Challenges for Air Pollution Monitoring: a Cyber-Physical Social Systems Approach

Marco Zappatore
Hesplora s.r.l. & University of Salento, Lecce, Italy
marco.zappatore@hesplora.it, marcosalvatore.zappatore@unisalento.it

Sergio Refolo
Dept. of Innovation Engineering
University of Salento
Lecce, Italy
sergio.refolo@studenti.unisalento.it

Antonella Longo
Dept. of Innovation Engineering
University of Salento
Lecce, Italy
antonella.longo@unisalento.it

ABSTRACT

Air pollution control plays a pivotal role today in urban contexts, as both citizens and public administrators are increasingly sensitive about it. Traditional air pollution sensing is performed and managed by public institutions with professional and expensive equipment, thus exhibiting a series of inherent limitations such as isolated monitoring campaigns, data heterogeneity, inconsistency and incompleteness, limited access to sensed data. Cyber-Physical-Social Systems (CPSS) promise to be a considerable step forward, as they promote the systematic involvement of citizens in monitoring processes and the provisioning of proactive services to end users. However, several elements hinders such a model. In this paper, we will discuss the challenges of applying cyber-physical paradigm to air pollution monitoring in smart cities, exemplifying the issues on the Italian case study and then we will show how CPSS will go over them and outline novel research directions.

KEYWORDS

Cyber-Physical Social Systems, Air Monitoring, Data Processing, Data Visualization, Mobile Crowd Sensing.

1 Introduction

Air quality monitoring is a strategic and long-term activity that gives experts the opportunity to make evaluations about air pollution, to study emission causes and sources as well as to develop corrective or mitigation plans. The air quality status cast several concerns amongst experts as well as citizens due to its related health risks [1], [2]. Therefore, it is evident how air pollution control is necessary to prevent human diseases and to protect ecosystems. That is why it must be addressed by local authorities and policy makers, as well as it should be a responsibility for the stakeholders in the industrial sector. Similarly, citizens must be made aware of air quality status and how to be more actively involved in actions aimed at improving daily life quality conditions.

The attention on environmental issues is continuously increasing and it involves more and more people: this results in a rising number of in-domain researches. However, people hardly can take any conclusion on the topic by themselves and usually the correct interpretation of research findings is troublesome for non-professional recipients [3]. In the most common scenario (usually known as *institutional monitoring*), professional and expensive sensors are placed in the close proximity of few significant areas (e.g., airports, hospitals, congested roads, etc.) by authorized agencies or public bodies devoted to environmental control on national, regional or even smaller scale. Raw data are collected and published online by the same agencies. This approach is worldwide adopted and falls under the definition of air quality assessment [4]–[8]. Published data come directly from sensors (i.e., *raw data*) or from simple data manipulation processes and usually no inferred knowledge is provided in a simple and effective way, especially when data sources and data formats differ significantly. It is, therefore, widely accepted that dedicated data processing solutions are needed in order to clean data from unwanted noise, thus focusing on what really matters [9], [10]. This implies the need of monitoring outcomes effectively presented to final users, in order to provide meaningful insights to the different involved actors, as citizens' needs differ from those exhibited by city administrators.

Cyber-Physical-Social Systems (CPSSs) promise to be a valuable solution for urban monitoring scenarios as they leverage on the availability of scores of heterogeneous sensors whose readings are collected, aggregated and analyzed by cyber processes and profitably merged to real-time, city-related data provided and shared by complementary social sources in order to be presented as relevant information to citizens and authorities [11]. However, this paradigm is far to be applied on a large scale. In this paper we will focus on the Italian scenario, by examining the existing solution and by proposing a first step towards the adoption of CPSSs. Currently, the Italian situation features traditional air pollution assessments based on local sensing stations that, even if reliable and properly manned, do not guarantee a wide coverage of monitoring campaigns (due to high costs and lack of skilled personnel) and expose several data heterogeneity issues, thus

1st Workshop on Cyber-Physical Social Systems (CPSS2019).
October 22, 2019, Bilbao, Spain.

making difficult any data comparison and aggregation on a wider scale.

For such a reason, in this paper we will thoroughly examine air pollution monitoring data provided by Italian regional agencies for environmental protection. A proper data model will be devised in order to aggregate data coherently. Data manipulation pipelines will be applied to collected data in order to aggregate and to visualize them properly with the help of business intelligence tools. This procedure highlights data incompleteness and heterogeneity coming from institutional sources. For partly overcoming the issue we propose a CPSS that is currently under development in the framework of an Italian regional research project, aimed at large-scale, low-cost urban environmental pollution monitoring.

The paper is structured as follows: section 2 introduces the domain of our investigation and the corresponding research questions. In section 2 the addressed scenario is described in detail. Section 3 deals with Cyber-Physical-Social Systems. Section 4 describes the addressed scenario while Section 5 shows our data analysis approach. Achieved results are discussed in Section 6, along with the proposed CPSS modelling. Finally, Section 7 draws conclusions.

2 Air Pollution: the Current Scenario

2.1 Legislation in Europe

Air pollution is an important environmental and societal issue that impacts on human health, ecosystems and climate changes.

Several official reports have addressed so far this topic, trying to propose regulations to be applied on large scale. For instance, the 2016 report of air quality in Europe [12] focuses on the scenario in the EU Member States. It shows that a large portion of the European population (as well as the ecosystems in the same region) is exposed to air pollution levels that exceed European standards and World Health Organization (WHO) Air Quality Guidelines (AQGs).

The most significant provenance of air pollutants is represented by anthropogenic sources. They encompass transportation systems, industry, power plants, agriculture machineries and household appliances.

Regardless of their origin, air pollutants can be divided into two main categories: primary and secondary ones. Primary pollutants are directly released into the environment from the processes that generate them. The main pollutants belonging to this class (e.g. CO, NO_x, SO_x) are the result of combustion processes.

