İSTANBUL AİR QUALİTY REPORT

PREPARED BY

STUDENT ID	NAME- SURNAME
528241005	AYSAN PAKMANESH
528241017	MELİS CİGER
528241041	SAMET ERKEK
528241003	SELİN BOZKURT

SUBMISSION DATE: 7/25/2025

CRN: 30046

COURSE: MACHİNE LEARNİNG WİTH BİG DATA VIA 509E

ADVİSOR: Dr Suha Tuna. Assistant Professor Of Computational Science & Engineering - Istanbul Technical University Informatics Institute

Executive Summary

This report presents a comprehensive data-driven analysis of Istanbul's air quality patterns using machine learning techniques. The study addresses fundamental questions: Is Istanbul truly one city from an air quality perspective, or does it function as multiple distinct zones requiring differentiated management approaches? What are some hidden patterns that we are missing? Can we predict how it will evolve? What do research say about Istanbul?

The analysis utilized over 9 million air quality measurements from monitoring stations across Istanbul, obtained from the Turkish Ministry of Environment, Urbanization and Climate Change's continuous monitoring system (https://sim.csb.gov.tr/), The dataset spans from 2020 to 2025 and includes measurements of six key pollutants: PM2.5, PM10, NO2, SO2, O3, and CO; and 236 academic papers from OpenAlex (https://openalex.org/) to benchmark our findings against global air quality research."

Chapter 1: Is Istanbul Really One City?

1.1 The Global Context and Local Question

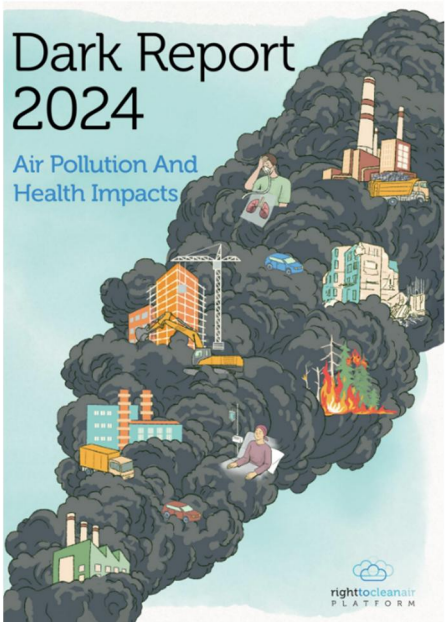
Air pollution represents one of the most pressing environmental challenges facing urban populations worldwide. Recent data from the World Health Organization indicates that over 90% of the global population breathes air exceeding recommended safety guidelines. Within this global context, Istanbul presents a particularly interesting case study as one of the world's largest transcontinental cities, home to over 16 million residents.

Recent headlines paint a stark picture of Istanbul's deteriorating air quality:

"Istanbul faces air quality challenges" - Daily Sabah

"The invisible killer threatens Istanbul" – Bianet

The Turkish Medical Association's 2024 "Dark Report" provides critical data about Turkey's air quality situation. The report documents that 92% of Turkey's population breathes air exceeding WHO standards, with Istanbul recording PM10 levels of 38.41 µg/m³ in 2022 - approximately 2.5 times the WHO guideline value of 15 µg/m³. Air pollution deaths approached 70,000 in 2022, representing a concerning increase from previous years. Additionally, every 10 µg/m³ increase in PM10 correlates with a 5% increase in breast cancer mortality risk.

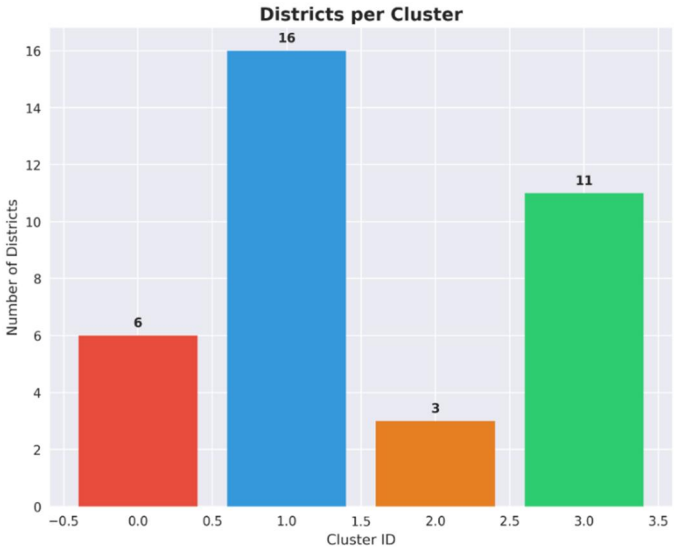


These statistics raise a fundamental question: Can Istanbul's air quality challenges be addressed through unified city-wide policies, or does the city's complexity require a more nuanced, differentiated approach?

