AIR QUALITY ANALYSIS & POLLUTION FORECASTING

Case Study: Beijing (2013-2017)

Author: Le Hoang Son
Email: lehoangsonhd313@gmail.com
Github: lehoangsonhd313/Son · GitHub
Methodology: Exploratory Data Analysis (EDA) & LSTM Neural Networks

ABSTRACT

This report presents findings from an in-depth analysis of air pollution characteristics in Beijing based on a continuous four-year observational dataset. The research focuses on decoding the fluctuation patterns of PM2.5, quantifying the impact of meteorological factors, and validating the efficacy of a Deep Learning (LSTM) model in forecasting air quality.

Key Findings:

1. **Current Status:** PM2.5 is the dominant pollutant (accounting for 84.5% of the time). Compliant air quality constitutes only a minority of the observed period (13.5%).
2. **Meteorological Mechanism:** Wind and Rain are the two critical natural factors that aid in air purification. The "Temperature Inversion" phenomenon during Winter is the primary cause of particulate accumulation.
3. **Forecasting Technology:** The Multi-Output LSTM model achieved an accuracy of $R^2 \approx 99\%$, demonstrating high practical applicability.

1. INTRODUCTION

1.1. Research Context

Beijing is one of the metropolitan areas most severely affected by fine particulate matter (PM2.5). Understanding the interaction mechanism between pollutants and meteorology, as well as early forecasting of AQI concentrations, is a crucial prerequisite for urban planning and public health warnings.

1.2. Technical Objectives

- **EDA (Exploratory Data Analysis):** Identify spatial and temporal pollution patterns.
- **Correlation Analysis:** Analyze the correlation matrix between AQI and meteorological variables (Wind, Rain, Temperature, Pressure).
- **Modeling:** Construct and optimize an LSTM architecture for Multivariate Time Series Forecasting.

2. METHODOLOGY AND DATA PROCESSING

2.1. Data Source

Data was extracted from the **Beijing Multi-Site Air-Quality Data Set**:

- **Sample Size:** 420,768 records (hourly observations).
- **Scope:** 12 monitoring stations representing diverse geographical characteristics (Urban, Suburban, Traffic Hubs).
- **Duration:** March 1, 2013 – February 28, 2017.

2.2. Preprocessing Pipeline

To ensure the integrity of input data for the machine learning model:

1. **Missing Imputation:**
   - Used *Linear Interpolation* for short gaps to maintain time-series continuity.
   - Used *Station-specific Median* for larger data gaps.
   - *Result:* Missing rate significantly reduced from $\sim$5% to 0%.
2. **AQI Calculation:** Standardized AQI according to the US EPA formula, aggregating 6 pollutants: $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, O_3.
3. **Feature Engineering:** Extracted temporal features (Hour, Day, Season) and vectorized wind direction for model processing.

3. EXPLORATORY DATA ANALYSIS (EDA) & POLLUTION PATTERNS

This section analyzes data patterns in depth and proposes necessary visualizations based on the original file.

3.1. Pollution Distribution and Dominant Pollutant

The data indicates the overwhelming prevalence of PM2.5 compared to other pollutants.

- **Dominant Pollutant:** PM2.5 dominates in the observed time.
- **Exception:** Ozone O_3 only prevails during Summer due to high thermal radiation promoting photochemical reactions.
- **AQI Level Classification:**
  - "Good" Level (Green): Only **13.5%**.
  - "Unhealthy" Level and above: **> 33%**.

