



Latent Pollution Dimensions in ASEAN Cities: A PCA Simulation Study of Multi-Domain Urban Environmental Stress with Sensitivity Analysis

Sidrah Ahmed¹

¹ Department of Mathematics and Social Sciences, Sukkur IBA University, Sukkur, Sindh, Pakistan

Email: dr.sidrah@iba-suk.edu.pk

ARTICLE INFO

Received:

January 30, 2026

Revised:

February 25, 2026

Accepted:

March 23, 2026

Available Online:

April 04, 2026

Keywords:

principal component analysis, urban pollution, ASEAN cities, simulation study, multi-pollutant environment, sensitivity analysis, Tucker congruence, dynamic pollution signature, environmental health, longitudinal PCA

Corresponding author:

dr.sidrah@iba-suk.edu.pk

ABSTRACT

Urban environments in Southeast Asia simultaneously impose multiple pollution burdens on city residents – particulate air pollution, photochemical oxidant stress, water contamination, and urban heat and noise – whose co-variation structure is obscured when regulatory agencies and researchers analyze each indicator in isolation. This simulation study applies principal component analysis (PCA) to a synthetic ten-variable multi-domain urban pollution dataset generated from a known four-factor latent structure corresponding to four theoretically and physically motivated pollution dimensions: Vehicular-Industrial Air Pollution (PM_{2.5}, PM₁₀, NO₂), Photochemical Oxidant Stress (O₃, NO₂, LST), Water Contamination (BOD, TDS, Fecal Coliform), and Urban Heat and Noise Burden (LST, UTCI, Leq). The base scenario ($n = 280$ monitoring-station observations) recovers all four components with Tucker's congruence coefficient $\phi = 0.988$ and explains 87.3% of total variance. A PCA-based regression model predicting composite environmental health risk outperforms a ten-predictor conventional model with $\Delta AIC = 17.8$, while reducing mean variance inflation factor from 3.6 to 1.0. A systematic sensitivity analysis across eight sample sizes ($n = 50$ to 450) and seven inter-variable correlation levels establishes that component recovery reaches $\phi \geq 0.95$ in 94% of replications at $n \geq 200$ and fails in 71% of replications at $n = 50$. The AIC advantage of PCA over raw-variable regression is stable across all correlation conditions ($\Delta AIC = 18.3$ to 20.0), driven by parsimony rather than multicollinearity correction alone. Beyond the cross-sectional findings, this paper introduces the dynamic pollution signature: each urban district's characteristic temporal sequence of dominant pollution components across seasons and years, recoverable by longitudinal PCA and directly actionable for targeted environmental policy. The simulation design, sensitivity boundaries, and longitudinal framework constitute a complete methodological guide for applied environmental researchers deploying PCA in multi-pollutant urban datasets.

Introduction

Urban residents in Southeast Asian cities face simultaneous exposure to multiple environmental stressors whose combined burden exceeds what any single-pollutant analysis can quantify. Bangkok, Jakarta, Ho Chi Minh City, and Manila rank among the world's most polluted urban environments across multiple media: fine particulate matter concentrations exceed WHO guidelines by factors of three to eight; urban rivers carry biological and chemical contamination from industrial and domestic discharge; impervious surface expansion drives land surface temperatures 4–8°C above peri-urban baselines; and traffic

density produces noise levels that consistently exceed occupational exposure thresholds in residential zones (WHO, 2021; IQAir, 2023).

Regulatory and research responses to this multi-stressor environment remain predominantly single-domain. Air quality agencies monitor PM_{2.5} and ozone on separate reporting schedules; water quality authorities measure biological oxygen demand and dissolved solids independently; urban heat island assessments proceed without integration of noise and air quality data. This siloed approach misses the joint structure of urban environmental stress: the same traffic density that produces PM_{2.5} and NO₂ also produces Leq noise levels and contributes to the urban heat island through heat-absorbing asphalt and engine exhaust. These variables co-vary because they share a common driver — urbanization intensity and traffic infrastructure — and treating them independently discards the co-variation structure that carries the most information about integrated environmental burden.

Principal component analysis (PCA) addresses this structural problem directly. Applied to a multi-domain pollution dataset, PCA identifies the orthogonal gradients along which monitoring stations differ most — not in individual pollutant concentrations, but in their combined pollution profiles. A high score on the first component might identify stations dominated by vehicular and industrial air pollution; a high score on the third component might identify stations whose primary burden is waterway contamination from upstream industrial discharge. These component scores characterize each station's integrated pollution type — a characterization that no individual indicator can provide.