Chapter 2: The Four Istanbuls - A K-Means Clustering Analysis

2.1 Methodology and Approach

To understand Istanbul's complex pollution landscape, we employed K-means clustering analysis on multiple variables including population density, traffic patterns, industrial activity, and infrastructure metrics. The clustering algorithm identified four distinct zones, each representing what we term a different "Istanbul."



The technical implementation involved several key preprocessing steps. First, we applied a unit correction for CO measurements, dividing values by 1000 to standardize units to mg/m³. We then employed winsorization at the 95th percentile to handle outliers, preventing single extreme values from dominating cluster assignments. The data was processed using PySpark's distributed computing framework, with features standardized using StandardScaler before clustering.

2.2 Cluster Analysis Results

The K-means algorithm tested multiple values of K (3, 4, and 5), with K=4 producing the highest silhouette score of 0.432, indicating well-separated and cohesive clusters. The four identified clusters represent distinct "Istanbuls":

2.2 Cluster Characteristics

Cluster 1: PM10-Heavy Districts (6 districts) includes Mecidiyeköy, Sancaktepe, Yenibosna, Üsküdar, Şirinevler, and Göztepe D-100. These areas show the highest particulate matter concentrations with PM10 averaging 47.6 µg/m³ and CO levels at 0.6 mg/m³. The elevated PM10 suggests a combination of traffic emissions, construction activities, and possible industrial sources. The D-100 highway's presence in this cluster name indicates the significant impact of major traffic arteries on local air quality.

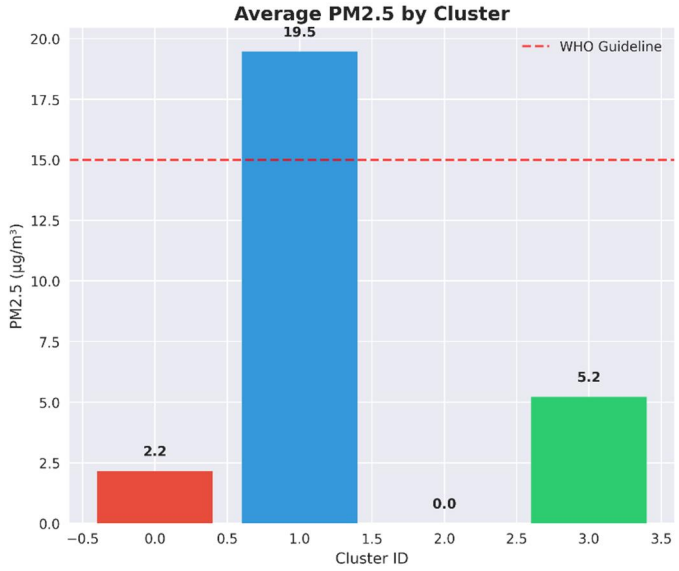
Cluster 2: NO2-Dominant Urban Core (16 districts) encompasses the heart of Istanbul's urban activity, including Ümraniye, Selimiye, Şişli Eğitim, Arnavutköy, Aksaray, Kadıköy, Kağıthane, Beşiktaş, Kartal, Avcılar, Sultangazi, Alibeyköy, Tuzla, Esenler, Çatladıkapı, and Bağcılar. These districts are characterized by intense traffic flow and dense urban development, with NO2 levels averaging 40.6 µg/m³ and CO at 0.5 mg/m³. The dominance of nitrogen dioxide clearly indicates traffic as the primary pollution source, reflecting these areas' role as major transportation corridors and commercial centers.

Cluster 3: The Sultangazi Anomaly (3 stations) consists of three monitoring stations within Sultangazi (Sultangazi 1, 2, and 3). This cluster presents an intriguing pattern with the highest PM10 levels at 62.1 µg/m³ but surprisingly shows 0.0 mg/m³ CO readings. This unusual combination suggests either specific local PM10 sources without accompanying traffic emissions, possible measurement anomalies, or unique topographical conditions that trap particulate matter while allowing gaseous pollutants to disperse.

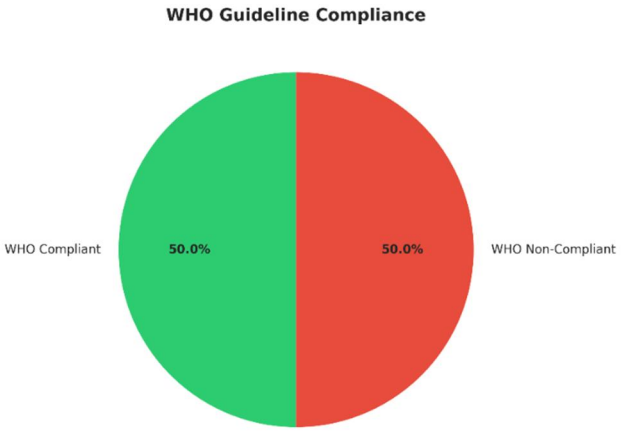
Cluster 4: O3-Dominant Coastal/Peripheral (11 districts) comprises Sultanbeyli, Silivri, Sarıyer, Esenyurt, Beylikdüzü, Şile, Kumköy, Başakşehir, Büyükada, Kandilli, and Maslak. These districts benefit from their peripheral locations and coastal proximity, showing O3 as the dominant pollutant at 37.3 µg/m³ with minimal CO levels of 0.1 mg/m³. The ozone dominance indicates these areas receive transported pollutants that undergo photochemical transformation rather than experiencing direct emission sources, a pattern typical of suburban and coastal zones downwind from urban centers.