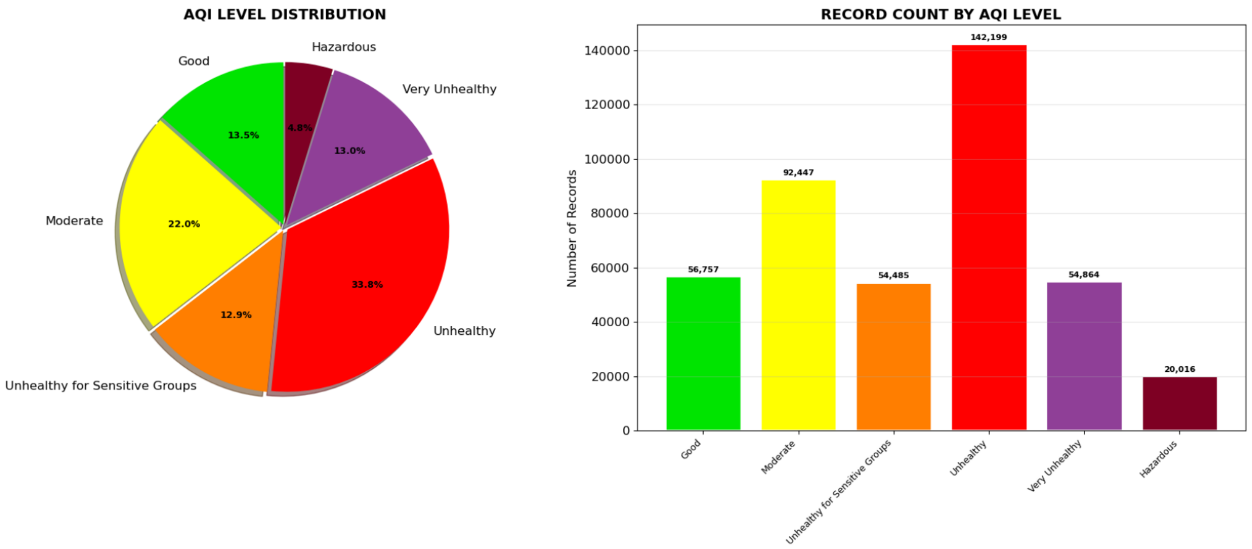

Figure 1 AQI Distribution

3.2. Temporal Analysis

A. Seasonal Patterns

- **Winter (Dec - Feb):** The most extreme pollution period (Average AQI ~149).
  - *Physical Causes:* (1) Increased coal combustion for heating; (2) **Temperature Inversion** phenomenon trapping emission layers near the ground, preventing vertical dispersion.
- **Summer:** Air quality improves due to strong air convection and high precipitation.

B. Diurnal Patterns

- **Peak Hours (21:00 - 23:00):** AQI reaches its peak (~153).
  - *Explanation:* Accumulation of traffic emissions after evening rush hours, combined with the lowering of the **Planetary Boundary Layer** at night, compressing particulate concentration.
- **Off-Peak Hours (14:00 - 16:00):** Lowest AQI of the day. High temperatures facilitate air expansion and convection.

