PUBLIC HEALTH
OpenJournal

Openventio
PUBLISHERS

Original Research Article

Statistical Appraisal of the Global Air Pollution Dataset: A Review and Empirical Re-analysis of City-Level Air Quality Indices

Raj Shekar M[1,2], Bhavana Kabber[3], Sujith Arokia Swamy B[1,2], Harish Nayak[4,*] and Naveen Kumar P[3]

[1]The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India – 110012
[2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India – 110012
[3]University of Agricultural Sciences, Raichur, Karnataka, India – 584104
[4]University of Agricultural Sciences, Dharwad, Karnataka, India – 580005

Corresponding Author
Harish Nayak
Email: harishnayak626@gmail.com

Article Information:
Received: 07 December 2025 | Revised: 09 January 2026 | Accepted: 13 January 2026 | Published: January 15, 2026

Cite this article:
Raj Shekar M, Bhavana Kabber, Sujith Arokia Swamy B, Harish Nayak and Naveen Kumar P (2025). Statistical Appraisal of the Global Air Pollution Dataset: A Review and Empirical Re-analysis of City-Level Air Quality Indices. *Public Health Open Journal*. 10(2):34–44. https://doi.org/10.17140/PHOJ.10.02.34

Abstract

Ambient air pollution is a leading environmental risk factor for premature mortality, and city-level air quality indices (AQI) are widely used to communicate exposure. The Global Air Pollution Dataset records the overall AQI and four pollutant sub-indices (PM2.5, ozone, nitrogen dioxide and carbon monoxide) for 23,463 cities in 175 countries, yet its statistical structure and the validity of analyses built on it are rarely examined together. This review combines a narrative synthesis of the literature with a full statistical re-analysis of the supplied file. We audit data quality and internal consistency, describe distributions, quantify pollutant dominance, correlation and relative importance, measure between-country heterogeneity, and evaluate predictive models under two validation designs. Our synthesis finds that (i) the overall AQI equals the maximum sub-index in 99.4% of records, and the 141 exceptions all occur at extreme PM2.5 levels; (ii) PM2.5 is the highest sub-index in 78.9% of cities and in every city rated Hazardous, and alone explains 96.9% of AQI variance; (iii) country of location explains about half of the variation in log AQI (intraclass correlation 0.54; Kruskal-Wallis $\epsilon^2 = 0.49$), with median AQI ranging from 20 (Bolivia) to 169 (Pakistan); and (iv) when country information is allowed to leak across cross-validation folds, the AUC for predicting unhealthy air from non-PM2.5 pollutants rises from 0.84 to 0.97, an artefact of the validation design. We propose a short reporting checklist and six priority research directions.

Keywords: Air quality index (AQI), PM2.5, ozone, between-country heterogeneity, Kruskal-Wallis test, intra-class correlation, grouped cross-validation, narrative review

1. Introduction

1.1 Motivation

Exposure to ambient air pollution, particularly fine particulate matter (PM2.5), is linked to cardiovascular and respiratory disease and to a very large global burden of premature death (Pope and Dockery, 2006; Cohen et al., 2017; Burnett et al., 2018; Landrigan et al., 2018). To communicate exposure to the public, monitoring agencies convert pollutant concentrations into unit-free indices whose categories range from "Good" to "Hazardous" (US EPA, 2018). Because the index is a transformation of concentration, any statistical analysis of AQI data must respect its construction: the overall AQI is the maximum of the pollutant sub-indices, the scale is capped, and values are strongly clustered by geography.

The Global Air Pollution Dataset is a compact, city-level table that is easy to download and analyse, and it is often used in teaching, dashboards and introductory machine-learning examples. Its simplicity hides several statistical features that influence conclusions: extreme skewness, near-constant sub-indices for two of the four pollutants, a deterministic relationship between the overall AQI and its components, strongly unequal country sample sizes, and clustering of cities within countries that can inflate the apparent accuracy of predictive models if ignored. A statistical appraisal is therefore useful before the data are used for inference or prediction.

1.2 Scope and Contribution

This article is a narrative review supported by a complete statistical re-analysis of the supplied dataset. Our contributions are fourfold:

A structured description of the dataset, including a quantitative audit of internal consistency between the overall AQI, its sub-indices and its category labels.

A battery of statistical tools suited to heavily skewed, hierarchical data (Jarque-Bera, Spearman correlation, Kruskal-Wallis, intraclass correlation, Wilson intervals, robust regression and Shapley-type relative importance).

A demonstration, using grouped and random cross-validation side by side, of how ignoring the country hierarchy inflates predictive performance.

A reporting checklist and six priority research directions for future work with this dataset.

1.3 Review Questions

RQ1: What is the structure and internal consistency of the dataset?

RQ2: How are the pollutant sub-indices distributed, and which pollutant determines the overall AQI?

RQ3: How strongly are the sub-indices associated, and what is each pollutant's relative contribution to the overall AQI?

RQ4: How much of the variation in AQI lies between countries, and which countries are at the extremes?

RQ5: Can unhealthy air be predicted from the other pollutants, and how much does the validation design matter?

2. Background and Theoretical Foundations

2.1 The Air Quality Index and the Structure of the Dataset

The file has 23,463 rows, one per city, and 12 columns: country, city, and for each of the overall AQI and four pollutants (CO, ozone, NO_2 and PM2.5) a numeric index value and a category label. The category labels correspond to the US EPA scheme (Table 2.1; US EPA, 2018). BP = [(I<sub>hi < / sub> I < sub>lo < / sub>) / (BP<sub>hi < / sub> - BP < sub>lo < / sub>)] (C - BP < sub>lo < / sub>) - BP < sub> + I_{lo}, and the overall AQI is the largest sub-index. The file does not document its data source, measurement date or averaging period, which limits any claim about the time period or the representativeness of the readings.

