

Article

Not peer-reviewed version

A Multi-Source Sensor Dataset for Spain: Integrating Air Quality, Meteorological, Mobility and Calendar Records

[Juan Bonastre-Egea](#) , Andrés Bueno-Crespo , [Juan Morales-García](#) *

Posted Date: 28 May 2026

doi: 10.20944/preprints202605.1976.v1

Keywords: open dataset; air quality monitoring; multi-source sensor integration; mobile network mobility data; meteorological sensors; environmental monitoring; nationwide; Spain



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

A Multi-Source Sensor Dataset for Spain: Integrating Air Quality, Meteorological, Mobility and Calendar Records

Juan Bonastre-Egea ¹ , Andrés Bueno-Crespo ¹  and Juan Morales-García ^{2,*} 

¹ Escuela Politécnica Superior. Universidad Católica de Murcia (UCAM), 30107, Murcia, Spain

² Department of Software and Computing Systems. Universidad de Alicante (UA), 03690, Alicante, Spain

* Correspondence: juan.morales@ua.es

Abstract

Air quality forecasting and environmental health research at urban and regional scales depend on the combination of measurements from heterogeneous sensor networks, yet the construction of integrated multi-source datasets is rarely described or released as a self-contained deliverable. This paper presents an open dataset that combines four sensor-derived sources covering the whole of Spain over the period 2022 to 2024: hourly air quality observations from the 588 stations of the national network operated by the *Ministerio para la Transición Ecológica y el Reto Demográfico* (MITECO), daily meteorological records from the *Agencia Estatal de Meteorología* (AEMET), daily mobility indicators derived from anonymised mobile telephony events published by the *Ministerio de Transportes y Movilidad Sostenible* (MITMA) at the municipality level, and a calendar of national and Autonomous-Community public holidays. The processing pipeline harmonises sources that differ in temporal resolution, spatial codification and quality regime into a tidy hourly table indexed by station and timestamp, with a fixed feature schema of 56 variables per record. Air quality stations are paired with their nearest AEMET station through a three-tier distance rule, and the daily exogenous features are aligned to the air quality time axis through a two-variant temporal-alignment scheme (lag-and-expand to the hourly grid for the hourly release, same-calendar-day join for the daily release). A complementary daily-resolution variant of the dataset is also released, with 72 columns and the same feature schema except for the air quality block, which is aggregated to daily mean, minimum and maximum. The integrated dataset contains approximately 14 million hourly records across the 588 stations and is released on Zenodo (DOI 10.5281/zenodo.20196221) under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. It is intended as a substrate for research on air quality forecasting, environmental epidemiology and multi-source data fusion at nationwide scale.

Keywords: open dataset; air quality monitoring; multi-source sensor integration; mobile network mobility data; meteorological sensors; environmental monitoring; nationwide; Spain

1. Introduction

Air pollution remains one of the most consequential environmental risks to public health in Europe. The European Environment Agency attributes hundreds of thousands of premature deaths each year to long-term exposure to fine particulate matter, nitrogen dioxide and ground-level ozone, and the 2021 update of the World Health Organization (WHO) air quality guidelines lowered the recommended thresholds for nearly every regulated pollutant, leaving most European cities far from compliance [1,2]. Accurate short-term forecasts of pollutant concentrations are not merely an academic exercise: they support the activation of low-emission zones, the planning of school activities, the issuance of health advisories for sensitive populations, and the calibration of regulatory air quality models operated by national agencies [3]. The quality of such forecasts, and of the analyses that underpin air quality policy more generally, depends in turn on the quality of the data that feed them.

The dataset described in this paper integrates four sources that jointly characterise the state of the atmosphere at each monitoring site. The first and core source is the air quality data itself, consisting of hourly concentrations of the regulated atmospheric pollutants (NO_2 , NO_x , O_3 , SO_2 , PM_{10} , $\text{PM}_{2.5}$, CO and benzene) measured at the 588 stations of the Spanish national air quality network operated by the *Ministerio para la Transición Ecológica y el Reto Demográfico* (MITECO). The second source is meteorology, since temperature, humidity, wind and precipitation drive both the emission and the atmospheric dispersion of pollutants, and the daily climatological summaries of the *Agencia Estatal de Meteorología* (AEMET) network provide the corresponding observations for the whole country. The third source is human mobility at the municipality level, since road traffic and other activity-related movements are among the largest proximate sources of urban air pollution; the daily mobility indicators published by the *Ministerio de Transportes y Movilidad Sostenible* (MITMA) are derived from anonymised mobile telephony records. The fourth source is the public holiday calendar, both national and regional. Each of these sources can be regarded as a sensor system in its own right: a regulated network of in-situ atmospheric analysers in the case of MITECO and AEMET, and a derived sensor network in the case of MITMA, where mobile telephony events generated by the network of base stations act as proxy observations of human movement.

The link between air quality and these three contextual sources is well established in the literature. Road traffic is the dominant source of urban NO_x emissions in most European cities and a substantial contributor to primary PM and to combustion-related species such as CO and benzene; the bimodal diurnal cycle observed at urban-traffic stations follows the commuting load with a morning and a stronger evening peak, and shifts in commuting intensity translate directly into shifts in the peak concentrations of these pollutants [4–6]. The public holiday calendar acts through the same mechanism but on a different temporal axis: weekends and public holidays cut commuting and freight traffic, depress combustion-related emissions and produce a measurable reduction in NO_2 and NO_x concentrations, while the reduction in fresh NO emissions weakens ozone titration and elevates ozone peaks (the so-called weekend ozone effect). National holidays, regional festivities and the summer-vacation period further modulate the baseline of mobility and emissions, in ways that a calendar-aware feature set captures naturally.

The relevance of these complementary sources has been documented repeatedly in the air quality literature, and was made unusually visible at a continental scale by the COVID-19 lockdowns of 2020, which provided the cleanest natural experiment available so far of the mobility-to-emissions link. NO_2 concentrations dropped by 50% to 70% in major Spanish cities in close correlation with reductions in vehicle mobility, while PM_{10} responded more weakly because of its broader range of emission sources [7–10]. Comparable reductions were observed across the European Union, with NO_2 drops of 30% to 60% reported in Italian, French and German urban areas during the strictest containment weeks, and a clear partial recovery once mobility restrictions were relaxed. The lockdown evidence demonstrated that the relationship between mobility, meteorology and pollutant concentrations can be quantified at city and regional scale when the inputs are available in a sufficiently structured form; under normal conditions the relationship is harder to isolate because the variation in mobility is smaller and the other drivers (meteorology, secondary chemistry, episodic regional transport) compete for the same variance. The study period covered by the dataset released here (2022 to 2024) is immediately post-lockdown and reflects the operational regime under which forecasting and policy decisions are normally made, while building on the integration methodology that the lockdown experience showed to be valuable.

The four sources are heterogeneous along several axes. They operate at different temporal resolutions (hourly air quality, daily meteorology, daily mobility, daily holidays), use different spatial codifications (national air quality station codes, AEMET station identifiers, MITMA mobility zones derived from *Instituto Nacional de Estadística* (INE) census districts), exhibit different patterns of missingness, and follow different quality control regimes. Most published air quality studies handle this by working with a single city or a small number of stations, and by using either weather variables alone [11–13] or coarse mobility proxies such as traffic counters or closed-circuit television (CCTV)

flows [5,6,14]. Even when richer multi-source combinations are explored at the city level [4,15,16], the integrated dataset is rarely released as a self-contained product: it lives inside a particular model training pipeline and has to be rebuilt from scratch by anyone wishing to reproduce or extend the work. Nationwide releases that combine air quality, weather and mobile-network-derived mobility into a tidy, station-level table covering several years of data, to the best of our knowledge, do not exist.

This paper addresses that gap. We describe and release an integration pipeline together with the resulting dataset, which covers the period 2022 to 2024. The lower bound of the temporal coverage is constrained by the availability of the MITMA mobility study, which was launched in early 2022 and does not provide municipality-level records for earlier years; the upper bound is the end of the last complete calendar year at the time of the release. The pipeline performs spatial association between air quality and meteorological stations through a three-tier distance rule, temporal alignment of records sampled at different rates (with one alignment scheme for the hourly release and one for the daily release), and a per-station enrichment join that produces a tidy table with a fixed schema across the network. Two temporal variants are released side by side: an hourly variant with 56 feature columns and approximately 14 million records, and a daily variant with 72 feature columns and approximately 600 000 records that uses an unlagged join of the daily exogenous features and is intended for users whose downstream analysis already operates at daily resolution. The resulting dataset is published on Zenodo under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence [17] and is intended for research on air quality forecasting, environmental epidemiology and multi-source data fusion at nationwide scale. The source code of the pipeline is also made available, so that the dataset can be regenerated, extended in time, or adapted to other countries with comparable open data infrastructure.

The rest of the paper is organised as follows. Section 2 reviews the literature on existing air quality datasets, on multi-source data fusion at the urban and nationwide scale, and on the use of mobility as a contextual predictor. Section 3 describes the four sensor-derived data sources and details the integration pipeline. Section 4 specifies the structure, format and contents of the released data records. Section 5 reports the technical validation checks. Section 6 offers guidance for users of the dataset, including a brief account of preliminary modelling experiments that informed the design. Section 7 closes the paper.

2. Related Work

This section reviews work relevant to the dataset described in this paper. We organise the discussion around three threads: (i) the landscape of open air quality data and the comparative absence of integrated multi-source releases, (ii) approaches that combine air quality with meteorological, mobility or other contextual sources, mostly at the urban scale, and (iii) the use of mobile-network-derived data as a sensor for human activity, both in general urban research and specifically in air quality studies.

2.1. Open Air Quality Data and the Integration Gap

Open air quality data have become considerably more accessible over the last decade, mostly through agency portals and through community platforms that harmonise heterogeneous national sources. The OpenAQ platform [18] aggregates real-time and historical observations from more than eighteen thousand monitoring locations across over a hundred countries, exposing a uniform API for measurements of the main regulated pollutants. At the European level, the e-Reporting infrastructure of the European Environment Agency [2] centralises the validated air quality data submitted by Member States and feeds the agency's annual status assessments. National administrations distribute their own raw observations in parallel, in formats that range from per-station REST APIs to bulk CSV downloads such as those of the Spanish MITECO network [19]. A complementary line of work explores how to densify these reference networks at urban scale through low-cost sensor platforms: Felici-Castell et al. [20], for instance, document an AI-enabled low-cost wireless sensor network deployed

in the Valencian Community alongside the official regulatory stations, with deep neural network calibration of the raw readings and a 24-hour forecasting horizon.

These resources share a common scope, however: they expose air quality measurements alone, or air quality plus locally co-located meteorology in the case of the urban sensor networks. Researchers wishing to combine them with mobility, land use or socio-demographic data must obtain each additional source from a different provider, harmonise temporal resolutions and spatial codifications by hand, and decide how to handle gaps and missingness across sources. The resulting integrated tables are typically internal to a single project's pipeline and are not redistributed; reproducing a paper's input data therefore requires reproducing the integration work, often from incomplete descriptions. A small number of dataset releases pair air quality measurements with weather or traffic counters at the city level, but to our knowledge there is no public release that combines air quality, meteorology and mobile-network-derived mobility at the nationwide scale across multiple years.

2.2. Multi-Source Data Fusion for Air Quality

A substantial body of work demonstrates that combining heterogeneous data sources can improve air quality estimation and forecasting, particularly in urban contexts. Early work by Yi et al. [15] introduced a distributed fusion network combining air quality stations, meteorology and points of interest at the city level in China. Subsequent contributions have extended this idea with attention-based encoders that align modalities at fine temporal resolution [21], with transformer-style spatio-temporal architectures that operate over multi-source cyber-physical sensor streams [16], and with hybrid supervised and self-supervised models that adapt feature importance during training [22]. Recent multimodal models incorporate emission inventories alongside meteorology and pollutant observations [23], or fuse remote sensing imagery with ground-based measurements [24]. In all of these works the input to the model is an integrated multi-source representation; in none of them is that integrated representation released as a self-contained dataset for reuse.

City-scale work in the same line illustrates how even modest source combinations improve forecasting accuracy and reveal interpretable patterns. Arsov et al. [25] report multi-horizon PM_{10} forecasts for Skopje using an LSTM-based recurrent architecture fed by historical air quality measurements from a network of city sensors and by daily meteorological summaries, showing systematic gains over an ARIMA baseline at the 6-, 12- and 24-hour horizons. Graça et al. [26] deploy a network of low-cost sensor stations together with co-located meteorological stations in the city of Aveiro and characterise the spatial and temporal patterns of NO_2 , PM and ozone across a full year, with traffic-hotspot stations exhibiting markedly different daily profiles from background sites. Cukjati et al. [27] combine ground-based IoT sensors with Sentinel-5P satellite retrievals through a Voronoi-cell ensemble of regression models, and produce gridded NO_2 fields whose spatiotemporal resolution exceeds that of either input alone.

At the European and Spanish level, several studies illustrate the value of combining air quality with related sources at the urban scale. Folgado et al. [28] develop a deep learning model fed by air quality, meteorology and traffic counters for the city of Valencia. González-Enrique et al. [12] use exogenous meteorological variables to forecast NO_2 in the Bay of Algeciras. Terroso-Sáenz et al. [29] extend the scope to the whole national network through a heterogeneous graph neural network that exploits inter-station relations. Earlier hybrid deep learning frameworks combining pollution and weather inputs established the methodological baseline at the city scale [11,13]. The pattern across this literature is consistent: data fusion is well established as a useful direction, the model architectures continue to evolve, and the integrated datasets on which these models are trained remain largely undocumented and unreleased.

2.3. Mobile-Network Data as a Sensor for Human Activity

Mobile telephony has been studied for over a decade as a passive sensor for human movement. Calabrese et al. [30] provide an early and influential survey of urban sensing applications based on mobile phone network data, covering individual mobility, collective flows, presence estimation and

event detection. The methodological backbone of this line of work is the extraction of origin–destination matrices from Call Detail Records (CDRs), where the relevant design choices (time-based versus routine-based aggregation, scaling to population totals, projection onto the road network) determine how faithfully the resulting flows reproduce reference traffic counts. Mamei et al. [31] compare prototypical implementations of the two main families of CDR-derived OD-matrix construction and evaluate them against ground-truth mobility surveys, providing a useful checklist of the assumptions that any downstream user of such matrices inherits. The Spanish national mobility study published by MITMA from 2022 onwards [32] belongs to this lineage: it is derived from anonymised events of the Orange Spain mobile network and reports daily aggregates of trips, overnight stays and population counts at the level of mobility zones that map to INE municipalities.

The use of mobility information as a covariate in air quality models is more recent, but several Spanish groups have already explored it. Morales-García et al. [4] introduce mobility-derived features into city-level pollution prediction models and report improvements over weather-only baselines, and subsequent work on the same line [33] examines how synthetic mobility data can supplement scarce real records. Hameed et al. [5] fuse air quality with traffic analytics in a multimodal deep learning framework, and Ibrahim and Lyons [6] show that even repurposed CCTV imagery can be converted into a city-scale NO₂ signal. Xu et al. [14] propose interpretable multi-source pipelines that include mobile-derived flows alongside meteorology and emissions. Work on mobility forecasting itself with region-based flows and hierarchical spatial tessellation [34] is complementary to the use of mobility as an air quality predictor.

Multi-source sensor integration is also an active topic in adjacent environmental domains. Recent work on near-real-time integration of heterogeneous seismic feeds [35] illustrates how a similar pipeline philosophy can be applied to a different domain, with comparable challenges of temporal alignment and spatial association across independent agencies.

3. Materials and Methods

This section describes the four data sources that feed the integration pipeline and the procedure that combines them into the released dataset. Section 3.2 presents each source in turn, with summary statistics computed at its native resolution. Section 3.3 details the spatial association, temporal alignment and per-station enrichment operations that produce the final tidy table.