2.3 Infrastructure Gap Analysis

A critical finding emerged when examining monitoring infrastructure distribution across clusters. The analysis revealed significant disparities in PM2.5 monitoring capabilities.



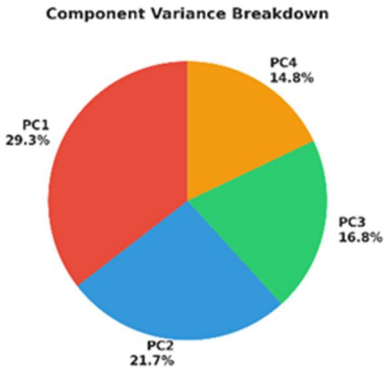
Clustering revealed that Cluster 3 (Sultangazi 1,2,3) represents districts with incomplete PM2.5 monitoring infrastructure, not a pollution pattern. This identifies critical sensor deployment needs.



While the original dataset contains 1.45 million PM2.5 measurements from 23 monitoring stations (21 districts), this represents only 58% district coverage across Istanbul's +36 administrative areas. The K-means algorithm identified this pattern by clustering 15 districts with missing PM2.5 sensors, which artificially inflated WHO guideline compliance from the actual 14.3% to an apparent 50% when zero-values were incorrectly counted as compliant. Among the 21 districts with functional PM2.5 monitoring, 85.7% exceed WHO annual guidelines (>15 µg/m³), with the highest concentrations in Kartal and Bağcılar (22.4 µg/m³). The 15 unmonitored districts, including the entire Sultangazi region which shows elevated PM10 levels, represent a significant public health blind spot requiring immediate sensor deployment. This analysis demonstrates that apparent "missing data" often reflects real infrastructure deficiencies rather than technical artifacts, transforming a data quality issue into actionable policy recommendations of better PM2.5 network coverage.

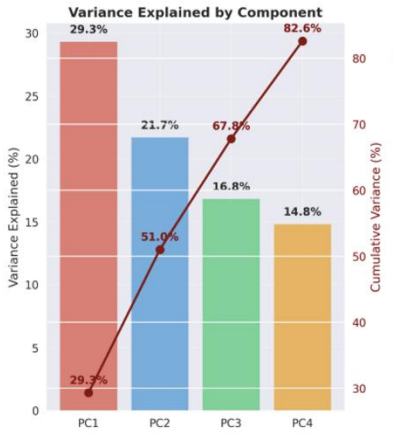
Chapter 3: What Drives Pollution? Principal Component Analysis of Istanbul Pollution

3.1 Technical Approach to Dimensionality Reduction



To understand the underlying forces driving Istanbul's pollution patterns, we applied Principal Component Analysis to the six-dimensional pollution dataset. The analysis used hourly pollution vectors created from averaged measurements across districts, with proper handling of missing values through sampling rather than zero-filling to avoid artificial correlations.

The PCA implementation utilized PySpark's ML library with standardized features to ensure equal weighting across pollutants with different measurement scales. We tested models with 2, 3, and 4 components to determine optimal dimensionality reduction.