Table 2.1: AQI categories (US EPA scheme)

AQI range	Category	Health interpretation
0-50	Good	Air quality is satisfactory
51-100	Moderate	Acceptable; some concern for very sensitive people
101-150	Unhealthy for Sensitive Groups	Sensitive groups may experience effects
151-200	Unhealthy	General population may be affected
201-300	Very Unhealthy	Health alert; everyone may be affected
301-500	Hazardous	Emergency conditions

2.2 Health and Policy Context

Long-term PM2.5 exposure has been associated with excess mortality across a wide concentration range, and global burden estimates attribute millions of deaths a year to ambient particulate pollution (Cohen et al., 2017; Burnett et al., 2018). Reflecting this evidence, the World Health Organization (2021) tightened its recommended guideline values, with the PM2.5 guideline

now far below the level at which the EPA scheme first leaves the "Good" category. Index-based categories are therefore a coarse communication tool and not a substitute for concentration data in exposure analysis.

2.3 Statistical Tools for Skewed and Hierarchical Data

Normality was assessed with skewness, excess kurtosis and the Jarque-Bera test (Jarque and Bera, 1987); the Shapiro-Wilk test is not suitable for samples above about 5,000. Because the sub-indices are highly skewed and contain many ties, monotone associations were measured with Spearman's rank correlation (Spearman, 1904), with approximate 95% confidence intervals from Fisher's z transformation. Agreement between category labels was quantified with Cohen's kappa (Cohen, 1960), interpreted with the benchmarks of Landis and Koch (1977). Differences across countries were tested with the Kruskal-Wallis test (Kruskal and Wallis, 1952), with epsilon-squared (ϵ^2) as an effect size (Tomczak and Tomczak, 2014). The share of variation lying between countries was expressed as the intra-class correlation (ICC) from a one-way random-effects decomposition (Shrout and Fleiss, 1979; Gelman and Hill, 2007), and proportions of unhealthy cities were given with Wilson score intervals (Wilson, 1927).

2.4 Regression and Relative Importance

The relation between the overall AQI and its sub-indices was examined with ordinary least squares using heteroscedasticity-consistent (HC3) standard errors (MacKinnon and White, 1985); heteroscedasticity was checked with the Breusch-Pagan test (Breusch and Pagan, 1979) and collinearity with variance inflation factors. Each predictor's share of the explained variance was obtained by the LMG (Shapley-value) decomposition, which averages a predictor's incremental R^2 over all orderings of the predictors (Lindeman et al., 1980; Grömping, 2006).

2.5 Validation with Clustered Data

Observations from the same country are not independent. When random cross-validation places cities from one country in both the training and the test folds, a flexible model can learn country-specific signatures and report performance that will not carry over to new countries (Roberts et al., 2017). Grouped cross-validation, which holds out entire countries, is the appropriate check when the goal is generalisation to unseen locations. A related hazard is target encoding of a high-cardinality variable such as country (Micci-Barreca, 2001), which is legitimate only if it is estimated within the training fold. Model comparisons were tested with the corrected resampled t-test of Nadeau and Bengio (2003).

3. Methodology of the Review

3.1 Review Design

The study has two components. The literature component is a selective narrative review of foundational work on air quality indices, health effects and the statistical methods applied; it does not claim the exhaustive search and screening counts of a formal PRISMA systematic review. The quantitative component is a full re-analysis of the supplied file, in which every figure and table in Sections 4 and 5 was computed directly from the data.

3.2 Data and Pre-processing

All 23,463 records were retained. No values were imputed or removed. 427 records (1.8%) have no country and one record has no city name; the former were excluded from country-level analyses only. For country comparisons, countries with at least 30 cities were retained (85 countries, 21,972 cities, 93.6% of the file), with thresholds of 20 and 50 cities used as sensitivity checks. A city was classified as having unhealthy air for sensitive groups or worse when its overall AQI exceeded 100 (4,244 cities, 18.1%), following the EPA ranges in Table 2.1.

3.3 Statistical Analysis Plan

The analysis proceeded in six stages: (i) consistency audit of the AQI, its sub-indices and its labels; (ii) descriptive statistics and distributional tests; (iii) pollutant dominance and category agreement; (iv) correlation and regression with relative-importance decomposition; (v) between-country heterogeneity (Kruskal-Wallis, ICC, chi-square on the unhealthy share); and (vi) predictive comparison under random and country-grouped validation.

3.4 Validation of Predictive Models

Two prediction tasks were defined. In the first, the PM2.5 sub-index was predicted from the CO, ozone and NO_2 sub-indices by linear regression, ridge regression, k-nearest neighbours (k = 15), random forest (200 trees) and gradient boosting, using ten-fold random cross-validation (seed 42), and R^2, RMSE and MAE were reported. In the second, unhealthy air (overall AQI above 100) was predicted from the same three non-PM2.5 sub-indices by logistic regression, random forest and gradient boosting, using ten-fold stratified random cross-validation and ten-fold grouped cross-validation by country (cities without a country forming their own group). PM2.5 was withheld from the second task because it almost completely determines the outcome. To demonstrate leakage, gradient boosting was also fitted with a smoothed country target encoding esti-

mated inside each training fold. Analyses used Python with SciPy, statsmodels and scikit-learn (Seabold and Perktold, 2010; Pedregosa et al., 2011), and reporting follows the spirit of the TRIPOD statement (Collins et al., 2015).

