CompSys Lab 2025 - Yilmaz

DATA AUGMENTATION STRATEGIES FOR IMPROVED PM2.5 FORECASTING USING TRANSFORMER ARCHITECTURES

Phoebe Pan

INTRODUCTION

Air pollution causes almost 7 million premature deaths every year. Research has shown that breathing in particulate matter with diameters less than 2.5 micrometers (PM2.5) greatly increases an individual's risk of respiratory and cardiovascular conditions from asthma to heart disease. PM2.5 comes from many sources around us including car emissions, forest fires, power plants, and construction sites.

RESEARCH PROBLEM

Recent trends have shown an alarming increase in PM2.5 emissions due to wildfires, exacerbated by climate change. Climate projections suggest that the area affected by wildfires in the western U.S. could expand by 54% between 2046 and 2055 compared to 1996-2005 (Spracklen et al., 2009). Thus, air pollution will only become a larger problem in the future

During severe wildfire events, PM2.5 level can spike to hazardous levels. Extreme events like these are underrepresented in the training dataset, leading to underestimated PM2.5 concentrations when they exceed 60 µg/m³ (Li et al., 2017; liu et al., 2022). Therefore, data imbalance makes it difficult for models to predict critical conditions accurately, as they are rare in the data.

The study focuses on major urban areas in the northeastern United States, specifically New York City, Philadelphia, and Washington, D.C., from 2021 - 2023.

DATA SOURCES

Satellite aerosol optical depth (AOD) was sourced from Multi-Angle Implementation of Atmospheric Correction (MAIAC), which retrieves aerosol levels based on observations from Moderate Resolution Imaging Spectroradiometer (MODIS).

Modern-Era Retrospective Analysis for Research and Applications - version 2 (MERRA-2) provided assimilated aerosol species.

Meteorological variables were sourced from European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5).

Ground-level hourly PM2.5 measurements were obtained from the U.S. Environmental Protection Agency (EPA)'s AirNow program

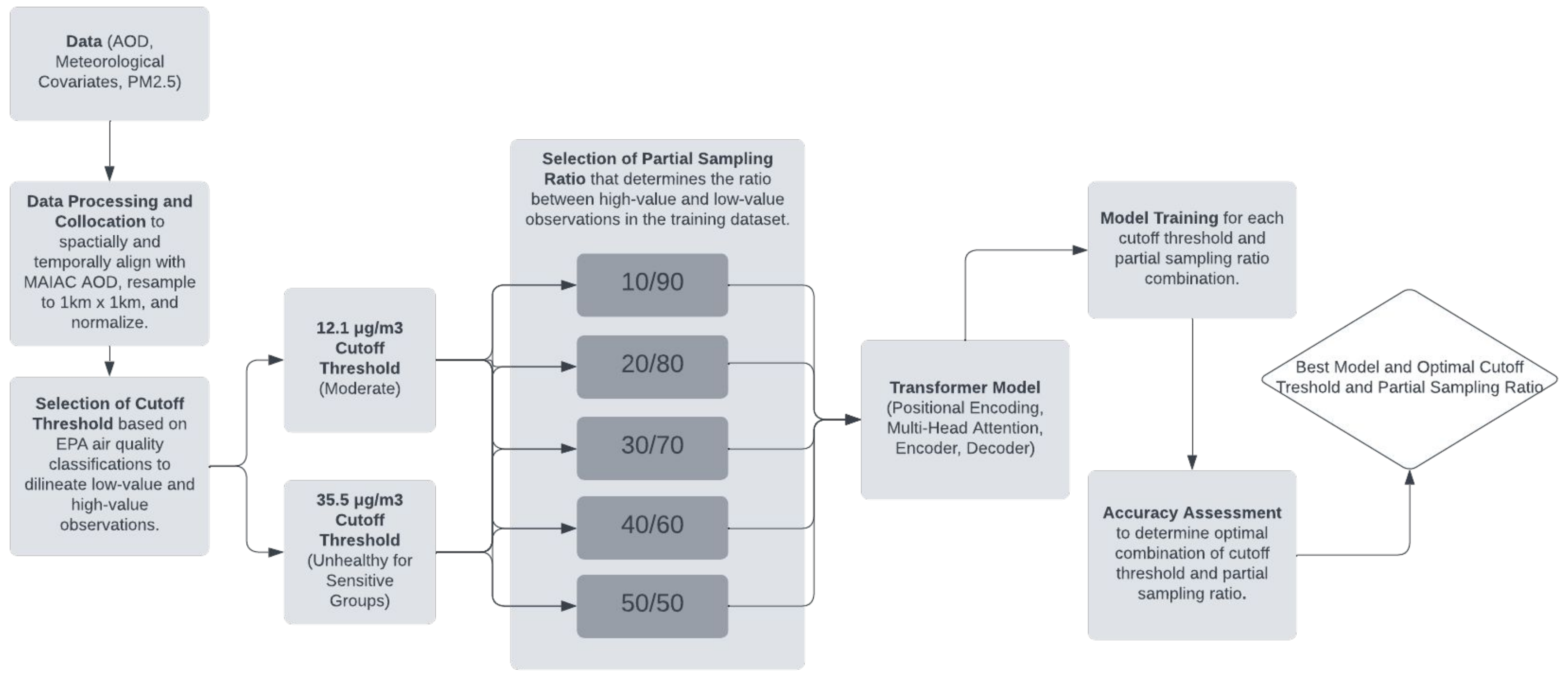
Data (AOD, Meteorological Covariates, PM2.5)
Data Processing and Collocation to spactially and temporally align with MAIAC AOD, resample to 1km x 1km, and normalize.
Selection of Cutoff Threshold based on EPA air quality classifications to dilineate low-value and high-value observations.
12.1 µg/m3 Cutoff Threshold (Moderate)
35.5 µg/m3 Cutoff Threshold (Unhealthy for Sensitive Groups)
Selection of Partial Sampling Ratio that determines the ratio between high-value and low-value observations in the training dataset.
10/90
20/80
30/70
40/60
50/50
Transformer Model (Positional Encoding, Multi-Head Attention, Encoder, Decoder)
Model Training for each cutoff threshold and partial sampling ratio combination.
Accuracy Assessment to determine optimal combination of cutoff threshold and partial sampling ratio.
Best Model and Optimal Cutoff Treshold and Partial Sampling Ratio

