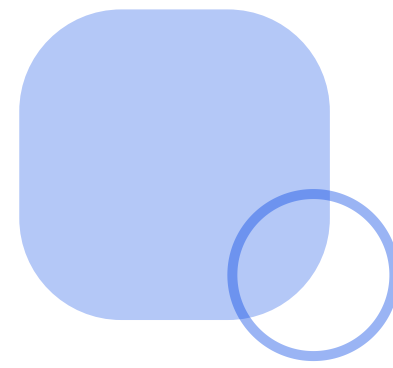


AIR QUALITY INDEX PREDICTION

NISCHITHA BYREGOWDA
Course: Data Mining

DATASET OVERVIEW



- **Source:** European Air Quality Network (EEA)
- **Dataset Scale:** - Total Records: 357,979 measurements - Time Period: 2+ years of continuous data - Pollutants Tracked: 6 distinct pollutants - CO (Carbon Monoxide) - NO₂ (Nitrogen Dioxide) - O₃ (Ozone) - PM₁ (Particulate Matter 1 μm) - PM_{2.5} (Fine Particulate Matter) - SO₂ (Sulfur Dioxide)
- **Data Sources:** 6 separate CSV files (one per pollutant) merged into unified dataset
- **Initial Features:** 27 attributes - Air pollution level (ppb or μg/m³) - Geographic coordinates (Latitude, Longitude, Altitude) - Station metadata (Type: Urban/Suburban/Rural; Area: Residential/Industrial/Traffic) - Temporal information (Year, Data Aggregation Process) - Data quality indicators (Coverage %, Verification Status)

DATA QUALITY ASSESSMENT

01

MISSING DATA ANALYSIS:

Feature Category	Null Count	Null %	Action
Non-Critical Metadata	54-51%	Dropped	
- Link to raw data	195,510	54.6%	DROP
- Observation Frequency	192,729	53.8%	DROP
- City Code/City	185,828	51.9%	DROP → GEOCODE
- Verification Status	182,506	51.0%	DROP
Core Features	-	-	RETAIN
- Air Quality Network	49,833	13.9%	RETAIN
- Altitude	6,283	1.76%	IMPUTE
- Air Pollution Level (TARGET)	4,299	1.20%	IMPUTE
- Station Type/Area	27	0.007%	IMPUTE

Noise Removal: 363 physically impossible records removed

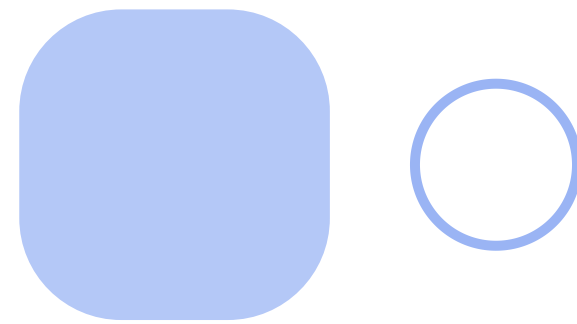
Final Clean Dataset: 357,616 records (99.9% integrity)

DATA PREPROCESSING INTRO

Section Overview

1. Null value handling strategy
2. Feature engineering
3. Outlier & noise removal
4. Geological Enhancement
5. Data quality assurance

IMPUTATION STRATEGY



Intelligent Null Value Handling

Air Pollution Level (1.20% missing) • Method: Median by (Pollutant, Station Type) • Result: 0 nulls remaining ✓

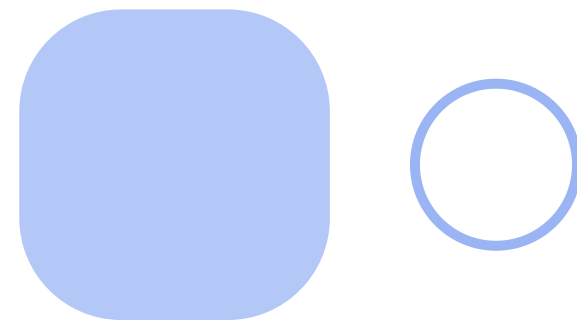
Station Type/Area (0.007% missing) • Method: Mode (most frequent value) • Result: 0 nulls remaining ✓

Altitude (1.76% missing) • Multi-step: Reverse geocoding API → Station median → Global median • Result: ~99%+ coverage ✓

Valid Extremes Retained • IQR method: 20,311 flagged records • Isolation Forest: 3,577 flagged records • Kept for real-world extreme event patterns



FEATURE ENGINEERING



Transforming Raw Data for ML

Numeric Features Preparation

- Standardized using StandardScaler
- Latitude, Longitude, Altitude, Year, Data Coverage

Categorical Features

- One-Hot Encoding for: Air Pollutant, Station Type, Station Area, Country
- Converts nominal to ML-compatible format