3.1. Overview of the Integration Pipeline

The integration pipeline produces, for each station of the MITECO air quality network, a tidy table indexed by hourly timestamp and containing a fixed feature schema. Figure 1 summarises the pipeline, which is organised in four stages:

1. **Metadata consolidation.** The yearly station inventory files published by MITECO are read and deduplicated by station code, producing a single station registry that contains coordinates, province and municipality codes, the area-type classification and the dominant emission source label. The same operation is applied to the AEMET station inventory obtained from the OpenData API. The remaining two sources do not require a consolidation step at this stage: the MITMA mobility records are keyed directly by the INE 5-digit municipality code, which is constructed from the province and municipality codes already present in the MITECO registry, so no separate metadata file is needed and no acquisition of additional inventory data is performed; the public holiday calendar is keyed by date and Autonomous Community code, both of which are derived from the same MITECO registry through the official INE province-to-CCAA mapping.
2. **Per-source acquisition and cleaning.** Each of the four sources is acquired and processed independently through its own data interface. MITECO air quality data are downloaded as bulk annual CSV files from the agency’s open data portal; AEMET meteorological data are obtained from the OpenData REST API, with paginated calls respecting the maximum six-month window and the rate limit of approximately 50 requests per minute; MITMA mobility data are downloaded as

monthly archives of compressed CSV files from the agency's open mobility portal; and the public holiday calendar is constructed through the Python `holidays` package, which encodes both the national list and the Autonomous-Community lists for the study period. The cleaning operations performed on each source are described in detail in Sections 3.2.1 to 3.2.4. The output of this stage is a set of cleaned per-station and per-municipality time series with continuous datetime indices, one row per native time unit and one column per published variable.

3. **Spatial association.** Each MITECO air quality station is linked to the relevant record in every other source through identifiers derived from its inventory metadata. For meteorology, the MITECO and AEMET networks use independent station codes with no shared identifier, so each MITECO station is paired with the nearest AEMET station through a three-tier rule based on great-circle distance, computed by the Haversine formula [36]. The first tier accepts the nearest match if it lies within a 50-kilometre threshold, the second tier restricts the search to the same INE province if the first tier fails, and the third tier accepts the nationwide nearest station as an unconditional fallback; the rule and the rationale for the 50-kilometre cut-off are described in detail in Section 3.3.1, and the resulting empirical distribution of pair distances is reported in Section 5, where it is shown that every MITECO station in the network is matched under the first tier and the median pair distance is below 5 km. The output of this stage is a lookup table that records, for each MITECO station, the assigned AEMET station, the Haversine distance between the two and the tier under which the match was made. For mobility, the five-digit INE municipality code of each MITECO station is constructed by concatenating its province and municipality codes, and the MITMA records of that municipality are then attached to the station; all stations within the same municipality share the same mobility features. For holidays, the Autonomous Community code derived from the station's province (through the official INE province-to-CCAA mapping) is used to retrieve the applicable regional holiday calendar, in addition to the national one which is the same for every station.
4. **Per-station integration.** For each MITECO air quality station the pipeline assembles the integrated table by joining the air quality series with the meteorological, mobility and holiday records. The exact procedure differs slightly between the two temporal variants of the released dataset. In the hourly variant the air quality observations remain at their native hourly resolution, and the daily meteorological and mobility records are first lagged by one day and then expanded to the 24 hours of the corresponding day (the lag-and-expand procedure described in Section 3.3.2). In the daily variant the air quality observations are previously aggregated to daily mean, minimum and maximum values, and the daily meteorological and mobility records are merged at the same calendar day without temporal shift. The holiday flags are assigned at time t without lag in both variants. The result is one file per station with a uniform schema across the network, in either hourly or daily resolution.

Raw physical values are preserved throughout the pipeline. No normalisation, imputation or station filtering is applied at this stage, since these choices are properly the user's downstream of the released dataset. The pipeline produces two temporal variants of the integrated dataset side by side. The hourly variant retains the native resolution of the air quality observations and uses the lag-and-expand alignment of the daily exogenous features described above. The daily variant first aggregates the hourly air quality series to daily mean, minimum and maximum statistics, and then joins the daily meteorological and mobility records at the same closed calendar day without any temporal shift, on the rationale that once a day has closed its full weather and mobility summary is part of the same closed-day snapshot as the air quality aggregates. The aggregation requires at least 18 valid hourly readings out of 24 (75% coverage) for a daily value to be computed, following the data-capture criterion that Annex I of Directive 2008/50/EC [37] prescribes for valid daily averages of regulated atmospheric pollutants; days that fall below this threshold are marked as missing rather than aggregated from partial data. Both variants are released (Section 4).

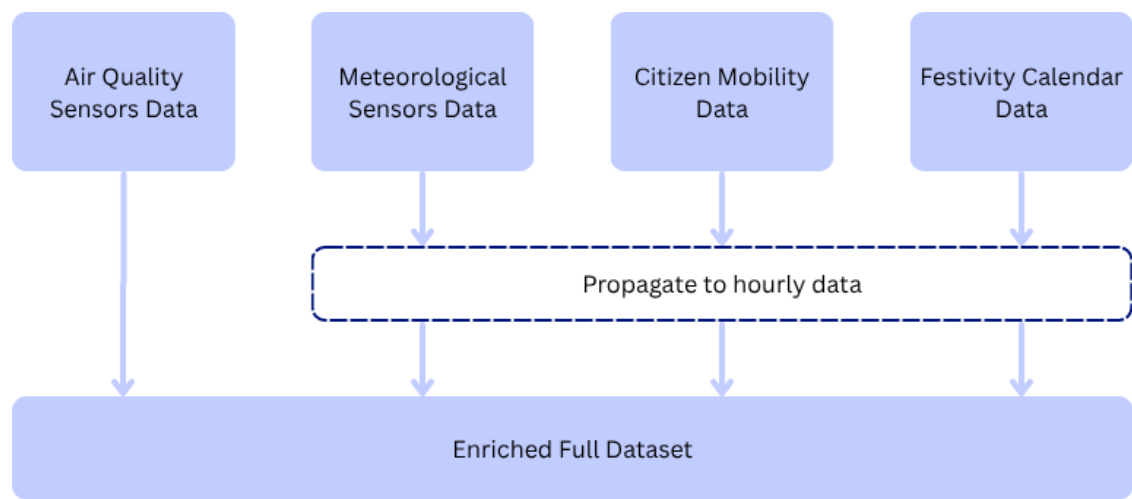


Figure 1. Overview of the four-stage multi-source data integration pipeline. See Section 3.3 for a description of each stage.

3.2. Data Sources

The four data sources that feed the integration pipeline are described below. Each operates under a different institutional mandate, follows a different publication schedule, and uses a different spatial and temporal encoding; the operations that reconcile them are described in Section 3.3.

3.2.1. Air Quality Data: The MITECO Sensor Network

The Spanish national air quality network is operated under the coordination of MITECO and brings together stations managed by the Autonomous Communities and several local administrations. The network reports hourly concentrations of the regulated pollutants (NO₂, NO_x, O₃, SO₂, CO, PM₁₀, PM_{2.5}, C₆H₆) following the European Air Quality Framework Directive, and publishes the data through annual CSV files with a uniform schema [19]. Each station is identified by a national code (COD_LOCAL) constructed from the INE province code, the INE municipality code and a station identifier. The metadata associated with each station include geographic coordinates (latitude, longitude, altitude), an area-type classification (urban, suburban, rural) and a dominant emission source classification (traffic, industrial, background). Hours are reported in UTC and assigned to the end of the sampling interval.

For this work we retrieved hourly data covering the years 2022 to 2024 for the 588 stations that were active for at least part of the period. The data are published as annual CSV files that can be downloaded from the MITECO web portal without API authentication. The native CSV format encodes one record per day, with one value column per hour (H01 to H24). The acquisition step pivots these wide records into a long hourly time series, joins the station metadata, and stores one file per station with a uniform schema. The processed dataset contains 14,944,488 station-hour records over the three-year window, an average of 25,416 valid records per station out of a theoretical maximum of 26,304 (96.6% temporal coverage on average). All eight pollutants are retained in the release, with substantially different coverage profiles across the network. NO₂ and NO_x are the most widely deployed and are present at almost every station; O₃, SO₂ and PM₁₀ are also broadly available; CO and PM_{2.5} are reported at roughly half the network; and C₆H₆ is the least common, present at a small subset of stations. Table 1 reports per-variable summary statistics for the eight pollutants.

Table 1. Per-variable summary statistics for the eight pollutants of the MITECO block, ordered by station coverage (most widely reported first). *N* is the number of station-hour records reported by stations that measure the variable, out of a theoretical maximum of 15,466,752 (588 stations × 26,304 hours); *missing* is the share of those records that are NaN. Concentrations are in $\mu\text{g}/\text{m}^3$ except for CO, which is in mg/m^3 .

Variable	Unit	N	Missing	Mean	Std	Median	Min	Max
NO2	$\mu\text{g}/\text{m}^3$	14,087,160	6.4%	12.84	14.29	8.0	−1.0	403
NOx	$\mu\text{g}/\text{m}^3$	13,886,592	33.9%	18.83	29.68	10.5	−2.0	2,785
O3	$\mu\text{g}/\text{m}^3$	11,364,600	4.3%	59.21	27.86	59.9	0.0	339
SO2	$\mu\text{g}/\text{m}^3$	11,016,048	6.4%	3.44	4.46	3.0	0.0	583
PM10	$\mu\text{g}/\text{m}^3$	10,433,952	9.9%	20.30	28.27	15.0	0.0	3,607
PM25	$\mu\text{g}/\text{m}^3$	6,318,768	11.7%	9.02	9.80	7.0	−0.3	1,024
CO	mg/m^3	6,148,536	11.8%	0.31	0.23	0.23	0.0	13.1
C6H6	$\mu\text{g}/\text{m}^3$	2,561,712	21.9%	0.46	1.01	0.2	0.0	200.9

The small negative minima for NO₂, NO_x and PM_{2.5} correspond to sensor calibration drift near the detection limit, which is a normal artefact of in-situ pollutant analysers. The values are retained in the release because the dataset preserves raw physical readings; a user wishing to remove them can do so trivially with a positivity filter.

Figure 2 shows the daily median concentrations of the eight pollutants of the MITECO block across the network. The seven-day rolling median is overlaid on the cross-station interquartile range, and reference lines are drawn for the WHO 2021 Air Quality Guideline, the current EU annual limit under Directive 2008/50/EC and the revised EU 2030 limit under Directive (EU) 2024/2881 where applicable.

The seasonal structure of the network differs markedly across pollutants. NO₂ exhibits a clear seasonal cycle, with daily medians of 12 to 18 $\mu\text{g}/\text{m}^3$ during autumn and winter (when stagnant atmospheric conditions, intensified domestic heating and sustained traffic emissions combine) and 5 to 10 $\mu\text{g}/\text{m}^3$ during much of the warm season, hovering close to the WHO guideline of 10 $\mu\text{g}/\text{m}^3$. The network-wide median stays well below the EU annual limit of 40 $\mu\text{g}/\text{m}^3$ throughout the period, but the upper end of the cross-station distribution regularly approaches the new EU 2030 limit of 20 $\mu\text{g}/\text{m}^3$ during cold-season episodes. NO_x follows a similar seasonal pattern at higher absolute concentrations.

PM₁₀ has a flatter baseline near the WHO guideline of 15 $\mu\text{g}/\text{m}^3$, with the network-wide median oscillating between 10 and 20 $\mu\text{g}/\text{m}^3$. The series is punctuated by short, intense spikes that correspond to specific events: the most prominent peak in the record, in mid-March 2022, coincides with a major Saharan dust intrusion that affected the entire Iberian Peninsula and pushed the network-wide daily median above 50 $\mu\text{g}/\text{m}^3$. The interquartile range across stations widens sharply during these episodes, indicating that not all areas are affected uniformly, and that PM₁₀ exposure is governed as much by episodic regional transport phenomena as by chronic local emissions. PM_{2.5} follows a similar profile at lower absolute concentrations (typical median between 5 and 10 $\mu\text{g}/\text{m}^3$), routinely exceeding the WHO guideline of 5 $\mu\text{g}/\text{m}^3$ and frequently approaching the EU 2030 limit of 10 $\mu\text{g}/\text{m}^3$.

Ozone shows the opposite seasonal pattern, with the highest concentrations in spring and summer (median between 70 and 90 $\mu\text{g}/\text{m}^3$) driven by photochemical production under intense solar radiation, and the lowest in winter (40 to 50 $\mu\text{g}/\text{m}^3$). The WHO peak-season threshold of 60 $\mu\text{g}/\text{m}^3$ is exceeded by the network-wide median throughout the warm half of each year, although the EU and EU 2030 limit of 120 $\mu\text{g}/\text{m}^3$ is rarely reached at the network-wide level. CO, SO₂ and benzene remain comfortably below their respective regulatory thresholds across the entire period: CO oscillates around 0.2 to 0.4 mg/m^3 against a WHO 24-hour guideline of 4 mg/m^3 ; SO₂ stays close to 2 to 3 $\mu\text{g}/\text{m}^3$, two orders of magnitude below the EU daily limit; and benzene varies between 0.1 and 0.6 $\mu\text{g}/\text{m}^3$ against an EU annual limit of 5 $\mu\text{g}/\text{m}^3$, with a clear seasonal modulation that elevates winter concentrations in line with NO₂ and CO.

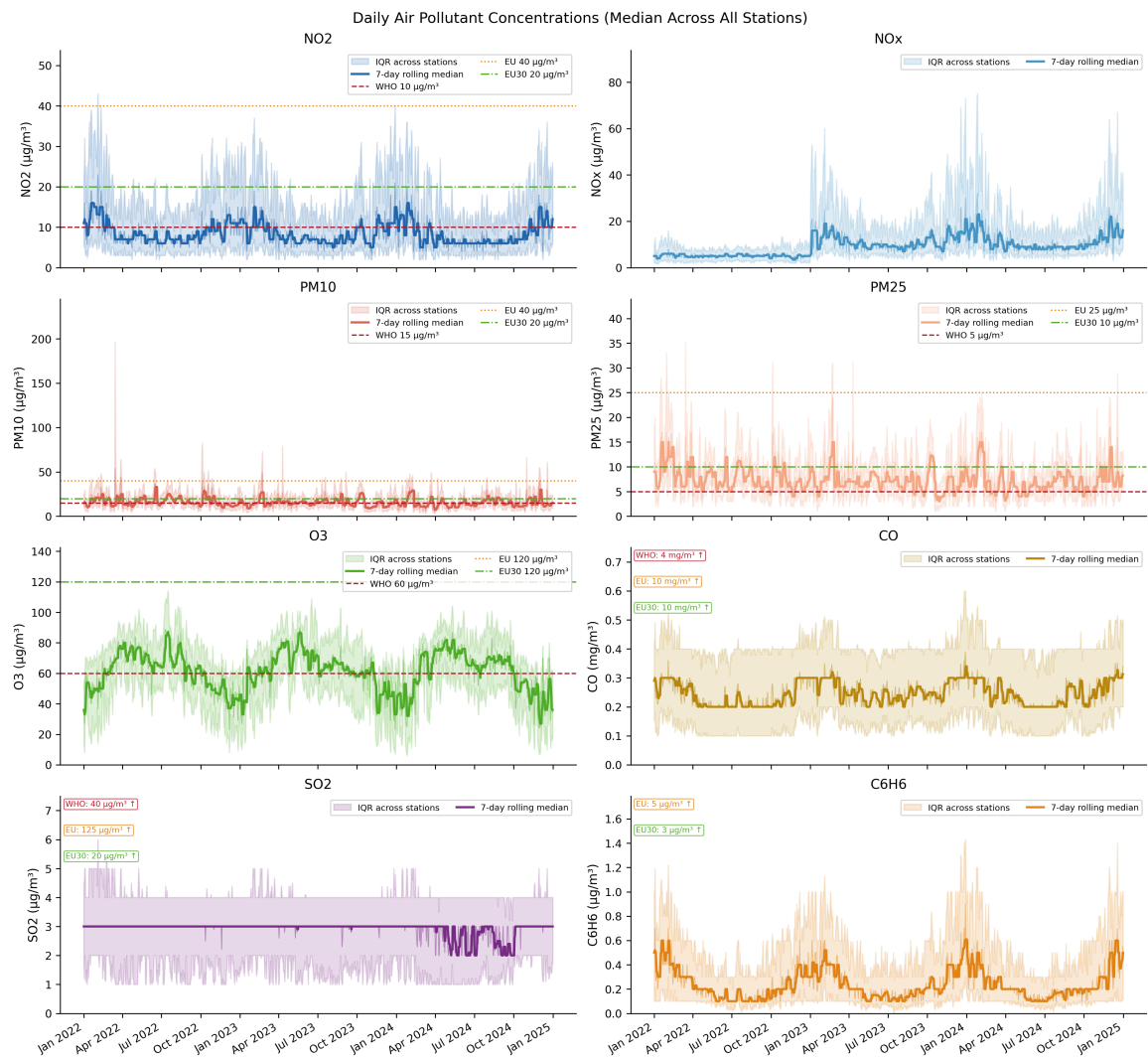


Figure 2. Daily median concentrations of the eight regulated air pollutants across the MITECO network over the 2022–2024 study period. For each pollutant, the solid line is the seven-day rolling median of the daily station-level medians, and the shaded band is the interquartile range across stations. Reference lines mark the WHO 2021 Air Quality Guideline (dashed), the current EU annual limit under Directive 2008/50/EC (dotted) and the revised EU 2030 limit under Directive (EU) 2024/2881 (dash-dot) where applicable. For pollutants whose regulatory limits are expressed on time bases other than the annual mean (CO, SO₂, C₆H₆), the corresponding thresholds are annotated as text rather than as horizontal lines. Concentrations are in µg/m³ except for CO, which is in mg/m³.