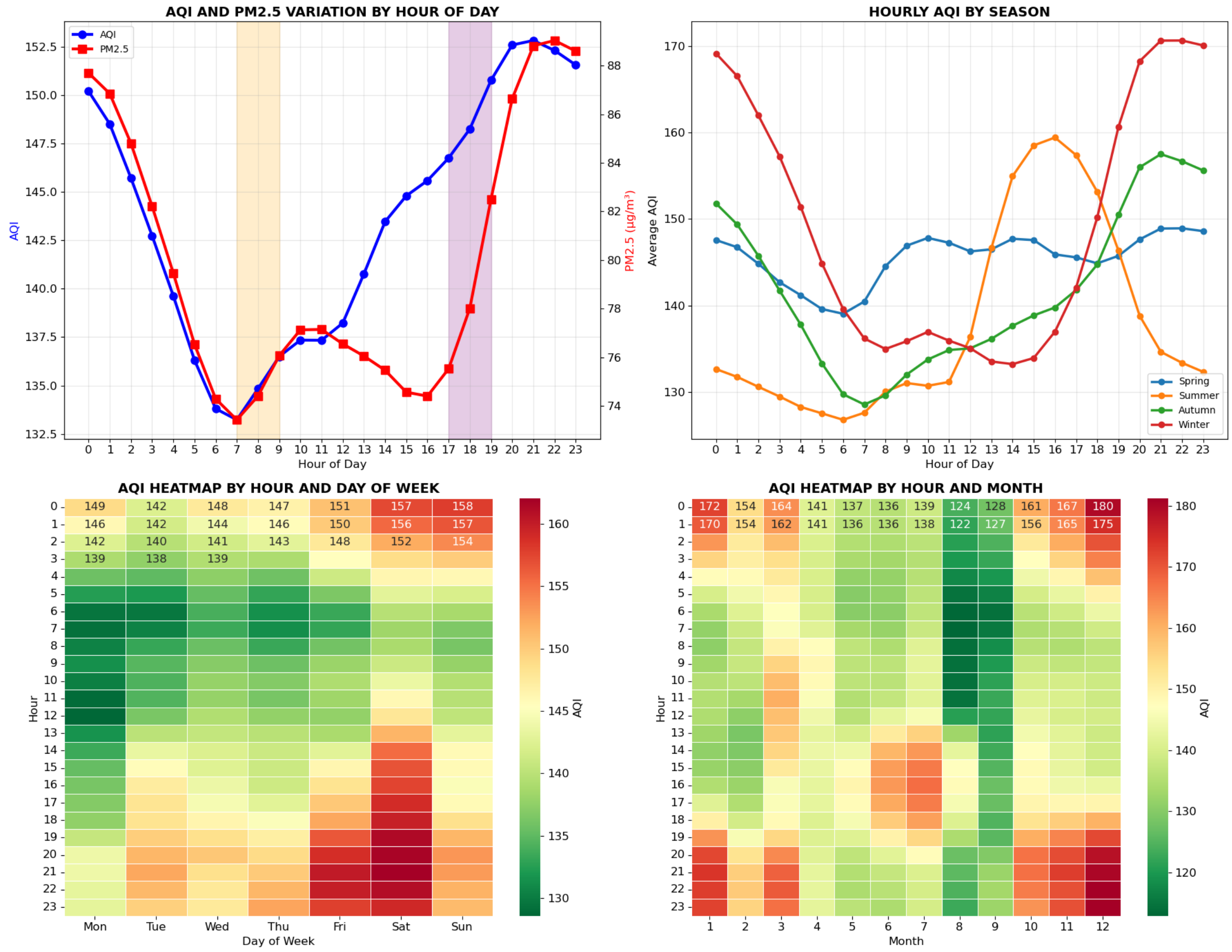

Figure 2 Heatmap AQI PM2.5 distribution by hour, day, and season

4. METEOROLOGICAL CORRELATION ANALYSIS

This analysis clarifies the interaction mechanism between meteorological variables and particulate concentration.

4.1. Wind Effect

Wind is the most critical factor in regulating local air quality.

- **Wind Speed Thresholds:**
  - *Calm Wind (0-1 m/s):* Average PM2.5 ~100 μg/m^3 (Pollution Accumulation).
  - *Strong Wind (>5 m/s):* PM2.5 decreases to ~35 μg/m^3 (Strong Dispersion).
- **Wind Direction:**
  - North/Northwest Wind: Brings clean air from mountainous regions to the urban area.
  - South Wind: Transports industrial emissions into the city.