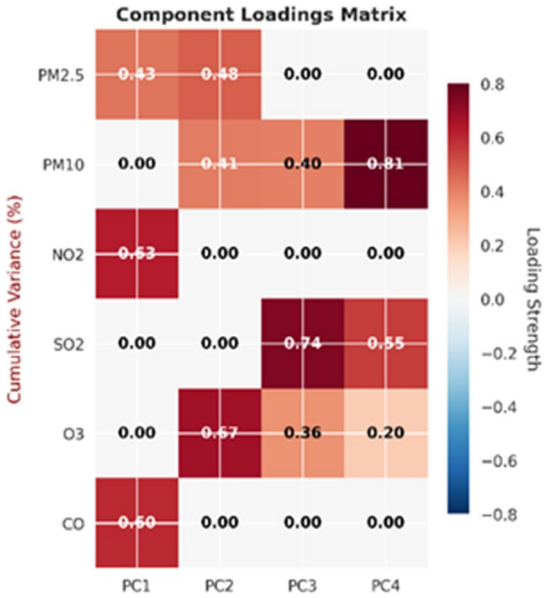
3.2 The Three Principal Components

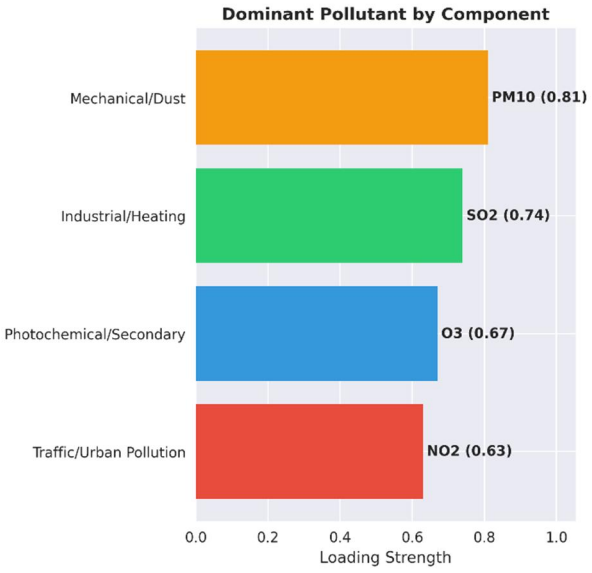
Our analysis identified three principal components that collectively explain 67.8% of the total variance in Istanbul's air quality:

Principal Component 1: Traffic-Urban Pollution (29.3% variance) The first component shows high loadings for NO2 (0.63), CO (0.60), and PM2.5 (0.43). This combination clearly indicates traffic-related emissions, as these pollutants are co-emitted from vehicle exhaust. The technical interpretation suggests this component captures the morning and evening rush hour patterns that dominate urban air quality.

Principal Component 2: Photochemical Activity (21.7% variance) With dominant loadings for O3 (0.67) and secondary contributions from PM2.5 (0.48) and PM10 (0.41), this component represents photochemical processes. The positive O3 loading combined with particulate matter indicates secondary aerosol formation through atmospheric chemistry, particularly during sunny afternoons.

Principal Component 3: Industrial Signatures (16.8% variance) The third component isolates industrial emissions with SO2 showing the highest loading at 0.74, accompanied by PM10 at 0.40. This signature matches the emission profile of coal combustion and industrial processes, distinguishing these sources from traffic-related pollution.





3.3 Variance Decomposition and Interpretation

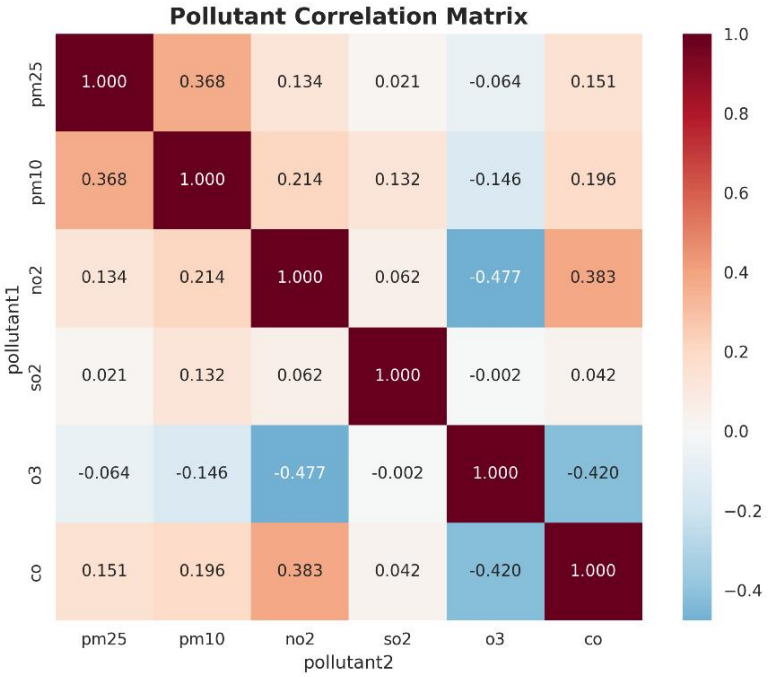
The remaining 32.2% of unexplained variance likely represents local meteorological effects, measurement noise, and other factors not captured by the main pollution sources. The clear separation of components into traffic, photochemical, and industrial sources validates our clustering results and provides a framework for targeted interventions.

Chapter 4: Pollution Pattern Recognition

4.1 Correlation Analysis Methodology

To validate our PCA findings and understand pollutant interactions, we conducted correlation analysis on concurrent measurements from the same locations and times. This approach avoided the common pitfall of correlating averaged values that can produce artificial relationships.

The technical implementation required careful data preparation, using only measurements where multiple pollutants were recorded simultaneously. This resulted in 1,827 high-quality concurrent measurements for correlation calculation.