Aggregating the hourly observations by season exposes the diurnal cycle of each pollutant (Figure 3). NO₂ and NO_x both show a clear bimodal pattern with a morning peak around 07:00–08:00 and a stronger evening peak between 19:00 and 21:00, both compatible with the temporal structure of the urban traffic load and amplified in winter by the lower mixing-layer height and reduced photolytic loss. The winter cycle of NO₂ spans from a minimum around 10 µg/m³ in the early hours to an evening peak of approximately 26 µg/m³, while in summer the same cycle contracts to roughly 7–16 µg/m³ as boundary-layer mixing and photochemical sinks weaken the diurnal accumulation. CO and benzene mirror the NO₂ profile, with rush-hour peaks that are most clearly defined in winter, in line with their shared origin in combustion processes.

Ozone exhibits the inverse pattern, with a single broad afternoon maximum centred around 14:00 in every season, reaching about 87 µg/m³ in spring and summer when photochemical production is at its highest, and a pronounced morning minimum coinciding with the NO₂ peak. The morning minimum reflects both the titration of ambient ozone by freshly emitted NO and the dominance of dry deposition during the nocturnal stable layer.

The particulate pollutants behave seasonally rather than uniformly. PM_{10} has its most pronounced diurnal cycle in winter, with a bimodal pattern that mirrors the combustion-driven shape of NO_2 : a morning shoulder around 09:00 near $24 \mu\text{g}/\text{m}^3$ and a stronger evening peak close to 20:00 reaching $27 \mu\text{g}/\text{m}^3$, against an early-hours minimum of about $18 \mu\text{g}/\text{m}^3$. In the other three seasons the cycle is markedly flatter, with a single morning peak around 07:00–08:00 and an overall amplitude roughly half of that observed in winter. $\text{PM}_{2.5}$ follows the same seasonal contrast in a more extreme form: in winter it traces a clear bimodal cycle with an evening peak of approximately $14 \mu\text{g}/\text{m}^3$ against a minimum near $9.5 \mu\text{g}/\text{m}^3$, whereas in spring, summer and autumn it remains nearly flat throughout the day, with intra-day variation within $1\text{--}2 \mu\text{g}/\text{m}^3$. SO_2 shows no significant intra-day structure in any season, consistent with its dominant industrial point sources and with the fact that most stations report values close to the detection limit. The takeaway for users of the dataset is that the within-day variation of every pollutant other than SO_2 is non-negligible in at least one season, whereas the daily-aggregated exogenous features described in the following subsections cannot describe within-day variation by construction.

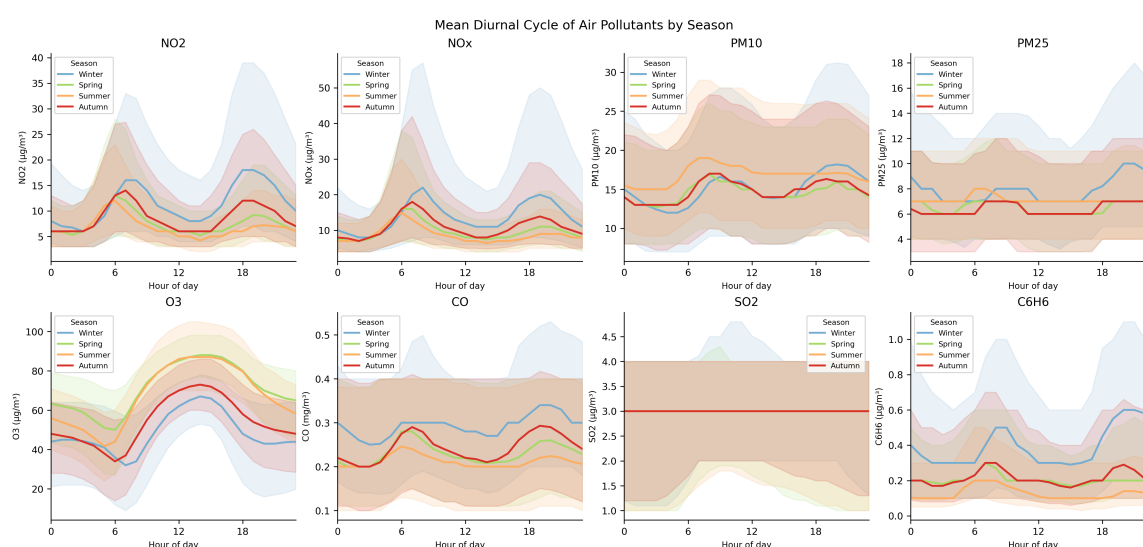


Figure 3. Mean diurnal cycle of the eight regulated pollutants by season, computed across the network for the years 2022–2024. Solid lines are seasonal means of the hourly station-level means; shaded bands span the 25th to 75th percentile of the cross-station distribution. CO is in mg/m^3 ; all other pollutants are in $\mu\text{g}/\text{m}^3$.

The same seasonal structure can be summarised compactly through the monthly distribution of each pollutant (Figure 4). NO_2 and NO_x peak in December and January, with monthly upper quartiles reaching $24 \mu\text{g}/\text{m}^3$ and $33 \mu\text{g}/\text{m}^3$ respectively, and trough in June through August at roughly half of those values. Ozone shows the opposite cycle, with the highest medians in April and May (close to $75 \mu\text{g}/\text{m}^3$) and a slightly lower summer plateau of 66 to $68 \mu\text{g}/\text{m}^3$ in June through August, declining sharply through autumn into a winter minimum; the upper whiskers occasionally approach the EU and EU 2030 threshold of $120 \mu\text{g}/\text{m}^3$ during the warm season. Benzene reproduces the NO_2 winter peak at much lower absolute concentrations (medians of $0.40 \mu\text{g}/\text{m}^3$ in December and January against $0.11 \mu\text{g}/\text{m}^3$ in June), consistent with their shared combustion origin. PM_{10} and $\text{PM}_{2.5}$ are comparatively flat across the calendar, with February the most variable month for both as judged by the interquartile spread, reflecting the combination of winter combustion peaks and the onset of Saharan dust transport. CO follows the NO_2 winter peak weakly. SO_2 shows no monthly structure, in line with its largely industrial sources.

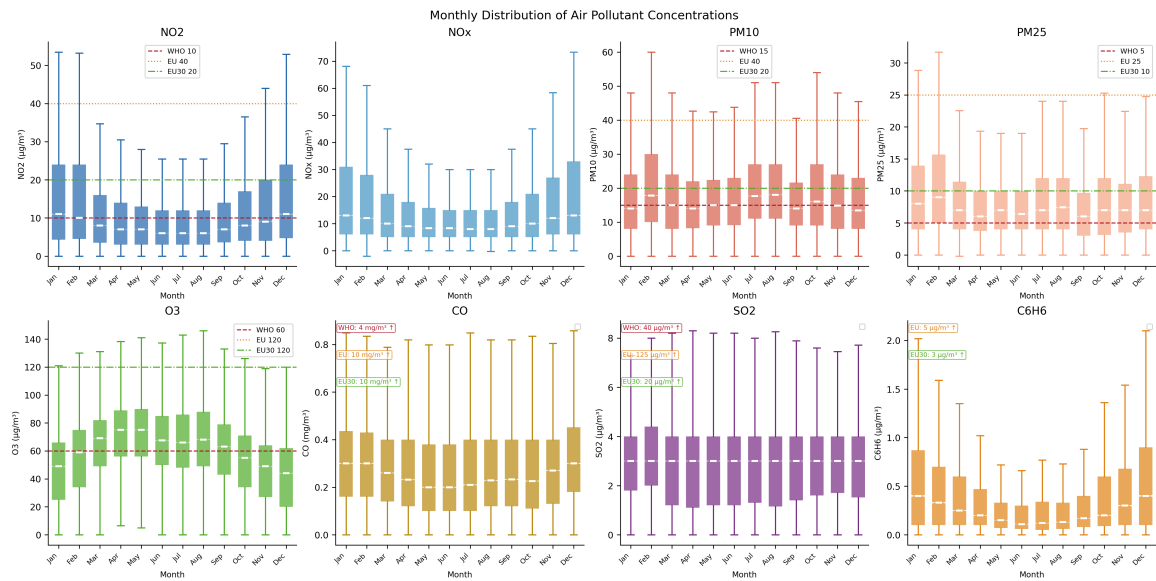


Figure 4. Monthly distribution of the eight regulated pollutants pooled across all station-hour records of the network over 2022–2024. Each box spans the interquartile range with the median marked; whiskers extend to the 1.5 IQR limits. Reference lines mark the WHO 2021 Air Quality Guideline (dashed), the current EU annual limit (dotted) and the revised EU 2030 limit (dash-dot) where applicable. CO is in mg/m^3 ; all other pollutants are in $\mu\text{g}/\text{m}^3$.

3.2.2. Meteorological Data: the AEMET Sensor Network

AEMET operates the official meteorological network of Spain and exposes its daily climatological summaries through the OpenData API [38]. The relevant product for this work is the daily climatological extract per station (*climatologías diarias*), which reports temperature (mean, minimum and maximum), precipitation, wind speed (mean and gust), wind direction, insolation, atmospheric pressure (maximum and minimum) and relative humidity (mean, minimum and maximum). The API enforces a maximum window of six months per request and a rate limit of approximately 50 requests per minute, both of which the acquisition module respects through paginated calls and throttling. The processed dataset contains 990,784 station-day records from 904 AEMET stations over the 1,096-day window, after deduplication of repeated identifiers. From this catalogue we retain the nine variables that constitute the meteorological block of the integrated dataset: mean, minimum and maximum daily temperature (*tmed*, *tmin*, *tmax*), daily precipitation (*prec*), mean and maximum wind speed (*velmedia*, *racha*), and mean, minimum and maximum relative humidity (*hrMedia*, *hrMin*, *hrMax*). Wind direction, solar radiation, atmospheric pressure and timing metadata are dropped at the download stage. The retained variables are uniformly well covered across the catalogue, with missingness of approximately 5% for all nine and a profile of mean values consistent with the climatic range of Spain. Table 2 reports per-variable summary statistics. The Spanish convention "Ip" (*indicios de precipitación*, trace amounts of rainfall) is converted to 0.0 at the acquisition stage. The integration of these daily records onto the hourly air quality time axis used by the merged release follows the lag-and-expand procedure described in Section 3.3.2: each daily summary is shifted forward by one day and replicated across the 24 hours of the corresponding day, in the hourly variant; the daily variant of the merged dataset uses the same daily values without temporal shift.

Table 2. Per-variable summary statistics for the nine retained AEMET variables. All variables share a denominator of approximately one million station-day records over the 1,096-day window; the exact *N* varies slightly between variables because not every station reports every channel for every day. The wind variables (*velmedia*, *racha*) are reported at a smaller subset of stations than the temperature, precipitation and humidity variables.

Variable	Unit	Missing	Mean	Std	Median	Min	Max
tmed	°C	4.8%	16.11	7.06	15.8	−15.0	40.4
tmin	°C	4.8%	10.60	6.86	10.4	−18.7	50.0
tmax	°C	4.8%	21.63	8.03	21.3	−50.0	46.8
prec	mm	4.7%	1.66	6.16	0.0	0.0	710.8
velmedia	m/s	5.1%	2.79	1.88	2.5	0.0	33.1
racha	m/s	5.3%	9.64	4.08	8.9	0.0	65.6
hrMedia	%	5.6%	64.25	18.38	64.0	1.0	100.0
hrMin	%	5.2%	46.36	19.48	43.0	0.0	100.0
hrMax	%	5.2%	86.45	13.96	91.0	1.0	111.0

The 111% maximum for *hrMax* and the −50°C minimum for *tmax* are present in the source records as published by AEMET, and the dataset preserves them as released.

3.2.3. Mobility Data: the MITMA Sensor-Derived Network

The MITMA mobility study is a continuously updated nationwide mobility dataset for Spain derived from mobile telephony records [32]. The study processes anonymised event records from approximately 13 million Orange Spain mobile lines through a hashing-based pseudonymisation function and aggregates the resulting trajectories into mobility indicators with daily temporal resolution. MITMA publishes results at multiple zonifications; we use the municipality-level aggregation, which covers 2,687 zones, each constructed from one or several INE census districts. The methodology combines passive network probe events with Call Detail Records, producing trip matrices that include an intra-day hourly breakdown, but the published aggregate files are daily studies that characterise mobility patterns for each calendar day as a whole.

Three indicator families per zone and per day are released by MITMA and incorporated into the integrated dataset. The *pernoctaciones* product counts the number of people who spent the night in a zone, distinguishing residents from visitors. The *personas* product counts people present in a zone segmented by age band and by number of trips made that day. The *viajes* product counts trips with origin or destination in the zone, segmented by trip purpose (home, work or study, frequent activity, infrequent activity), by household income level, by age band and by residency status. These products are published as monthly archives of daily files in compressed CSV format and were downloaded from the MITMA open data portal. After acquisition and pivoting, the processed mobility dataset contains 2,824,332 zone-day records for the *viajes* block and 2,781,949 zone-day records for each of the *personas* and *pernoctaciones* blocks over the 1,096 calendar days of the study period, with no missing values at the zone-day level. The raw trip matrices include an intra-day hourly breakdown (a period field with 24 *franja* values), but counts are aggregated to the day level so that the resulting indicators describe the daily context of each station rather than within-day traffic dynamics. Table 3 reports per-variable summary statistics for the three blocks. Mean values across the network are dominated by the largest urban zones; medians give a better sense of the typical zone-day, as the distributions are strongly right-skewed.

Table 3. Per-variable summary statistics for the 36 MITMA variables released across the three indicator blocks. The *pernoctaciones* and *personas* blocks are computed across 2,781,949 zone-day records; the *viajes* block is computed across 2,824,332 zone-day records. Missingness is 0% for every variable. The distributions are strongly right-skewed; medians are therefore a more representative summary of the typical zone-day than means. Values above 10⁶ are reported with the M suffix.