Secondary pollutants derive from primary ones, and are obtained

Table 1: European legislation about emissions

Policies		Pollutants*							
		PM	O_3	NO_2, NO_x, NH_3	SO_2, SO_x	CO	Heavy metals	BaP / PAH	VOCs
Directives regulating ambient air quality	2008/50/EC (EU, 2008)	PM	O_3	NO_2, NO_x	SO_2	CO	Pb		Benzene
	2004/107/EC (EU, 2004)						As, Cd, Hg, Ni	BaP	
Directives regulating emissions of air pollutants	(EU) 2015/2193 (EU, 2015)	PM		NO_x	SO_2				
	2001/81/EC (EU, 2001)			NO_x, NH_3	SO_2				NMVOC
	2010/75/EU (EU, 2010a)	PM		NO_x, NH_3	SO_2	CO	Cd, Tl, Hg, Sb, As, Pb, Cr, Co, Cu, Mn, Ni, V		VOC
	European standards on road vehicle emissions	PM		NO_x		CO			VOC, NMVOC
	2012/46/EU (EU, 2012)	PM	NO_x		CO				HC
	94/63/EC (EU, 1994)								VOC
	2009/126/EC (EU, 2009c)								VOC

***Pollutants**: **PM**: Fine particles; **O_3**: Ozone; **NO_2**: Nitrogen dioxide; **NO_x**: Nitrogen oxides; **NH_3**: Ammonia; **SO_2**: Sulphur dioxide; **SO_x**: Sulphur oxides; **CO**: Carbon monoxide; **CO**: Carbon monoxide; **BaP**: Benzo[a]pyrene; **PAH**: Polycylcic Aromatic Hydrocarbon; **VOC**: Volatile Organic Compound; **NMVOC**: Non-Methane VOC; **HC**: Hydrocarbons; **Pb**: Lead, **As**: Arsenic; **Cd**: Cadmium; **Co**: Cobalt; **Cr**: Chromium; **Cu**: Copper; **Hg**: Mercury; **Mn**: Manganese; **Ni**: Nickel; **Sb**: Antimony; **Tl**: Thallium; **V**: Vanadium.

from their transformation due to reactions usually involving oxygen and light: oxidation is therefore a phenomenon strictly correlated to this pollutant's category.

More specifically, PM (particulate matter), BaP (benzo[a]pyrene) and mercury (Hg) emissions come from the incomplete combustion of various fuels, while emissions of ammonia (NH3) or CH4 (methane) from agriculture. The current trend about PM foresees threshold exceedances even in 2020: PM with a diameter of about 10 μm (henceforth, PM10) exceeds the EU limit value in 21 of the 28 EU Member States, while PM 2.5 (i.e., particles whose diameter is nearly 2.5 μm) exceeds on average in 4 states [13].

The transport sector and the industry have been taking a considerable reduction of their emissions of air pollutants in Europe since 2000 (except for BaP and Cadmium, Cd, emissions in transports, and CH4 and BaP in industry). The trend of commercial, institutional and households' emissions is less positive, with a 3% increase in BaP from 2000 to 2014. Moreover, less significant reductions of air pollutants have been experienced in agriculture.

In Table 1 the most relevant European directives concerning air pollution are reported.

The main goal of monitoring campaigns is providing indicators to define emissions trend; the following list collects the main indicators used in national monitoring campaigns with the related reference directives [14], [15]:

1. Greenhouse gases (CO2, CH4, N2O) – Framework Convention on climate change (1992) ratified with L 65 of 15/01/94; Kyoto Protocol (1997) ratified with L 120 of 01/06/02; CIPE resolution 19/12/02; D.Lgs. 51/08; D.Lgs. n. 30 13/03/13
2. Acidifying substances (SOx, NOx, NH3) – Goteborg Protocol (1999); NEC (2001/81/CE) directive; D.Lgs. 171/04
3. Particulate – LCP 2001/80/CE directive; CE 715/2007 regulation; CE 595/2009 regulation
4. Carbon monoxide (CO) – D.Lgs. n. 152 of 03/04/2006; 97/68/CE directive; 98/77/CE directive
5. Benzene (C6H6) – L 413 of 04/11/97
6. Persistent organic pollutants (IPA) – Aarhus Protocol (1998); L 125/06
7. Heavy metals – Aarhus Protocol (1998)

Humans can be adversely affected by exposure to air pollutants in ambient air. In response, the European Union has produced an extensive body of legislation which establishes health-based standards and objectives for several air pollutants. These objectives are developed over different periods because pollutants impact human health in different ways according to exposure time (we refer the interested reader to the existing-legislation section related to air quality in the EC Web portal [16]).

2.2 Legislation and Environmental Control Agencies in the Italian scenario

In this paper our analysis is focused on the Italian situation: the 2008/50/CE directive, implemented in Italy with the legislative decree D.Lgs.155/2010, defines how to evaluate and manage air quality for human health defense and environment protection.

In Table 2 we summarize the currently enforced D.Lgs.155/2010: it presents pollutant concentration, reference averaging period, legal nature of the specific norm enlisted, permitted exceedances per year and limit values for each pollutant.

In Italy air quality monitoring is decentralized and performed autonomously by regional or local agencies for environmental protection: each agency deals only with its own territory.

These agencies, named ARPA (whose acronym stands for Regional Agency for Environmental Protection, in Italian) are public institutions that provide technical support to Italian regional administrations (except for Trentino-Alto Adige, which has been split into the two autonomous provinces of Trento and Bolzano) to perform environmental control and enforce regulations.

These agencies, born in 1993 and nationally coordinated by SNPA (The National System for Environmental Protection, in Italian), are nationwide dedicated to yearly environmental quality assessments. On the one hand, the decentralization in local agencies implies detailed control over a relatively limited portion of the national territory. On the other hand, however, this causes heterogeneity across the different regions due to the lack of shared data format and collection, management and publication policies. As a consequence, even if the agencies apply the same environmental control methodologies and comply with the same regulations, citizens experience different air-pollution-related monitoring services and tools depending on the agency they refer to. Moreover, different regions present different levels of detail about information offered by their environmental agencies and this makes difficult to compare directly data coming from different locations.

This scenario does not facilitate the analysis of the overall Italian pollution scenario. Indeed, it is not possible to carry out this task properly without any technical knowledge needed to overcome the technical issues briefly sketched above. The support of a software application capable to normalize and integrate different sources is, at present, fundamental in order to make readable and understandable huge amount of available data merged from several monitoring agencies. This, in addition to the possible presence of supporting and complementary data sources provided by citizens, would be the ideal scenario for the implementation of the CPSS paradigm. However, such a scenario is still far to come.

Table 2: D.Lgs. 155/2010

P*	C* [μg/m³]	T_{avg}*	TVED*/LVED*	AE*
PM 2.5	25	1Y‡	TVED: 1.1.2010 LVED: 1.1.2015	n/a
SO2	350	1h	LVED: 1.1.2005	24
	125	24h	LVED: 1.1.2005	3
NO2	200	1h	LVED: 1.1.2010	18
	40	1Y	LVED: 1.1.2010	n/a
PM10	50	24h	LVED: 1.1.2005	35

	40	1Y	LVED: 1.1.2005	n/a
Pb	0.5	1Y	LVED: 1.1.2005[**]	n/a
CO	0,010	Max 8h	LVED: 1.1.2005	n/a
Benzene	5	1Y	LVED: 1.1.2010	n/a
Ozone	120	Max 8h	TVED: 1.1.2010	25d/ 3Y

***P**: Pollutant name; **C**: Pollutant concentration; $\mathbf{T_{avg}}$: Averaging period; **TVED**: Target Value Enforcement date; **LVED**: Limit Value Enforcement Date; **AE**: Permitted exceedances each year.
[*]: **Y**: Year; **h**: Hour; **d**: Day; **Max 8h**: Maximum daily 8 hour mean.
[**]: or 1.1.2010 in the immediate vicinity of specific, notified industrial sources; 1.0 µg/m³ limit value applied from 1.1.2005 to 31.12.2009.

