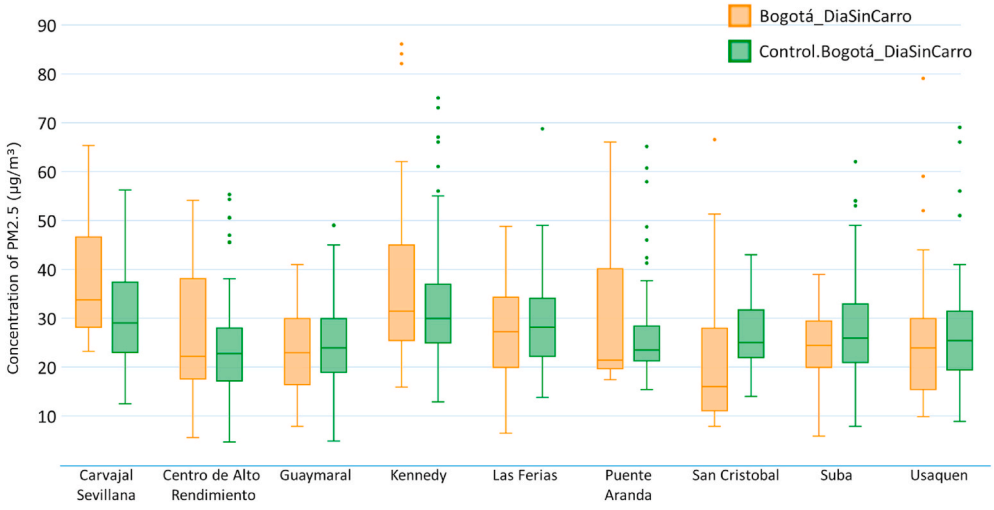

Fig. 7. Comparison between the $PM_{2.5}$ measured concentrations during the 2018 *Día* sin *carro* and the control days.

year. Results are shown for the nine stations that have valid data available for the analysis.

Without intending to enter into a detailed analysis, we observe that the visualization reported in Fig. 7 allows to identify several monitoring stations that reported appreciable reductions in $PM_{2.5}$ measured concentrations, while others did not show clear benefits from the traffic restrictions. Furthermore, the data collected at the two monitoring stations of Carvajal-Sevillana and Kennedy, both located in the South-Western part of the city, indicate that air quality worsened during the *Día sin carro*. This may be due to the increase in public transportation offers, deployed to satisfy the mobility demand. The Bogotá public transportation fleet has indeed been recognized as a major contributor to the city's air pollution of $PM_{2.5}$ or smaller particles (Franco et al., 2016; Morales Betancourt et al., 2019; Castillo-Camacho et al., 2020).

3.4. Time-series analysis

A common simple visualization method for pollutant historical data is a line chart. Line charts or time-series provide an immediate intuition on short-term trends, peak periods and variability. It is also the required format of visualization for all Colombian environmental agencies, to report to the Ministry of Environment. In Fig. 8 a time-series visualization is produced for a user-selected monitoring station, and a control is added for choosing an aggregation level for the data points, which can be displayed at the hour (no aggregation), day, week, month, quarter or annual frequency.

3.5. Reporting air quality indexes

Several indicators and indexes have been proposed and established to quantify and communicate air pollution emissions, concentrations, and human health impacts internationally (Franco et al., 2019). Such indicators help bridge the "science-policy" gap by synthesizing complex scientific information and presenting it to the citizens, stakeholders and policy-makers in understandable ways (Hsu et al., 2013). For example, the Air Quality Index (AQI) widely used by the United States Environmental Protection Agency (US EPA), presents daily measurements of several major air pollutants on a color-coded scale, so citizens and other stakeholders can easily correlate one color with an air quality condition (i.e., good, moderate, bad), and furthermore identify spatial or temporal hot-spots of poor air quality in a city.