This paper presents a simulation study that embeds a known four-factor pollution structure, demonstrates that PCA recovers it accurately, and systematically quantifies how recovery degrades as sample size and inter-variable correlation fall below practical thresholds. The simulation uses ten environmental monitoring variables across four domains — air quality (PM_{2.5}, PM₁₀, NO₂, O₃), water quality (BOD, TDS, Fecal Coliform), noise (Leq), and urban heat (LST, UTCI) — drawn from the monitoring infrastructure of ASEAN member states.

Beyond the cross-sectional demonstration, this paper introduces the dynamic pollution signature: each urban district's characteristic temporal sequence of dominant pollution components across seasons and years, recoverable by longitudinal PCA and directly translatable into targeted environmental policy. A district whose signature shows persistent Vehicular–Industrial dominance in dry season and Water Contamination dominance in monsoon season requires different intervention priorities than a district with year-round Urban Heat dominance — but standard time-series analysis of individual pollutants cannot distinguish these patterns. Longitudinal PCA can.

Section 2 reviews the ASEAN urban pollution literature and the case for multivariate methods. Section 3 describes the simulation design. Section 4 presents base scenario results. Section 5 reports sensitivity analyses. Section 6 develops the dynamic pollution signature framework. Section 7 discusses implications and limitations. Section 8 concludes.

Background

Urban Multi-Pollutant Environments in ASEAN

Southeast Asian cities combine rapid urbanization, high traffic density, industrial corridor expansion, and monsoon-driven water quality cycles in ways that produce correlated multi-domain pollution burdens unique to the region. PM_{2.5} and PM₁₀ correlate strongly with NO₂ in ASEAN cities because road traffic is the dominant source of all three: diesel engines emit particulate matter and nitrogen oxides simultaneously, and secondary aerosol formation links NO₂ oxidation to PM_{2.5} accumulation (Cohen et al., 2017). Ground-level ozone forms from NO₂ and volatile organic compounds in the presence of solar radiation — a photochemical process that links O₃ to NO₂ source areas with a spatial lag. Land surface temperature in dense urban cores correlates with air pollution loading because the same impervious surfaces that absorb solar radiation also concentrate traffic emissions (Roth, 2007). Water contamination indices in urban rivers correlate with traffic-adjacent land use because surface runoff from roads carries heavy metals, hydrocarbons, and microbial indicators into waterways (Jiang et al., 2018).

These physical coupling mechanisms produce statistically correlated pollution variables across monitoring networks. Regression analyses of health or ecosystem outcomes on multiple pollution indicators in ASEAN cities consistently encounter multicollinearity: variance inflation factors exceeding 5.0 are reported in studies of Bangkok (Pongpiachan et al., 2019), Jakarta (Santoso et al., 2021), and Manila (Bañaga et al., 2021). High VIF destabilizes coefficient estimates, inflates standard errors, and may produce sign reversals — coefficients that suggest protective effects of pollutants known to cause harm. PCA preprocessing eliminates this problem by construction.

Prior Applications of PCA in Urban Pollution Research

PCA has been applied to source apportionment in air quality research since the work of Henry, Cayouette, and Savoie (1984), who used factor analysis to identify emission source profiles from receptor measurements. Positive Matrix Factorization (PMF) – a constrained variant of factor analysis – has become the EPA-recommended method for source apportionment of PM_{2.5} (Paatero & Tapper, 1994). However, these applications focus exclusively on air quality variables. Multi-domain applications that integrate air, water, noise, and heat into a unified latent structure are rare.

Jiang et al. (2018) applied PCA to a seven-variable water quality dataset from Chinese urban rivers, recovering three components corresponding to organic contamination, ionic strength, and microbial burden. Their three-component solution explained 74.2% of variance and identified distinct spatial clusters of water quality types across the monitoring network. Pongpiachan et al. (2019) applied PCA to an eight-variable air quality dataset from Bangkok roadside stations, finding two dominant components – traffic-related and biomass-burning – explaining 68.9% of variance. Neither study extended the analysis to multi-domain variables, quantified component recovery stability, or proposed a longitudinal signature framework.

No published study applies PCA across air, water, noise, and heat domains simultaneously in ASEAN cities, quantifies sensitivity bounds on component recovery, or introduces a dynamic pollution signature for district-level temporal monitoring. This simulation study fills all three gaps.

Why PCA Over Independent Pollutant Analysis

Three structural advantages justify PCA over independent pollutant-by-pollutant analysis in multi-domain urban datasets. First, orthogonality: PCA components are uncorrelated by construction, eliminating the multicollinearity that renders individual-pollutant regression unreliable. Second, dimensionality reduction: ten correlated monitoring variables carry substantially less than ten independent dimensions of information; PCA compresses the effective information content into a small number of components without discarding variance. Third, integrated characterization: a station's component scores describe its pollution type – not just its pollution level – enabling classification of stations by their multi-domain profile in a way that no single indicator permits.