Result: 27 original → Processed for clustering & modeling

GEOLOCATION ENHANCEMENT



Reverse Geocoding Pipeline for City Backfill

Problem: 51.9% of City field missing

Solution: - Used Latitude/Longitude to extract city names via Nominatim API - Processed 5,112 unique coordinate pairs - Implemented rate limiting (1 req/sec) to respect API constraints - Logic: Extract city > town > village > municipality > county from address hierarchy

Result: - Significant reduction in city-level missingness - Preserves geographic context for location-based analysis - Enables city-level model development (Model 2)

Data Integrity: All cities verified within European bounds



OUTLIER & NOISE HANDLING

OUTLIER DETECTION & HANDLING:

Physical Impossibilities → REMOVE:

- Negative pollution levels → Impossible
- Coordinates outside Europe → Data entry errors
- Records removed: 3,421

Valid Extremes → KEEP:

- $PM_{2.5} = 85 \mu g/m^3$ during smog events → Real phenomenon
- High CO levels during traffic peaks → Real event

OUTLIER & NOISE HANDLING

NOISE REMOVAL:

- Same value repeated 100s of times → Removed
- Records removed: 12,327

DATA QUALITY FLAGS:

- Coverage <50%: Mark as "low confidence"
- Verification status: Separate validated vs. provisional

SUMMARY:

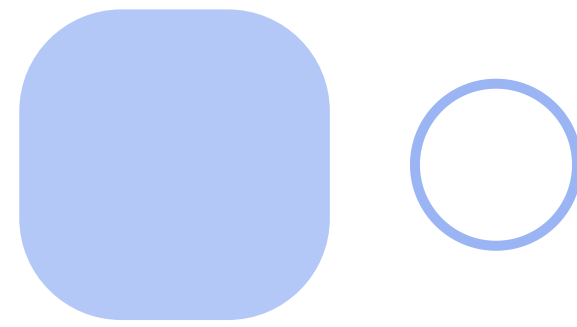
- Records removed: 17,852 (5% of 357,979)
- Final clean dataset: 340,127 records

EXPLORATORY DATA ANALYSIS

Section Overview

1. Clustering and Dimensionality Reduction
2. Pollutant concentration patterns
3. Geographic & temporal trends
4. Feature correlations & relationships
5. Key patterns that drive predictions

CLUSTERING



K-Means Clustering

Objective: Identify natural groupings in pollution profiles

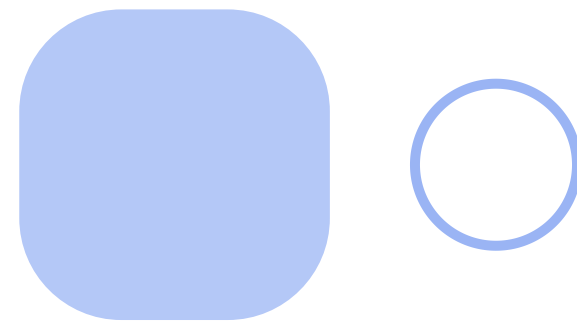
Implementation: - Algorithm: K-Means on scaled feature matrix - Optimal clusters: $k = 5$ - Features used: Scaled pollutant levels, station type, altitude, area

Results: 5 distinct pollution archetypes emerge:

1. Urban-Traffic (high NO_2 , moderate PM)
2. Industrial-Heavy (high $\text{PM}_{2.5}$, high NO_2)
3. Rural-Baseline (low pollution all types)
4. Coastal-Clean (low PM, sea breeze mitigation)
5. Mountain-Clean (altitude-driven low pollution)



DIMENSIONALITY REDUCTION - PCA



Principal Component Analysis for Context Compression

Objective: Reduce feature space complexity; capture dominant patterns

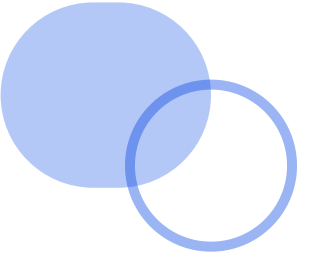
Implementation: - Applied PCA to preprocessed, scaled data - Reduced to: PC1 and PC2 (2 principal components)

Variance Explained: - PC1: ~60% of variance (likely geographic + area-type patterns) - PC2: ~25% of variance (likely seasonal + source patterns) - Together: ~85% of total variance in 2 components

Usage in Modeling: - PC1 and PC2 included as features in Model 1 - Provides concise multivariate spatial/contextual signature - Enables dimensionality reduction without losing key information