Colombia has developed her own air quality index (ICA, for its acronym in Spanish). Both ICA's calculation methodology and ICA cut-off points' definitions are established in the Colombian National Air Quality Standard document (Resolution 2254 of 2017) (Ministerio de Ambiente y Desarrollo Sostenible, 2017). Table 2 shows the ICA scale for $PM_{2.5}$ and PM_{10}. ICA's methodology is based on the guidelines given by the United States Environmental Protection Agency US-EPA, through its Technical Assistance Document for the Reporting of Daily Air Quality (version of September 2018) (US Environmental Protection Agency and Technical Assistance Document for the Reporting of Daily Air Quality, 2018).

Given the relevance of the air quality indices, aiRe includes the function of visualizing such indexes through heat-maps. In Fig. 9) the horizontal axis is the time, and on the vertical axis a categorical variable (ICA) is reported for each monitoring station. Each cell of the heat-map reports the average ICA color for the day at a given station. When there is not enough data during a day to reliably calculate an average value, the cell is given a null value, which is graphically rendered using the gray color.

This visualization serves several purposes, such as to find patterns in the behavior of PM concentrations over time. For instance, Fig. 9 shows that the months in the middle of the year, i.e. June and July, tend to be cleaner than the other months in Bogotá. Also, it allows to compare among various stations. For example, it is easy to spot that the stations of Kennedy and Sevillana are the most polluted in the city. Finally, it permits identifying particular events that affect the city, when most stations simultaneously report very high concentration of pollutants. For instance, we can appreciate in Fig. 9 one of such events occurring in February 2016, when the smoke caused by a massive wildfire on the mountains bordering the Eastern side of Bogotá generated a surge in air pollution levels.

We can observe such an event in detail on the chart presented in Fig. 10, which is obtained by zooming in Fig. 9 over the specific time window of the wildfire. All the visualizations of aiRe are rendered using the Plotly library (Sievert et al., 2017), which provides several controls that are directly accessible to the user by clicking on the displayed images.

3.6. Analysis of daily pollution patterns

This analysis is of particular interest to the environmental agencies that can regularly register hourly data. It can be used to identify peak hours and changes in the daily patterns of contamination over time, for instance as a result of traffic circulation restrictions. Fig. 11 shows the line chart that aiRe produces for the daily pollution pattern, for the concentration of PM_{10} in Bogotá. The x-axis shows the hours of the day, and the y-axis the concentration of the pollutant. Each line shows the hourly average computed for the days of a specific year.

The plot also shows a 95% confidence interval for each hourly point. As pointed out in other studies (e.g. (Mura et al., 2020),), this kind of plot makes it easy to observe that the concentration of PM_{10} has steadily diminished through the years in Bogotá.

4. Validation of aiRe

In this section we describe the process of validating the implemented functionalities of the software by experts and intended users. Besides loading, cleaning and data analysis functionalities, aiRe took into account requirements related to usability and intuitiveness of

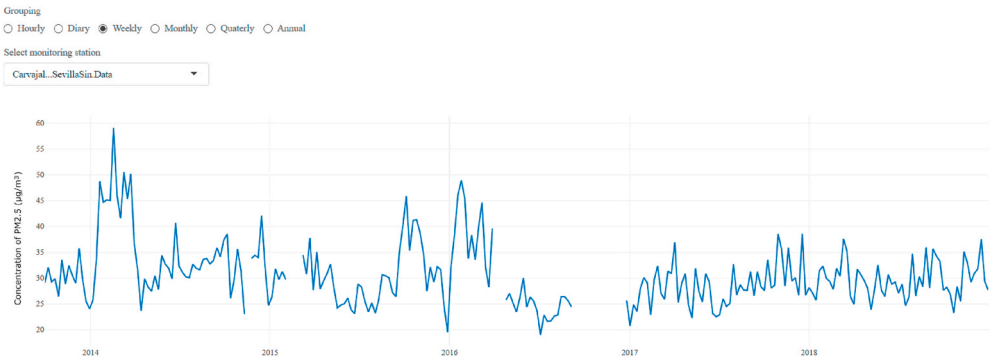

Fig. 8. Time-series of $PM_{2.5}$ measured concentration for the Carvajal-Sevillana monitoring station in Bogotá, weekly aggregation level.