4. Structure and Quality of the Dataset

4.1 Variables

Table 4.1: Variables in the dataset

Variable	Description	Type
Country	Country name (175 distinct, 427 missing)	Categorical
City	City name (23,462 distinct, 1 missing)	Identifier
AQI Value	Overall AQI (maximum of sub-indices)	Integer
AQI Category	EPA category of the overall AQI	Ordinal (6)
CO / Ozone / NO_2 / PM2.5 AQI Value	Pollutant sub-index values	Integer
CO / Ozone / NO_2 / PM2.5 AQI Category	EPA category of each sub-index	Ordinal

4.2 Consistency and Quality Audit

Table 4.2: Data-quality and consistency audit

Check	Result	Share of records
Records × columns	23,463 × 12	-
Duplicate rows	0	0.0%
Missing country	427	1.8%
Missing city	1	0.004%
Overall AQI equals maximum sub-index	23,322	99.4%
Overall AQI exceeds all sub-indices	141	0.6%
Overall AQI below maximum sub-index	0	0.0%
Boundary values labelled in higher class (AQI = 100, 150, 300)	148	0.6%
NO_2 sub-index equal to zero	8,242	35.1%
CO sub-index equal to zero	3,110	13.3%

The file contains no duplicates and no missing values in any numeric column. Two internal inconsistencies deserve documentation. First, the overall AQI never falls below the largest sub-index, but in 141 records (0.6%) it exceeds all four, by 56 points on average and up to 166. Every such record has a PM2.5 sub-index of at least 208 (median 386) and 135 of the 141 are rated Hazardous, so the discrepancy is confined to extreme particulate readings. It is consistent with the overall AQI being computed or capped differently from the displayed sub-index (the AQI reaches 500 in 103 cities while the PM2.5 sub-index reaches 500 in only 31), but the file does not allow this to be verified. Second, 148 records whose AQI is exactly 100, 150 or 300 (52, 95 and 1 respectively) are labelled with the higher category, whereas the EPA ranges in Table 2.1 place them in the lower one. This affects only 0.6% of rows but would change counts of "unhealthy" cities if labels were used instead of values.

4.3 Descriptive Statistics and Distributional Shape

Table 4.3: Descriptive statistics of the AQI values (n = 23,463)

Variable	Mean	SD	Median	Q1-Q3	Range	Skew	Kurt.	Skew of log(1+x)
Overall AQI	72.0	56.1	55	39-79	6-500	3.29	17.6	0.59
PM2.5 AQI	68.5	54.8	54	35-79	0-500	2.82	13.6	0.05
Ozone AQI	35.2	28.1	31	21-40	0-235	3.10	12.1	0.91
NO_2 AQI	3.1	5.3	1	0-4	0-91	3.79	22.4	0.72
CO AQI	1.4	1.8	1	1-1	0-133	23.08	1289.1	0.80

The Jarque-Bera test rejected normality for every variable (all $p < 0.001$), and the departures are large: skewness ranges from 2.82 (PM2.5) to 23.1 (CO), and the CO sub-index has an excess kurtosis above 1,000 because 99.99% of cities sit in the "Good" category with a handful of outliers up to 133. The NO_2 and CO sub-indices are almost constant on the index scale (median 1 for both; 35.1% of NO_2 values are zero). A log(1+x) transformation brings the skewness of PM2.5 to 0.05 and of the overall AQI to 0.59 (Figure 1), which supports using the log scale for parametric summaries of AQI and rank-based methods otherwise.

Table 4.4: Number (%) of cities in each AQI category

Category	Overall AQI	PM2.5	Ozone	NO_2	CO
Good	9,936 (42.3%)	10,208 (43.5%)	21,069 (89.8%)	23,448 (99.9%)	23,460 (100.0%)
Moderate	9,231 (39.3%)	9,075 (38.7%)	1,445 (6.2%)	15 (0.1%)	2 (0.0%)
Unhealthy for Sensitive Groups	1,591 (6.8%)	1,624 (6.9%)	491 (2.1%)	0 (0.0%)	1 (0.0%)
Unhealthy	2,227 (9.5%)	2,129 (9.1%)	405 (1.7%)	0 (0.0%)	0 (0.0%)
Very Unhealthy	287 (1.2%)	255 (1.1%)	53 (0.2%)	0 (0.0%)	0 (0.0%)
Hazardous	191 (0.8%)	172 (0.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Overall, 42.3% of cities are rated Good, 39.3% Moderate and 18.1% fall above 100. The category tables show that the CO and NO_2 sub-indices essentially never leave the Good category (23,460 and 23,448 cities respectively), whereas PM2.5 closely mirrors the overall AQI and ozone contributes a smaller but real tail (2,394 cities at Moderate or worse).