Variable	Unit	Mean	Std	Median	Max
Pernoctaciones (overnight presence)					
pernoc_total	persons	18,534	79,878	6,201	3.4M
pernoc_residentes	persons	6,553	8,776	4,625	128,276
pernoc_visitantes	persons	11,981	80,244	640	3.4M
pernoc_ratio_visitantes	ratio (0–1)	0.258	0.341	0.096	1.000
Personas (daytime presence)					
pers_edad_0_25	persons	4,478	18,279	1,458	763,209
pers_edad_25_45	persons	4,804	22,554	1,518	961,954
pers_edad_45_65	persons	5,563	23,049	1,908	980,847
pers_edad_65_100	persons	3,681	16,171	1,461	696,233
pers_viajes_0	persons	5,078	18,619	2,521	1.2M
pers_viajes_1	persons	819	3,482	252	371,659
pers_viajes_2	persons	3,803	16,367	1,309	759,272
pers_viajes_2plus	persons	8,827	42,286	2,425	2.0M
pers_ratio_moviles	ratio (0–1)	0.640	0.138	0.684	1.000
Viajes — aggregate trip counts and kilometres					
viajes_entrantes	trips	48,507	238,943	12,590	11.5M
viajes_salientes	trips	48,507	238,933	12,588	11.5M
viajes_internos	trips	27,728	194,811	2,358	9.7M
viajes_km_entrantes	km	463,566	1,971,094	188,015	163.4M
viajes_km_salientes	km	463,566	2,004,148	188,051	154.4M
<i>Inbound trips by destination activity</i>					
viajes_ent_act_casa	trips	18,345	83,060	5,152	3.9M
viajes_ent_act_frecuente	trips	17,606	88,784	3,971	5.0M
viajes_ent_act_no_frecuente	trips	7,230	43,030	1,855	2.9M
viajes_ent_act_trabajo_estudio	trips	5,325	27,225	1,371	1.6M
<i>Inbound trips by household income</i>					
viajes_ent_renta_alta	trips	9,308	111,662	444	6.1M
viajes_ent_renta_media	trips	30,918	130,580	8,002	5.3M
viajes_ent_renta_baja	trips	8,281	31,438	544	1.9M
<i>Inbound trips by traveller age band</i>					
viajes_ent_edad_0_25	trips	8,924	52,224	1,026	2.5M
viajes_ent_edad_25_45	trips	10,388	69,358	989	3.4M
viajes_ent_edad_45_65	trips	11,886	72,336	1,245	3.6M
viajes_ent_edad_65_100	trips	5,500	35,834	575	1.8M
viajes_ent_edad_na	trips	11,810	17,061	7,316	339,867
<i>Inbound trips by traveller residency</i>					
viajes_ent_residentes	trips	46,023	230,571	11,480	11.1M
viajes_ent_externos	trips	2,484	9,263	853	691,003
<i>Outbound trips by origin activity</i>					
viajes_sal_act_casa	trips	17,961	81,724	5,089	4.0M
viajes_sal_act_frecuente	trips	18,093	90,741	4,100	5.1M
viajes_sal_act_no_frecuente	trips	7,173	42,759	1,799	2.9M
viajes_sal_act_trabajo_estudio	trips	5,279	26,804	1,361	1.5M

The integration step assigns the mobility indicators of the INE 5-digit municipality code corresponding to each station, constructed by concatenating the province and municipality codes from the station inventory. An official cross-reference file (*relacion_ine_zonificacion_mitma*) published by MITMA is used to translate from internal MITMA zone identifiers to INE codes. As a consequence, all air quality stations located within the same municipality share identical mobility features. This is a known approximation: intra-municipal variability of mobility patterns is not captured, although the spatial heterogeneity of large urban municipalities should be small relative to the inter-municipal differences that the indicators are designed to express. Each daily mobility record is then aligned to the air quality time axis following the same convention used for the meteorological features: in the hourly variant the record is lagged by one day and expanded to the 24 hours of the corresponding day, while in the daily variant it is merged directly at the same calendar day without temporal shift. The two alignments are described in Section 3.3.2.

Figure 5 summarises the temporal structure of the mobility data at the national level. The aggregate weekly trip volume is remarkably stable around 1.4 to 1.5 billion trips per week throughout the three years of the study, with sharp drops to fewer than 600 million on the two weeks of late December that contain Christmas Day and New Year (covering the transitions from 2021 to 2022 and from 2024 to 2025), and on a single week at the end of October 2023 whose unusually low aggregate sits on the boundary of two MITMA monthly batches and most likely reflects a partial-data effect rather than a real mobility collapse. The day-of-week breakdown reveals a steady weekday baseline of approximately 215 million daily trips and a noticeable contraction on weekends (around 187 million on Saturdays and 166 million on Sundays, the latter representing a roughly 23% reduction relative to the weekday mean). The monthly profile is comparatively flat, with a modest dip in August reflecting the Spanish summer-vacation convention.

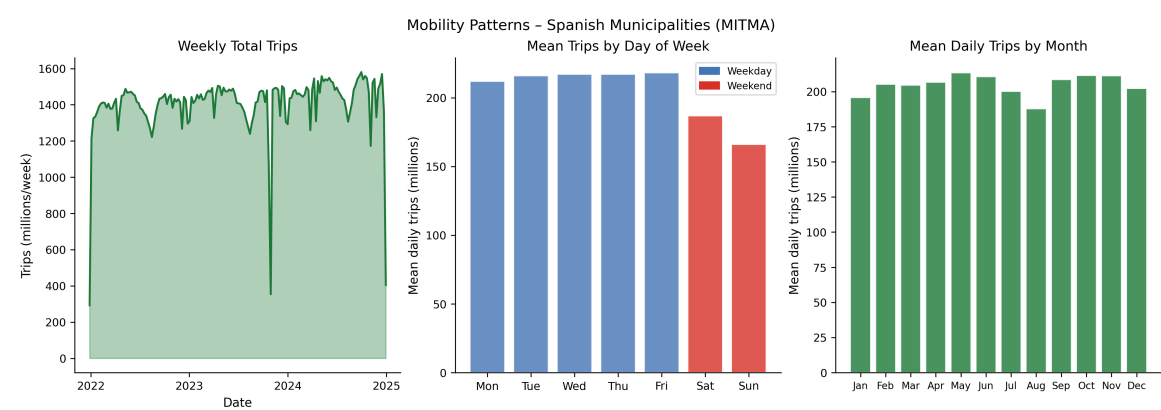


Figure 5. Aggregate mobility patterns over the 2022–2024 period, summed across all 2,687 zones of the municipal-level MITMA zonification. Left: total weekly trips. Centre: mean daily trip count by day of week, with weekdays (blue) and weekends (red) distinguished. Right: mean daily trip count by month of the year.

3.2.4. Public Holiday Calendar

National and regional public holidays are obtained from the open-source Python package `holidays` [39], which compiles official holiday calendars per country and per subnational region. For Spain, the package covers national holidays and the calendars of all 17 Autonomous Communities plus the autonomous cities of Ceuta and Melilla, identified by their two-letter ISO codes (for example AN for Andalusia, MD for Madrid, CT for Catalonia, PV for the Basque Country, MC for the Region of Murcia); the finest spatial granularity available is the Autonomous Community level, not the province. Across the three-year window we record 26 unique national holiday dates (8 in 2022, 9 in each of 2023 and 2024) and 83 unique regional dates that, when expanded across the Autonomous Communities to which they apply, yield 204 region-specific holiday entries. Most communities contribute around ten distinct regional dates over the period.

The integration step generates two binary indicators per station and per timestamp: *fiesta_nacional*, which equals one if the date is a national public holiday, and *fiesta_ccaa*, which equals one if the date is a holiday specific to the Autonomous Community in which the station is located. Both indicators are retained in the release. Although their content overlaps in part (every national holiday is also a holiday in every Autonomous Community), the regional indicator captures dates that the national one does not, and the marginal cost of carrying it as an additional binary column is negligible. In contrast to the meteorological and mobility features, the holiday indicators are assigned at time *t* without any temporal shift, since the calendar of public holidays is known in advance.

Figure 6 displays the calendar of national and regional holidays for the three years of the study period. National holidays (orange) cluster on a recurring set of civil and religious dates (New Year, Epiphany, Good Friday, Labour Day, the Assumption, Spain’s National Day, All Saints, Constitution Day, the Immaculate Conception and Christmas), while the regional component (yellow) is denser and irregularly distributed across the calendar, with each Autonomous Community contributing a small

set of dates that reflect local historical or religious observances. The figure confirms that the regional indicator is not redundant with the national one for a large share of the year, particularly during late winter, spring and early autumn.

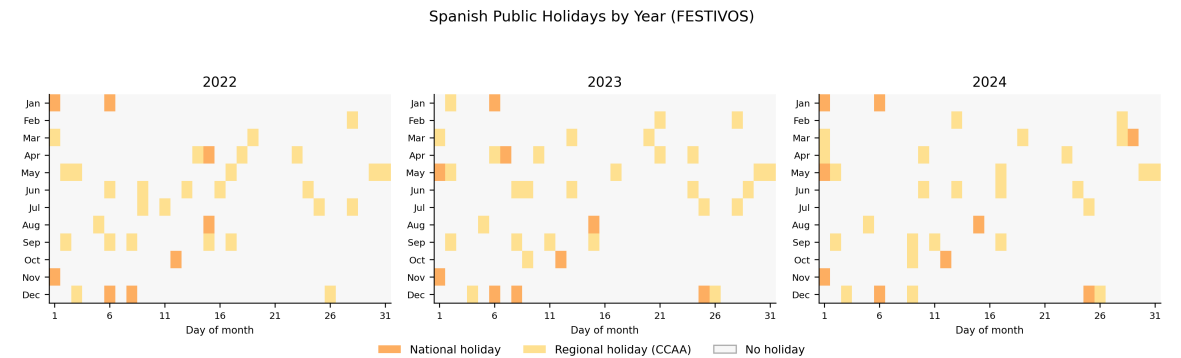


Figure 6. Calendar of Spanish public holidays for 2022, 2023 and 2024. Orange cells mark national holidays; yellow cells mark days that are public holidays in at least one Autonomous Community.

3.2.5. Source-Level Summary

The four sources described above produce datasets with substantially different shapes and granularities. Table 4 summarises the high-level properties of each, as observed at its native resolution after acquisition and per-source cleaning, before the integration step that brings everything onto the common hourly station grid described in Section 3.3.

Table 4. Source-level summary of the four processed datasets. Counts and resolutions refer to the data as published by each provider, prior to the integration step. The missing-data column for AIRQ is the range observed across the eight pollutants; for AEMET it is the range across the nine retained variables. The MITMA record count corresponds to the *viajes* block; the *personas* and *pernoctaciones* blocks cover 2,781,949 zone-day records each.

Dataset	Provider	Resolution	Spatial units	Records	Missing	Download
Air quality (AIRQ)	MITECO	hourly	588 stations	14,944,488	4.3%–33.9%	CSV
Meteorology (AEMET)	AEMET	daily	904 stations	990,784	4.7%–5.6%	Bulk API
Mobility (MITMA)	MITMA	daily	2,687 zones	2,824,332	0%	CSV
Public holidays	Gov. ES	daily	19 CCAAs	230	—	Library

The asymmetry in spatial and temporal granularity is the central characteristic that the integration pipeline must reconcile: hourly station data on one side, daily zone or station data on the other, and a calendar of dates on the third. Section 3.3 describes how the pipeline brings these levels into a single tidy table.

3.3. Integration Procedure

The integration procedure takes the per-source cleaned tables of Section 3.2 and produces one CSV file per air quality station, indexed by hourly timestamp and containing the full feature schema. Three operations compose the procedure: the spatial association of each MITECO station with the other sources (meteorology by distance-based pairing, mobility and holidays by administrative code lookup), the temporal alignment of the daily exogenous features to the hourly grid, and a per-station enrichment join that brings all the blocks together.

3.3.1. Distance-Based Pairing of Air Quality and Meteorological Stations

Among the three spatial associations performed by the pipeline, the link between MITECO and AEMET stations is the only one that does not reduce to a lookup by administrative code: the mobility and holiday linkages use the INE municipality code and the province-to-CCAA mapping respectively, and are described together with the corresponding source in Sections 3.2.3 and 3.2.4. The

meteorological linkage, in contrast, must reconcile two independent networks that share no identifier. Each MITECO station is paired with exactly one AEMET station rather than with a weighted average of several neighbours; the one-to-one assignment keeps the feature schema fixed across the network and avoids the sparse missing-value patterns that an averaging scheme would introduce when different MITECO stations sample different sets of AEMET sources. Both station lists are pre-filtered to include only stations with a data file on disk, preventing assignments to AEMET stations with no usable record.

The pairing follows a three-tier rule based on Haversine distance. For each MITECO station, the closest AEMET station is identified; if it lies within a 50-kilometre threshold, the match is accepted (priority 1, label *distancia*). If no AEMET station lies within 50 km, the search is restricted to stations in the same province (matched by accent-stripped name) and the closest one is selected (priority 2, label *provincia*). If no AEMET station exists in the same province, the closest one nationwide is assigned regardless of distance (priority 3, label *fallback_absoluto*). The 50-kilometre threshold is a compromise between proximity and coverage: lower thresholds leave stations in sparsely instrumented interior provinces unmatched, while higher thresholds allow pairs that we judged too distant to share a similar mesoscale meteorological state. In practice, the AEMET network is dense enough that every MITECO station in the release is matched under the first tier; the empirical distribution of pair distances is reported in Section 5 (Figure 8) and has a median of 4.8 km, a mean of 6.3 km and a maximum of 30.1 km, so the two fallback tiers are not exercised in the present release although they remain part of the pipeline for robustness against future MITECO additions in sparsely instrumented areas. The result is published alongside the dataset as a lookup table that records, for each MITECO station, the assigned AEMET station, the Haversine distance between the two, and the tier label of the assignment. Figure 7 shows the geographic distribution of the matched pairs.

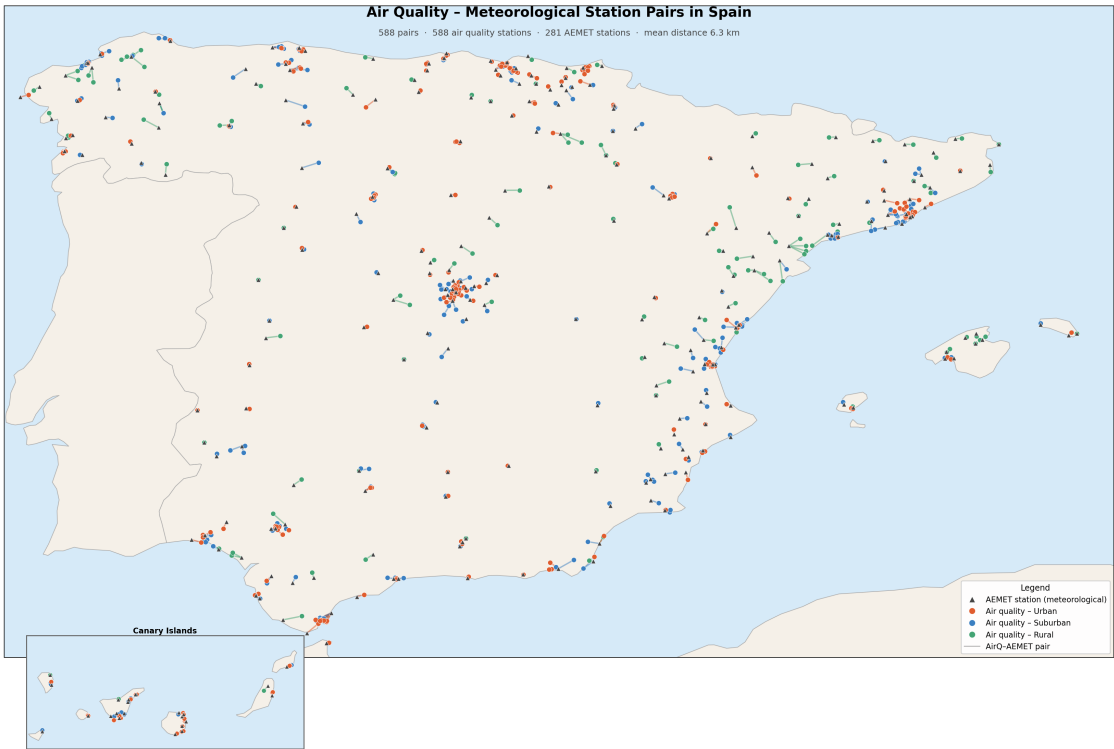


Figure 7. Geographic distribution of the 588 MITECO air quality stations, classified by area type (urban, suburban, rural). The lines connect each air quality station to the AEMET meteorological station assigned by the three-tier spatial association procedure described in Section 3.3.1.

3.3.2. Temporal Alignment of Daily Exogenous Features

Air quality is reported hourly while meteorology, mobility and holidays are reported daily. The integration brings the three daily sources onto the air quality time axis through a procedure that depends on the temporal resolution of the released variant.

In the hourly variant, the daily meteorological and mobility records are first shifted forward by one day (lag d-1) and then expanded to the 24 hours of the corresponding day by repetition. The lag reflects the operational forecasting constraint that, at the moment of issuing a forecast for time t , the most recent complete daily summary of weather and mobility available is the one that closed at the end of the previous day; the shift therefore prevents future information from leaking into the hourly record at time t . The combination of lag and expansion preserves the operational validity of every record and produces a uniform hourly schema, at the cost of the redundant repetition of each daily value across the 24 hours of its day.

In the daily variant, the air quality series is first aggregated to the daily grid (per-pollutant daily mean, minimum and maximum, with the 75% data-capture rule introduced in Section 3.3), and the daily meteorological and mobility records are then merged at the same calendar day as the air quality aggregates without temporal shift. The operational rationale for the lag does not apply at this resolution: once a calendar day has closed, its complete weather and mobility summary is part of the same closed-day snapshot as the air quality aggregates, so applying a d-1 shift would discard one day of valid information without operational justification. The daily variant therefore aligns every feature block on the same closed-day timestamp.