Table 2
Daily split points and color coding for particulate matter concentrations in the ICA index.

ICA range	Color	Classification	$PM_{2.5}$ $(\mu g/m^3)$ 24 hours	PM_{10} $(\mu g/m^3)$ 24 hours
0 - 54	Green	Good	0 - 12	0 - 54
51 - 100	Yellow	Moderate	13 - 37	55 - 154
101 - 150	Orange	Unhealthy for sensitive groups	38 - 55	155 - 254
151 - 200	Red	Unhealthy	56 - 150	255 - 354
201 - 300	Purple	Very unhealthy	151 - 250	355 - 424
301 - 500	Brown	Hazardous	251 - 500	425 - 604

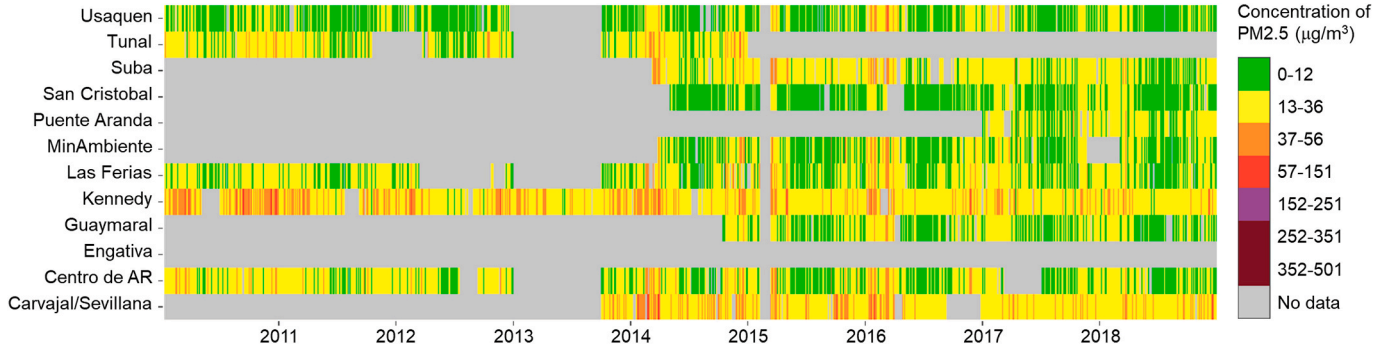

Fig. 9. Heat-map visualization of ICA index for all stations in Bogotá of PM2.5 concentration between 2010 and 2018.