5. Empirical Evidence from the Dataset

5.1 Pollutant Dominance and Category Agreement

Table 5.1: Pollutant with the highest sub-index, by overall AQI category

Overall AQI category	Cities (n)	PM2.5 highest, n (%)	Ozone highest, n (%)
Good	9,936	5,898 (59.4%)	4,037 (40.6%)
Moderate	9,231	8,799 (95.3%)	432 (4.7%)
Unhealthy for Sensitive Groups	1,591	1,446 (90.9%)	145 (9.1%)
Unhealthy	2,227	1,953 (87.7%)	274 (12.3%)
Very Unhealthy	287	236 (82.2%)	51 (17.8%)
Hazardous	191	191 (100.0%)	0 (0.0%)
All cities	23,463	18,523 (78.9%)	4,939 (21.1%)

Ties (247 cities in which two or more sub-indices share the maximum) are assigned to PM2.5 first. NO_2 is the highest sub-index in one city and CO in none.

PM2.5 is the highest sub-index in 78.9% of cities and in all 191 Hazardous cities, and its dominance grows with severity (from 59.4% of Good cities to 87.7% of Unhealthy cities). Ozone sets the overall AQI mainly where air quality is good or moderate (40.6% of Good cities), where PM2.5 is itself low. The category labels confirm this: PM2.5 and overall categories agree in 97.6% of cities (Cohen's $\kappa = 0.963$, almost perfect agreement), compared with 48.8% for ozone ($\kappa = 0.136$, slight agreement).

5.2 Correlation Structure

Table 5.2: Spearman rank correlations (95% CI) between AQI values

Pair	Spearman ρ	95% CI	Pearson r
Overall AQI - PM2.5 AQI	0.968	[0.967, 0.969]	0.984
Overall AQI - Ozone AQI	0.246	[0.234, 0.258]	0.405
Overall AQI - NO_2 AQI	0.258	[0.246, 0.270]	0.232
Overall AQI - CO AQI	0.601	[0.592, 0.609]	0.431
PM2.5 AQI - Ozone AQI	0.146	[0.134, 0.159]	0.340
PM2.5 AQI - NO_2 AQI	0.289	[0.277, 0.301]	0.260
PM2.5 AQI - CO AQI	0.580	[0.572, 0.588]	0.439
Ozone AQI - NO_2 AQI	0.257	[0.269, 0.245]	0.182
Ozone AQI - CO AQI	0.064	[0.052, 0.077]	0.146
NO_2 AQI - CO AQI	0.502	[0.493, 0.512]	0.488

The overall AQI is almost perfectly rank-correlated with PM2.5 ($\rho = 0.968$), moderately with CO ($\rho = 0.60$), and weakly with ozone and NO_2 ($\rho \approx 0.25$ and 0.26). The CO-PM2.5 ($\rho = 0.58$) and CO-NO_2 ($\rho = 0.50$) associations are plausible given shared combustion sources, whereas ozone is negatively correlated with NO_2 ($\rho = 0.26$), consistent with ozone titration by nitrogen oxides near emission sources. Because the sample is large, all intervals are narrow, and the differences between Spearman and Pearson coefficients (for example for CO-AQI, 0.43) reflect the influence of tied and extreme values rather than sampling error. Dropping the 191 cities with AQI of 300 or more changes the Pearson correlation of the AQI with PM2.5 only slightly (0.984 to 0.982) but raises that with CO from 0.43 to 0.57 and with ozone from 0.41 to 0.46, showing that a few extreme cities suppress the linear association of the secondary pollutants.

5.3 Regression and Relative Importance

Table 5.3: Regression of the overall AQI on the four sub-indices (n = 23,463)

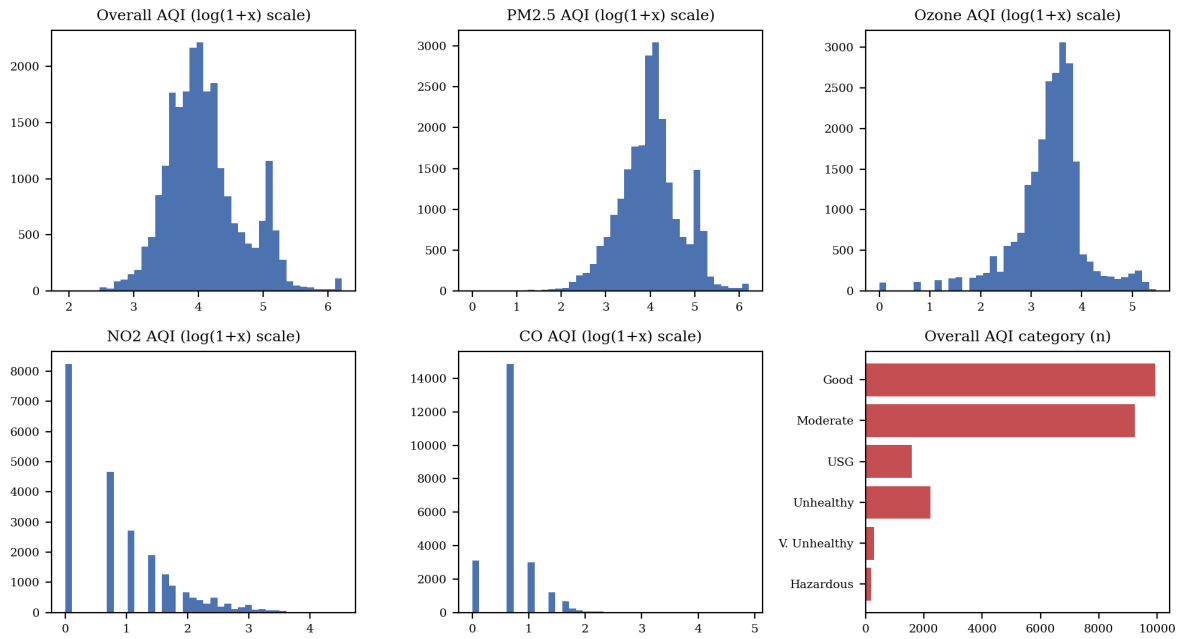


Figure 1. Distributions of the AQI values on the log(1+x) scale, and the number of cities per overall AQI category.