2.3 Monitoring Networks and Data Availability

Air monitoring is a long-term activity and it requires necessarily careful studies. Usually, a monitoring network (i.e., a set of monitoring stations positioned in places of interest which provides some measures) is required. Monitoring stations record data about pollutants concentration in the lower atmosphere: through specific tools they perform measurements summarized in indicators, which are useful to make comparisons with limit values defined by directives and to know whether the situation is safe or not.

In [17] the EU scenario in terms of air quality monitoring is reported: monitoring campaigns are usually performed all year long with urban/local or regional scope. Monitoring stations are categorized into traffic, urban industrial or rural industrial locations. While there is a substantial homogeneity in these aspects amongst EU countries, data availability and data reporting differ significantly amongst Member States. As for data availability, the following categories can be identified: 1) validated data available for authorities only; validated data available for the public after a time delay (normally 1 day for data validation procedures); non-validated data available for the public in real-/near-time. Data reporting is also variegated: in some countries it is not performed on a nationwide scale, in some others, instead, annual reports are published by environment control agencies.

However, data are sometimes incomplete and not certain. For instance, 15 EU Member States reported uncertainty in their emission estimations and, in 2014, nearly 33% of data was incomplete [18], [19]. In this context, therefore, proper data cleaning and management operations become essential in order to make data usable and to minimize errors [20]. As a consequence, existing approaches to air pollution monitoring leverage significantly on big data and data mining solutions.

Several actions are underway in order to cope with this scenario, such as the Copernicus Atmosphere Monitoring Service (CAMS), implemented by the EU Centre for Medium-Range Weather Forecast (ECMWF) [21], aimed at reducing air pollution effects and the concentration of toxic breathable elements.

In Italy, monitoring campaigns are performed in sensitive locations (e.g., high-density traffic hotspots, airports, schools, downtown areas, industrial sites, etc.) by positioning fixed monitoring stations for long time periods (at least 6 months). These stations are sometimes relocated to other sites, due to their limited number. Large amounts of collected raw data are made openly available as daily or annual datasets in (semi-structured) text formats such as .csv, .xls(x) or .json.

Data heterogeneities affect the Italian scenario as well: regional environment control agencies do not share a common data publication format and do not comply with a unified template for publishing data. Each agency publishes validated data on a daily/weekly basis on its own Web portal but adopts different data visualization strategies and offers a variable set of tools for data manipulation, ranging from simple data filtering to customized chart composition. Data granularity is inconsistent as well, as in some cases users can access single-day datasets while larger datasets are available in other cases, thus determining critical gaps in user experience.

The lack of a common standard hinders the chance of joint analysis: inconsistency between data formats, data structure or detection metrics affect research potentials and limit non-professionals from acquiring environmental awareness.

However, as pointed out throughout the text, the most significant issue affecting the Italian scenario is represented by the absence of an institutional unified platform allowing users to access, navigate and manage monitoring data on a national scale.

From a legislative perspective, a federal council of Italian regional environment control agencies has been established in 2016 and a national air information system (SINAnet) [22] has been established. However the council only promotes administrative cooperation amongst agencies and the national information system is not open to the public yet. Indeed, at the moment of writing this paper, the system is accessible only by authorized personnel from regional agencies (i.e., ARPAs).

3 CPSSs for environmental monitoring

Cyber-Physical Social Systems (CPSS) are rooted into Cyber-Physical Systems (CPS) and Cyber-Social Systems (CSS) [11]. Therefore, CPSS are made up of multiple layers of sensors and actuators capable of monitoring physical phenomena and people's actions and of cyber components capable of receiving sensor data and generate digital representations of the monitored world (i.e. the digital twins), so that specific actions can be implemented accordingly. Sensing layers are usually populated by IoT (Internet of Things) sensors, mobile devices, and WSNs (Wireless Sensor Networks) that provide time-referenced and geo-referenced datasets. In addition to them, social data streams are managed, as well. Therefore, CPSS represent an evolution of IoT applications and are based on the integration of physical, cyber and social spaces, so that new knowledge can be inferred and the interactions with humans can easier happen. The core idea is that heterogeneous data sources from the physical world are fed to data processing and analytics processes, thus enabling further data fusion procedures whose output can be used by end-user applications, as described in the so-called data-oriented CPSS functional architectural model [23], where a CPSS solution for a

urban scenario is described as a set of "*data sourcing, collection and analysis mechanisms in order to obtain city intelligence*". More specifically, in [23], the authors consider a CPSS as built on top of three core elements. The first one is represented by *collaborative sensing sources*, operating according to multiple sensing paradigms but sensing the same physical contexts. This element, therefore, not only consists of traditional WSNs and IoT nodes but also of "*smartphone-carrying citizens*" who become "*valuable sensing resources*". The second core element is given by *data analysis tools*, needed in order to highlight any existing spatial/temporal or content-related pattern (or correlation) amongst datasets from different sources in order to increase context awareness. The third element is provided by *cross-spatial data fusion tools*, which are in charge of mining collected multimodal datasets and cope with heterogeneous measurement scales, combination of quantitative variables and qualitative classifications, etc.

Several CPSS solutions based on this model have been proposed in the recent years, addressing a wide range of applications. The studies that specifically tackled urban environmental monitoring can be clustered depending on the targeted application. For instance, the urban noise mapping problem has been addressed in [24] by adopting a fixed and mobile sensing infrastructure, enriched via participatory sensors by users, but no data fusion solutions have been proposed. The air quality assessment has been analyzed in [25], considering social data sources only (as the adopted CPSS infrastructure was fed by tweets from citizens about perceived air pollution levels), and in [26], distributing sensors only across communities of people, rather than to a large portion of citizens. Other CPSS approaches have been applied in Santander, Spain [27], where large IoT networks were deployed for environmental participatory sensing and car parking management, but no advanced data processing and data fusion solutions were proposed.

In the following sections, we will talk about the case of Italy, which has allowed us to identify the most significant challenges in managing environmental monitoring data on national scale hindering the adoption of a CPSS approach and, subsequently, we will introduce a proposal for a CPSS platform dedicated to urban pollution control.

4 Case Study

In order to identify current challenges in environmental monitoring in Italy, as introduced in Section 1, we have defined a nationwide case study about air pollution and, subsequently, we have developed a solution for collecting, managing and visualizing data. We have analyzed a 5-year range (from 2013 to 2017) by referring to standardized pollutants only (i.e., C6H6, CO, NO2, O3, SO2, PM10, PM2.5). The following subsection will deal with the dataset.