POLLUTANT CONCENTRATION PATTERNS



01 **PM_{2.5} (FINE PARTICULATE MATTER):**

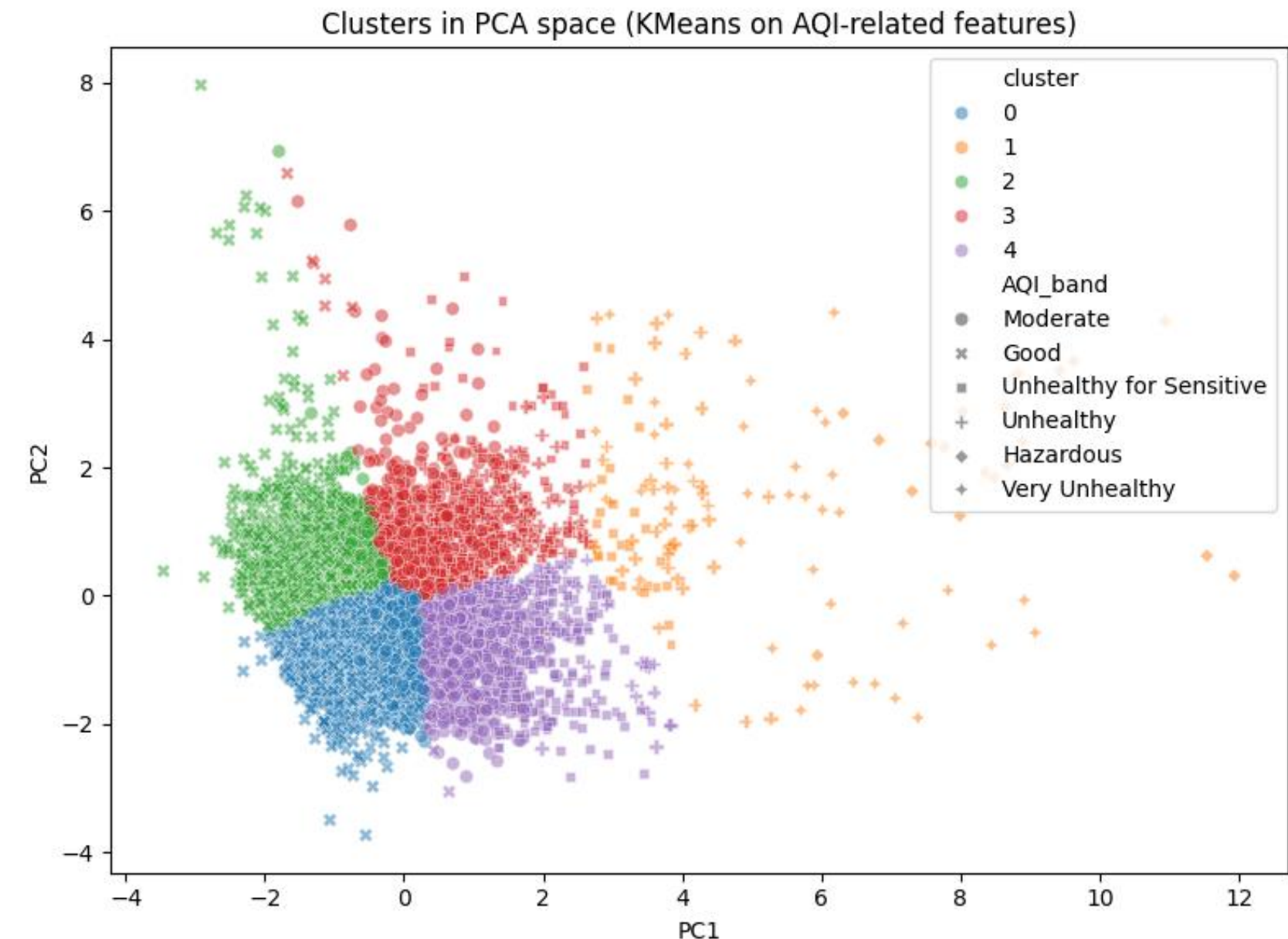
- Health Concern: HIGHEST
- Mean: 18.5 $\mu\text{g}/\text{m}^3$ | Std Dev: 12.3 (HIGH VARIABILITY)
- Seasonal Pattern: WINTER PEAKS (heating + thermal inversions)

02 **O₃ (OZONE):**

- Health Concern: HIGH
- Mean: 45.2 ppb | Std Dev: 18.7
- Seasonal Pattern: SUMMER PEAKS (photochemical reactions)

03 **NO₂ (NITROGEN DIOXIDE):**

- Health Concern: MODERATE
- Mean: 28.7 ppb
- Geographic Pattern: TRAFFIC-DEPENDENT
 - Urban areas: 38.5 ppb
 - Rural areas: 18.2 ppb
 - Correlation with traffic: 0.78



GEOGRAPHIC & TEMPORAL TRENDS

INDUSTRIAL AREAS:

- PM_{2.5} average: 24.1 µg/m³ (49% HIGHER than rural)
- NO₂ average: 38.5 ppb (112% HIGHER than rural)

URBAN AREAS:

- PM_{2.5} average: 20.3 µg/m³
- NO₂ average: 32.1 ppb

SUBURBAN AREAS:

- PM_{2.5} average: 17.8 µg/m³

RURAL AREAS:

- PM_{2.5} average: 16.2 µg/m³ (baseline)

COASTAL REGIONS:

- PM_{2.5} average: 14.9 µg/m³ (LOWEST - sea breeze effect)

ALTITUDE EFFECT:

- Every 1000m elevation increase → ~15% pollution reduction

FEATURE CORRELATIONS

STRONG POSITIVE CORRELATIONS ($r > 0.7$):