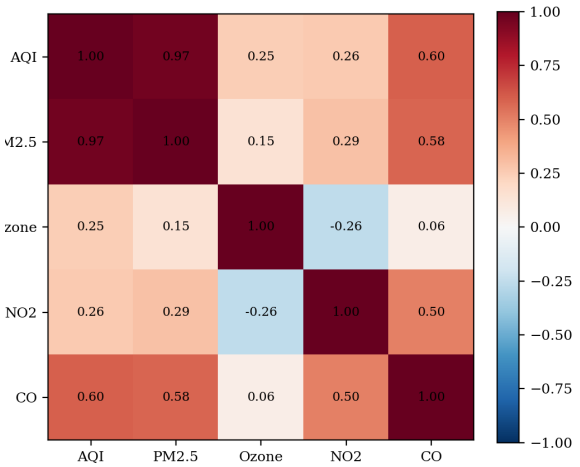


Figure 2. Spearman rank correlation matrix of the AQI values.

Predictor	β	HC3 SE	95% CI	p	Std. β	LMG share of R^2
PM2.5 AQI	0.9801	0.0048	[0.971, 0.990]	<0.001	0.958	81.6%
Ozone AQI	0.1575	0.0058	[0.146, 0.169]	<0.001	0.079	8.4%
NO_2 AQI	0.0362	0.0158	[0.067, 0.005]	0.022	0.003	2.4%
CO AQI	0.0171	0.0764	[0.133, 0.167]	0.823	0.001	7.6%

Intercept = 0.604. R^2 = 0.9746, F = 224,721 on 4 and 23,458 df, p < 0.001. Breusch-Pagan LM = 2182, p < 0.001. Maximum VIF = 1.55.

The four sub-indices explain 97.5% of the variance of the overall AQI. PM2.5 has a coefficient close to 1 (0.980) and ozone a small positive coefficient (0.158), whereas the NO_2 coefficient is slightly negative (0.036, p = 0.022) and the CO coefficient is indistinguishable from zero (p = 0.82). The LMG decomposition attributes 81.6% of the explained variance to PM2.5, 8.4% to ozone, 7.6% to CO and 2.4% to NO_2. The CO share arises from its correlation with PM2.5 and not from a unique effect, since its regression coefficient is null. PM2.5 on its own gives R^2 = 0.969. Collinearity is not a problem (maximum VIF 1.55), but the Breusch-Pagan test indicates strong heteroscedasticity, which is why robust standard errors were used. This model should be read as a description of how the maximum-of-sub-indices rule appears in a linear fit, and not as a causal or predictive model, because the overall AQI is a deterministic function of the sub-indices.

5.4 Between-Country Heterogeneity

Table 5.4: Kruskal-Wallis test of overall AQI across countries

Minimum cities per country	Countries	Cities	H	p	ϵ^2
20	107	22,518	11,167	<0.001	0.494
30	85	21,972	10,894	<0.001	0.494
50	62	21,110	10,342	<0.001	0.488

Country is a dominant source of variation. Across the 85 countries with at least 30 cities, the Kruskal-Wallis test gives H = 10,894 (p < 0.001) with ϵ^2 = 0.494, and the result is stable to the inclusion threshold (ϵ^2 between 0.488 and 0.494; Table 5.4). A one-way decomposition of log AQI attributes 52.6% of variance to country (45.2% on the raw scale), with an intraclass correlation

of 0.537. The share of cities exceeding AQI 100 also differs strongly by country (χ^2 = 10,273, 84 df, p < 0.001, Cramér's V = 0.68), ranging from zero in 22 countries to 97% in the worst case. In practical terms, knowing a city's country tells us about as much about its AQI as all remaining information combined.

Table 5.5: Ten countries with the highest median AQI (at least 30 cities)

Country	Cities	Median (IQR)	Mean	% AQI > 100	95% Wilson CI
Pakistan	307	169 (159-184)	178.8	96.4%	[93.7%, 98.0%]
Senegal	33	157 (146-169)	152.4	97.0%	[84.7%, 99.5%]
India	2488	153 (89-172)	153.0	70.9%	[69.1%, 72.7%]
China	795	137 (82-164)	127.0	65.5%	[62.2%, 68.8%]
Nepal	33	120 (68-150)	116.4	60.6%	[43.7%, 75.3%]
Uzbekistan	58	120 (89-146)	119.1	63.8%	[50.9%, 74.9%]
Iran	179	105 (74-148)	108.3	51.4%	[44.1%, 58.6%]
Dominican Republic	38	104 (80-129)	108.1	57.9%	[42.2%, 72.1%]
Bangladesh	80	100 (68-159)	113.5	48.8%	[38.1%, 59.5%]
El Salvador	33	98 (75-110)	97.4	48.5%	[32.5%, 64.8%]

Table 5.6: Ten countries with the lowest median AQI (at least 30 cities)