Holiday flags are assigned at time t without lag in both variants, since the public holiday calendar is known in advance and is therefore available to any forecaster issuing a prediction at any time during the day.

3.3.3. Per-Station Enrichment Join and Feature Schema

For each MITECO station, the pipeline assembles the integrated table as follows. The hourly air quality series is loaded and reindexed to a continuous hourly `DatetimeIndex` over the study period, with gaps filled with NaN to preserve the time axis. The AEMET counterpart of the station, looked up in the spatial pairing table, supplies the daily meteorological record, which is shifted by d-1 and expanded to hourly resolution as described above. The municipality-level MITMA mobility record (pernoctaciones, personas and viajes blocks) supplies the daily mobility indicators, which undergo the same lag-and-expand operation; if no MITMA file exists for the municipality (a few very small municipalities are not covered by MITMA), the corresponding columns are filled with NaN. The holiday flags are computed for each timestamp from the calendar of national and Autonomous-Community holidays, using the province-to-CCAA mapping derived from the station inventory. Finally, the station's area-type label (*tipo_area*) is appended as a constant categorical column. The first calendar day of the series is discarded from the output because the d-1 lag of the mobility and meteorological features would require a record from the day before the series starts, which is NaN by construction.

The resulting per-station table contains 56 feature columns indexed by hourly timestamp, grouped as summarised in Table 5. The same schema is produced for every station, regardless of which subset of pollutants the station reports; pollutants not measured at a given station appear as columns of NaN, which keeps the schema uniform across files and allows downstream consumers to scan stations by column without conditional logic.

Table 5. Feature schema of the integrated per-station table. The total of 56 columns reflects the 8 air quality variables, 9 meteorological variables, 4 overnight-stay indicators, 9 population indicators, 23 trip-related indicators, 2 holiday flags and 1 area-type categorical column. Daily-resolved blocks (AEMET, MITMA) are lagged by one day (d-1) and expanded to hourly resolution; the air quality block is at native hourly resolution; the holiday flags are assigned without lag.

Source / block	Cols	Variables
Air quality (MITECO)	8	NO ₂ , NO _x , O ₃ , SO ₂ , CO, PM ₁₀ , PM _{2.5} , C ₆ H ₆ . Hourly concentrations in µg/m ³ (CO in mg/m ³).
Meteorology (AEMET)	9	tmed, tmin, tmax, prec, velmedia, racha, hrMedia, hrMin, hrMax. Daily values, lagged d-1, expanded to hourly.
Mobility, pernoctaciones (MITMA)	4	Total overnight stays, residents, visitors, visitor ratio. Daily, lagged d-1, expanded.
Mobility, personas (MITMA)	9	Population counts by age band (0–25, 25–45, 45–65, 65–100), by number of trips per day (0, 1, 2, 2+) and mobile-population ratio. Daily, lagged d-1, expanded.
Mobility, viajes (MITMA)	23	Inbound and outbound trip counts and kilometre aggregates, segmented by trip purpose, household income, traveller age band (including unknown) and residency status; plus internal trips. Daily, lagged d-1, expanded.
Holiday flags	2	Binary indicators for national and Autonomous-Community holidays, assigned at time <i>t</i> without lag.
Station metadata	1	Area-type categorical label (<i>tipo_area</i>): <i>urbana de tráfico</i> , <i>urbana de fondo</i> , <i>suburbana</i> , <i>rural de fondo</i> , etc. Constant per station.

The resulting dataset comprises approximately 14 million hourly records distributed across the 588 stations of the network, released as a set of per-station CSV files. The full released artefact, including all per-station files, the spatial pairing lookup table and the source-level metadata files, is described in Section 4.

4. Data Records

The dataset described in this paper is released as an open record on the Zenodo repository, under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence [17]. The persistent identifier of the record is 10.5281/zenodo.20196221, and the record can be reached at <https://doi.org/10.5281/zenodo.20196221>. Future revisions of the dataset are deposited as new versions of the same Zenodo record; the cited concept DOI always resolves to the latest published version. The deposit contains all per-station files of the integrated dataset, the corresponding cleaned source files prior to integration, the spatial and metadata lookup tables, the dataset-wide statistics used in this paper, and a README that summarises the contents and serves as the primary entry point for new users.

4.1. Repository Structure

The released artefact follows the directory structure shown below. The two top-level data directories (*merged* and *segregated*) correspond to two complementary forms of the dataset: the *merged* directory contains the integrated per-station tables described in Section 3.3, ready to be consumed directly by downstream applications; the *segregated* directory contains the cleaned per-source tables prior to integration, so that users wishing to apply different temporal alignment or spatial association choices can build their own integration from the same starting material.

```
multisource_aq_dataset_spain_2022_2024/
|-- README.md
|-- merged/
|   |-- hourly/           588 CSV files, hourly resolution
|   '-- daily/           588 CSV files, daily resolution
|-- segregated/
|   |-- airq_hourly/      588 CSV files
```

```
| |-- airq_daily/                588 CSV files
| |-- aemet/                    904 CSV files
| |-- personas/                 ~2,569 CSV files
| |-- pernoctaciones/           ~2,569 CSV files
| |-- viajes/                   ~2,687 CSV files
| '-- festivos_spain.csv
|-- lookups/
| |-- airq_stations.csv
| |-- aemet_stations.csv
| '-- aemet_station_pairs.csv
'-- stats/
    '-- dataset_statistics.csv
```

All data files are plain CSV with a comma as field separator, a period as decimal separator, UTF-8 encoding and a header row. Missing values are encoded as empty fields. The first column of every time series file is the temporal index; the exact name of this column varies across the segregated source files because each follows the convention of its original provider, while the merged files use a uniform `datetime` index. The naming convention of all station-level files is `{COD_LOCAL}.csv` for the MITECO-anchored files and `{idema}.csv` or `{ine_muni_code}.csv` for the AEMET- and MITMA-anchored files respectively.

4.2. Integrated Per-Station Files

The integrated tables in `merged/hourly/` are the primary deliverable of this work. Each of the 588 files contains the full hourly time series for one MITECO air quality station over the 2022–2024 study period, joined with the lagged daily exogenous features described in Section 3.3.3. The 56 feature columns plus the `datetime` index follow the schema summarised in Table 5. The column names are stable across files, so a downstream consumer can iterate over the network without conditional logic. Pollutants not measured at a given station appear as columns of NaN; the categorical column `tipo_area` carries a station-specific constant such as *urbana de tráfico* or *rural de fondo*.

The directory `merged/daily/` contains the daily-aggregated variant of the integrated dataset. It is produced from the same pipeline but with the hourly air quality block collapsed to daily mean, minimum and maximum statistics prior to integration, as described in Section 3. The schema is identical to the hourly variant except that the 8 hourly pollutant columns are replaced by 24 daily columns (one mean, one minimum and one maximum per pollutant), bringing the total to 72 feature columns plus the `datetime` index at daily resolution. The variant is provided as a convenience for users for whom daily resolution is sufficient and a smaller file footprint is preferable; for any application that requires sub-daily resolution, the hourly variant is the authoritative product.

4.3. Segregated Per-Source Files

The directory `segregated/` contains the cleaned per-source tables produced by stage 2 of the pipeline (per-source acquisition and cleaning), before any spatial pairing or temporal alignment is applied. These files reproduce the form in which each source can be obtained from its provider after the preprocessing operations described in Section 3.2. They are released alongside the merged files for users who wish to apply their own integration choices, for example combining the air quality data with mobility at a different aggregation level, or aligning the daily blocks without the d-1 lag convention used here.

Air quality files are released in two resolutions. `segregated/airq_hourly/{COD_LOCAL}.csv` contains the schema `datetime`, `NO2`, `NOx`, `PM10`, `O3`, `CO`, `SO2`, `C6H6`, `PM25`. `segregated/airq_daily/{COD_LOCAL}.csv` contains `datetime` followed by `{pollutant}_mean`, `{pollutant}_min` and `{pollutant}_max` for each of the eight pollutants in the same order, for a total of 24 numeric columns; the daily aggregation follows the 75% hourly-coverage rule described in Section 3.

Meteorological files are released in segregated/aemet/, with one CSV file per AEMET station at daily resolution and the schema fecha, tmed, prec, tmin, tmax, velmedia, racha, hrMedia, hrMax, hrMin.

Mobility files are stored in three directories, with one CSV file per INE 5-digit municipality code in each. All files are at daily resolution and follow the products defined in the MITMA methodology document [32]. The schemas are: segregated/personas/{ine_muni}.csv with date, pers_edad_0_25, pers_edad_25_45, pers_edad_45_65, pers_edad_65_100, pers_viajes_0, pers_viajes_1, pers_viajes_2, pers_viajes_2plus, pers_ratio_moviles; segregated/pernoctaciones/{ine_muni}.csv with date, pernoc_total, pernoc_visitantes, pernoc_residentes, pernoc_ratio_visitantes; and segregated/viajes/{ine_muni}.csv with date followed by the 23 trip-related variables documented in Table 3. The MITMA municipality-level zonification covers 2,687 zones [32]: all 2,687 zones are present in the *viajes* block, while the *personas* and *pernoctaciones* blocks cover 2,569 zones each, with the 118-zone gap explained by MITMA's privacy-driven suppression of indicators for zones where the sample size would compromise the anonymity of the underlying mobile telephony records. These privacy-driven exclusions affect predominantly very small rural municipalities. The *viajes* files released here are obtained from the original MITMA origin–destination matrices, in which each row describes trips between a pair of zones, by aggregating to the destination zone for inbound trips and to the origin zone for outbound trips; the intra-day hourly breakdown present in the source matrices (a *franja* field with 24 values) is collapsed to the daily level, so that each row of the released file represents the totals for a single municipality on a single day.

The *public holiday calendar* is released as a single file segregated/festivos_spain.csv, covering the study period with the schema fecha, año, nombre, ambito, ccaa_codigo, ccaa_nombre, codauto, es_nacional. The ambito column distinguishes national from regional holidays, and the ccaa_codigo column carries the two-letter ISO code of the Autonomous Community when applicable. The integration step described in Section 3.3.3 reduces this calendar to the two binary indicators fiesta_nacional and fiesta_ccaa that appear in the merged files.

4.4. Supporting Files and Source Code

The lookups/ directory holds three reference tables used by the pipeline and useful for any downstream geographic operation. airq_stations.csv contains the consolidated inventory of the 588 MITECO air quality stations, with COD_LOCAL, name, geographic coordinates, altitude, area-type classification, dominant emission source, and the INE province and municipality codes used by the pipeline to derive the mobility and holiday linkages described in Section 3.3. aemet_stations.csv contains the consolidated inventory of the 904 AEMET meteorological stations, with the AEMET station identifier (idema), name, geographic coordinates, altitude and province. aemet_station_pairs.csv contains the output of the three-tier spatial pairing procedure described in Section 3.3.1, with one row per MITECO station; columns include the MITECO COD_LOCAL, the assigned AEMET idema, the Haversine distance between the two in kilometres, and the tier label of the assignment (distancia, provincia, or fallback_absoluto).

The file stats/dataset_statistics.csv contains the per-variable summary statistics used to populate Tables 1–3 of this paper. The file is in long format, with one row per (dataset, variable, metric) triple, and the supported metrics include n_total, n_missing, pct_missing, mean, std, min, q25, median, q75 and max. The file is provided so that users can reproduce the per-variable summaries reported in this paper and verify any custom statistics computed on the released files against an external reference.

The Python source code that implements the full pipeline (per-source acquisition, cleaning, spatial pairing, temporal alignment and per-station integration) is released as an open repository on GitHub at <repository_url>. The repository tag matching the release version of the dataset is also archived on Zenodo and cross-referenced from the data record. The code is provided to allow users to regenerate

the dataset, extend it in time as new monthly batches of source data become available, or adapt it to other countries with comparable open data infrastructure.

5. Technical Validation

This section documents the technical checks applied to verify that the integrated dataset behaves as expected given the characteristics of the four source systems. The validation is organised in three threads: the spatial quality of the MITECO–AEMET station pairing, the internal consistency of each source over time and across stations, and the coherence of the cross-source signal.

5.1. Spatial Pairing Quality

The three-tier pairing rule described in Section 3.3.1 assigns to each MITECO station the nearest available AEMET meteorological station. A key question for the usability of the meteorological block is how close the assigned pairs actually are in practice. Figure 8 shows the empirical distribution of Haversine distances across the 588 pairs.

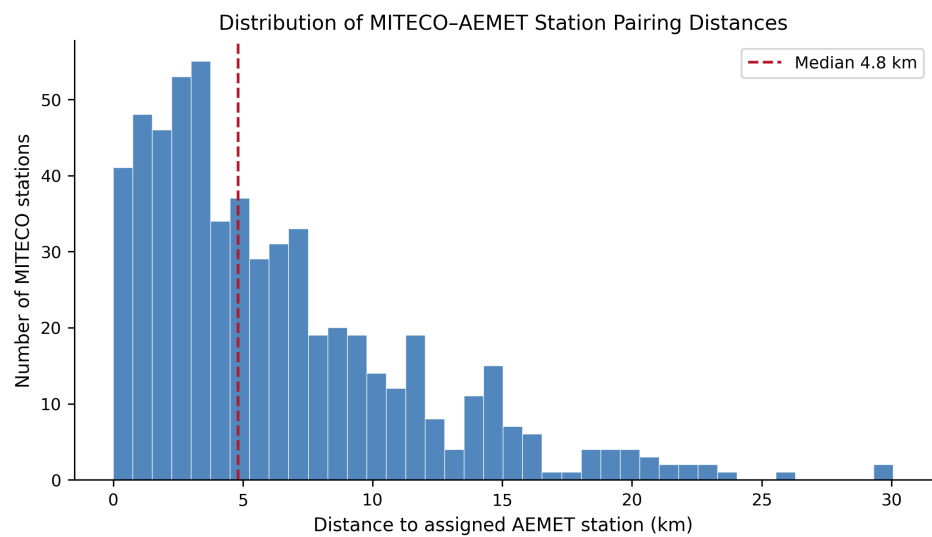


Figure 8. Distribution of Haversine distances (km) between paired MITECO air quality stations and their assigned AEMET meteorological stations, across the 588 pairs in the network. The dashed line marks the median distance of 4.8 km.

All 588 MITECO stations were matched under the first-priority tier (nearest within 50 km), so the two fallback tiers were not required for any station in the network. The median pairing distance is 4.8 km and the mean is 6.3 km, reflecting the density of the Spanish AEMET network relative to the MITECO monitoring sites. The 90th percentile lies at 14.0 km and the 99th percentile at 22.2 km; the single most distant pair in the dataset is 30.1 km apart. No pair exceeds the 50 km threshold. These distances are consistent with the mesoscale meteorological homogeneity expected within typical urban and suburban airshed zones, where temperature, humidity and wind conditions a few kilometres away can reasonably approximate those at the monitoring site. Users applying the dataset to analyses that require very precise local meteorology should consult the pairing table (released in `lookups/aemet_station_pairs.csv`) and may wish to exclude pairs beyond a stricter distance threshold for their specific use case.

5.2. Internal Consistency of Source Data

Figure 9 shows the monthly distribution of four AEMET variables (mean, min and max temperature, daily precipitation, mean wind speed and mean relative humidity) across the network over the 2022–2024 study period. The patterns are consistent with the climatic characteristics of the Iberian Peninsula: median temperatures follow a smooth sinusoidal annual cycle from approxi-

mately 9°C in January to 25°C in July–August; mean daily precipitation is highest in October (around 3.1 mm/day with a long upper tail of comparable magnitude across stations) and lowest in July (around 0.3 mm/day), with a secondary autumn-onset rise in September and a sustained wet period from late autumn through early spring; humidity is in clear anti-phase with temperature, peaking in late autumn and winter at around 75% and falling to around 52% in mid-summer; wind speed varies more weakly across the year, with a slight maximum in March around 3.0 m/s. No anomalous distributional patterns that would indicate a systematic acquisition or processing error are visible in any of the variables.