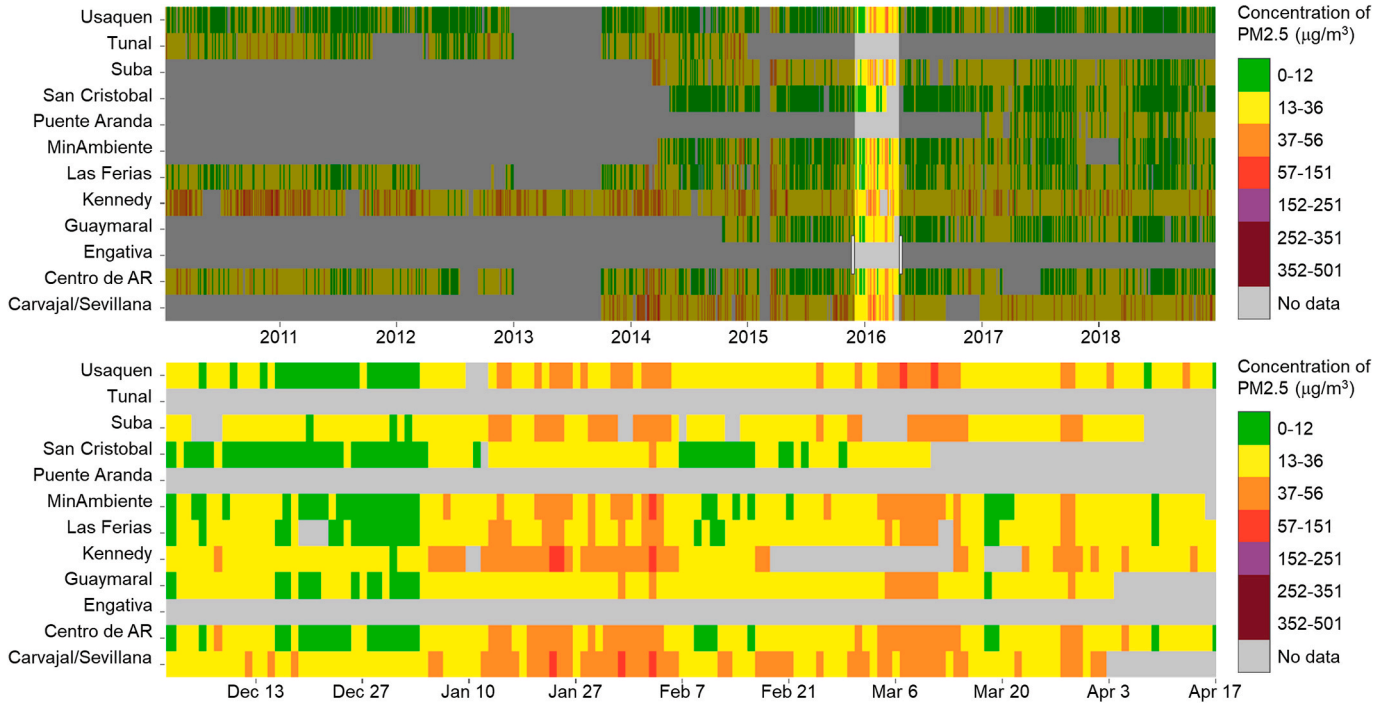

Fig. 10. On top: Fig. 9 with a selected region for zooming. On bottom: picture of the magnified area, corresponding to high contamination events in early 2016.

visualizations. Therefore, actors with different expertise participated in the overall development process. Fig. 12 depicts a summary of the contributions and roles that the different actors played in the verification and validation process of aiRe.

aiRe was developed by an inter-disciplinary research team that included undergraduate and graduate students, researchers and professors at Universidad de los Andes. The team joined forces from the Department of Industrial Engineering, whose skills revolve mainly around data science, and the Department of Civil and Environmental Engineering, which mostly contributed on the technical air quality facet of the project. This combination of complementary competences ensured that diverse perspectives were represented, generated a stimulating and challenging learning environment, and provided a high degree of criticism instrumental to sustaining continuous verification of the project outcomes.