Country	Cities	Median (IQR)	Mean	% AQI > 100	95% Wilson CI
Bolivia	33	20 (14-29)	23.8	0.0%	[0.0%, 10.4%]
Argentina	68	22 (17-31)	28.2	1.5%	[0.3%, 7.9%]
Madagascar	65	25 (22-41)	31.0	0.0%	[0.0%, 5.6%]
Tanzania	229	27 (21-36)	31.6	0.9%	[0.2%, 3.1%]
Australia	85	27 (23-32)	33.6	3.5%	[1.2%, 9.9%]
Peru	128	32 (25-44)	36.4	3.1%	[1.2%, 7.8%]
Finland	81	34 (30-36)	38.4	3.7%	[1.3%, 10.3%]
Myanmar	59	34 (26-51)	40.2	1.7%	[0.3%, 9.0%]
Lithuania	38	35 (34-40)	38.1	0.0%	[0.0%, 9.2%]
Portugal	153	35 (29-40)	34.6	0.0%	[0.0%, 2.4%]

Among the largest samples, India (2,488 cities) has a median AQI of 153 with 70.9% of cities above 100, and China (795 cities) a median of 137 (65.5% above 100). The United States (2,872 cities) has a median of 55 and 5.3% above 100, and Germany (1,345 cities) 49 and 0.1%. The Wilson intervals for countries with about 30 to 40 cities are wide (for example Nepal 43.7%-75.3%), so the ordering among countries with small samples and similar medians is not reliable. Because the data contain one row per city with no population weights, country summaries describe the sampled cities, not the people living in a country.

5.5 Predictive Modelling and the Effect of Validation Design

Table 5.7: Predicting the PM2.5 sub-index from CO, ozone and NO_2 (ten-fold random cross-validation)

Model	R^2 (SD)	RMSE	MAE
Linear regression	0.281 (0.052)	46.4	28.0
Ridge	0.281 (0.052)	46.4	28.0
k-NN (k=15)	0.335 (0.028)	44.7	26.5
Random forest	0.371 (0.030)	43.4	25.0
Gradient boosting	0.374 (0.030)	43.3	25.0

SD of PM2.5 sub-index = 54.8. A constant median prediction has MAE = 34.1.

Table 5.8: Difference in fold R^2 relative to linear regression (corrected resampled t-test)

Model	Mean R^2 difference	Corrected p
Ridge	0.000	0.710
k-NN (k=15)	0.053	0.084
Random forest	0.089	0.008
Gradient boosting	0.093	0.004

The co-pollutants explain only a modest part of the PM2.5 sub-index (R^2 = 0.28 for linear regression and 0.37 for gradient boosting), which is consistent with the weak-to-moderate correlations in Table 5.2. Random forest and gradient boosting improve significantly on the linear model (corrected p = 0.008 and 0.004), while ridge regression is identical to it and k-NN is not significantly better. Even the best model leaves an RMSE of 43.3 index points (MAE 25.0, against 34.1 for a constant prediction), so PM2.5 cannot be reliably inferred from the other pollutants. Part of the nonlinear gain may reflect country-specific combinations of sub-indices, which the classification experiment below examines directly.

Table 5.9: Predicting AQI > 100 from CO, ozone and NO_2 under random and country-grouped cross-validation

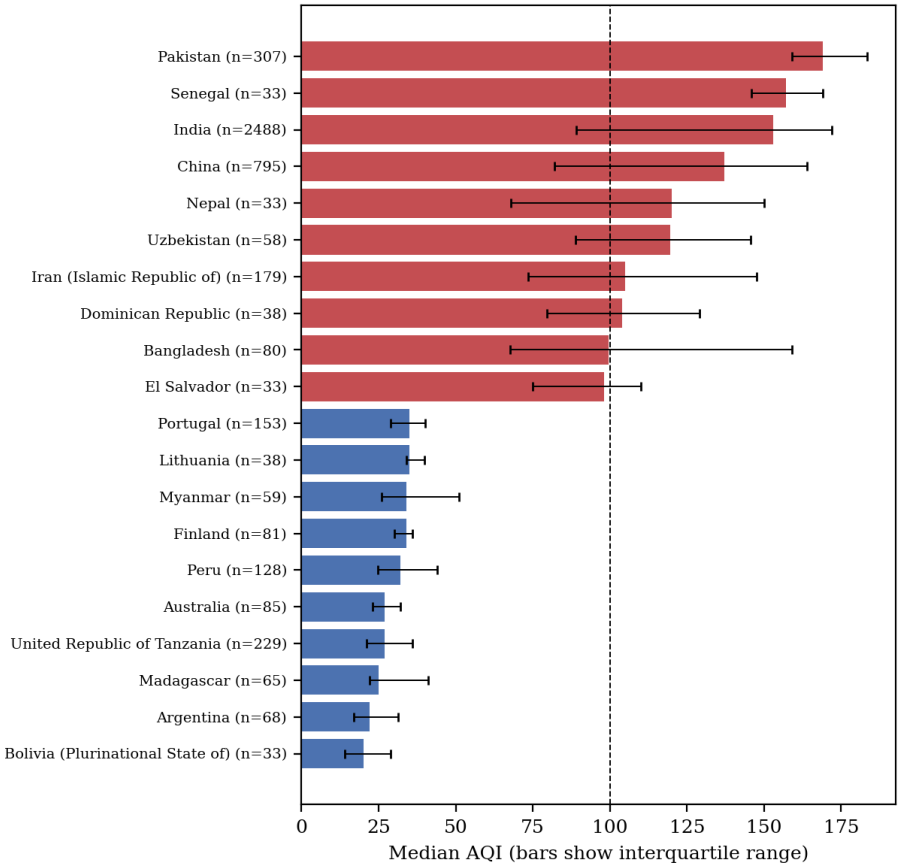


Figure 3. Median overall AQI (and interquartile range) for the ten lowest and ten highest countries; the dashed line marks AQI 100.