4.2 Key Correlation Findings

The correlation matrix reveals several important relationships:

Traffic Signature Correlations:

- NO2 ↔ CO: 0.383 (strong positive correlation indicating common vehicular source)
- NO2 ↔ O3: -0.477 (negative correlation showing NO2's role in ozone depletion)
- CO ↔ O3: -0.420 (confirming traffic emissions suppress ozone formation)

Particulate Matter Relationships:

- PM2.5 ↔ PM10: 0.368 (moderate correlation indicating PM2.5 is a subset of PM10)
- PM10 ↔ NO2: 0.214 (weak correlation suggesting different primary sources)

Industrial Indicators:

- SO2 shows weak correlations with most pollutants (0.021-0.132), confirming its distinct industrial source

- The low SO2 correlations validate its use as an industrial tracer in our PCA analysis

4.3 Chemical Interpretation

The correlation patterns confirm three distinct pollution regimes operating in Istanbul:

1. **Traffic-Urban Regime:** High NO2-CO correlation with anti-correlation to O3
2. **Photochemical Regime:** O3 formation suppressed by primary pollutants
3. **Industrial Regime:** SO2 operating independently of traffic pollutants

These correlations aren't just numbers - they're validation that mathematical modelling captures real atmospheric chemistry. The negative NO2-O3 correlation proves the titration effect where traffic emissions actually destroy ozone locally, explaining why our coastal districts (Cluster 4) have higher ozone despite lower overall pollution.

Chapter 5: Predicting PM2.5 Levels

5.1 Objective and Approach

Building on our understanding of Istanbul's pollution patterns, we developed a predictive model for PM2.5 concentrations. The goal is to forecast PM2.5 levels using historical data from 2020-2025, enabling authorities to issue timely warnings and implement preventive measures. The ability to predict PM2.5 levels even a few hours in advance transforms reactive pollution management into proactive public health protection. When we can forecast that PM2.5 levels will exceed safe thresholds tomorrow afternoon, authorities can implement traffic restrictions, issue health advisories, and vulnerable populations can take protective measures.

The prediction framework uses the same air quality dataset analyzed in previous chapters but focuses specifically on PM2.5 as the target variable. The technical implementation leveraged PySpark's distributed computing capabilities to handle the temporal complexity of the data.

```
%pyspark
df_openaq_conv.filter("value < 0").show()
```

```
+-----------+----------+--------------------+-------------------+------+-------+---------+-----+------+----+-----+---+
|location_id|sensors_id|            location|           datetime|   lat|    lon|parameter|units| value|year|month|day|
+-----------+----------+--------------------+-------------------+------+-------+---------+-----+------+----+-----+---+
|        100|   4275786|Badhoevedorp-Slot...|2024-01-11 12:00:00|52.334|4.77401|       no|µg/m³|-999.0|2024|    1| 11|
|        100|   4275786|Badhoevedorp-Slot...|2024-01-28 18:00:00|52.334|4.77401|       no|µg/m³|-999.0|2024|    1| 28|
|        100|      4231|Badhoevedorp-Slot...|2024-02-17 12:00:00|52.334|4.77401|     pm10|µg/m³|-995.0|2024|    2| 17|
|        100|      4231|Badhoevedorp-Slot...|2024-02-24 18:00:00|52.334|4.77401|     pm10|µg/m³|-995.0|2024|    2| 24|
|        100|      4147|Badhoevedorp-Slot...|2024-01-28 21:00:00|52.334|4.77401|     pm25|µg/m³|-999.0|2024|    1| 28|
|        100|   4275786|Badhoevedorp-Slot...|2024-02-10 21:00:00|52.334|4.77401|       no|µg/m³|-999.0|2024|    2| 10|
|        100|      4147|Badhoevedorp-Slot...|2024-02-25 16:00:00|52.334|4.77401|     pm25|µg/m³|-999.0|2024|    2| 25|
|        100|      4147|Badhoevedorp-Slot...|2024-01-21 02:00:00|52.334|4.77401|     pm25|µg/m³|-999.0|2024|    1| 21|
|        100|      4147|Badhoevedorp-Slot...|2024-02-10 00:00:00|52.334|4.77401|     pm25|µg/m³|-999.0|2024|    2| 10|
|        100|       162|Badhoevedorp-Slot...|2024-01-28 10:00:00|52.334|4.77401|       co|µg/m³|-999.0|2024|    1| 28|
|        100|   4275786|Badhoevedorp-Slot...|2024-01-24 21:00:00|52.334|4.77401|       no|µg/m³|-999.0|2024|    1| 24|
|        100|       162|Badhoevedorp-Slot...|2024-01-25 10:00:00|52.334|4.77401|       co|µg/m³|-999.0|2024|    1| 25|
|        100|       162|Badhoevedorp-Slot...|2024-01-13 13:00:00|52.334|4.77401|       co|µg/m³|-999.0|2024|    1| 13|
|        100|       162|Badhoevedorp-Slot...|2024-02-10 08:00:00|52.334|4.77401|       co|µg/m³|-999.0|2024|    2| 10|
```

Took 7 sec. Last updated by anonymous at April 09 2025, 3:33:25 PM.

5.2 Feature Engineering Highlights

The predictive model relies on engineered features that capture temporal patterns in PM2.5 behavior. These features were created using window functions to track how pollution levels change over time.