We developed aiRe with the continuous accompaniment of the

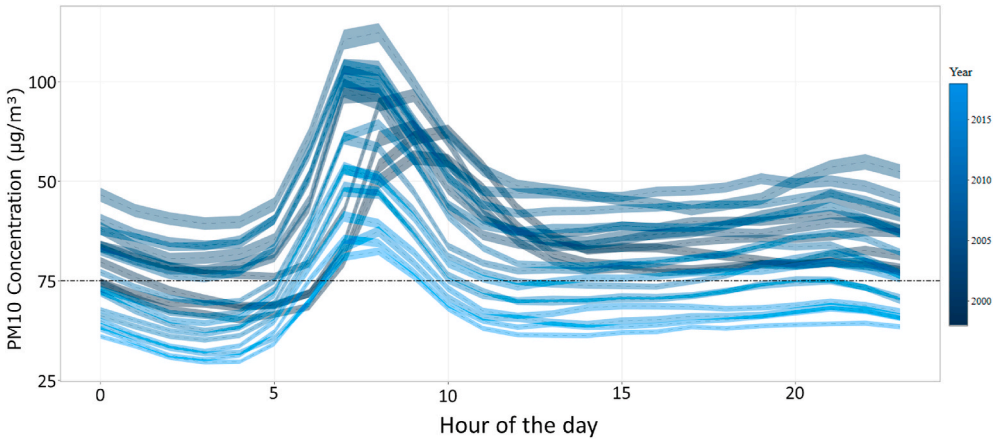

Fig. 11. PM10 concentration daily profile across all years in all measuring stations in Bogotá.

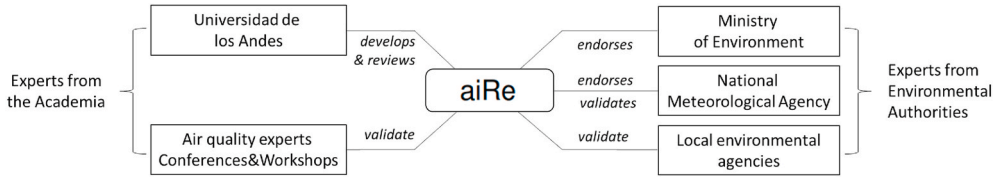

Fig. 12. Diagram of the validation of aiRe.

Colombian Ministry of Environment (MinAmbiente), the main governmental entity responsible for the development of public policies that have an impact on the environment. Although MinAmbiente itself is not involved in the monitoring of air quality, it has the role of defining norms and guidelines, recommending best practices and disseminating knowledge in order to improve the performance of local environmental agencies. When starting the development of aiRe, we contacted MinAmbiente to present the early software prototypes. The possibility of distributing an open-source tool for air quality to local environmental authorities and to foster its use for the sake of a more consistent data analysis, proved to be very attractive to MinAmbiente. Hence, MinAmbiente sponsored our project, and based on their profound knowledge of the practices and needs of air quality monitoring, we jointly defined a roadmap for functionality addition. This gave aiRe an institutional endorsement.

We developed a first version of aiRe, and then liaised with IDEAM, the national meteorological agency, which maintains the protocols for data collection, cleaning, aggregation and analysis, to realize software validation. We carefully reviewed with IDEAM the data cleaning and analysis functionality of aiRe, pinpointing any non-compliance with the defined protocols. The fix of these defects and the development of a further set of visualizations defined the requirements of the second version of aiRe.

In the final stages of aiRE development, MinAmbiente held an event that brought together representatives of all Colombian local environmental agencies, target users of the software. We were granted a space to show a live demo of aiRe. Although this may not be considered a formal acceptance test, the reception of the software was mostly positive. We could additionally gather requirements on new desirable features to facilitate the identification of outliers, and counting the number of days in each interval of the ICA index (which is now implemented in the final version of aiRe, see visualization in Fig. 4). In conjunciton with these activities, Bogotá's central environmental agency (*Secretaría Distrital de Ambiente*) made important contributions in separate meetings, by validating the box-plot visualizations that aiRe generates to show the results of comparative analyses.

The implementation of data handling and visualization requirements were also validated by academia experts outside Universidad de los Andes. We presented aiRe at a scientific conference organized by the *Colombian Association for Operational Research (ASOCIO)* held in 2017 (Díaz Baquero et al., 2017a). We discussed the algorithmic details of the implementation, and improvements to the visualizations, including controls that would allow zooming into compact graphical renderings (now implemented in Fig. 10), and smoothing to facilitate the identification of trends within the data (not yet implemented).