Model	AUC, random CV (SD)	AUC, grouped CV (SD)	Sensitivity (grouped)	Specificity (grouped)
Logistic regression	0.835 (0.017)	0.855 (0.051)	0.492	0.980
Random forest	0.881 (0.013)	0.847 (0.065)	0.503	0.972
Gradient boosting	0.880 (0.014)	0.850 (0.056)	0.493	0.978

Prevalence of AQI > 100 = 0.181; threshold 0.5. For comparison, the AUC of PM2.5 alone is 0.998, CO 0.781, ozone 0.653 and NO_2 0.600.

Table 5.10: Effect of a country target encoding on gradient boosting

Validation design	AUC (SD)	Sensitivity	Specificity
Random CV, country encoding	0.968 (0.004)	0.764	0.975
Grouped CV, country encoding	0.841 (0.059)	0.499	0.965
Grouped CV, no country information	0.850 (0.056)	0.493	0.978

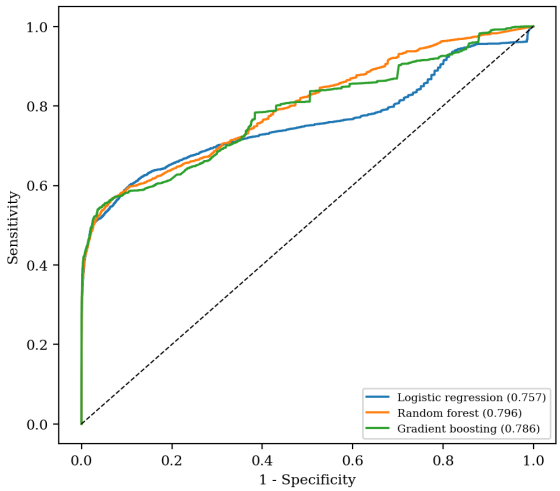


Figure 4. ROC curves for predicting AQI > 100 without PM2.5 under country-grouped cross-validation (AUC in parentheses).

Without PM2.5, the other three sub-indices still separate unhealthy from healthier cities reasonably well (grouped-CV AUC 0.86-0.85), driven mostly by CO and ozone, but sensitivity is only about 0.49 at the default threshold because 82% of cities are healthy. The validation design matters. Under random cross-validation, random forest and gradient boosting appear clearly

better than logistic regression (AUC 0.881 and 0.880 against 0.835). Under country-grouped cross-validation the advantage disappears and the ordering reverses (0.847 and 0.850 against 0.855; corrected p = 0.71 and 0.77). The flexible models were therefore exploiting country-specific patterns that are unavailable for a new country. The target-encoding experiment makes this explicit (Table 5.10): adding a country encoding raises the AUC to 0.968 under random cross-validation but leaves it at 0.841, no better than the model with no country information, once whole countries are held out. Fold-to-fold variation is also three to four times larger under grouped cross-validation (SD about 0.05 against 0.017), reflecting heterogeneity between countries.

6. Comparative Analysis and Discussion

6.1 RQ1: Structure and consistency of the dataset

The dataset is complete for all numeric fields and has no duplicates, but it is not fully internally consistent. The overall AQI equals the maximum sub-index in 99.4% of rows, as the EPA rule requires, but exceeds it in 141 extreme-PM2.5 cities, and 148 labels place boundary values in the higher category. Country is missing for 1.8% of rows. None of these issues is large, but all should be documented, and analyses that use labels instead of values will give slightly different counts.

6.2 RQ2: Distributions and the dominant pollutant

All sub-indices are strongly right-skewed, and CO and NO_2 are nearly constant on the index scale, so they carry little information about the overall AQI. PM2.5 is the highest sub-index in 78.9% of cities and in every Hazardous city, whereas ozone matters mainly where air is already good. In an index-based dataset, therefore, the overall AQI is in practice a PM2.5 indicator with an ozone floor. Users who need information on gaseous pollutants should work with concentrations, not with this index.

6.3 RQ3: Association and relative importance

The overall AQI is almost perfectly rank-correlated with PM2.5 ($\rho = 0.97$), and the LMG decomposition attributes 81.6% of the explained variance to it. The secondary pollutants matter through ozone alone; CO and NO_2 have no unique contribution. Because the overall AQI is a deterministic maximum, these results describe the index construction and should not be taken as evidence about the relative health importance of the pollutants.

6.4 RQ4: Between-country heterogeneity

Table 6.1: Summary of evidence by review question

Review question	Principal finding	Evidence
RQ1 Consistency	AQI = maximum sub-index in 99.4%; 141 exceptions at extreme PM2.5	Table 4.2
RQ2 Dominant pollutant	PM2.5 highest in 78.9% of cities; κ = 0.96	Table 5.1
RQ3 Contribution	PM2.5 explains 81.6% of AQI variance explained	Table 5.2, 5.3
RQ4 Heterogeneity	ICC = 0.54; ϵ^2 = 0.49	Table 5.4-5.6
RQ5 Prediction	Grouped CV AUC 0.86; no gain over logistic regression	Table 5.9, 5.10

Country explains about half of the variation in log AQI (ICC 0.54), and median AQI ranges from 20 to 169 across countries with at least 30 cities. The ten highest-median countries are concentrated in South and Central Asia, West Africa and parts of the Middle East and Caribbean, whereas the lowest are in South America, Oceania, Northern and Southern Europe and parts of Africa. The sampled-city design, small samples for many countries and the absence of population weights mean that these figures describe monitored or reported cities and not national populations (an ecological caveat).