4.1 Referred Dataset

Initially, all Italian regions were considered for the analysis: this allowed us to sketch the overall scenario and to identify differences in the way regions perform the same task. The very first aspect is that, despite the availability of the federal council and of SINAnet platform (see Section 3), the accessibility and availability of monitoring data vary depending on the region, thus making troublesome to perform analyses and comparisons on a national scale. We used data available online via the regional ARPA portals.

For this reason, data integration is crucial, in order to merge files from different sources and define a shared and common data format.

At the starting point the count of overall data spanned across a time window from 2010 to 2018 and amounted nearly 71M records. The overall dataset exhibited a significant heterogeneity in terms of data granularity, format and structure. Therefore, for the sake of this cases, we selected a subset of sources in order to skim raw data before cleaning and to keep only the most homogeneous ones. This decision consisted in selecting only those regions that provide records referred to: 1) the five-year period 2013-2017 (because other years had less available data): 2) fundamentals pollutants (i.e., the standardized ones: C6H6, CO, NO2, O3, SO2, PM10, PM2.5). Moreover our analysis evaluates only regions whose measurements have been collected in compliance with regulatory limits. For instance, in the case of Lazio region, data were available in annual metrics, while pollutant metrics must be computed daily, according to regulations. This aspect poses a severe incompatibility among sources having different record granularity.

Such a preliminary record filtering operation has reduced significantly, the size of the initial dataset, by moving from 71M to 32M records. Selected regions are nine out of twenty: Basilicata, Campania, Emilia-Romagna, Lombardia, Marche, Puglia, Sicilia, Toscana, Valle D'Aosta.

4.2 Data quality issues

According to the definition of data quality dimension clusters depicted in [28], the main issues faced in managing data coming from regional agencies have been:

- Completeness: data format heterogeneity and sparsity in terms of time (i.e., regional datasets cover different time periods) and types (i.e., regional datasets refer to different subsets of pollutants). Therefore we have been forced to focus on 2013-2017 time period and on 7 standardized pollutants only. Moreover records outside the considered time range and not referring to the considered subset of standardized pollutants were discarded, totally or in part. For instance, regional datasets like those from Apulia and Lombardy regions were skimmed in order to keep only those subsets of data that complied with referenced intervals. A special concern was related to CO: for this pollutant, the majority of the analyzed regions have used an hourly average metric, while the regulation requires an 8-hour moving average instead: it has been decided to use for this pollutant a metric functional to datasets, namely an hourly average metric. Therefore, CO records with 8-hour moving average metric were removed through proper filtering
- Redundancy: some datasets included multiple versions of the same type of record (i.e., same measurement with different type of metrics): therefore we had to select only one metric per record set (according to the corresponding regulation) in order to cope with data redundancy. Datasets from Campania, Sicily and Valle D'Aosta were the ones affected by data redundancy the most.
- Accessibility (and the corresponding access time): some regions provide datasets through the website of their environmental control agencies and other regions provide data via open data portals. In addition, for regions like Abruzzo, Liguria or Sardinia, it is very hard to compose a 1-year dataset that refers to all monitored pollutants. Indeed, it is possible either to download records on a daily basis, referring to all the pollutants or to download annual datasets referring to a single pollutant at a time. This leads to download scores of different files manually. In some cases datasets were not accessible directly and data owners (i.e., the corresponding environment control agency) were contacted with no official answer.

5 The Analysis Platform

From the issues presented above non-professional users are prevented to extract meaningful insights from such little-comparable [29] data without any technical help. To make this data effective researchers need platforms supporting the analysis without dedicating excessive time and computational resources to data preparation and non-professional users must be supported even in accessing data and then guided across data visualization options, as publicly accessible and easy-to-understand data on environmental pollution can significantly improve environmental awareness across the population. For such reasons, we have designed and implemented a platform capable of merging heterogeneous data about air pollution, cleaning them and visualizing them in a meaningful and effective way, by using a few dashboards.

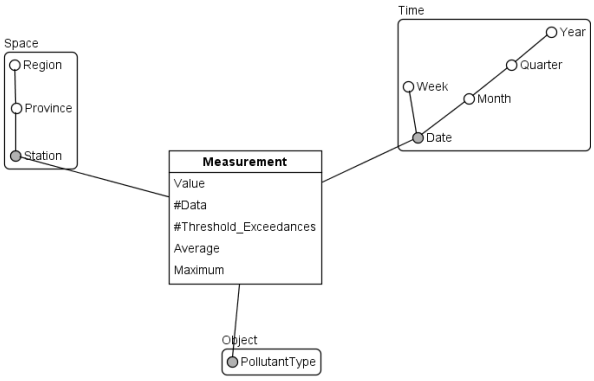

Figure 1: Dimensional Fact Model

5.1 Data model

A specific data model is behind the tool, so that data processing and visualization tasks can be performed rigorously and coherently. We refer to the Dimensional Fact Model (DFM), which has been proposed by Golfarelli et al. [30] specifically to support data mart design. This conceptual representation consists of a set of *fact schemata* that basically model the analyzed domain in terms of *facts* (i.e., any concept describing a time-evolving entity relevant to decision-making processes), *dimensions* (i.e., any qualitative description of a fact, composed by *dimensional attributes*), *measures* (i.e., any numerical property or calculation about a fact) and *hierarchies* (i.e., any directed tree made up of dimensional attribute).

In our case study, the measurement fact is chosen as the most significant one (Figure 1). The combination of three dimensions (time, location and pollutant type) results in multiple potential views of the same fact, so that it can be examined from multiple perspectives. Several measures can be associated to the defined fact (e.g., number of threshold exceedances for each parameter, average value sensed during a given time window for a given parameter in a given province, etc.), so that effective numerical indicators can be then derived and implemented into visualization dashboards.

5.3 Data Processing

Transformations of raw data are fundamental in order to reconcile data provided by regional environment control agencies.

Technically ETL pipeline has been developed using Pentaho Community Edition, an open source ETL (Extraction, Transformation and Loading) platform [31].