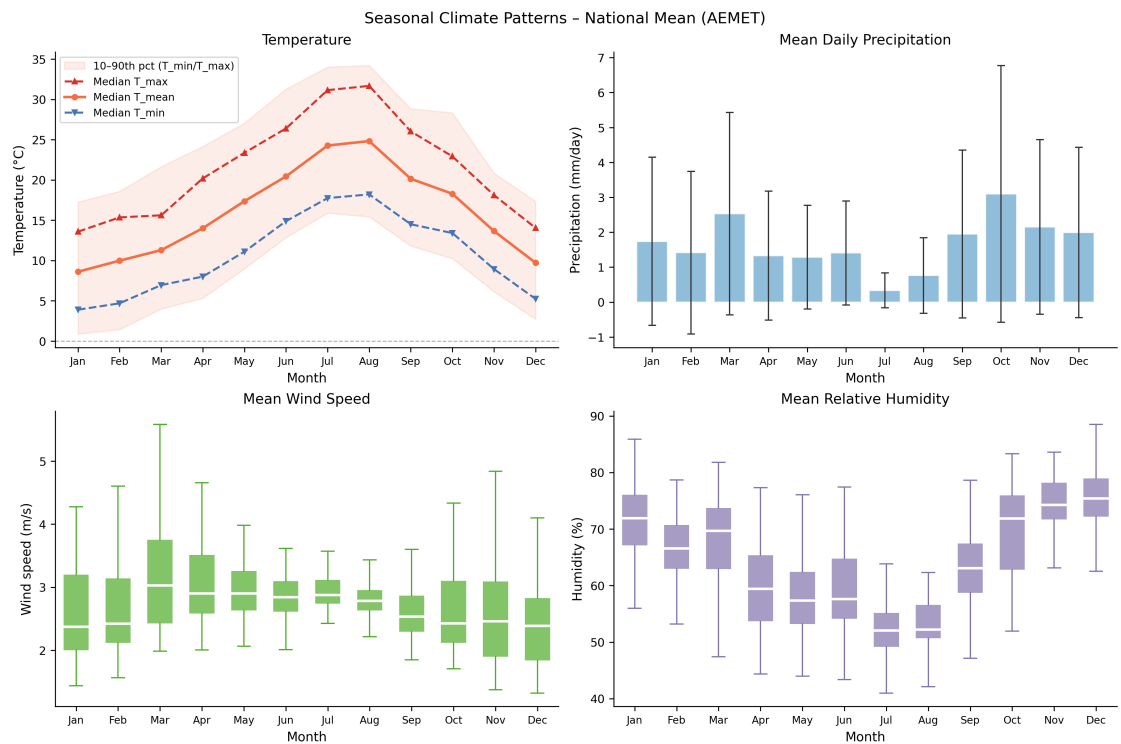


Figure 9. Monthly distributions of four AEMET variables across the network: median daily mean, min and max temperature with the 10th–90th percentile band (top-left); mean daily precipitation with error bars showing the spread across stations (top-right); mean wind speed as monthly box plots (bottom-left); and mean relative humidity as monthly box plots (bottom-right). Each panel pools all station-day records of the corresponding month over 2022–2024.

Figure 10 characterises the per-station mean concentration across the network for the eight pollutants. All eight distributions are right-skewed, reflecting the typology mix of the network: rural background and remote suburban locations populate the lower end, while urban-traffic stations in large cities populate the upper tail. The per-station median is 11.4 $\mu\text{g}/\text{m}^3$ for NO_2 and 16.1 $\mu\text{g}/\text{m}^3$ for NO_x , with NO_x exhibiting a noticeably longer upper tail. Particulate matter has medians of 18.8 $\mu\text{g}/\text{m}^3$ for PM_{10} and 9.2 $\mu\text{g}/\text{m}^3$ for $\text{PM}_{2.5}$, both close to their respective WHO guidelines and broadly compliant with the new EU 2030 annual limits at the median station. Ozone has the most symmetric station-level distribution (median 58.2 $\mu\text{g}/\text{m}^3$, very close to the WHO peak-season threshold of 60 $\mu\text{g}/\text{m}^3$), consistent with its secondary photochemical origin, which homogenises concentrations across station typologies more strongly than primary-emission pollutants do. CO, SO_2 and benzene have low absolute concentrations (medians of 0.3 mg/m^3 , 3.1 $\mu\text{g}/\text{m}^3$ and 0.3 $\mu\text{g}/\text{m}^3$ respectively) with right-skewed tails attributable to a handful of industrially exposed stations.

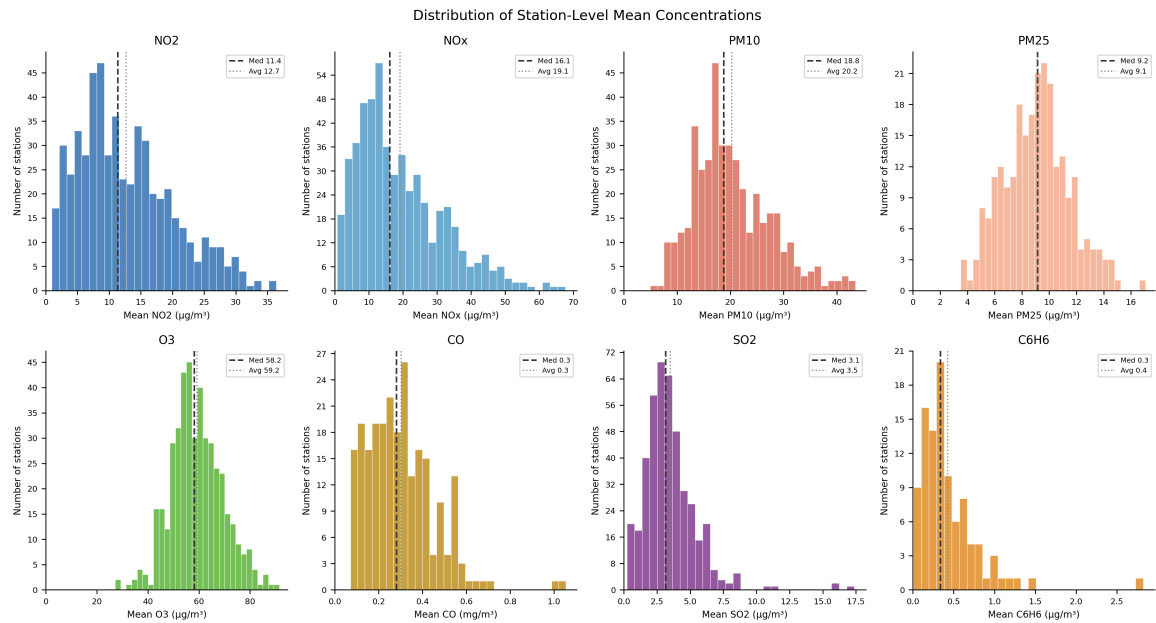


Figure 10. Distribution of station-level mean concentrations across the 588 MITECO stations, for each of the eight regulated pollutants. The dashed and dotted vertical lines mark the across-station median and mean, respectively. CO is in mg/m^3 ; all other pollutants are in $\mu\text{g}/\text{m}^3$.

The companion distribution of station-level missing-data rates (Figure 11) makes the coverage tiering of the network visually explicit. NO_2 is the only pollutant that combines a low typical missingness (median around 3% among the stations that report it) with a small mass at 100% (only 36 of the 588 stations do not measure it). O_3 and SO_2 show similarly low typical missingness on the lower cluster (medians of 2.7% and 2.6% respectively), but the mass at 100% is substantial in both cases: about 24% of the network does not measure O_3 and about 27% does not measure SO_2 . NO_x has a higher typical missingness, with the bulk of the distribution clustered around 35% rather than near zero; this reflects the partial-period coverage already visible in Figure 2, where the network-wide NO_x series only stabilises towards the end of 2022, indicating that many stations began reporting this pollutant part-way through the study period. PM_{10} shows a central peak around 3% missingness on the lower cluster, with an upper quartile near 7%, together with a large secondary mass at 100% (177 stations, about 30% of the network, do not measure it). The remaining three pollutants ($\text{PM}_{2.5}$, CO and C_6H_6) display the most pronounced bimodality, with a station-level median of 100% and a heavy concentration of stations at the upper end: 58%, 59% and 83% of the network respectively report no data at all for these pollutants. Users intending to work with $\text{PM}_{2.5}$, CO or benzene should therefore expect to encounter a substantial number of stations with no measurement available for the pollutant of interest, regardless of how the data are subsequently processed.

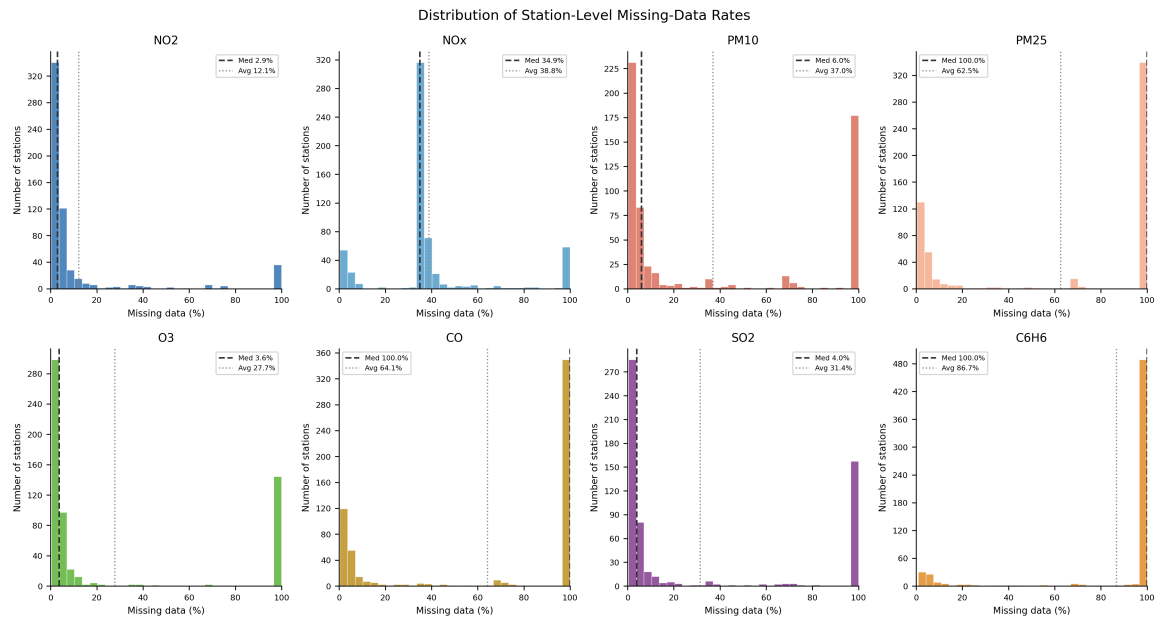


Figure 11. Distribution of station-level missing-data rates across the 588 MITECO stations, for each of the eight regulated pollutants. The peak at 100% in each panel groups the stations that do not measure the pollutant at all; the lower cluster groups the stations that do measure the pollutant, with their typical share of NaN records over the study period. The dashed and dotted vertical lines mark the across-station median and mean.

5.3. Cross-Source Consistency

Figure 12 shows the Spearman rank correlation matrix between the eight air quality pollutants and the full set of predictors of the integrated dataset: the air quality block itself (for the pollutant-to-pollutant relationships), the AEMET meteorological block, the MITMA mobility block at the municipal level, and the holiday indicators. Correlations are computed at daily resolution across all station-day records of the network. The rank-based measure is preferred over the linear (Pearson) coefficient because several variables (precipitation, trip counts, PM₁₀) are heavily right-skewed and contain large episodic outliers that would otherwise dominate the linear coefficient.

The sign and approximate magnitude of every meaningful correlation are consistent with established physical and behavioural mechanisms. Within the air quality block, the strongest associations are the structural coupling between particulate matter fractions (PM₁₀–PM_{2.5}, $\rho = 0.91$) and between the combustion-related pollutants, which share emission sources and meteorological controls (NO₂–CO at $\rho = 0.73$, NO₂–C₆H₆ at $\rho = 0.78$, CO–C₆H₆ at $\rho = 0.84$). Ozone is negatively correlated with all the combustion-related pollutants (ρ between -0.28 and -0.59), as expected from the titration of freshly emitted NO and from the inverse seasonality of photochemical production and combustion emissions; its association with the particulate fractions is much weaker (close to zero for PM₁₀ and around -0.12 for PM_{2.5}), reflecting the fact that mineral dust and resuspended particles do not participate in the photochemical balance that drives the NO–O₃ relationship.

The cross-correlations with the meteorological block reproduce the physically expected patterns. NO₂ is strongly negatively correlated with wind speed ($\rho = -0.66$) and wind gust ($\rho = -0.58$), reflecting the ventilation effect that disperses locally emitted pollutants. Benzene shows the same dispersion signature ($\rho = -0.47$ with wind), reinforced by a strong negative correlation with temperature (ρ between -0.69 and -0.74) that captures the winter accumulation of combustion pollutants under cold-stagnation regimes. Ozone shows the inverse meteorological signature: positively correlated with temperature and wind, and negatively correlated with humidity ($\rho = -0.54$). PM₁₀ correlates positively with daily maximum temperature ($\rho = 0.39$) and negatively with precipitation ($\rho = -0.41$), consistent with soil resuspension during dry warm conditions and with wet deposition of coarse particles.

The mobility block contributes weaker correlations whose sign depends on which family of mobility indicators is considered. Trip counts (*entrantes*, *salientes* and *internos*) correlate positively with the locally emitted combustion pollutants NO₂ and NO_x (ρ between 0.28 and 0.41), in line with their direct origin in road traffic. Trip kilometres (km_in and km_out), by contrast, correlate negatively with CO and C₆H₆ (ρ between -0.34 and -0.49): municipalities with longer typical incoming or outgoing trips are predominantly the less densely populated ones, which have systematically lower baseline concentrations of these combustion pollutants regardless of their daily mobility load. The same spatial gradient also produces a positive correlation of similar magnitude between ozone and trip kilometres ($\rho \approx 0.32$), since rural municipalities accumulate less NO and therefore titrate less of the available ozone. These correlations are further diluted by the daily-mean aggregation and by the fact that all stations within a municipality share the same mobility record. The national-holiday indicator yields only marginal coefficients in the matrix-level view, although the day-type analysis in Figure 14 (which compares conditional distributions rather than rank correlations) reveals a much clearer structure once the working-day baseline is isolated.

The absence of anomalous correlations in the matrix, in particular the lack of any sign inversion against expectation for the meteorological and pollutant-to-pollutant blocks, provides confidence that the spatial pairing procedure and the daily-to-hourly expansion have not introduced systematic distortions in the link between the four sources.

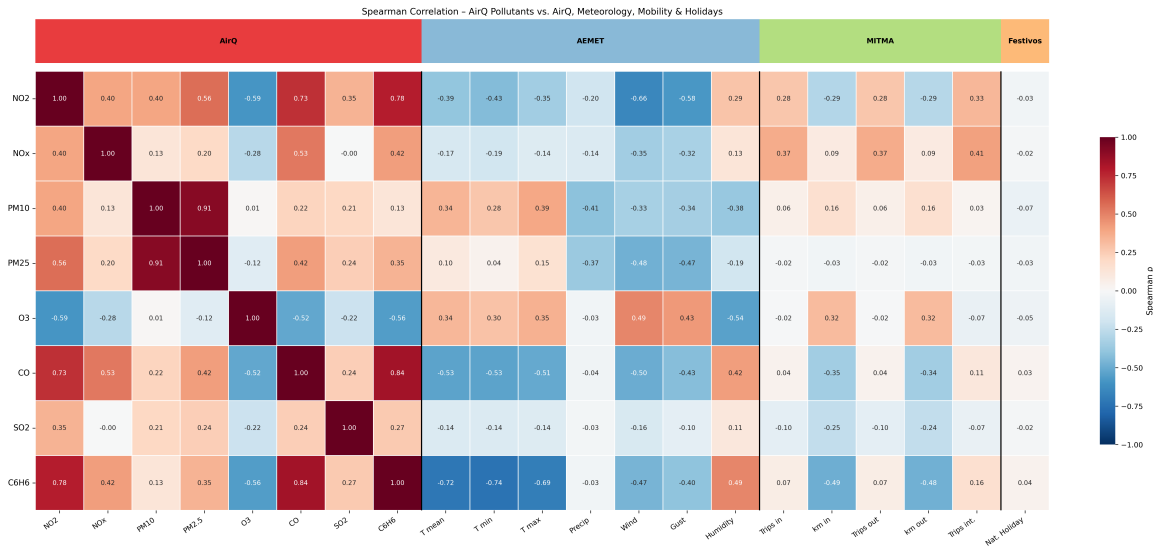


Figure 12. Spearman rank correlation matrix between the eight air quality pollutants (rows) and the integrated set of predictors (columns), grouped by source block: AirQ (pollutant-to-pollutant relationships), AEMET (meteorology), MITMA (mobility), and Festivos (national-holiday indicator). Correlations are computed at daily resolution across all station-day records of the network over 2022–2024.