6.5 RQ5: Prediction and the role of validation

The most important methodological finding of this review concerns validation. Random cross-validation lets cities from the same country appear in training and test folds, and flexible models and target encoding then exploit country signatures. When whole countries are held out, the apparent gain from nonlinear models and from country encoding disappears (Tables 5.9 and 5.10). Any study that reports high accuracy on this dataset should state whether its validation separates countries. This parallels the look-ahead and leakage problems that have been documented in other forecasting fields.

7. Challenges and Limitations

7.1 Methodological Challenges

Deterministic outcome: the overall AQI is the maximum of the sub-indices, so regression of AQI on its components is structural, not explanatory.

Index versus concentration: sub-indices are nonlinear transformations of concentration; correlations and regressions on the index scale can differ from those on concentrations.

Extreme values: a few hundred very high observations

influence linear summaries (Section 5.2); rank-based and log-scale methods were used for this reason.

Untuned models: default hyperparameters were used, so model comparisons refer to default configurations.

7.2 Data Challenges

Undocumented provenance: the data source, measurement date and averaging period are not stated in the file.

Single snapshot: there is no time dimension, so seasonality, trends and episodes cannot be studied.

Uneven sampling: country sample sizes range from 1 to 2,872 cities (median 28), and city coverage is not population-weighted.

No coordinates: the absence of latitude and longitude prevents any analysis of spatial autocorrelation.

Limited pollutants: PM10 and SO_2 are not included, and CO and NO_2 are nearly constant on the index scale.

7.3 Reporting Challenges

Accuracy reported without stating whether validation separates countries.

Use of category labels instead of values, with inconsistent treatment of boundary values.

Country-level rankings published without sample sizes or confidence intervals.

The literature component of this review is selective and narrative, not an exhaustive systematic search.

8. Priority Research Directions

Concentration-level data. Replace index values with raw pollutant concentrations and compare them with the WHO (2021) guideline values.

Time series. Add repeated measurements per city to study seasonality, trends and pollution episodes with appropriate time-series models.

Hierarchical and spatial models. Fit multilevel models with country and city random effects, and use coordinates to model spatial autocorrelation, with spatially blocked validation.

Population weighting. Weight cities by population to move from descriptions of sampled cities to estimates of population exposure.

Uncertainty quantification. Report bootstrap or Bayesian intervals for country-level summaries and prediction intervals for models.

Reporting checklist. Adopt a minimal checklist for studies on this dataset: (i) state the source and date of the data; (ii) specify whether AQI values or labels are used; (iii) state how boundary values and missing countries are handled; (iv) report sample sizes with every country summary; (v) use country-grouped validation for any predictive claim; (vi) compare against a simple baseline such as logistic regression.

9. Conclusion

The Global Air Pollution Dataset is a convenient resource whose statistical structure is easy to misread. Three conclusions are robust. First, the overall AQI is effectively a PM2.5 indicator: PM2.5 is the highest subindex in 78.9% of cities and accounts for 81.6% of the explained variance in AQI, while CO and NO_2 are nearly constant. Second, geography dominates: country explains roughly half of the variation in AQI (ICC 0.54), with median AQI differing about eight-fold between the cleanest and most polluted countries. Third, predictive performance reported on this dataset depends critically on the validation design: models that appear to improve on logistic regression under random cross-validation do not do so when whole countries are held out. Future studies should document the data source, analyse concentrations rather than indices where possible, report uncertainty for country summaries, and validate predictive models across countries.

References

1. Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5), 1287–1294.
2. Burnett, R., Chen, H., Szyszkowicz, M., et al. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences*, 115(38), 9592–9597.
3. Cohen, A. J., Brauer, M., Burnett, R., et al. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, 389(10082), 1907–1918.
4. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.
5. Collins, G. S., Reitsma, J. B., Altman, D. G. and Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*, 350, g7594.
6. Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

7. Grömping, U. (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of Statistical Software*, 17(1), 1–27.
8. Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, 55(2), 163–172.
9. Kruskal, W. H. and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
10. Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159–174.
11. Landrigan, P. J., Fuller, R., Acosta, N. J. R., et al. (2018). The Lancet Commission on pollution and health. *The Lancet*, 391(10119), 462–512.
12. Lindeman, R. H., Merenda, P. F. and Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman, Glenview, IL.
13. MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305–325.
14. Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, 3(1), 27–32.
15. Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, 52(3), 239–281.
16. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
17. Pope, C. A. and Dockery, D. W. (2006). Health effects of fine particulate air pollution: lines that connect. *Journal of the Air & Waste Management Association*, 56(6), 709–742.
18. Roberts, D. R., Bahn, V., Ciuti, S., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929.
19. Seabold, J. and Perktold, J. (2010). statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, 92–96.
20. Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428.
21. Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15(1), 72–101.
22. Tomczak, M. and Tomczak, E. (2014). The need to report effect size estimates revisited: an overview of some recommended measures of effect size. *Trends in Sport Sciences*, 21(1), 19–25.
23. US Environmental Protection Agency (2018). Technical Assistance Document for the Reporting of Daily Air Quality - the Air Quality Index (AQI). EPA-454/B-18-007, Research Triangle Park, NC.
24. Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.
25. World Health Organization (2021). *WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*. WHO, Geneva.