# Data Preprocessing Pipeline

Raw monitoring data needed preprocessing through an essential analytical pipeline because this process converted the information into an organized structure for machine learning applications. The stage used systematic methods to fix data quality problems throughout the process, which protected vital environmental signals from measurement data.

## 4.1 Missing Value Treatment

The presence of three separate types of missing data in the dataset required different approaches for filling in these gaps. The strong hourly measurement autocorrelation (ρ=0.78 for PM2.5 at lag=1) led to selecting temporal linear interpolation as an optimal method for continuous variables such as pollutants and weather conditions. The applied approach maintained data temporal coherency through filling 5.2% of PM2.5 values, together with 0.1 to 4.9% of weather observation gaps. The categorical wind direction values were imputed through modal completion that distinguished between seasons because wind directions change according to seasonal preferences (Alzubaidi *et al*. 2023). The preprocessing stage enabled the creation of full cases throughout all variables while producing undetectable distortions as shown by distributional testing opreprocesseded and postprocessed variables (Kolmogorov-Smirnov p-value>0.05 for all measured variables).

## 4.2 Feature Engineering

The model obtained improved periodic pattern detection for air quality data through the implementation of temporal feature extraction methods. The synthetic datetime index allows detailed operation on time-series data by combining year, month, day and hour columns. The trigonometric conversion of temporal data using sine and cosine functions encoded the recurring pattern of daily and yearly shifts, which improved upon numerical encodings because they failed to maintain December-January sequence continuity. The one-hot encoding technique applied to wind direction led to a 16-interval variable before specialists determined the northwest sector (NW, NNW) as the primary effective factor in PM2.5 dispersion patterns. Meteorological calculations were performed to enhance the representation of atmospheric stability by determining virtual temperature and saturation vapor pressure deficit (Ahmed *et al*. 2023).

## 4.3 Outlier Management

The IQR method located extreme values but protected genuine pollution events through its implementation. The algorithm measured threshold values as Q1-1.5×IQR and Q3+1.5×IQR per station to accommodate different baseline pollution levels in each geographic area. Through its conservative method, the system maintained 97.8% of original data points as it removed impermissible values like negative concentrations along with sensor error indications (Hajjaji *et al*. 2021). The treatment approach to outliers normalized PM2.5 skewness from 2.34 to 1.87 without distorting actual pollution data distributions because it remained consistent with historical Beijing environmental bulletin records.