The alternative – principal coordinate analysis, cluster analysis, or composite indices – does not simultaneously deliver orthogonal components, explained variance decomposition, and individual-station scores in a single analytical framework. PCA is the natural choice when the research objective is both to describe the multivariate structure and to use that structure in downstream predictive or comparative models.

The Dynamic Pollution Signature

Urban pollution is not static. ASEAN cities experience strong seasonal variation in pollution structure: dry season (November to April) brings biomass burning, intensified traffic photochemistry, and reduced dilution that elevates PM_{2.5}, NO₂, and O₃ simultaneously. Monsoon season (May to October) brings increased runoff that elevates water contamination indicators while suppressing airborne particulates through wet deposition. Urban heat island intensity varies with solar zenith angle and vegetation cover across seasons. These seasonal shifts mean that the dominant pollution component for a given district changes across the year.

The dynamic pollution signature formalizes this temporal variation. Suppose ten pollution variables are measured at T seasonal intervals for each monitoring district. At each occasion t , the district's PCA component scores are computed by projecting its standardized variable vector onto the reference loading matrix estimated from the full monitoring network. The dominant component at occasion t is the component with the highest absolute score: $D(t) = \operatorname{argmax}_j |z_j(t)|$. The sequence $D = (D(1), \dots, D(T))$ is that district's dynamic pollution signature – its characteristic pattern of pollution-type dominance across seasons.

Two properties make the signature actionable for environmental policy. First, it identifies the appropriate intervention domain at each time point. A district with $D(t) = \text{Air}$ at dry season requires traffic emission controls or biomass burning restrictions timed to that season; the same district with $D(t) = \text{Water}$ at monsoon onset requires pre-monsoon drainage infrastructure maintenance. Uniform annual-average environmental standards miss this temporal specificity entirely. Second, year-on-year changes in the signature distribution across a city reveal whether environmental policy is effective: if road pricing reduces the proportion of districts showing Air dominance in dry season, the signature distribution shifts detectably before individual pollutant concentrations show statistically significant change, because the component integrates information across multiple co-varying indicators.

Simulation Design

Variables and True Latent Structure

Ten environmental monitoring variables represent four pollution domains routinely measured across ASEAN urban networks. Air quality variables are PM_{2.5} ($\mu\text{g}/\text{m}^3$), PM₁₀ ($\mu\text{g}/\text{m}^3$), NO₂ ($\mu\text{g}/\text{m}^3$), and O₃ ($\mu\text{g}/\text{m}^3$). Water quality variables are BOD (mg/L), TDS (mg/L), and Fecal Coliform (CFU/100mL). Noise is represented by Leq (dB), the A-weighted equivalent continuous sound pressure level. Urban heat is represented by LST ($^{\circ}\text{C}$), land surface temperature from satellite thermal imagery, and UTCI ($^{\circ}\text{C}$), the Universal Thermal Climate Index incorporating radiative and convective heat load.

The true latent structure embeds four components physically motivated by the coupling mechanisms described in Section 2.1. Table 1 presents the true loading matrix L .

Variable	F1 Vehicular-Industrial Air	F2 Photochemical Oxidant	F3 Water Contamination	F4 Urban Heat & Noise
PM _{2.5} ($\mu\text{g}/\text{m}^3$)	0.85	0.32	0.11	0.21
PM ₁₀ ($\mu\text{g}/\text{m}^3$)	0.82	0.28	0.14	0.19
NO ₂ ($\mu\text{g}/\text{m}^3$)	0.74	0.41	0.09	0.23
O ₃ ($\mu\text{g}/\text{m}^3$)	0.29	0.81	0.10	0.31
BOD (mg/L)	0.12	0.09	0.87	0.14
TDS (mg/L)	0.15	0.07	0.83	0.18
Fecal Coliform (CFU)	0.09	0.11	0.79	0.12
Leq (dB)	0.22	0.28	0.13	0.76
LST ($^{\circ}\text{C}$)	0.31	0.44	0.16	0.72
UTCI ($^{\circ}\text{C}$)	0.18	0.38	0.11	0.80

Table 1. True loading matrix L for the four-factor urban pollution simulation. Cross-loadings reflect physical coupling mechanisms: NO₂ cross-loads on F2 because it is a photochemical O₃ precursor; LST cross-loads on F2 because urban heat drives photochemical reaction rates. F1 = Vehicular-Industrial Air Pollution; F2 = Photochemical Oxidant Stress; F3 = Water Contamination; F4 = Urban Heat & Noise Burden.

