The dataset for air quality forecasting

Prepared by Mohammad Taghi Abbasi

DOWNLOAD (54.7 MB)

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

Dataset Characteristics	Subject Area	Associated Tasks
Tabular	Environmental Science	Regression

Feature Type	# Instances	# Features
Real	35064	9

Dataset Information

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 Missing Values?

No

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 and the Department of Environment (DOE). 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.

To incorporate meteorological variables in Tehran, we utilized data from the city's meteorological stations, established by the Tehran Meteorological Organization. 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 demonstrated that the pollutants O_3, CO, and SO_2 exhibited consistent patterns across AQMSs in Tehran, showing minimal variation due to changes in station location. In contrast, the pollutants NO_2, PM_{10}, and $PM_{2.5}$ 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**
NO_2	Feature	Continuous	parts per billion (ppb)	0.565	301.055	48.597	22.942
PM_{10}	Feature	Continuous	$\mu g/m^3$	0.677	697.977	76.780	46.340
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
$PM_{2.5}$	Target	Continuous	$\mu g/m^3$	0.167	249.724	30.680	20.309