Figure 1. Systems Architecture

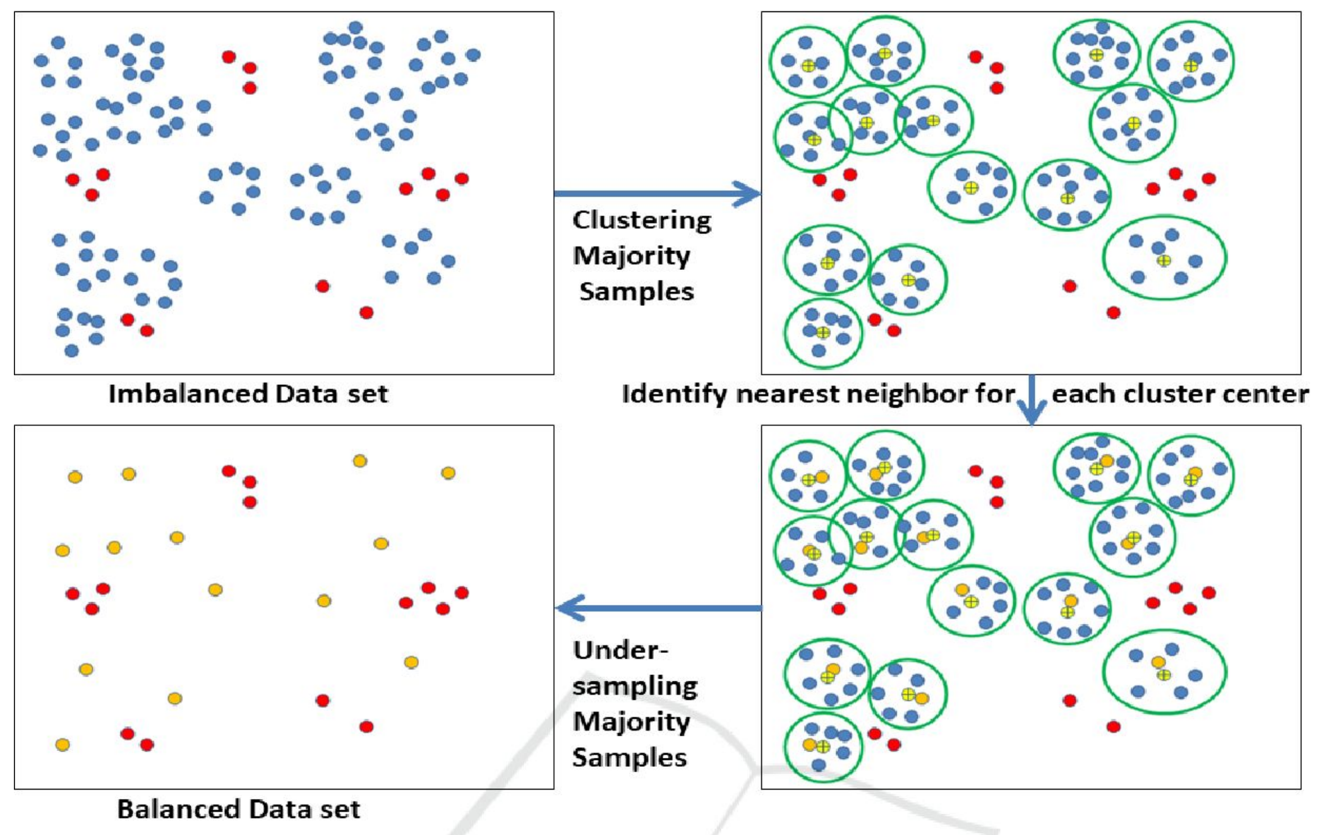
Imbalanced Data set
Clustering Majority Samples
Identify nearest neighbor for each cluster center
Under-sampling Majority Samples
Balanced Data set