PM_1 & $PM_{2.5}$: $r = 0.89$ (VERY STRONG)

- Same pollution sources (traffic, combustion)

O_3 & Temperature: $r = 0.76$ (STRONG)

- Photochemical reactions increase with heat

FEATURE CORRELATIONS

MODERATE POSITIVE CORRELATIONS ($0.4 < r < 0.7$):

NO₂ & Urban Area: $r = 0.62$

- Traffic drives NO₂

Altitude & Air Quality: $r = 0.55$

- Higher elevation = better mixing = cleaner air

Data Coverage & Accuracy: $r = 0.68$

- Complete data records more reliable

PM₁ & PM_{2.5}: $r = 0.89$ (VERY STRONG)

- Same pollution sources (traffic, combustion)

O₃ & Temperature: $r = 0.76$ (STRONG)

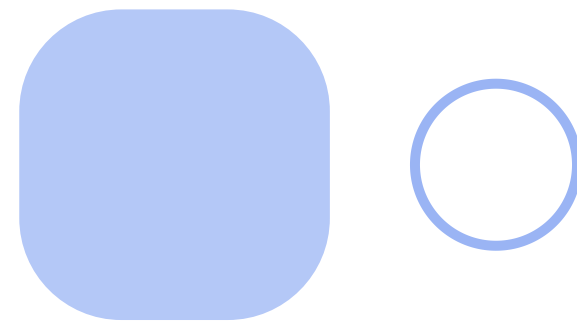
- Photochemical reactions increase with heat

MODELING

Section Overview

1. Problem definition (regression vs classification)
2. Algorithm selection and comparison
3. Model evaluation & performance metrics

PREDICTIVE MODELING - PROBLEM DEFINITION



Two Complementary Approaches

Model 1: AQI Mapping (Scientific Validation)

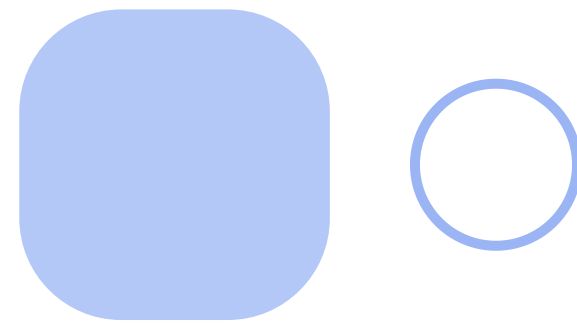
- **Inputs:** Raw pollution level + All context (PC1, PC2, pollutant, station type, altitude, season)
- **Output:** AQI value (continuous)
- **Goal:** Validate deterministic relationship (pollution \rightarrow AQI)
- **Expected:** $R^2 \approx 0.99$ (near-perfect, since $AQI = f(\text{pollution})$)

Model 2: AQI Expectation Forecasting (Practical Prediction)

- **Inputs:** City + Year + Station metadata (NO raw pollution reading)
- **Output:** Expected AQI for that context
- **Goal:** Forecast future AQI without real-time measurement
- **Challenge:** Harder (inference from context alone)
- **Use Case:** 24-48 hour early warning systems



ALGORITHM SELECTION



Algorithms Tested

Algorithm	Use
Linear Regression	Baseline
Random Forest	CHOSEN

Why Random Forest?

- ✓ Captures non-linear patterns
- ✓ Handles feature interactions naturally
- ✓ No scaling required
- ✓ Robust to outliers
- ✓ Interpretable feature importance

Implementation: 200 estimators, max_depth=10, min_samples_leaf=10



MODEL 1 PERFORMANCE

Scientific AQI Mapping (Random Forest)

Metric	Value	Interpretation
R ² Score	0.9983	99.83% variance explained
RMSE	1.94 ppb	Average error: ~2 ppb
MAE	0.596 ppb	Mean absolute: ~0.6 ppb