Lag features capture PM2.5 values from 1, 3, 6, and 24 hours before each timestamp. These represent the recent history of pollution levels and help identify short-term dependencies. Rolling statistics, including 3-hour, 6-hour, and 24-hour moving averages, smooth out fluctuations to reveal underlying trends. Standard deviations over 6 and 24-hour windows quantify volatility in air quality.

Temporal features extract patterns from timestamps, including hour of day, day of week, month, and season. A binary weekend indicator distinguishes between weekday and weekend pollution patterns. Differential features calculate changes in PM2.5 over 1-hour and 24-hour periods, capturing acceleration or deceleration in pollution accumulation.

The model also includes a PM2.5 to PM10 ratio feature, which provides insights into the types of particles present and their likely sources. Data quality was ensured through outlier removal using z-score normalization, retaining only observations within 4 standard deviations of the mean.

5.3 Model Implementation

A Random Forest Regressor was selected for the prediction task, balancing accuracy with interpretability. The model pipeline processes categorical features like season through encoding steps before combining all features for model training.

Metric	Value
RMSE	3,61
MAE	1,15
R2	0,94

The data was split using a time-aware approach, with 80% used for training and 20% for testing, ensuring the model learns from past data to predict future values. This mimics real-world deployment scenarios where predictions must be made for unseen future timepoints.

5.4 Results and Performance

The model achieved strong performance metrics:

- R^2 score of 0.94, explaining 94% of variance in PM2.5 levels
- RMSE of 3.61 $\mu g/m^3$
- MAE of 1.15 $\mu g/m^3$

These metrics indicate the model can accurately predict PM2.5 concentrations within ranges useful for public health applications. Validation confirmed no overfitting, with training and test

performance remaining nearly identical. This robustness stems partly from our outlier handling approach, where z-score filtering removed extreme values that could dominate model training.

5.5 Feature Importance Analysis

The Random Forest model identified the most influential features for prediction:

Feature	Importance
rolling_mean_3h	0,21
diff_1h	0,17
diff_24h	0,12
lag_1h	0,11
rolling_mean_6h	0,11
rolling_std_6h	0,1
lag_3h	0,03
PM10	0,03
rolling_std_24h	0,03
lag_24h	0,02

Short-term features dominate the importance rankings. The 3-hour rolling mean (21.4%) and 1-hour differential (17.4%) are the strongest predictors, indicating PM2.5 levels are highly autocorrelated over short time periods. This aligns with atmospheric behavior where pollution accumulates and disperses gradually.

Interestingly, the PM2.5 to PM10 ratio feature, while having lower importance, offers insights into pollution sources. A rising ratio often indicates fresh combustion emissions (traffic) rather than resuspended dust, helping operators identify appropriate response strategies.

The combined importance of lag features and rolling statistics (over 60% total) confirms that recent pollution history is the primary driver of near-term predictions. Temporal features like hour and day of week contribute moderately, capturing regular daily and weekly cycles in emissions.

Chapter 6: Literature Analysis Using Machine Learning

6.1 Objective

This chapter analyzes academic literature on air quality research to understand where our Istanbul project fits within existing work. We processed 236 research papers from the OpenAlex database using TF-IDF, t-SNE, and anomaly detection methods.

6.2 Data Processing

Starting with 368 air quality papers which contained the word “Istanbul” in anywhere in their Full - Text, we successfully extracted text from 236 documents. The preprocessing removed page

numbers, URLs, citations, and common academic terms. Standard NLP techniques including stemming and stopword removal prepared the text for analysis.

6.3 TF-IDF Results

The TF-IDF analysis used 5,000 features with bigram support. Top terms across all papers were:

- pollut (pollution) - 17.25
- air qualiti - 11.90
- concentr (concentration) - 11.27
- model - 11.15
- data - 10.40

These terms confirm that pollution monitoring and modeling dominate the research focus.