For readability of the strongest signals, Figure 13 represents the same matrix as a bipartite network in which the eight pollutants are connected to the meteorological, mobility and holiday predictors whenever the absolute Spearman coefficient exceeds 0.3. Edge width is proportional to $|\rho|$ and edge colour distinguishes positive (red) from negative (blue) correlations. The network view exposes three groups of nodes whose behaviour is qualitatively different. The combustion pollutants (NO₂, NO_x, CO, C₆H₆) share a dense bundle of negative edges to wind speed, wind gust and temperature, in line with their dispersion-driven nature, but split apart in their mobility coupling: NO₂ and NO_x link positively to trip counts (internal, inbound and outbound), whereas CO and C₆H₆ link negatively to trip kilometres, for the reasons discussed in the heatmap analysis above. Ozone shows the meteorological mirror image (positive edges to wind speed, wind gust and temperature, a strong negative edge to humidity, and negative edges to the three combustion pollutants that pass the 0.3 threshold), reflecting its photochemical secondary origin; it also exhibits positive edges to incoming and outgoing trip kilometres of similar magnitude to the CO and C₆H₆ ones but of opposite sign, consistent with the

urban-to-rural spatial gradient already identified. PM₁₀, PM_{2.5} and SO₂ are more selectively integrated into the network: PM₁₀ connects to precipitation, temperature, humidity and wind speed at modest strength; PM_{2.5} is tightly coupled to PM₁₀ ($\rho = 0.91$) but also shows substantial edges to NO₂, CO, wind speed and wind gust; and SO₂ has no edge above the 0.3 threshold to any external predictor, consistent with its dominant industrial-source profile. The national-holiday node remains disconnected at the 0.3 threshold, in line with the matrix observation that its rank-correlation signature is too diffuse to surface as a network edge despite the visible day-type effect documented below.

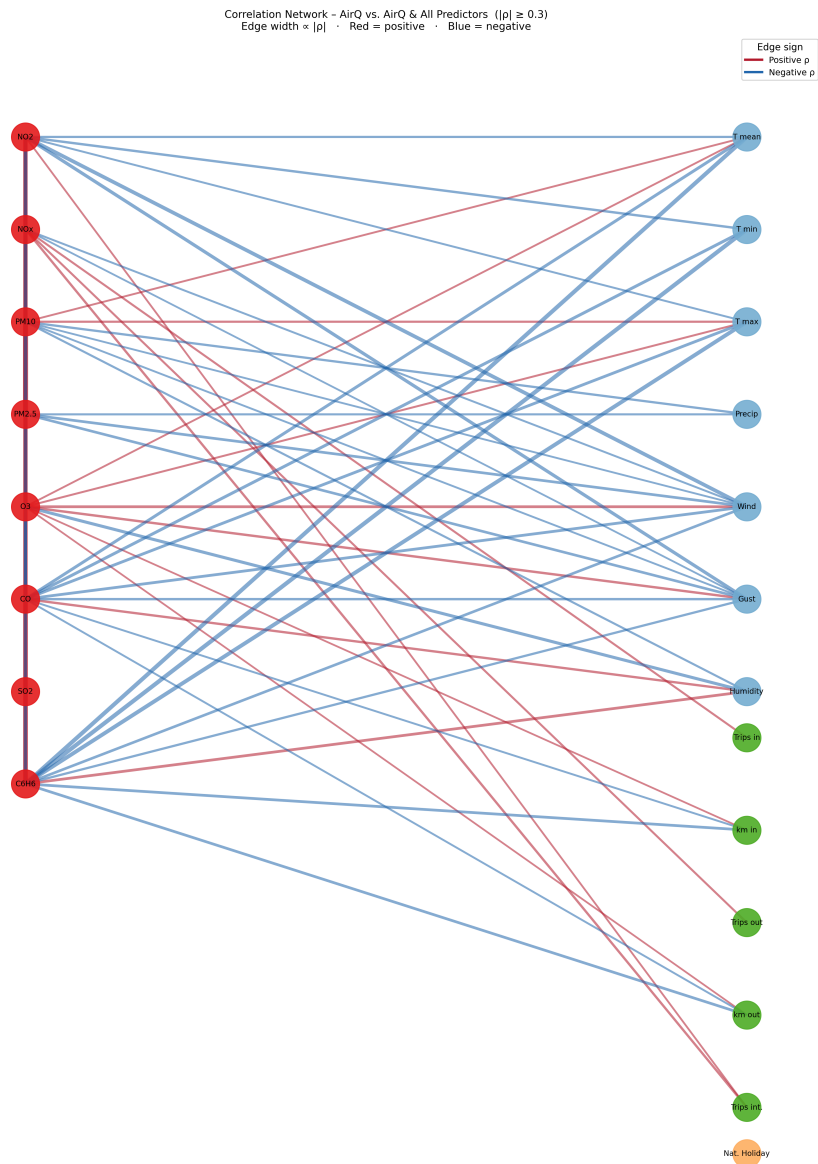


Figure 13. Bipartite correlation network between the eight air quality pollutants (left) and the AEMET, MITMA and holiday predictors (right). An edge is drawn whenever the absolute Spearman rank correlation exceeds 0.3; edge width is proportional to $|p|$ and edge colour indicates the sign of the correlation (red: positive, blue: negative).

Figure 14 shows the distribution of the eight pollutants stratified by day type: working weekday, weekend, regional public holiday (a holiday in at least one Autonomous Community but not nation-wide) and national public holiday. The percentage above each non-weekday box reports the change in median concentration relative to the weekday baseline.

The combustion-related pollutants display the clearest signal. For NO₂, the weekend reduction is the largest (−23.5%), followed by the national-holiday reduction (−17.6%) and the regional-holiday reduction (−11.8%); NO_x follows the same ordering (−22.7%, −18.2% and −9.1%). The signal is consistent with the dominant role of commuting and commercial road traffic in NO_x emissions, which weakens visibly on non-working days. Particulate matter responds more weakly to day type (PM₁₀: −12.5% on weekends, −18.8% on national holidays; PM_{2.5}: −0.9% on weekends, −7.1% on national holidays), reflecting its broader and less commuter-driven source mix. Carbon monoxide shows a small weekend reduction (−5.7%) and benzene a small regional-holiday reduction (−5.0%), consistent with their lower absolute concentrations and noisier station coverage. SO₂ is essentially flat across day types, in line with its industrial-source dominance and detection-limit-bounded distribution.

Ozone moves in the opposite direction (+3.1% on weekends, +8.0% on regional holidays, −9.2% on national holidays). The positive weekend and regional-holiday signal is the well-documented weekend ozone effect: the reduction in fresh NO emissions weakens the titration of ambient ozone and allows higher peak concentrations to persist. The negative national-holiday signal most likely reflects the seasonal concentration of national holidays around late autumn and Christmas, when low temperatures and weak photochemistry depress baseline ozone irrespective of the working-day calendar.

The signal is interpretable on two levels. First, the magnitude and direction of every day-type effect are consistent with the dominant role of road traffic in primary urban pollution and with the chemistry of secondary ozone formation. Second, and more importantly for the dataset itself, the distinct reduction observed on regional holidays (days that are public holidays in the Autonomous Community of the station but not nationwide) demonstrates that the province-to-CCAA mapping described in Section 3.3 correctly assigns the relevant regional calendar to each station: had the regional flag been miscoded or assigned to the wrong stations, the regional-holiday box would be statistically indistinguishable from the weekday box. This subsection therefore validates both the calendar-source integration and the spatial association of holidays to stations.

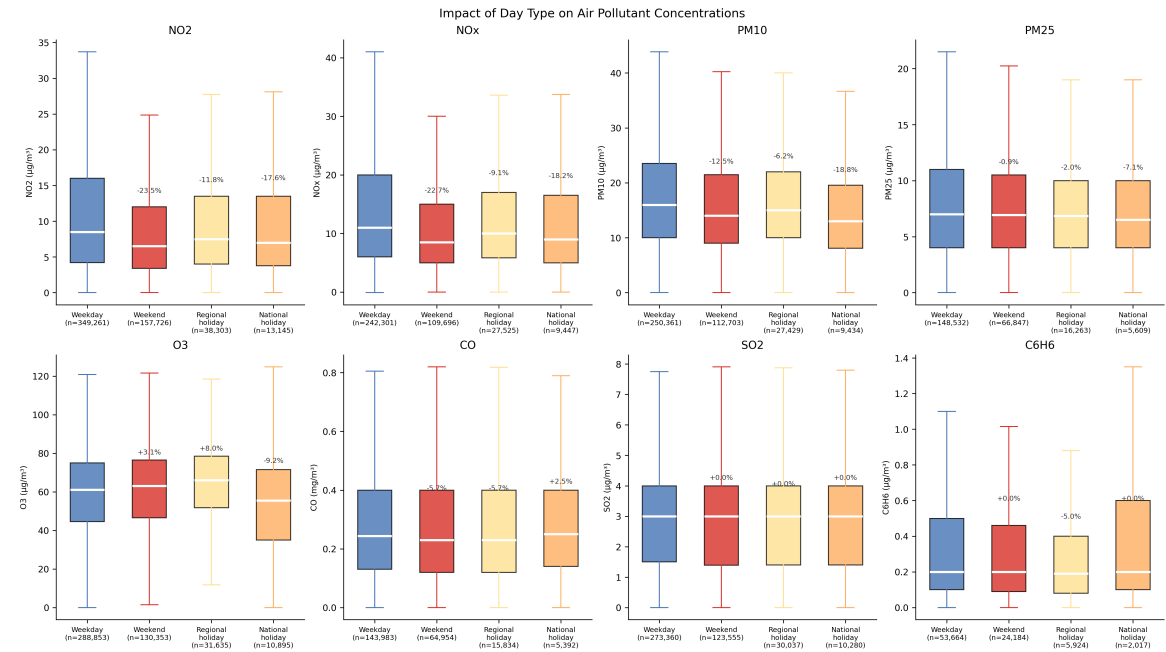


Figure 14. Distribution of hourly concentrations of the eight regulated pollutants stratified by day type: working weekday, weekend, regional public holiday (a holiday in at least one Autonomous Community but not nationwide), and national public holiday. The percentage above each non-weekday box reports the change in median concentration relative to the weekday baseline. Box plots summarise all station-hour records of each type over 2022–2024. CO is in mg/m³; all other pollutants are in µg/m³.

A final cross-source consistency check is provided by the most prominent episodic feature of the air quality record. The PM₁₀ spike on 15 March 2022 visible in Figure 2 reaches a network-wide daily median of 54 $\mu\text{g}/\text{m}^3$, more than three times the typical baseline of around 15 $\mu\text{g}/\text{m}^3$, and coincides with a well-documented Saharan dust intrusion over the Iberian Peninsula that generated air quality alerts across multiple Autonomous Communities. PM_{2.5} shows a smaller but coincident peak on the same dates (reaching 16 $\mu\text{g}/\text{m}^3$ against a baseline near 9 $\mu\text{g}/\text{m}^3$), consistent with the fine-particle fraction of the transported dust load. The appearance of the event as the single most pronounced network-wide particulate exceedance in the three-year record, affecting stations across all area types and both the Mediterranean and Atlantic sides of the Peninsula, is consistent with the origin and spatial extent of the event as reported at the time. None of the gaseous pollutants in the dataset (NO₂, NO_x, O₃, SO₂, CO, C₆H₆) shows a corresponding spike on the same dates, as expected given that mineral dust transport has no direct relationship to local combustion or to photochemical production; this confirms that the eight pollutant series in the dataset are not cross-contaminated or otherwise spuriously coupled, and that the integrated time axis preserves the event in the pollutants where it is physically expected.

6. Usage Notes

This section provides practical guidance for users of the dataset. It covers loading and typical applications, recommended preprocessing for common downstream tasks, known characteristics that may affect specific analyses, and a brief account of the preliminary modelling experiments that informed the design of the release.

6.1. Loading and Suggested Applications

The dataset is distributed as plain CSV files, which can be loaded with standard data analysis libraries. A single station of the merged hourly variant can be read with pandas through a one-line call such as `pd.read_csv(f"merged/hourly/{cod_local}.csv", parse_dates=["datetime"], index_col="datetime")`, where `cod_local` is the MITECO station code listed in `lookups/airq_stations.csv`. Users wishing to assemble a panel across the network can iterate over the file list of the corresponding directory.

The combination of sources released here supports several research lines that would require substantial duplicated acquisition work to reproduce from the four providers independently. The most direct application is operational air quality forecasting at the station level under a multi-source feature set. Network-level comparative analyses are also natural, for example, the systematic comparison of pollutant levels across area types (urban traffic, urban background, suburban, rural background) or across Autonomous Communities, and the climatological characterisation of pollutant concentrations under different meteorological regimes. The mobility block is particularly suited to the study of activity-related drivers of urban air pollution at the municipality scale, including weekend and holiday effects, and the joint analysis of mobility patterns with environmental health outcomes. The dataset also provides a substrate for cross-validation of dispersion or chemistry-transport models against a dense network of in-situ measurements over three full years.

6.2. Recommended Filtering and Preprocessing

Not all stations are equally usable for every purpose. For forecasting applications that train per-station models, we recommend filtering out stations with insufficient coverage on the target pollutant before splitting the data into training and evaluation periods. A coverage threshold of approximately 4,000 valid hours over the three-year period (around 50% of the total) is a reasonable default for NO₂ and PM₁₀; stricter thresholds may be appropriate for shorter training windows or for the less widely reported pollutants.

Users for whom hourly resolution is not required may prefer the daily-aggregated variant in `merged/daily/`, which has approximately one twenty-fourth of the file size of the hourly variant. It replaces the eight hourly pollutant columns with their daily mean, minimum and maximum aggregates

(24 columns in total) and keeps every other block of the feature schema unchanged. The daily variant offers two further advantages for users whose downstream model already operates at daily resolution. First, it does not contain the stepwise pattern that the daily-to-hourly expansion of the exogenous features introduces in the hourly tables, since every feature block in the daily variant is sampled at the same temporal resolution. Second, it is built without the one-day lag of the meteorological and mobility blocks, so those records refer to the same closed calendar day as the air quality aggregates and no information is discarded for the sake of operational realism.

The one-day lag (d-1) applied to the meteorological and mobility blocks in the hourly variant reflects the operational forecasting convention that exogenous information from day d only becomes available the following day. Users with access to nowcast data, or who wish to study contemporaneous rather than predictive relationships between mobility and air quality at hourly resolution, can recover the unlagged daily values from the segregated source files in `segregated/` and re-join them without the temporal shift. The daily variant of the integrated dataset already provides the unlagged join for users whose downstream analysis does not require hourly granularity.

A small number of NO_2 , NO_x and $\text{PM}_{2.5}$ records carry slightly negative values, typically within 1 to $2 \mu\text{g}/\text{m}^3$ of zero, which correspond to analyser calibration drift near the detection limit. These values are released as-is to preserve the raw physical readings. Users computing distributional statistics or feeding the data to scale-sensitive models may wish to apply a positivity filter, either clipping the affected values to zero or discarding them; we did not impose either choice at the dataset level.

6.3. Caveats and Known Limitations

Several characteristics of the released dataset are worth flagging explicitly. The first concerns the spatial granularity of mobility. The MITMA indicators are released at the municipality level, so all air quality stations within the same INE municipality share identical mobility features. In large urban municipalities such as Madrid or Barcelona, this approximation hides the intra-municipal heterogeneity of mobility patterns and should be borne in mind when interpreting station-level mobility-air-quality relationships within a single large city.

The second concerns the coverage profile of the air quality pollutants. NO_2 and NO_x are reported at almost every station, but the remaining six pollutants are reported by progressively smaller subsets of the network, with C_6H_6 available at only a small fraction of stations (Table 1). Users should verify coverage on the specific pollutant of interest before scoping a study.

The third concerns the meaning of the `viajes_ent_edad_na` column. This is not an additional age band but a fallback category that aggregates trips for which the traveller's age could not be inferred from the underlying mobile telephony records. Its distribution is qualitatively different from the four genuine age-band columns (a much smaller variance and no extreme outliers), and it should not be treated as comparable.