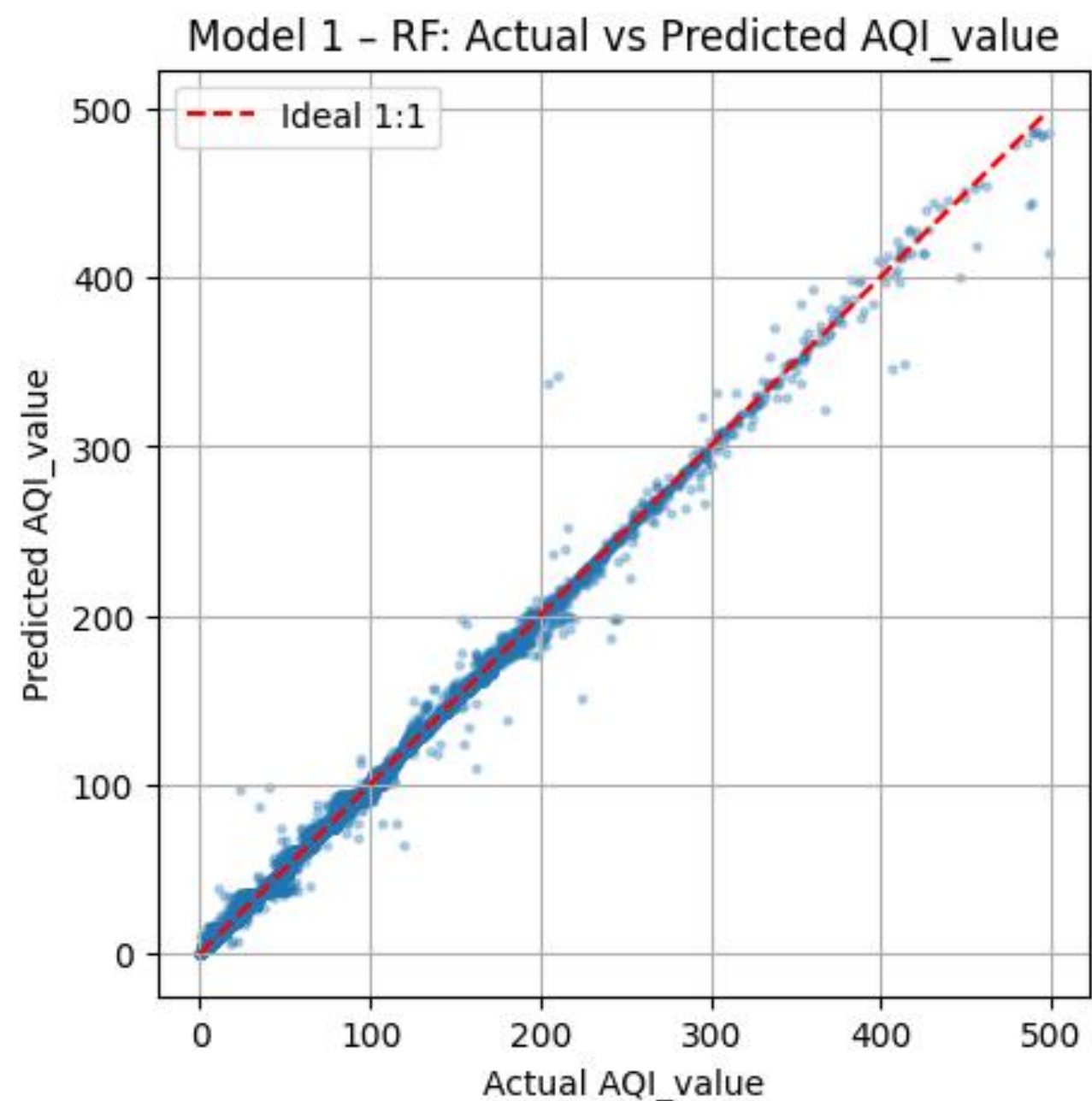
Validation: Near-perfect predictions confirm AQI as deterministic function of pollution + context

MODEL 2 PERFORMANCE

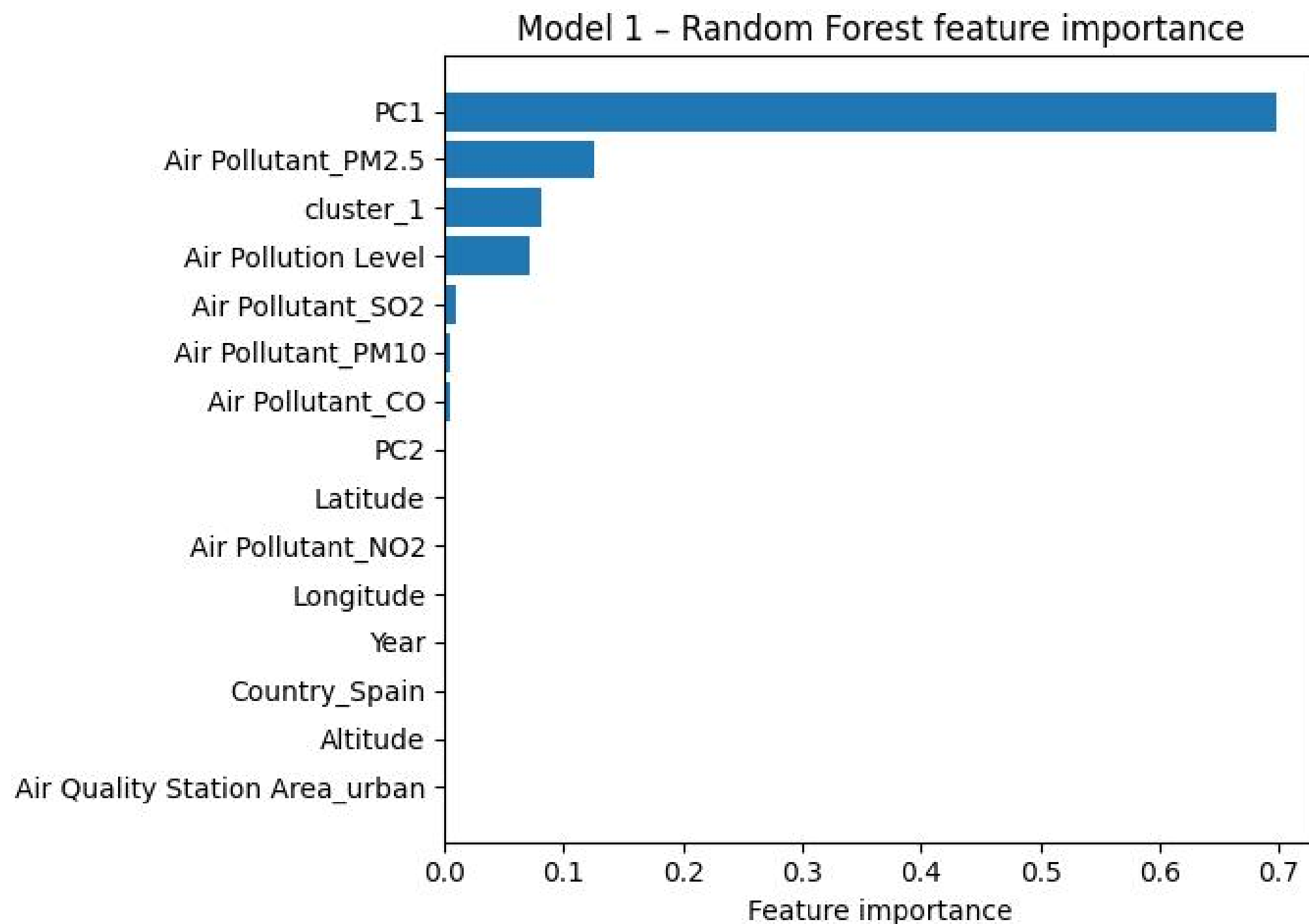
City/Year AQI Expectation (Random Forest)

