**DOWNLOAD (54.7 MB)**

Prepared by Mohammad Taghi Abbasi

**The dataset for air quality forecasting**

A well-preprocessed dataset for implementing and evaluating air pollutant forecasting methods.

Regression

**Associated Tasks**

Environmental Science

Tabular

**Dataset Characteristics**

**Subject Area**

**Feature Type**

9

35064

Real

**# Features**

**# Instances**

**Dataset Information**

**Has Missing Values?**

No

**What do the instances in this dataset represent?**

Each instance is an air pollutant or meteorological variable.

**How are air quality and meteorological data integrated?**

Inverse distance weighted (IDW) interpolation

**Has Outliers?**

No

**Description of the Dataset for Air Quality Forecast**

Tehran, the capital of Iran, is equipped with an extensive network of Air Quality Monitoring Stations (AQMSs) established by [the municipality](https://airnow.tehran.ir/) and [the Department of Environment (DOE)](https://aqms.doe.ir/). At the time of preparing this dataset in October 2023, there were 37 AQMSs in Tehran, 24 of which belonged to the municipality, with the rest operated by the DOE. We collected data from these AQMSs from 00:00 on January 1, 2019, to 23:00 on December 31, 2022.

One of the main challenges of this dataset was the presence of missing and outlier data, which rendered it unsuitable for deep neural network-based forecasting tasks. To address this, we imputed missing values but first excluded AQMSs with over 50% missing data or more than 90 consecutive days of missing records. It should be noted that some AQMSs do not meet these criteria for all 6 pollutants, and we used the measurements from those AQMSs for the pollutants that did not meet the criteria. This is why the number of AQMSs is not the same for each pollutant (See Data from the AQMSs section). To gain a comprehensive understanding of the methods for removing and managing missing values in this dataset, please read the [paper](https://www.nature.com/articles/s41598-024-75678-6).

To incorporate meteorological variables in Tehran, we utilized data from the city's meteorological stations, established by [the Tehran Meteorological Organization](https://www.tehranmet.ir/fa/). Notably, during the preparation of this dataset, three stations—Shemiranat, Mehrabad, and Geophysics—were selected, and their data was collected for the same time period as the air quality data. This dataset, like air quality data, faced the issue of missing values, which were imputed using linear interpolation. Another challenge was its 3-hour sampling frequency, which was not aligned with the 1-hour frequency of the air quality data. To address this, linear interpolation was applied to adjust it to a 1-hour sampling frequency.

[The recent study](https://www.nature.com/articles/s41598-024-75678-6) demonstrated that the pollutants O3, CO, and SO2 exhibited consistent patterns across AQMSs in Tehran, showing minimal variation due to changes in station location. In contrast, the pollutants NO2, PM10, and PM2.5 showed distinct spatial variability in their behavior and concentration levels, influenced by the location of the AQMSs.

Since the goal of this dataset is spatiotemporal forecasting, only location-dependent pollutants are retained, and meteorological variables for valid AQMSs (11 AQMSs) are calculated at the location of each station using the inverse distance interpolation method.

**Variables Table**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Role** | **Type** | **Units** | **Statistical characteristics** | | | |
| **Min** | **Max** | **Mean** | **Std** |
| NO2 | Feature | Continuous | parts per billion (ppb) | 0.565 | 632.534 | 48.617 | 23.082 |
| PM10 | Feature | Continuous |  | 0.677 | 1985.000 | 77.888 | 56.521 |
| Humidity | Feature | Continuous | Percentage (%) | 2.479 | 99.147 | 36.656 | 20.766 |
| Temperature | Feature | Continuous | Degrees Celsius (°C) | -7.631 | 40.888 | 17.770 | 10.191 |
| Pressure | Feature | Continuous | Millibar (mbar) | 956.452 | 1037.462 | 1011.236 | 8.928 |
| Dew point temperature | Feature | Continuous | Degrees Celsius (°C) | -26.819 | 24.372 | 0.061 | 5.308 |
| Wind\_ x | Feature | Continuous | Kilometer per hour (km/h) | -11.653 | 7.269 | -0.874 | 1.408 |
| Wind\_ y | Feature | Continuous | Kilometer per hour (km/h) | -18.558 | 9.383 | -0.188 | 1.936 |
| PM2.5 | Target | Continuous |  | 0.167 | 979.956 | 30.825 | 21.515 |