Finally, this work was also presented to representatives of academia with expertise in air quality. A demo of aiRe was given to a panel composed by academics, using data of the monitoring network of Bogotá, at the *2017 Colombian Congress and International Conference on Air Quality and Public Health (CASAP)* (Díaz Baquero et al., 2017b). Experts and researchers suggested adding to aiRe visualizations on the amount of available data before analysis, which has now been implemented in several analyses. In other visualizations, we added slide control bars to allow the user choose a tolerance level for the amount of missing data (see for instance Fig. 4). We also presented aiRe at the *23rd International Sustainable Development Research Society (ISDRS)* conference (Franco et al., 2017). The importance of making results reproducible for all analyses upon which high-level decisions would be made was remarked, as well as the opportunity of endowing aiRe with the functionality to export cleaned data (now implemented) and tables and plots into reports (not yet implemented).

5. Conclusion and future work

This work presents aiRe, a web-based open source software package in R and Shiny for air quality data cleaning and authentication, analysis, visualization, and reporting. The software implements data loading, cleaning and analysis, while providing various visualization functionalities, aiming at strengthening the process of analysis and reporting of air quality data by entities in charge of managing air pollution. aiRe was developed and validated in collaboration with the Colombian environmental authorities and academia experts. The code of the tool can be freely accessed from the Gitlab page of aiRe (see Appendix A).

The authors believe that the process of developing aiRe was

extremely valuable, through gathering the public sector and academia around the same table, with the ultimate purpose of supporting cities in air quality management while strengthening local capabilities to improve urban air pollution. We believe that aiRe can offer a useful complement to other open-source initiatives for the analysis and interpretation of air pollution data.

The future work on aiRe includes the implementation of new functionalities, such as the analysis of meteorological variables and spatial distribution of air pollution in a specific geographical location. Several local environmental agencies have also suggested adding spacial analysis capabilities. Visualizing the locations of monitoring stations in a city together with aggregated data would be of great help to the authorities, even-more-so if a spacial distribution of contaminants would be added by using a spacial interpolation model. We also believe that an important step would be to reach the general public as the audience of aiRe visualizations, thus raising awareness and creating participation. Finally, the addition of predictive and prescriptive analysis capabilities, as well as generating a historical report of all operations performed and results obtained during a session, will be considered as discussed earlier.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Software Availability

Developer: Juan José Díaz Contact Address: Cra 1 N ° 18A - 12 Dep of Industrial Engineering, Universidad de los Andes, Bogotá, Colombia Contact Emails: jj.diaz1067@uniandes.edu.co, ivan.mura@dukekunshan.edu.cn Tested browsers: Firefox and Google chrome Software Required: R, rStudio R-Packages required: shiny, ggplot2, plotly, openair, RColorBrewer, scales, grid, shinythemes Programming Language: R Available Since: 2018 Availability: The aiRe source code package is stored in a freely accessible Gitlab repository hosted by Duke University: https://gitlab.oit.duke.edu/im90/aire Cost: Free.

References

Air pollution in the world's megacities, 1994. Environment: science and policy for sustainable development 36(2), pp. 4–37. https://doi.org/10.1080/00139157.1994.9929147.

Baklanov, A., Molina, L.T., Gauss, M., 2016. Megacities, air quality and climate. Atmos. Environ. 126, 235–249 doi:10.1016/j.atmosenv.2015.11.059. http://www.sciencedirect.com/science/article/pii/S1352231015305665.

Carslaw, D.C., Ropkins, K., 2012. Openair — an R Package for Air Quality Data Analysis. Environmental Modelling & Software 27–28(0), pp. 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008.

Castillo-Camacho, M.P., Tunarrosa-Grisales, I.C., Chacón-Rivera, L.M., Guevara-Luna, M. A., Belalcázar-Cerón, L.C., 2020. Personal exposure to PM2.5 in the massive transport system of Bogotá and Medellín, Colombia. Asian Journal of Atmospheric Environment 14, 210–224.

Chang, W., Cheng, J., Allaire, J., Xie, Y., McPherson, J., et al., n.d.. Shiny: Web Application Framework for R, R Package Version 1 (5).