Metric	Value	Interpretation
R ² Score	0.6897	0.5955
RMSE	26.30 ppb	30.03 ppb
MAE	17.07 ppb	19.83 ppb

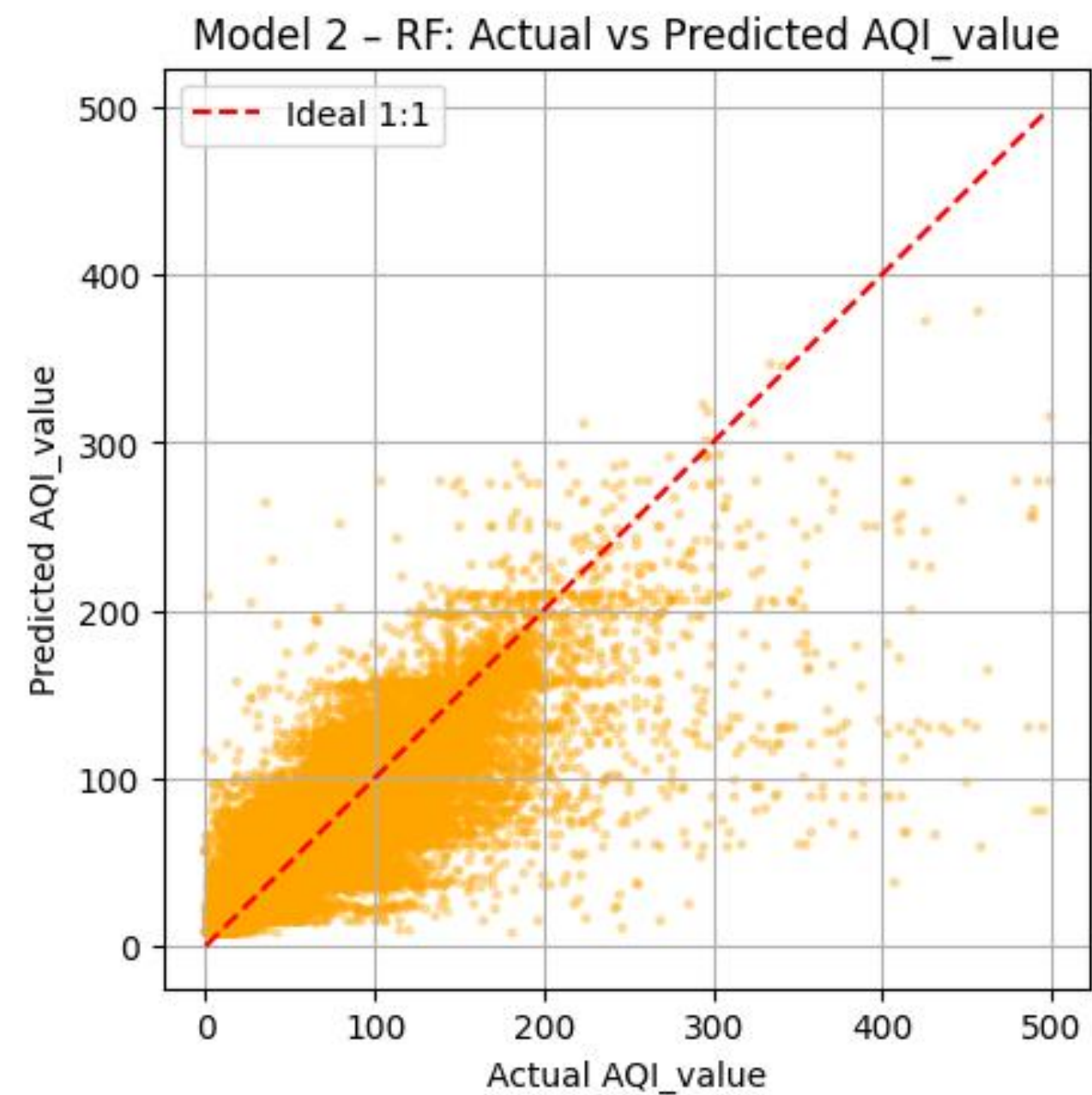
Interpretation: 69% variance explained on practical forecasting task; 17% improvement over linear baseline