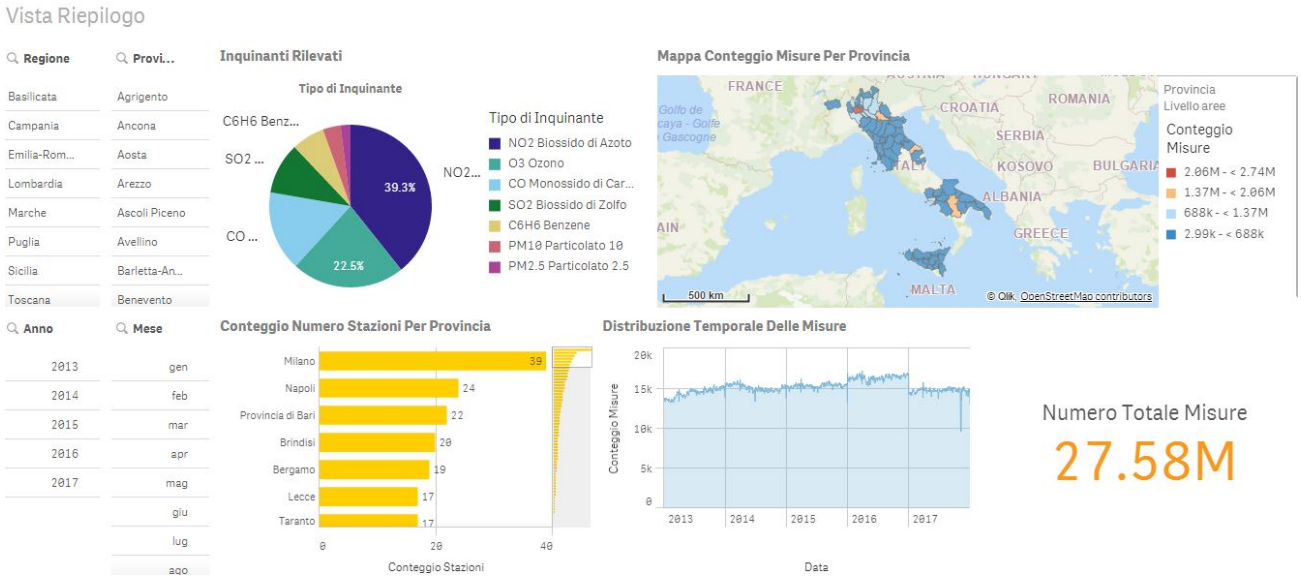

Figure 2: Data visualization – Qlik Sense (summary sheet)

Pentaho has been used for merging data sources and normalizing the corresponding datasets. Data normalization tasks have addressed data redundancy and data inconsistency amongst data sources and within the same source. In our case study, we specifically checked:

1) misspellings (e.g., station name and/or address, pollutant name, unit of measurement name, etc.),
2) data formats,
3) invalid values.

After this phase, regional datasets can be integrated as a shared destination format to whom all different sources must conform.

5.4 Data Visualization

After the ETL process, the dataset size was reduced to 27.58M records from the initial 32M. This dataset has been used as the input for data visualization. In order to achieve fast in-memory data loading and effective visualization options, we have adopted a widely-used, freely available, data analytics platform: Qlik Sense (Desktop Version) [32]. By using Qlik Sense, we have developed a set of dashboards dedicated to the different stakeholders for the examined case study (i.e., citizens, researchers, environment control agency personnel) with the aim

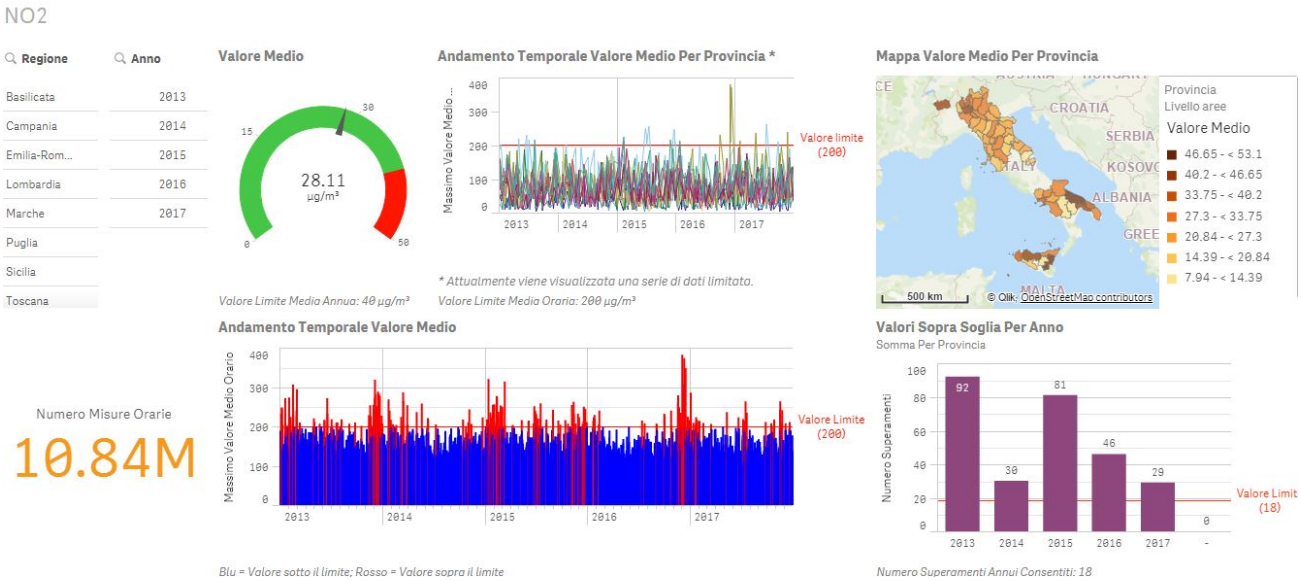

Figure 3: Data visualization – Qlik Sense (pollutant detail sheet, NO_2 case)

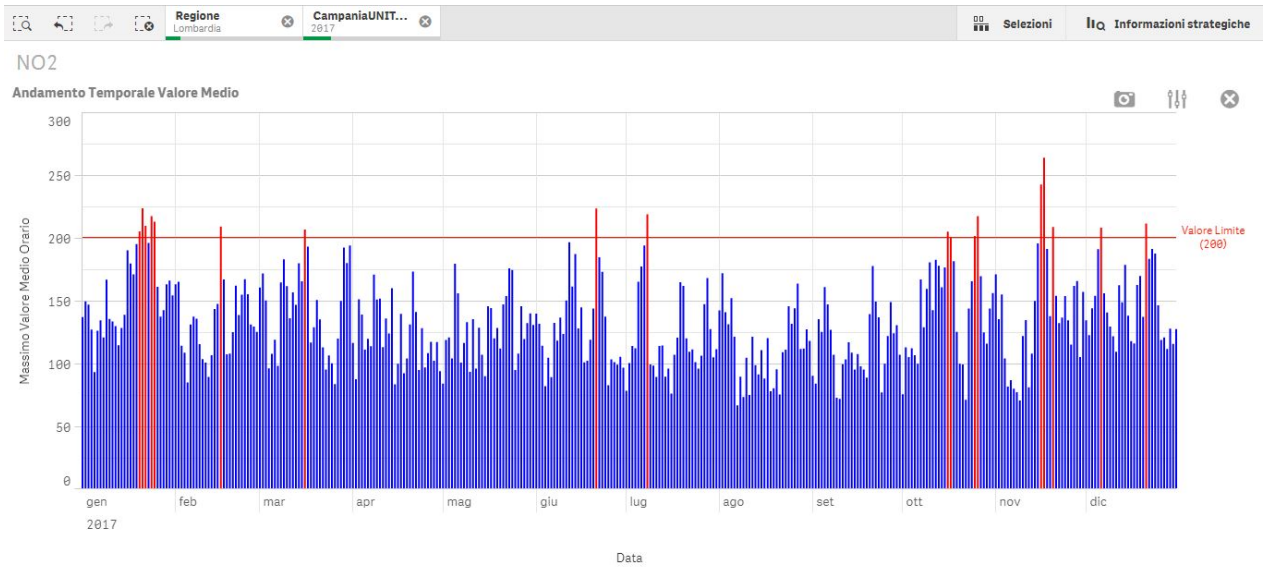

Figure 4: Detail of the vertical bar chart about the average value of the given pollutant (in this case, NO_2). Filters by year (2017) and by region (Lombardy) are applied.

of graphically explaining and effectively analyses performed on cleaned datasets. These dashboards are made up of several charts and filters. According to the Qlik Sense terminology, the developed solution is defined as Qlik Sense app, while each thematic group of charts represents an app sheet.