6.4 t-SNE Visualization

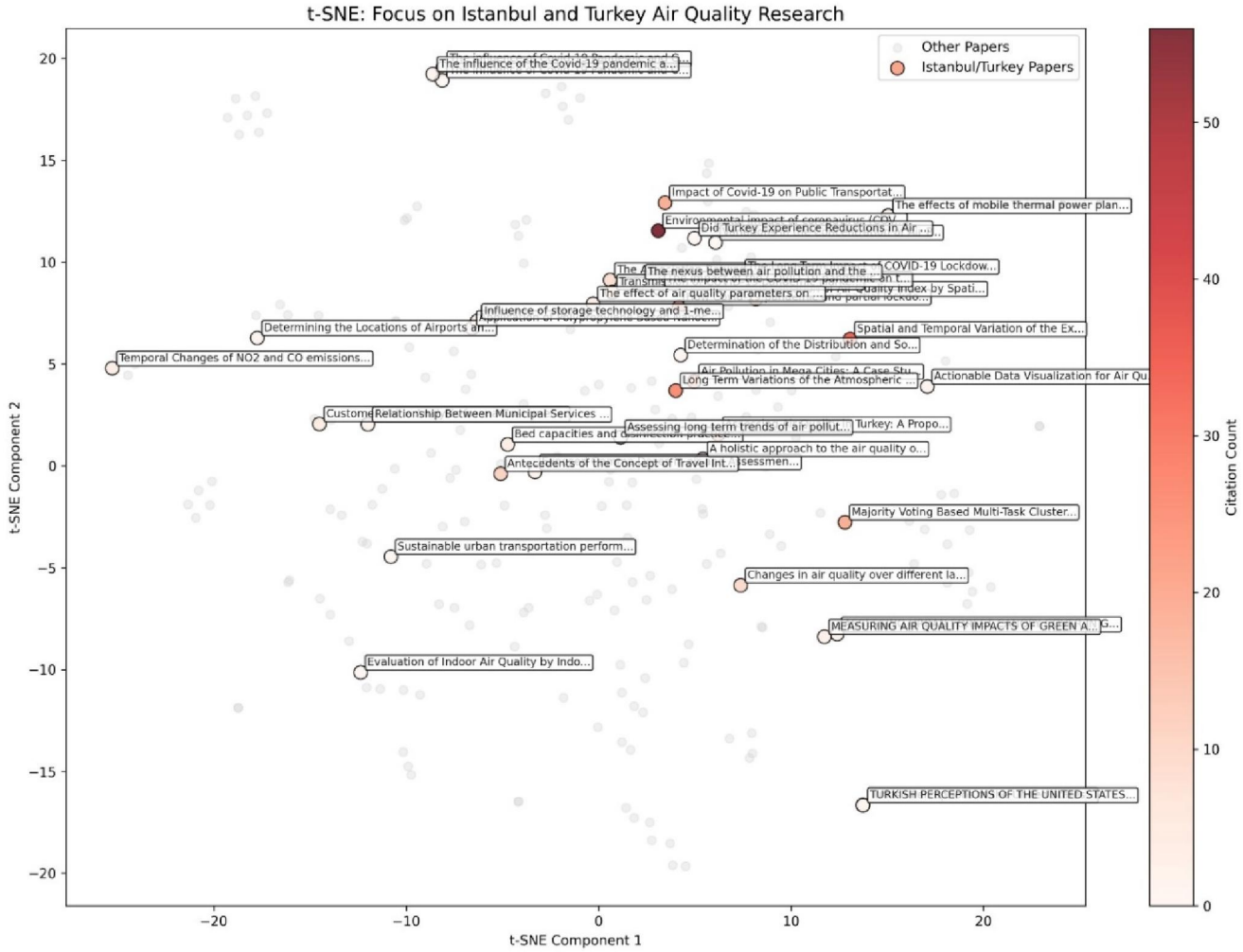
To visualize relationships between papers, we employed t-Distributed Stochastic Neighbor Embedding (t-SNE). The implementation first reduced dimensionality from 5,000 features to 50 components using PCA, capturing the primary variance before applying t-SNE for final 2D projection. The algorithm parameters included a perplexity of 30 and 1,000 iterations with learning rate set to auto for optimal convergence.

The visualization revealed several paper groupings:

- COVID-19 studies clustered in the upper region (Y ≈ +19)
- Turkey-specific papers in the right-upper area (X = +10 to +15)
- Technical methodology papers along the left side
- Policy studies in the lower region

6.5 Istanbul Research Coverage

42 papers (17.8%) focused on Istanbul or Turkey. These papers distributed across different research themes rather than clustering together, indicating diverse local research approaches.

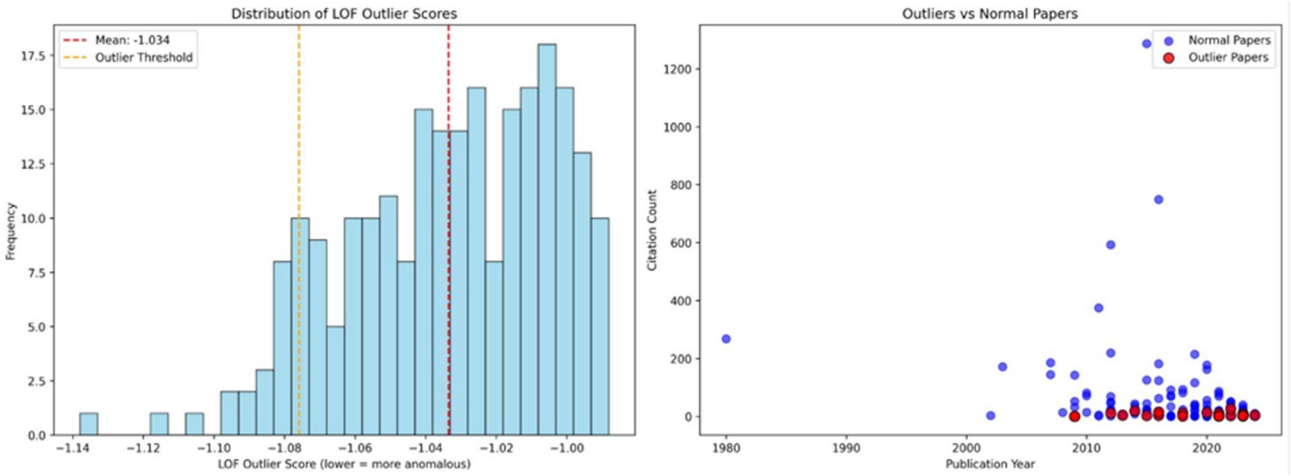


Istanbul papers achieved 30-45 citations on average, showing competitive research output. Topics ranged from traditional monitoring to emission modeling.