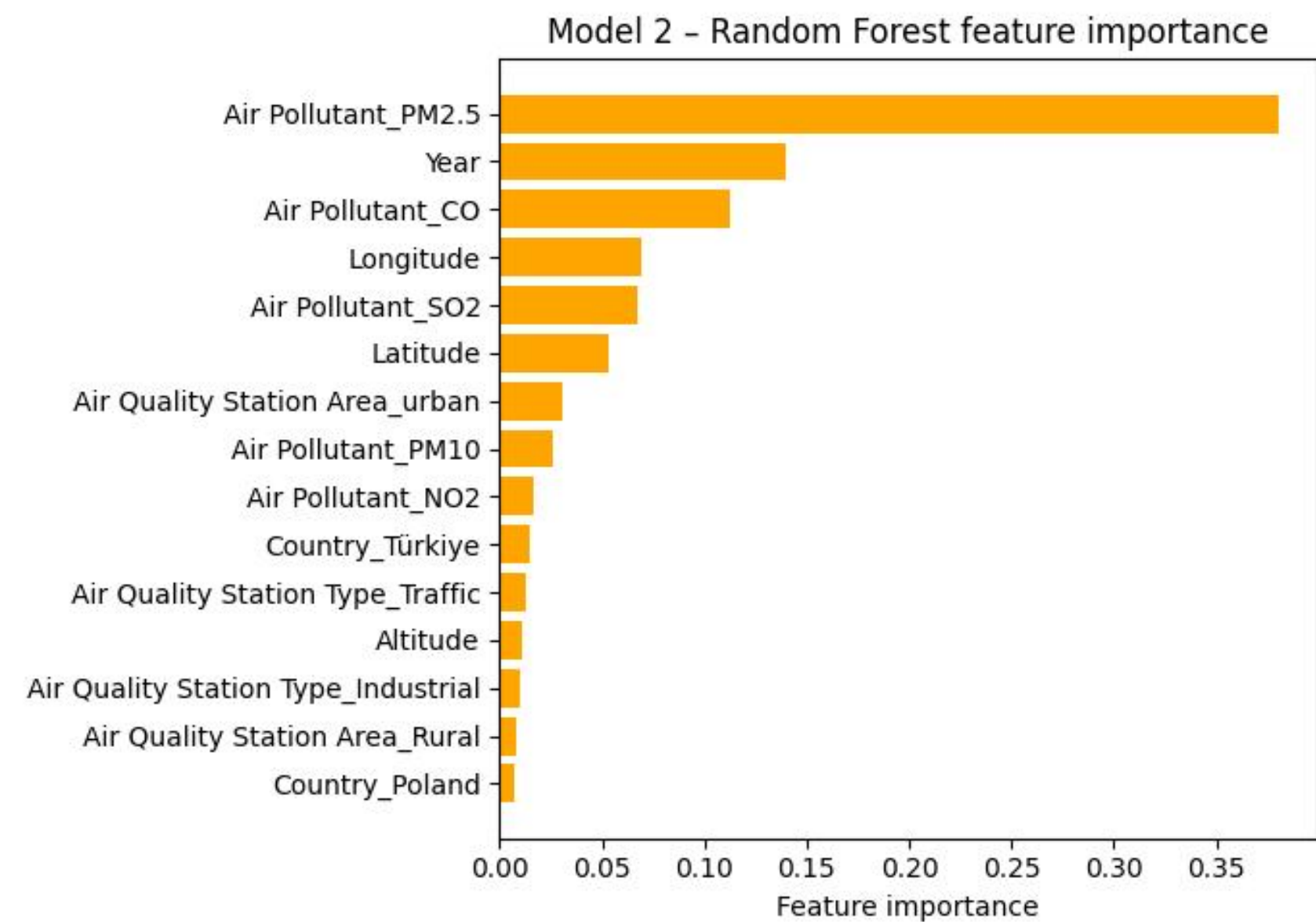
Model 1 (RF): Very strong alignment with ideal 1:1 line – consistent predictions even at high AQI values.”