Depending on the user role, indeed, different charts and views can be accessed. Overall, the developed Qlik Sense app consists of 8 sheets: the first one summarizes core details while the remaining sheets represent a specific set of specific analyses (according to the DFM presented in Section 5.2) on each considered pollutant. Proper time-based and location-based filters have been implemented, as well.

Let us now examine with more details the sheets composing the app.

The first app sheet (represented in Figure 2) is a summary view about all the processed records, in order to count them depending on various criteria.

This sheet is customizable thanks to several filters placed on the left side that allow the user to refine visualization by time period (by year, by month) and by location (by region, by province). The pie chart on the left shows the overall distribution of detected pollutant types. For instance, it can be seen that NO_2 amounts for the 39.3% of the available sensor readings (i.e, nearly 10.84M). The map on the right shows the number of records per province according to a gradient color-scale ranging from blue (less values) to dark red (more values). As it can be seen, the province of Milan has the largest number of records for the referred 5-year time period (it is worth to point out that Figure 2 shows the overall analysis with no filters applied).

The lower part of the sheet hosts, going from left to right, a horizontal a bar chart where monitoring stations per province are counted (also in this case, the province of Milan has 39 monitoring stations, which is the highest number on a per-province basis), a line chart where the daily amount of recordings is reported and an overall counter of the available data points in the referred dataset.

Each of the following sheets refer to a different pollutant. Figure 3 reports the one associated to NO_2. These sheets are aimed at underlying relationships between detected values and regulated thresholds, in order to identify potential sources of concern. Each sheet is formatted as specified below.

In the top left corner, two filters (by region, by year) are available. The speedometer on the right (i.e., the gauge-like chart) allows to compare the average detected value of the given pollutant type against its corresponding regulatory threshold (values beyond the limit are highlighted in red). The limit value is identified by a red line and explicitly mentioned in the footnote of the chart.

By moving towards the right, in the top section of the sheet, we have a line chart depicting the average value detected per province on a daily basis, with an explicit indication of the threshold exceedances. The chart is aimed at emphasizing existing differences amongst provinces. It expresses its maximum potential by selecting a single region via the dedicated filter on the left, so that all the provinces in the same region can be compared, while with no region selected it may be slightly chaotic.

In the top right corner, a map is available, where all Italian provinces are outlined. A gradient color scale is used for depicting average values per province, ranging from light brown (lower values) to dark brown (higher values). This type of chart is very useful to make a straight and effective comparison of values about different areas.

In the bottom left corner, a counter reports the number of readings for the pollutant under examination (i.e., the one the sheet is associated with) depending on the filtering options.
By proceeding towards the right, a vertical bar chart compares, on a daily basis, the average or maximum value (depending on the specific pollutant) against the corresponding limit value. In order to make the chart more effective, measurements are depicted in blue unless they exceed the threshold (in that case they are highlighted in red). Therefore, the proportion of measurements going beyond the threshold is immediately evident. A detail of this chart is reported in Figure 4. As it can be seen, by filtering by time and by region, average values passing the threshold are clearly identifiable.
Finally, in the bottom right corner, a vertical bar chart is locate. It is used for counting the number of measurements exceeding the corresponding limit value per year. Since Italian national regulations allow a given set of threshold exceedances per year per pollutant, this chart immediately shows whether in a given year that limit has been trespassed or not.

5.4 Discussion

Previous sections have highlighted the significant comparability issues in the regional environmental datasets.
In the Table 3, the number of available files from regional websites (related only to year/pollutant considered for the analysis), and the overall size of the sets of files are shown.

Table 3: Processed files and size per region

Region	No. of Files	Size
Basilicata	1295	135 MB
Campania	5	147 MB
Emilia-Romagna	868	191 MB
Lombardia	5	694 MB
Marche	10	26 MB
Puglia	1	42 MB
Sicilia	1	6 MB
Toscana	592	157 MB
Valle D'Aosta	5	5 MB
Total	**2782**	**1.4 GB**

The region with the largest number of files is Basilicata (1295), while the region whose files are the largest ones is Lombardy (694 MB). The last row shows that the total number of files used for this analysis are 2782, while the total weight of all these files is 1.4 GB.
As for memory consumption, the developed full Qlik Sense app has included 8 sheets and 65 charts (interactive elements) for an overall memory occupancy of nearly 1.6 GB.

As for the performed data analyses, several useful insights have been achieved. The following list points out the most relevant ones, per each pollutant.

1. PM10: Lombardy and Campania are the regions with the highest average value; moreover, PM10 is by far the pollutant with the greatest number of threshold exceedances.
2. PM2.5: Lombardy is still the region with the highest average value, with peak in Milano, Monza-Brianza and Cremona provinces. Overall, regions from Northern Italy have a higher average value than southern regions. This is due to the the combination of weather conditions and vehicle density.
3. CO: Sicily is the region with the highest number of limit exceedances, while Campania is the region with the highest average value.
4. NO_2: Apulia and Sicily are the regions with the highest average value; in addition, Barletta-Andria-Trani, Bari, Taranto, Palermo and Catania provinces have an average value greater than the corresponding limit value.
5. O_3: southern regions have a higher average value than northern regions, with the exception of Valle D'Aosta that also has a high average value. Enna and Lecce are the provinces with the highest average value.
6. SO_2: the situation is under control, since values are well below the allowed limit. Messina province is the one with highest average value.
7. C_6H_6: the situation is similar to the one found for SO_2, if not even better. The only province with high values is Siracusa.