Chen, P., 2019. Visualization of real-time monitoring datagraphic of urban environmental quality. Eurasip Journal on Image and Video Processing (1), 42.

Díaz Baquero, J.J., Mura, I., Franco, J.F., 2017a. Analtica aplicada a la calidad de aire, in: 2nd Congress of the Colombian Association for Operational Research (ASOCIO). Medellin, Colombia. http://docs.wixstatic.com/ugd/099b1e_008c7e4d262944668a7a96e0ffad1238.pdf.

Díaz Baquero, J.J., Mura, I., Franco, J.F., 2017b. Making air quality data analysis simple. Accessible and repeatable through R web-based applications, in: VI Colombian congress and international conference of air quality and public health (CASAP). Santiago de Cali, Colombia. https://gallery.mailchimp.com/d38af58c463e9270f937271a7/files/12ded4f8-4c54-4179-9013-9ef80cbd08b4/Memorias_CASAP_2017.pdf.

Feenstra, B., Collier-Oxandale, A., Papapostolou, V., Cocker, D., Polidori, A., 2020. The AirSensor open-source R-package and DataViewer web application for interpreting community data collected by low-cost sensor networks,. Environ. Model. Software 134, 104832.

Franco, J.F., Segura, J.F., Mura, I., 2016. Air pollution alongside bike-paths in bogotá-Colombia. Frontiers in Environmental Science 4, 77. https://doi.org/10.3389/fenvs.2016.00077.

Franco, J.F., Bernal, L., Melo, N., Díaz Baquero, J.J., Raha, A.-T., Mura, I., 2017. The role of data analytics for sustainable city development and the implications for the academia, in: proceedings for the 23th International Sustainable Development Research Society (ISRDS). Bogotá, Colombia.

Franco, J.F., Gidhagen, L., Morales, R., Behrentz, E., 2019. Towards a better understanding of urban air quality management capabilities in Latin america. Environ. Sci. Pol. 102, 43–53.

Gulia, S., Shiva Nagendra, S., Khare, M., Khanna, I., 2015. Urban air quality management-a review, Atmospheric Pollution Research 6(2), pp. 286–304. https://doi.org/10.5094/APR.2015.033.

Gulia, S., Prasad, P., Goyal, S., Kumar, R., 2020. Sensor-based wireless air quality monitoring network (SWAQMN)-a smart tool for urban air quality management, atmospheric pollution research 11(9), pp. 1588–1597.

Hsu, A., Reuben, A., Shindell, D., de Sherbinin, A., Levy, M., 2013. Toward the next generation of air quality monitoring indicators. Atmos. Environ. 80, 561–570.

IDEAM - Instituto de Hidrología, 2018. Meteorología y Estudios Ambientales, Subsistema de Información sobre Calidad del Aire – SISAIRE. http://sisaire.ideam.gov.co/ideam-sisaire-web/informacion.xhtml?de=que_es.

Kisters, A.G.. Ambient air quality monitoring [commercial software]. https://air.kisters.de/en/competences/ambient-air-quality-monitoring/.

Landrigan, P., Fuller, R., Acosta, N., Adeyi, O., Arnold, R., Basu, N., Baldé, A., Bertollini, R., Bose-O'Reilly, S., Boufford, J., Breysse, P., Chiles, T., Mahidol, C., Coll-Seck, A., Cropper, M., Fobil, J., Fuster, V., Greenstone, M., Haines, A., David Hanrahan, D., David Hunter, D., Khare, M., Krupnick, A., Lanphear, B., Lohani, B., Martin, K., Mathiasen, K., McTeer, M., Murray, C., Ndahimananjara, J., Perera, F., Potočnik, J., Preker, A., Ramesh, J., Rockström, J., Salinas, C., Samson, L., Sandilya, K., Sly, P., Smith, K., Steiner, A., Stewart, R., Suk, W., van Schayck, O., Yadama, G., Yumkella, K., Zhong, M., 2018. The lancet commission on pollution and health. THE LANCET COMMISSIONS 391, 462–512. https://doi.org/10.1016/S0140-6736(17)32345-0.