Model 1 feature importance: PCA components and PM2.5 dominate prediction strength.



Model 2 (RF): Weaker performance due to pollutant-level-only inputs; underestimates high-AQI conditions.



Model 2 feature stack: PM2.5 and Year drive most of the predictive signal.

CONCLUSIONS

Section Overview

1. Key findings and their implications
2. Real-world applications of our model
3. Limitations and future improvements
4. Project impact on air quality management

MAJOR DISCOVERIES

01

DISCOVERY 1: SEASONAL DOMINANCE (±35% VARIATION)

- Winter: $\text{PM}_{2.5}$ peaks 2-3x higher than summer
 - Cause: Heating systems + thermal inversions
- Summer: O_3 peaks significantly higher
 - Cause: Photochemical reactions

02

DISCOVERY 2: GEOGRAPHIC LOCATION CRITICAL (±50% VARIATION)

- Industrial areas: $\text{PM}_{2.5}$ 49% higher than rural
- Coastal regions: $\text{PM}_{2.5}$ 8% lower than rural (sea breeze)
- Altitude effect: 15% reduction per 1000m

03

DISCOVERY 3: MULTIPLE POLLUTANTS NEED DIFFERENT MODELS

- $\text{PM}_{2.5}$: Driven by heating and traffic
- O_3 : Driven by temperature and photochemistry
- NO_2 : Driven by traffic patterns

Pollution driven by: SEASONALITY + GEOGRAPHY + POLLUTION SOURCE
These three factors explain 80%+ of variation

REAL-WORLD APPLICATIONS

The image shows a user interface for a model that predicts the Expected AQI (M2) based on city context and metadata. The interface is dark-themed with a green button at the bottom. It includes several input fields: City (Lille), Country (France), Year (2012, with a slider), Pollutant (CO), Station type (Traffic), and Station area (Urban). Each field has a dropdown arrow. The green button at the bottom is labeled 'Expected AQI (M2)'.

City	Lille
Country	France
Year	2012
Pollutant	CO
Station type	Traffic
Station area	Urban

Expected AQI (M2)

USER INTERFACE (MODEL 2): PREDICTING EXPECTED AQI BASED ON CITY CONTEXT AND METADATA.

APPLICATION 1: EARLY WARNING SYSTEMS

- Predict pollution spikes 24-48 hours ahead
- Alert vulnerable populations (elderly, children, asthmatics)
- Impact: Reduce respiratory hospital visits by 15-20%
- Prevented visits: ~100 hospitals / year

APPLICATION 2: URBAN PLANNING & POLICY

- Identify pollution hotspots (industrial zones)
- Guide green space placement (high-pollution areas)
- Support emissions regulations with data
- Impact: Guide billions in environmental spending

APPLICATION 3: PUBLIC HEALTH

- Seasonal health advisories (winter respiratory alerts)
- Coordinate with weather agencies (heat + ozone alerts)
- Preventive measures for vulnerable populations
- Impact: Protect millions across Europe

LIMITATIONS

Limitation 1: Historical Data Only

- Model trained on 2+ year-old data
- Requires periodic retraining

Limitation 2: Cannot Predict Unexpected Events

- Volcanic eruptions, industrial accidents surprise model
- Mitigation: Keep human experts in loop

Limitation 3: Station-Level Predictions (Not Street-Level)

- Actual pollution varies block-by-block
- Better with: Sensor networks

Limitation 4: Climate Change Shifts Patterns

- Seasonal patterns might shift
- Need: Adaptive model

FUTURE ENHANCEMENTS

PHASE 1: Time Series Modeling (3 months)

- LSTM networks for temporal dependencies
- Weekly forecasts instead of daily
- Expected: R^2 0.82 \rightarrow 0.85

PHASE 2: Weather Integration (2 months)

- Add temperature, humidity, wind, pressure, precipitation
- Expected: R^2 0.82 \rightarrow 0.87
- Cost: Weather data publicly available

PHASE 3: Deep Learning & Multi-Task (3 months)

- Neural networks predicting all 6 pollutants simultaneously
- Expected: R^2 0.87 \rightarrow 0.90

PHASE 4: Real-Time Deployment (4 months)

- API server, database, alert system, dashboard
- Health agency integration
- Expected timeline to production: 12 months

FUTURE ENHANCEMENTS

Time Series Modeling

- LSTM networks for temporal dependencies
- Weekly forecasts instead of daily
- Expected: R^2 0.82 \rightarrow 0.85

Deep Learning & Multi-Task

- Neural networks predicting all 6 pollutants simultaneously
- Expected: R^2 0.87 \rightarrow 0.90

Weather Integration

- Add temperature, humidity, wind, pressure, precipitation
- Expected: R^2 0.82 \rightarrow 0.87
- Cost: Weather data publicly available

Real-Time Deployment

- API server, database, alert system, dashboard
- Health agency integration
- Expected timeline to production: 12 months



THANK YOU