6 A Proposal for a CPSS Platform: APOLLON

The challenges emerged so far in managing data from Italian regional environment control agencies and the promising approach disclosed by CPSSs in this research area have motivated the research project named APOLLON, which targets the large-scale, mobile-mediated sensing of pollutants in urban context, according to the CPSS principles. More specifically, the APOLLON Project [3] is a research initiative granted by Apulia Region (Italy) aimed at designing, developing and deploying a platform for urban environmental monitoring in terms of noise and air. Several data streams are gathered from heterogeneous sources (e.g., citizen-owned personal devices, city-managed monitoring stations, etc.). The project novelty relies on: 1) integrating low-cost sensors deployed in urban area; 2) involving citizens directly in monitoring campaigns according to citizen science principles; 3) sharing monitoring outcomes to city managers directly. One of the specific requirements of the platform is to build a monitoring network to integrate information flows gathered from sensors with other information sources thanks to semantic technologies and geo-referential data analysis utilities so that useful insights and high-level correlations can be achieved in near-time.
The architecture of the APOLLON system is organized into four layers (Figure 5). The *IoT layer* includes devices able to collect information on the environment (i.e., mobile and stationary environmental sensors). The *data layer* is devoted to process, integrate and store heterogeneous data sources (social data, sensors, climatic data, clinical data, open data, etc.). The *business layer* is a central processing layer that executes the business logic

and communicates with the persistence level. Finally, the *semantic Decision Support System* (sDSS) represents the interface level between the system and the end user that manages all services related to the interaction with the user (analyses, reporting, cartography, etc.). More specifically, the data layer is in charge to manage the acquired data according to specific ETL (Extraction, Transformation and Loading) procedures, by exploiting typical functionalities of Decision support Systems and a microservice-based architecture.

The architecture briefly described so far is compliant with core CPSS principles (see Section 3). Moreover, the platform backend features a set of components specifically dedicated to data management and included in the so-called *Hybrid Storage Layer (HSL),* made up of five elements: the *Data Management*, the *Data Processing* and the *Message Management* block plus a *Service Catalogue* that indexes and exposes available services.

The *HSL* allows to manage structured/semi-structured/unstructured data, and to manage all storage solutions provided in the APOLLON Data Lake (health data, open data, multidimensional data, user profile/community of interest data, semantic data, IoT sensor data, social data and urban geospatial data). The *HSL* contains relational, non-relational, multidimensional, and SFTP type storage systems. Specifically, we consider the following storage solutions:

- Staging area: temporary storage area for raw data collection to be exposed for processing and cleaning operations provided by the "Data Management" and "Data Processing" blocks;
- Health data storage: area health data provided by local health authorities and Ministry of Health Web portal (e.g., admissions for respiratory diseases, mortality data, etc.);
- Open data storage: area for the collection of the data streams coming from ARPA (i.e., Italian regional agency for the environmental protection) junction boxes and meteorological stations;
- User profile/community of interest storage: area for registering and managing users involved in the project;
- Multidimensional data storage: multidimensional analyses on collected data to highlight any existing correlation;
- Semantic storage: area for ontologies and linked data;
- IoT sensor data storage: area for collecting the data streams coming from sensors;
- Social data storage: the area required for the sentiment analysis phase on data coming from social networks;
- Urban Geospatial data storage: the area aimed at hosting the thematic cartography related to pollutants and weather-climatic stuffs.

At the moment of writing this paper, the APOLLON platform

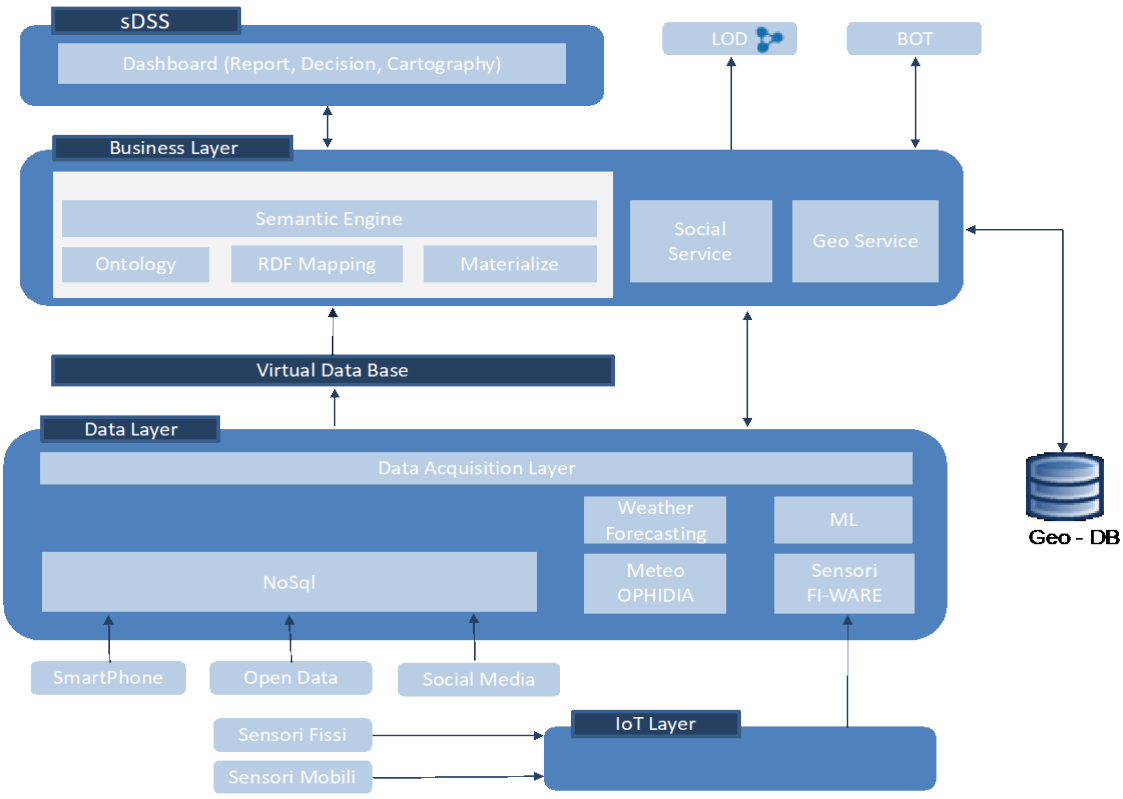

Figure 5: APOLLON platform (logical architecture).

deployment is currently under way and the first two pilot sites are providing the first datasets coming from citizens. Preliminary assessments have shown clearly the potential of the proposed CPSS-based approach, in terms of platform scalability, learning potential for end users, involvement of end users, engagement of policy makers and city managers, suitability to further integration with additional systems (such as analysis of population healthcare status). As for the integration of mobile sensed data with whose of official statistics, crucial aspects are the specialization of completeness considering the representativeness, selectivity and sparsity aspects, the trustworthiness in the security quality dimension and the specialization of accuracy, consistency and redundancy aspects [28].

A thorough analysis of the effectiveness of the proposed CPSS-based approach will be performed in the upcoming months.

7 Conclusions

In this paper, a thorough analysis of the current Italian scenario in terms of available and comparable institutional datasets for air pollution monitoring has been performed. By starting from available data sources (i.e., datasets published by Italian regional environment control agencies), a series of shortcomings has been identified, ranging from data heterogeneity, inconsistency and incompleteness to significant limitations in accessing monitoring data.