The fourth concerns the granularity of the holiday calendar. The `fiesta_ccaa` flag is assigned at the Autonomous Community level. Local festivities specific to a province, municipality or smaller administrative unit (which can have measurable effects on local activity and on air quality) are not captured by the released binary indicators.

Finally, the AEMET source records contain a small number of physically impossible values, such as relative humidity above 100% or daily maximum temperature of -50°C , which are present in the published data as released by AEMET. These have been preserved unchanged in the dataset. Users running anomaly-detection or scale-sensitive analyses may wish to filter them out before processing.

6.4. Preliminary Modelling Experiments

In the course of developing this dataset we ran preliminary forecasting experiments using a per-station bidirectional LSTM trained on the air quality block alone, compared against the same model trained on the full enriched feature schema (air quality, meteorology, mobility and holidays). Across the target pollutants and forecast horizons we tested, the enriched feature set did not yield a robust improvement over the air-quality-only baseline. Three factors plausibly explain this outcome

and are worth noting for future users. These results are not evidence against the value of the enriched feature set. They suggest that capturing the predictive content of the multi-source signal may require a model architecture that exploits the network structure across stations, for instance a graph neural network over the spatially associated stations, or a temporal resolution different from the hourly one we tested. The daily-aggregated variant of the dataset, in which the lag and the daily-to-hourly expansion of the exogenous features are both absent, is plausibly more directly amenable to this kind of experiment than the hourly variant. We release the dataset in the hope that this kind of experiment becomes more practical now that the integration work is no longer a prerequisite for it.

7. Conclusions

This paper has described the construction and release of an open dataset that integrates four sensor-derived data sources covering the whole of Spain over the period 2022 to 2024: hourly air quality observations from the 588 stations of the MITECO national network, daily meteorological summaries from the AEMET network, daily mobility indicators derived from anonymised mobile telephony records published by MITMA at the municipality level, and a calendar of national and Autonomous-Community public holidays. The four sources are reconciled by a pipeline that links each MITECO station to its meteorological counterpart through a distance-based pairing, attaches the corresponding mobility records through the INE municipality code, and aligns the daily exogenous features to the air quality time axis through a two-variant temporal-alignment scheme: a lag-and-expand procedure for the hourly variant of the dataset, and a same-calendar-day join for the daily variant. The resulting dataset comprises approximately 14 million hourly records distributed across 588 station files with a uniform feature schema of 56 columns, accompanied by a daily-resolution variant with 72 columns, and is published on Zenodo under a CC BY 4.0 licence [17]. The pipeline source code is released on GitHub at https://github.com/juanbonastre/airq_enriched_dataset so that the dataset can be regenerated, extended in time, or adapted to other countries with comparable open data infrastructure.

The contribution is intentionally a data contribution. Most published work on air quality forecasting that combines heterogeneous sources rebuilds the integration from scratch within a single project's pipeline, and very few of those integrated tables are ever released. By documenting and publishing the integration, we hope to lower the cost of entry for researchers wishing to work on multi-source air quality problems at nationwide scale, whether in forecasting, environmental epidemiology, mobility-aware air quality modelling, or the validation of dispersion and chemistry-transport models against in-situ measurements.

The technical validation reported in Section 5 indicates that the integrated dataset behaves as expected. The 588 MITECO stations are paired with AEMET stations at a median distance of 4.8 km and a maximum of 30.1 km, well within the 50 km threshold that defines the first tier of the pairing rule. The seasonal climatology of the meteorological block matches the climate of the Iberian Peninsula. The Spearman rank correlations between the eight air quality pollutants and the meteorological and mobility blocks agree in sign and approximate magnitude with established physical and behavioural mechanisms: strong negative correlations between the combustion-related species (NO_2 , NO_x , CO , C_6H_6) and wind speed, gust and temperature, a positive correlation between PM_{10} and dry warm conditions together with a negative correlation with precipitation, the mirror-image meteorological signature of ozone (positive with temperature and wind, negative with humidity), and the structural couplings within the particulate fractions and within the combustion-pollutant cluster. The mobility block splits in two parts that are individually coherent: trip counts correlate positively with the locally emitted NO_2 and NO_x , in direct correspondence with their traffic origin, while trip kilometres correlate negatively with CO and C_6H_6 and positively with ozone, a pattern that reflects the urban-to-rural spatial gradient rather than a direct mobility-emission link. The day-type effects are clearly distinguishable between weekdays, weekends and regional and national holidays for the main traffic-driven pollutants (NO_2 , NO_x and PM_{10}), with ozone showing the well-documented weekend effect of

opposite sign, confirming that the calendar source has been correctly assigned to the right stations through the province-to-CCAA mapping. The Saharan dust intrusion of mid-March 2022 appears in the network-wide PM₁₀ and PM_{2.5} series exactly where it should and does not contaminate any of the gaseous-pollutant series.

The dataset has limitations that users should keep in mind, and which are documented in Section 6. The most consequential are the municipality-level granularity of the mobility block, which hides intra-municipal heterogeneity in large urban areas, and the coverage tiering of the air quality pollutants, which leaves several regulated species available at only a fraction of the network. Future versions of the dataset deposited on Zenodo will extend the temporal coverage as new monthly batches of MITMA data are published, and may incorporate the finer-grained MITMA zonifications (census districts and large urban areas) for users who require sub-municipal mobility information. Adapting the pipeline to other countries that publish comparable air quality, meteorological and mobile-network mobility data through open interfaces is a natural extension that the released source code is intended to support.

Author Contributions: Conceptualization, J.M.G.; methodology, J.M.G. and J.B.E.; software, J.B.E.; validation, A.B.C. and J.M.G.; formal analysis, J.M.G. and J.B.E.; investigation, J.B.E., A.B.C. and J.M.G.; writing—original draft preparation, J.B.E., A.B.C. and J.M.G.; writing—review and editing, J.B.E., A.B.C. and J.M.G.; visualization, J.B.E.; supervision, J.M.G.; project administration, J.M.G.; funding acquisition, J.M.G.

Data Availability Statement: The integrated multi-source dataset described in this paper is released as an open record on the Zenodo repository under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence [17]. The persistent DOI of the record is 10.5281/zenodo.20196221, accessible at <https://doi.org/10.5281/zenodo.20196221>, and resolves to the latest published version. The Python source code of the integration pipeline is released as an open repository on GitHub at https://github.com/juanbonastre/airq_enriched_dataset; the tag corresponding to the release version of the dataset is mirrored on Zenodo and cross-referenced from the data record.

Acknowledgments: This work was supported by the SATRAI project (CIGE/2024/199), funded by the Conselleria de Educació, Cultura, Universidades y Empleo (Generalitat Valenciana); the AM-DS project (INREED/2024/7), funded by the Conselleria de Innovació, Indústria, Comerç y Turisme (Generalitat Valenciana) through the GVANEXT programme and the European Union through NextGenerationEU/PRTR; and the R&D project CPP2023-010458, funded by MICIU/AEI/10.13039/501100011033 and FEDER, UE.

References

1. World Health Organization. WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, 2021. Available at: <https://www.who.int/publications/i/item/9789240034228>.
2. European Environment Agency. Europe's air quality status 2024. Briefing 06/2024, European Environment Agency, 2024. <https://doi.org/10.2800/5970>.
3. Copernicus Atmosphere Monitoring Service. CAMS National Collaboration Programme Spain, 2024. Available at: <https://atmosphere.copernicus.eu/spain>.
4. Morales-García, J.; Ramos-Sorroche, E.; Balderas-Díaz, S.; Guerrero-Contreras, G.; Muñoz, A.; Santa, J.; Terroso-Sáenz, F. Reducing pollution health impact with air quality prediction assisted by mobility data. *IEEE Journal of Biomedical and Health Informatics* **2024**.
5. Hameed, S.; Islam, A.; Ahmad, K.; Belhaouari, S.B.; Qadir, J.; Al-Fuqaha, A. Deep learning based multimodal urban air quality prediction and traffic analytics. *Scientific Reports* **2023**, *13*, 22181.
6. Ibrahim, M.R.; Lyons, T. Transforming CCTV cameras into NO₂ sensors at city scale for adaptive policymaking. *Scientific Reports* **2025**, *15*, 3640.
7. Baldasano, J.M. COVID-19 lockdown effects on air quality by NO₂ in the cities of Barcelona and Madrid (Spain). *Science of the Total Environment* **2020**, *741*, 140353.
8. Donzelli, G.; Cioni, L.; Cancellieri, M.; Llopis-Morales, A.; Morales-Suárez-Varela, M. Relations between air quality and COVID-19 lockdown measures in Valencia, Spain. *International Journal of Environmental Research and Public Health* **2021**, *18*, 2296.
9. Cordero, J.M.; Narros, A.; Borge, R. True reduction in the air pollution levels in the community of madrid during the COVID-19 lockdown. *Frontiers in Sustainable Cities* **2022**, *4*, 869000.

10. Gangoiti, G.; de Blas, M.; Gómez, M.C.; Rodríguez-García, A.; Torre-Pascual, E.; García-Ruiz, E.; Saez de Camara, E.; Zuazo, I.; García, J.A.; Valdenebro, V. Impact of the COVID-19 lockdown in a European regional monitoring network (Spain): are we free from pollution episodes? *International Journal of Environmental Research and Public Health* **2021**, *18*, 11042.
11. Huang, C.J.; Kuo, P.H. A deep CNN-LSTM model for particulate matter (PM_{2.5}) forecasting in smart cities. *Sensors* **2018**, *18*, 2220.
12. González-Enrique, J.; Ruiz-Aguilar, J.J.; Moscoso-López, J.A.; Urda, D.; Deka, L.; Turias, I.J. Artificial neural networks, sequence-to-sequence LSTMs, and exogenous variables as analytical tools for NO₂ (air pollution) forecasting: a case study in the Bay of Algeciras (Spain). *Sensors* **2021**, *21*, 1770.
13. Du, S.; Li, T.; Yang, Y.; Horng, S.J. Deep air quality forecasting using hybrid deep learning framework. *IEEE Transactions on Knowledge and Data Engineering* **2019**, *33*, 2412–2424.
14. Xu, L.; Xu, B.; Sun, Y.; Cao, Y. Interpretable and scalable deep learning for urban NO₂ prediction via multisource data. *Transportation Research Part D: Transport and Environment* **2025**, *149*, 105039. <https://doi.org/https://doi.org/10.1016/j.trd.2025.105039>.
15. Yi, X.; Zhang, J.; Wang, Z.; Li, T.; Zheng, Y. Deep distributed fusion network for air quality prediction. In Proceedings of the Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 965–973.
16. Yin, P.; Deng, S.Y.; Zhu, Y.; Quan, G. Multi-source CPS data fusion model for air pollution prediction based on spatio-temporal Transformer. *Cyber-Physical Systems* **2026**, *12*, 203–224.
17. Bonastre-Egea, J.; Bueno-Crespo, A.; Morales-García, J. A Multi-Source Sensor Dataset for Spain: Integrating Air Quality, Meteorological, Mobility and Calendar Records (2022–2024), 2026. Licensed under CC BY 4.0, <https://doi.org/10.5281/zenodo.20196221>.
18. Rosales, C.M.; R. Bratburd, J.; Diez, S.; Duncan, S.; Malings, C.; Pant, P. Open air quality data platforms for environmental health research and action. *Current Environmental Health Reports* **2025**, *12*, 27.
19. Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO). Datos de Calidad del Aire, 2024. Red nacional de vigilancia de la calidad del aire. Datos horarios oficiales publicados en formato CSV.
20. Felici-Castell, S.; Segura-Garcia, J.; Perez-Solano, J.J.; Fayos-Jordan, R.; Soriano-Asensi, A.; Alcaraz-Calero, J.M. AI-IoT Low-Cost Pollution-Monitoring Sensor Network to Assist Citizens with Respiratory Problems. *Sensors* **2023**, *23*, 9585. <https://doi.org/10.3390/s23239585>.
21. Yin, H.; Zhang, Y.M.; Xu, J.; Chang, J.L.; Li, Y.; Liu, C.L. Air quality prediction with a meteorology-guided modality-decoupled spatio-temporal network. *IEEE Transactions on Artificial Intelligence* **2025**.
22. Xu, M.; Hu, W.; Han, K.; Ji, W. A multi-source data fusion model for air quality inference with hybrid supervised and self-supervised learning and adaptive feature importance estimation. *Information Processing & Management* **2026**, *63*, 104474.
23. Geng, Z.; Fan, X.; Lu, X.; Zhang, Y.; Yu, G.; Huang, C.; Wang, Q.; Li, Y.; Ma, W.; Yu, Q.; et al. FuXi-Air: Urban Air Quality Forecasting Based on Emission-Meteorology-Pollutant multimodal Machine Learning. *arXiv preprint arXiv:2506.07616* **2025**.
24. Kalaiselvi, S.; Anitha, V.; Manimaran, V.; Lawrence, T.S. Air quality prediction using multi-source remote sensing data integration with hybrid deep learning framework. *Scientific Reports* **2025**.
25. Arsov, M.; Zdravevski, E.; Lameski, P.; Corizzo, R.; Koteli, N.; Gramatikov, S.; Mitreski, K.; Trajkovik, V. Multi-Horizon Air Pollution Forecasting with Deep Neural Networks. *Sensors* **2021**, *21*, 1235. <https://doi.org/10.3390/s21041235>.
26. Graça, D.; Reis, J.; Gama, C.; Monteiro, A.; Rodrigues, V.; Rebelo, M.; Borrego, C.; Lopes, M.; Miranda, A.I. Sensors Network as an Added Value for the Characterization of Spatial and Temporal Air Quality Patterns at the Urban Scale. *Sensors* **2023**, *23*, 1859. <https://doi.org/10.3390/s23041859>.
27. Cukjati, J.; Mongus, D.; Rizman Žalik, K.; Žalik, B. IoT and Satellite Sensor Data Integration for Assessment of Environmental Variables: A Case Study on NO₂. *Sensors* **2022**, *22*, 5660. <https://doi.org/10.3390/s22155660>.
28. Folgado, M.G.; Sanz, V.; Hirn, J.; Lorenzo-Sáez, E.; Urchueguia, J. Deep learning for urban air quality: a traffic-based prediction and alert system for Valencia. *Neural Computing and Applications* **2025**, *37*, 15837–15854.
29. Terroso-Saenz, F.; Morales-García, J.; Munoz, A. Nationwide air pollution forecasting with heterogeneous graph neural networks. *ACM Transactions on Intelligent Systems and Technology* **2024**, *15*, 1–19.
30. Calabrese, F.; Ferrari, L.; Blondel, V.D. Urban sensing using mobile phone network data: a survey of research. *Acm computing surveys (csur)* **2014**, *47*, 1–20.

31. Mamei, M.; Bicocchi, N.; Lippi, M.; Mariani, S.; Zambonelli, F. Evaluating Origin–Destination Matrices Obtained from CDR Data. *Sensors* **2019**, *19*, 4470. <https://doi.org/10.3390/s19204470>.
32. Ministerio de Transportes y Movilidad Sostenible. Estudio de movilidad de viajeros de ámbito nacional aplicando la tecnología Big Data, 2024. Informe metodológico (Versión 8). Open Data Movilidad.
33. Morales-García, J.; Ramos-Sorroche, E.; Balderas-Díaz, S.; Guerrero-Contreras, G.; Muñoz, A.; Santa, J.; Terroso-Sáenz, F. Exploiting synthetic data generation to enhance pollution prediction. *Applied Soft Computing* **2025**, *175*, 113076.
34. Terroso-Saenz, F.; Morales-García, J.; Puig-Cabrera, M.; Martínez-del Vas, G.; Muñoz, A. Tourist Mobility Forecasting with Region-Based Flows and Hierarchical Spatial Tessellation. *International Journal of Information Technology & Decision Making* **2025**, pp. 1–33.
35. Melgarejo-Hernández, J.; García-Tapia-Mateo, P.; Morales-García, J.; Mazón, J.N. Near-Real-Time Integration of Multi-Source Seismic Data. *Sensors* **2026**, *26*, 451.
36. Robusto, C.C. The Cosine-Haversine Formula. *The American Mathematical Monthly* **1957**, *64*, 38–40. <https://doi.org/10.2307/2309088>.
37. European Parliament and Council. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. Official Journal of the European Union, L 152/1, 2008.
38. Agencia Estatal de Meteorología. AEMET OpenData, 2024. Sistema de acceso a datos meteorológicos y climatológicos mediante API REST.
39. Vacanza. python-holidays: Definitive list of public holidays per country, 2024. Python library for generating holidays.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.