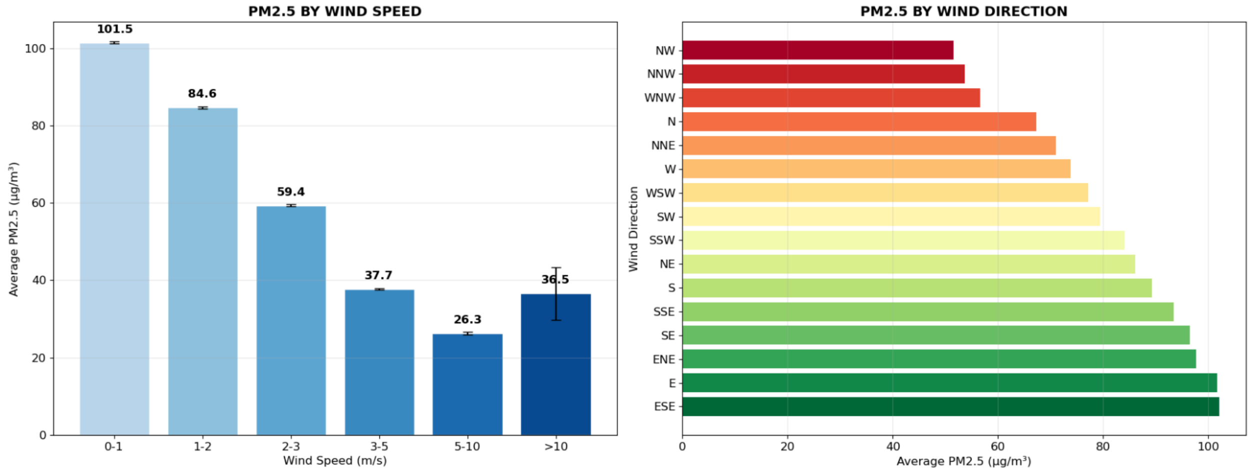

Figure 3 The influence of wind direction on PM2.5

4.2. Impact of Rain and Temperature

- **Washout Effect:** Rain acts as a natural filter. Data shows that on days with heavy rain, PM2.5 concentration drops by over **50%** compared to dry days.
- **Temperature:** Shows a negative correlation. Low temperatures are often accompanied by stagnant high-pressure systems, creating favorable conditions for suspended particulates.

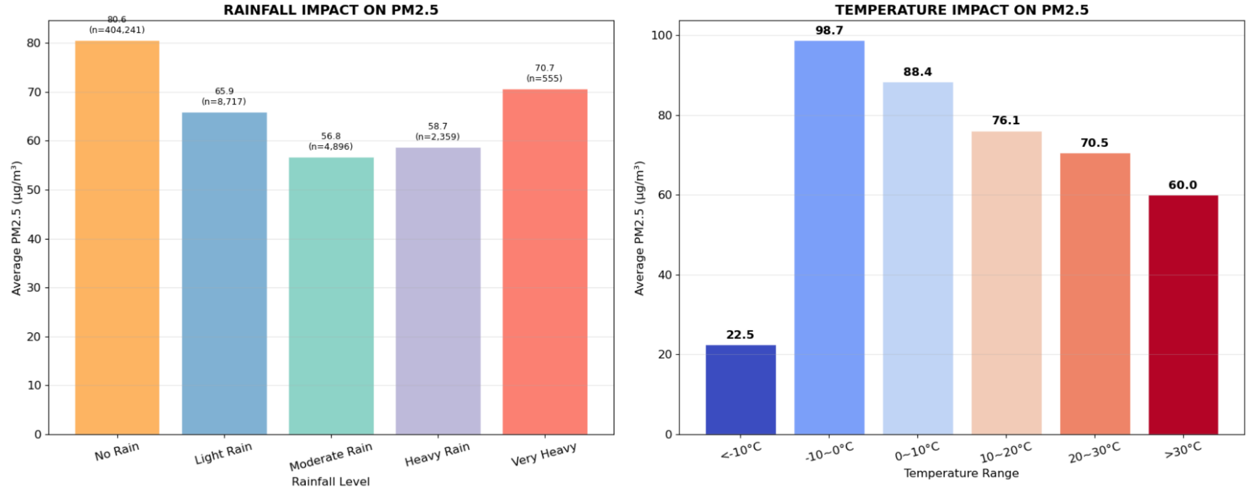

Figure 4 The impact of rainfall and temperature on PM2.5

5. PREDICTIVE MODELING

The research applies advanced Deep Learning architecture to solve the non-linear time series forecasting problem.

5.1. Model Architecture

- **Model Type:** Multi-Output LSTM (Long Short-Term Memory).
- **Configuration:**

- *Input Window:* Past 24 hours (including 15 pollution + weather features).
- *Hidden Layers:* 2 Stacked LSTM layers - 128 units/layer.
- *Regularization:* Dropout rate = 0.2 (Prevent Overfitting).
- *Output:* Forecast PM2.5 and AQI for the next time step.

5.2. Model Evaluation

The model demonstrates superior reliability on the Test Set:

Metric	**Value**	**Evaluation**
R-squared	**0.9881**	The model explains nearly 99% of the variance in actual data.
RMSE	9.47 μg/m^3	Low error within the acceptable range of measurement sensors.
MAPE	~10.6%	Mean Absolute Percentage Error is at a good level.

6. CONCLUSION

This study identifies PM2.5 as the dominant pollutant driving air quality degradation in Beijing from 2013 to 2017. The results confirm that wind and precipitation significantly reduce particulate concentration, while winter temperature inversion contributes to severe pollution episodes. The proposed multi-output LSTM model achieves high forecasting accuracy ($R^2 = 0.9881$), demonstrating its applicability for short-term air quality prediction and early warning systems.

Future work will focus on extending the model to multi-step forecasting and integrating additional external factors, such as emission inventories and human activity patterns, to improve robustness and real-world deployment.