Figure 2. Cluster-Based Undersampling; Source: Henein et al., 2018

DATA AUGMENTATION

In this study, we implemented cluster-based undersampling to address class imbalance in the training data. The first step involved grouping data points into clusters using the k-means algorithm. By applying the undersampling strategy within each cluster, we selected a subset of instances, effectively reducing the majority class without losing diversity within the data.

The selection of a cutoff value for distinguishing between majority and minority classes plays a crucial role in determining class distribution and, consequently, the model's performance. In the context of air quality in the United States, we used PM2.5 classifications by the EPA to determine a cutoff threshold of 35.5 µg/m³.

Many studies aim to achieve a perfect 50/50 balance between minority and majority class points. However, this idealized ratio is not always the most effective. Partial sampling involves adjusting the class ratio to values between the original class distribution and an equal 50/50 split (Kamalov et al., 2022). Following this insight, the present study applied various partial sampling ratios (10/90, 20/80, 30/70, 40/60, 50/50) to explore whether these findings hold for PM2.5 forecasting models.

PM2.5 FORECASTING

The Transformer model has revolutionized various domains of ML, including NLP and time series forecasting (Vaswani et al., 2017). In the context of PM2.5 forecasting, the Transformer model's ability to capture long-range dependencies and complex temporal patterns makes it a powerful tool for forecasting air pollution levels. A Transformer model differentiates itself from traditional convolutional and recurrent neural networks by employing a novel positional encoding mechanism to preserve temporal relationships.

An attention mechanism is often used during the encoding phase to help the model assign different weights to the input time series information, quantifying the dependencies between them.

In this study, the encoder consists of six identical layers where the input undergoes multi-head self-attention followed by LayerNorm, feed-forward computation, and subsequent normalization with dropout. Each decoder in the stack of six layers applies self-attention with a target sequence mask, followed by cross-attention to the encoder output with a source mask, and then passes the result through a feed-forward network with dropout.

RESULTS

Ratio	Whole			High-Value		
	RMSE	MAE	R²	RMSE	MAE	R²
Original	3.174	0.661	0.801	41.34	28.269	0.607
10/90	2.282	1.592	0.897	19.747	13.81	0.633
20/80	2.080	1.386	0.914	15.353	10.077	0.778
30/70	2.306	1.671	0.895	16.095	12.204	0.756
40/60	2.423	1.726	0.884	16.556	12.917	0.741
50/50	2.677	1.875	0.858	19.116	14.321	0.656

Figure 3. Accuracy metrics for experiments performed with a cutoff threshold of 35.5 µg/m³.