A solution for collecting, processing, aggregating and visualizing air pollution datasets from a subset of Italian regions, referring to a five-year time period and to a subset of standardized pollutants has been proposed. This approach highlighted data-related challenges to the adoption of Cyber-Physical Social Systems (CPSS) in this sector. Both these challenges and the analysis insights achievable in the visualization process have been presented in this paper. Moreover, by starting from the elements identified during the design and implementation steps of the proposed solution, a regional CPSS addressing noise and air pollution monitoring, has been devised, as the first step towards the adoption of CPSS for environmental monitoring nationwide. The CPSS platform, named APOLLON has been described in the final section of the paper.

In the next near future, challenges related to the integration of big data and official statistics will be investigated in order to properly exploit the potentials of mobile crowd sensing for urban environmental pollution monitoring. Main dimension clusters of data quality will be detailed and analyzed in the domain of big data from mobile crowd sensors and approaches to effectively include people as data scientists will be described.

ACKNOWLEDGMENTS

This work was supported in part by the research project “APOLLON - environmentAl POLLution aNalyzer”, within the “Bando INNONETWORK 2017” funded by Regione Puglia (Italy) in the framework of the “FESR - Fondo Europeo di Sviluppo Regionale”.

REFERENCES

[1] WHO (World Health Organization), “Ambient Air Pollution: A Global Assessment of Exposure and Burden of Disease,” 2016.

[2] WHO (World Health Organization), “How air pollution is destroying our health,” 2019. [Online]. Available: https://www.who.int/air-pollution/news-and-events/how-air-pollution-is-destroying-our-health.

[3] The Center for Public Integrity, “Most EPA Pollution Estimates Are Unreliable, So Why Is Everyone Still Using Them?,” *EcoWatch*, 2018. [Online]. Available: https://www.ecowatch.com/epa-emission-factors-2529636639.html.

[4] WHO (World Health Organization), *Monitoring ambient air quality for health impact assessment*. 2002.

[5] European Union, “Guidance on Assessment under the EU Air Quality Directives,” 2005.

[6] IAQM (Institute of Air Quality Management), “A guide to the assessment of air quality impacts on designated nature conservation sites,” London, UK, 2019.

[7] P. Barn, P. Jackson, N. Suzuki, and T. Kosatsky, “Air Quality Assessment Tools: A Guide for Public Health Practitioners,” Vancouver, Canada, 2011.

[8] EPA (Environment Protection Authority) - South Australia, “Ambient Air Quality Assessment,” 2016.

[9] Italian Ministry for the Environment Land and Sea, “Environmental Challenges - Summary of the State of the Environment in Italy,” Rome, Italy, 2009.

[10] T. A. J. Kuhlbusch, “Challenges and the Future of Urban Air Quality Monitoring in Europe,” 2014.

[11] J. Zeng, L. T. Yang, M. Lin, H. Ning, and J. Ma, “A survey: Cyber-physical-social systems and their system-level design methodology,” *Futur. Gener. Comput. Syst.*, 2016.

[12] EEA (European Environment Agency), “Air quality in Europe - 2018 Report,” Luxembourg, 2018.

[13] Down To Earth, “Air Pollution: PM Levels Continue to Exceed EU Limit in Large Parts of Europe,” 2016. [Online]. Available: https://www.downtoearth.org.in/news/air/air-pollution-pm-levels-continue-to-exceed-eu-limit-in-large-parts-of-europe-56427.

[14] ISPRA, “Qualità dell’Ambiente Urbano,” Rome, Italy, 2014.

[15] ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale), “Annuario dei Dati Ambientali (Environmental Data Yearbook) 2018,” 2019.

[16] European Commission, “Air Quality - Existing Legislation,” *Environment*. [Online]. Available: https://ec.europa.eu/environment/air/quality/existing_leg.htm.

[17] EEA (European Environment Agency), “The Air Quality Monitoring Situation in Europe: State and Trends,” 2016. [Online]. Available: https://www.eea.europa.eu/publications/92-9167-058-8/page010.html.

[18] C. B. B. Guerreiro, V. Foltescu, and F. de Leeuw, “Air quality status and trends in Europe,” *Atmos. Environ.*, vol. 98, pp. 376–384, 2014.

[19] EEA (European Environment Agency), “The air quality monitoring situation in Europe - State and trends,” 2016. [Online]. Available: https://www.eea.europa.eu/publications/92-9167-058-8/page010.html.

[20] S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, and B. Nath, “Real-time air quality monitoring through mobile sensing in metropolitan areas,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2013, p. 8.

[21] ECMWF, “Monitoring Air Pollution Across Europe,” 2019. [Online]. Available: https://atmosphere.copernicus.eu/monitoring-air-pollution-across-europe.

[22] ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale), “Sistema InfoAria SINAnet,” 2018. [Online]. Available: http://www.webinfoaria.sinanet.isprambiente.it/.

[23] B. Guo, Z. Yu, and X. Zhou, “A Data-Centric Framework for Cyber-Physical Social Systems,” *IT Prof.*, vol. 17, pp. 4–7, 2015.

[24] J. Jin, J. Gubbi, S. Marusic, and M. Palaniswami, “An information framework for creating a smart city through internet of things,” *IEEE Internet Things J.*, vol. 1, no. 2, pp. 112–121, 2014.

[25] X. Du, O. Emebo, A. Varde, N. Tandon, S. N. Chowdhury, and G. Weikum, “Air quality assessment from social media and structured data: Pollutants and health impacts in urban planning,” in *2016 IEEE 32nd International Conference on Data Engineering Workshops, ICDEW 2016*, 2016, pp. 54–59.

[26] S. Kuznetsov, “Authoring urban landscapes with air quality sensors,” in *Proceedings of Sigchi Conf. on Human Factors in Computing Systems*, 2011, pp. 2375–2384.

[27] L. Sanchez *et al.*, “SmartSantander: IoT experimentation over a smart city testbed,” *Comput. Networks*, vol. 61, pp. 217–238, 2014.

[28] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi, “From data quality to big data quality,” *J. Database Manag.*, vol. 26, no. 1, pp. 60–82, 2015.

[29] M. Ehling, “Harmonising Data in Official Statistics,” in *Advances in Cross-National Comparison*, Boston, MA: Springer US, 2003, pp. 17–31.
[30] M. Golfarelli and S. Rizzi, *Data Warehouse Design, Modern Principles and Methodologies*, 1st ed. McGraw-Hill, 2009.
[31] Hitachi Group, “Pentaho Community Edition (CE): Data Integration, Business Analytics and Big Data,” 2016. [Online]. Available: http://www.pentaho.com. [Accessed: 01-Nov-2016].
[32] QlikTech International AB, “Qlik Sense - Data Analytics Platform,” 2019. [Online]. Available: https://www.qlik.com/us/products/qlik-sense.