Larman, C., Basili, V., 2003. Iterative and incremental developments. a brief history, Computer 36(6) 47–56. https://doi.org/10.1109/MC.2003.1204375.

Lozano, N., 2004. Air pollution in bogotá, Colombia: a concentration-response approach, revista Desarrollo y sociedad (54), pp. 133–177. https://doi.org/10.13043/dys.54.4.

ENVITECH Ltd, 2007. Envista air resources manager [commercial software]. https://www.environmental-expert.com/downloads/envistaarm-brochure-pdf-294-mb-32690.

Lu, W., Ai, T., Zhang, X., He, Y., 2017. An interactive web mapping visualization of urban air quality monitoring data of China. Atmosphere 8(8) 148.

Ministerio de Ambiente y Desarrollo Sostenible, 2008. Protocolo para el Monitoreo y seguimiento de la calidad del aire. http://www.ideam.gov.co/documents/51310/527391/Protocolo+para+el+Monitoreo+y+seguimiento+de+la+calidad+del+aire.pdf/6b2f53c8-6a8d-4f3d-b210-011a45f3ee88.

Ministerio de Ambiente y Desarrollo Sostenible, 2017. Resolución n. 2254: por la cual se adopta la calidad del aire ambiente y se dictan otras disposiciones. https://www.minambiente.gov.co/images/normativa/app/resoluciones/96-res%202254%20de%202017.pdf.

Morales Betancourt, R., Galvis, B., Rincón-Riveros, J., Rincón-Caro, M., Rodriguez-Valencia, A., Sarmiento, O., 2019. Personal exposure to air pollutants in a bus rapid transit system: impact of fleet age and emission standard. Atmos. Environ. 202, 117–127. https://doi.org/10.1016/j.atmosenv.2019.01.026.

Munzner, T., 2014. Visualization analysis and design. CRC press.

Mura, I., Franco, J.F., Bernal, L., Melo, N., Díaz, J.J., Akhavan-Tabatabaei, R., 2020. A decade of air quality in Bogotá: a descriptive analysis,. Frontiers in Environmental Science 8, 65. https://doi.org/10.3389/fenvs.2020.00065.

Ofoegbu, E., Fayemiwo, M., Omisore, M., 2014. Data mining industrial air pollution data for trend analysis and air quality index assessment using a novel back-end aqms application software, International Journal of Innovation and Scientific Research. ISSN 2351–8014.

Oh, Y., Park, E., 2015. Data Visualization of Airquality Data Using R Software. Journal of the Korean Data and Information Science Society 26(2), pp. 399–408.

Osseiran, N., Chriscaden, K., 2016. Air pollution levels rising in many of the world's poorest cities. http://www.who.int/news-room/detail/12-05-2016-air-pollution-levels-rising-in-many-of-the-world-s-poorest-cities.

Prados, A.I., Leptoukh, G., Lynnes, C., Johnson, J., Rui, H., Chen, A., Husar, R.B., 2010. Access, Visualization, and Interoperability of Air Quality Remote Sensing Data Sets via the Giovanni Online Tool. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 3(3), pp. 359–370.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., Despouy, P., 2017. Plotly: Create Interactive Web Graphics via 'plotly. Js', R Package Version 4, p. 110.

Tomić, J., Kušljević, M., Vidaković, M., Rajs, V., 2014. Smart scada system for urban air pollution monitoring. Measurement 58, 138–146.

US Environmental Protection Agency, Technical Assistance Document for the Reporting of Daily Air Quality, 2018. https://www.airnow.gov/publications/air-quality-index/technical-assistance-document-for-reporting-the-daily-aqi/.