6.6 Anomaly Detection

Local Outlier Factor analysis (20 neighbors, 10% contamination) identified 24 unusual papers. Top outliers included:

1. Rice Straw Open Burning Environmental Effects (-1.138)
2. Tunnel Projects Environmental Assessment Model (-1.117)
3. Gamma Radiation in Gebze Region (-1.104)



Most anomalous papers appeared after 2015, suggesting increasing methodological diversity in recent years.

6.7 Clustering Analysis

K-means clustering (K=5) identified research themes:

- Cluster 0: Turkish language papers (6 papers)
- Cluster 1: COVID-19 and lockdown studies (58 papers)
- Cluster 2: Modeling and emissions - largest cluster (115 papers)
- Cluster 3: Aerosol and sampling methods (50 papers)
- Cluster 4: Off-topic international business papers (7 papers)

6.8 Technical Details

The analysis pipeline used standard configurations:

- TF-IDF: min_df=2, max_df=0.95, ngram_range=(1,2)
- PCA: First 50 components for t-SNE preprocessing
- t-SNE: Perplexity adjusted to dataset size
- LOF: Euclidean distance with MinMax scaling

Key implementation choices included using PCA before t-SNE due to high dimensionality and selecting K=5 for clustering based on coherent topic separation.

6.9 Findings for Istanbul Project

The literature analysis shows:

- 42 existing Istanbul studies provide local context
- Modeling approaches dominate current research (115 papers)
- Limited work combining IoT sensors with big data for Turkish cities
- Recent innovation trends favor interdisciplinary approaches

Conclusion

This analysis examined whether Istanbul functions as one city or multiple zones from an air pollution perspective. Using machine learning techniques on 9 million air quality measurements, we identified four distinct pollution zones through K-means clustering: a NO2-dominant urban core, PM10-heavy districts, the Sultangazi anomaly, and O3-dominant coastal areas.

Principal Component Analysis revealed three main pollution drivers explaining 67.8% of variance: traffic (29.3%), photochemical processes (21.7%), and industrial sources (16.8%). Correlation analysis confirmed expected atmospheric chemistry relationships, with strong NO2-CO correlation indicating traffic sources and negative NO2-O3 correlation explaining coastal ozone patterns.

A key finding was the infrastructure gap - 21 districts lack PM2.5 monitoring, limiting our understanding of pollution exposure in these areas. Among monitored districts, only 14.3% meet WHO guidelines. Our predictive model achieved 94% accuracy for PM2.5 forecasting, with short-term features proving most important for predictions.

Literature analysis of 236 papers showed 42 Istanbul-specific studies, though few integrate IoT sensors with big data approaches. This gap represents an opportunity for future work.

The results suggest Istanbul requires zone-specific policies rather than uniform city-wide approaches. Priority actions include expanding monitoring infrastructure to cover all districts and implementing the predictive system for proactive pollution management. Each of the four zones needs tailored interventions based on their distinct pollution profiles while working toward common air quality goals.

References

1. World Health Organization. (2021). *WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide.* Geneva: World Health Organization. World Health Organization
2. Turkish Statistical Institute (TÜİK). (2025, February 6). *Results of Address-Based Population Registration System 2024* (ABPRS). TÜİK Data Portal
3. Turkish Medical Association & Right to Clean Air Platform. (2024). *Dark Report 2024: Air Pollution and Health in Turkey.* Temiz Hava Hakkı
4. Daily Sabah. (2025, January 7). *Istanbul faces air quality challenges amid rising pollution.* Daily Sabah
5. Bianet. (2025, February 3). *Istanbul grapples with high air pollution levels.* Bianet
6. Carslaw, D. C., & Beevers, S. D. (2005). Evidence of an increasing NO_2/NO_x emissions ratio from road-traffic emissions. *Atmospheric Environment, 39*(3), 705–715. ScienceDirect
7. Turner, M. C., Krewski, D., Diver, W. R., et al. (2017). Ambient air pollution and cancer mortality in the American Cancer Society cohort. *Environmental Health Perspectives, 125*(8), 087013. ehp.niehs.nih.gov
8. Ministry of Environment, Urbanization and Climate Change of Türkiye. (2023). *Ulusal Hava Kalitesi İzleme Ağı (National Air Quality Monitoring Network).* e-Government of Turkey
9. OpenAlex. (2025, January 18). *OpenAlex API Technical Documentation — API Overview.* docs.openalex.org