The 20/80 resampling ratio emerges as the optimal configuration overall, achieving the lowest RMSE (2.080) and MAE (1.386), alongside the highest R² value of 0.914. This strong performance suggests that a 20/80 ratio balances the trade-off between capturing minority and majority points while minimizing error.

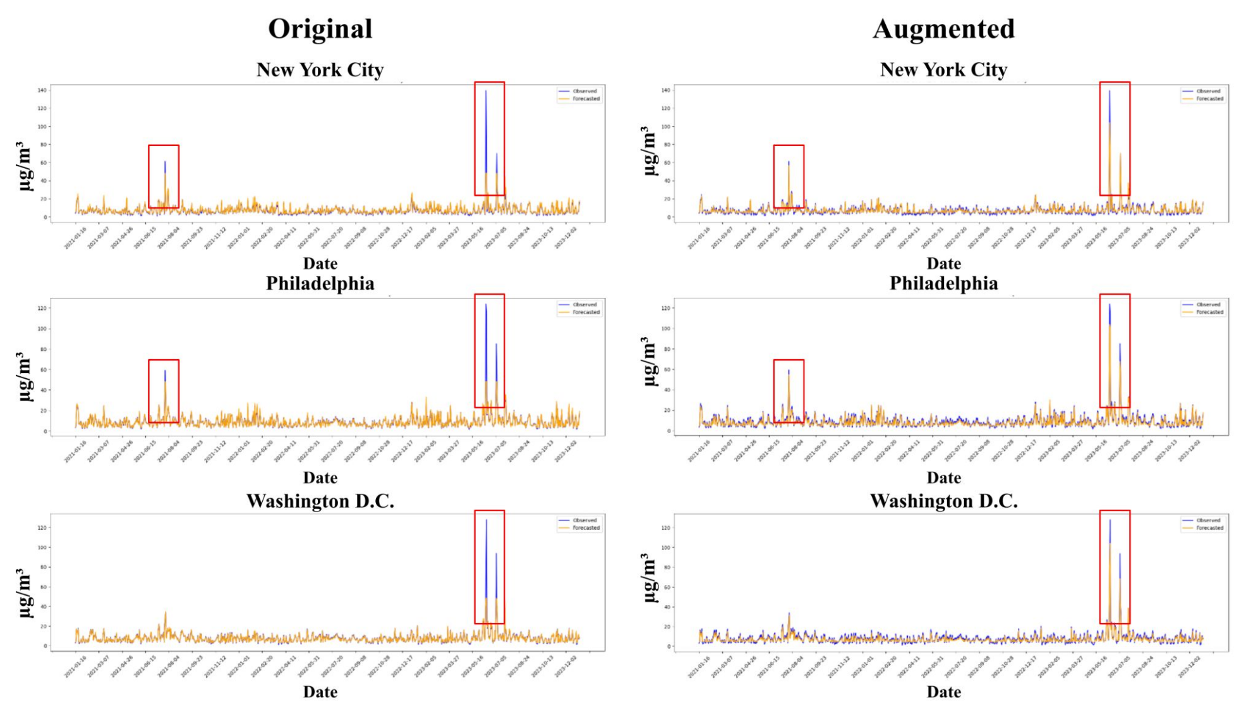

Figure 4. Time series of observed (blue) and forecasted (orange) PM2.5 concentrations from 2021 to 2023 in New York City (top), Philadelphia (middle), and Washington D.C. (bottom).

The model trained on the original dataset shows strong accuracy for lower concentrations but struggles to predict extreme pollution events due to the imbalanced dataset. In contrast, models trained on an augmented dataset with a 35.5 µg/m³ cutoff and a 40/60 partial sampling ratio demonstrate improved performance in forecasting high PM2.5 concentrations, though at the cost of slightly reduced accuracy for lower levels.

FUTURE WORK

For future work, we plan to expand the size and scope of our dataset to improve the model's generalizability across diverse conditions and locations. Incorporating transfer learning or developing a robust base model that can be fine-tuned with smaller, region-specific datasets may enhance performance in data-scarce environments. Additionally, integrating uncertainty quantification techniques will help assess the bias introduced by data augmentation and multistep forecasting. This will allow us to better interpret the model's confidence and improve reliability in high-stakes applications.