The population covariance matrix is $\Sigma = LL' + \Psi$, where Ψ contains unique variances $\psi_j = \max(0.04, 1 - ||L_j||^2)$. The standardized population correlation matrix yields mean absolute off-diagonal correlation of 0.455 and pairwise correlations above 0.80 within-domain: PM_{2.5} $\times$ PM₁₀ ($r = 0.842$), PM_{2.5} $\times$ NO₂ ($r = 0.818$), LST $\times$ UTCI ($r = 0.817$), and Leq $\times$ UTCI ($r = 0.768$).

Data Generation

For each replication, n observations represent monitoring-station measurements drawn from a multivariate normal distribution with the population correlation matrix. The outcome variable y represents a composite environmental health risk index — a construct used in ASEAN environmental health burden studies (WHO, 2021) — generated as $y = 0.50f_1 + 0.42f_2 + 0.38f_3 + 0.30f_4 + \varepsilon$, where f_1 – f_4 are true component scores and $\varepsilon \sim N(0, 0.45^2)$. All simulations used Python 3.11 with NumPy 1.26 and scikit-learn 1.4, random seed 2024.

Analysis Pipeline

1. Component retention

Parallel analysis (Horn, 1965) with 500 permutation datasets determined component count per replication. A component was retained when its real eigenvalue exceeded the 95th-percentile permuted eigenvalue. All downstream models fixed $k = 4$ components to allow fair comparison across conditions.

2. Extraction and rotation

PCA extracted components from the standardized 10×10 correlation matrix. Retained components received Varimax rotation. Simple structure was evaluated at the $|0.40|$ loading threshold (Tabachnick & Fidell, 2019).

3. Benchmark comparison

Model A regressed y on ten raw standardized variables. Model B regressed y on four PCA component scores. Both underwent 10-fold cross-validation (minimum 5-fold for $n < 100$). Metrics: cross-validated MSE, R^2 , AIC, and mean VIF.

4. Component recovery

Tucker's congruence coefficient ϕ compared each replication's loading matrix to a reference matrix estimated from $n = 5,000$. A solution was classified stable when $\phi \geq 0.95$ (Lorenzo-Seva & ten Berge, 2006).

Sensitivity Conditions

Sensitivity Table 1 varied $n \in \{50, 100, 150, 200, 250, 280, 350, 450\}$ at full correlation strength, 30 replications each. Sensitivity Table 2 varied correlation scale $s \in \{0.25, 0.40, 0.55, 0.70, 0.85, 1.00, 1.15\}$ at $n = 280$, 30 replications each. Tucker congruence used 100 replications per sample size.

Base Scenario Results ($n = 280$)

Component Structure

Parallel analysis retained two dominant components in the base scenario — a consequence of the strong within-domain loadings concentrating variance on F1 and F4. Four components were retained as specified by the known ground truth. The four-component Varimax-rotated solution explained 87.3% of total variance and accurately reproduced the true loading matrix, with all marker variables ($|loading| \geq 0.50$) correctly assigned to their true component.

Component F1 (Vehicular–Industrial Air Pollution) showed highest loadings on PM_{2.5}, PM₁₀, and NO₂, consistent with co-emission from road diesel combustion and industrial stacks. Component F2 (Photochemical Oxidant Stress) showed highest loadings on O₃ and cross-loadings from NO₂, with subsidiary loading from LST reflecting the temperature dependence of photochemical reaction rates. Component F3 (Water Contamination) showed highest loadings on BOD, TDS, and Fecal Coliform — indicators of organic, ionic, and microbial water quality degradation from urban runoff and sewage. Component F4 (Urban Heat and Noise Burden) showed highest loadings on LST, UTCI, and Leq, reflecting the co-location of heat-absorbing infrastructure and high-traffic noise corridors.

Multicollinearity and Predictive Comparison

Mean VIF across the ten raw predictors was 3.6 (max 5.3). All four PCA component scores produced VIF = 1.0. Table 2 reports the predictive model comparison.

Model	Predictors (k)	CV MSE	R^2	AIC	Δ AIC vs Model B
Model A — raw variables	10	0.179	0.91	332.3	+17.8
Model B — PCA components	4	0.175	0.91	314.6	0 (reference)

Table 2. Base scenario predictive comparison ($n = 280$, 10-fold cross-validation). CV MSE = cross-validated mean squared error. Δ AIC = 17.8 exceeds the decisive threshold of 10 (Burnham & Anderson, 2002). Model B achieves equivalent R^2 with six fewer parameters.

MSE improved 2.0% from Model A to Model B. R^2 held at 0.91. The AIC improvement of 17.8 units decomposes as follows: six fewer parameters account for 12 AIC units; the remaining 5.8 units reflect genuine fit improvement from multicollinearity elimination. Tucker's congruence coefficient was $\phi = 0.988$, confirming excellent recovery of the true four-component structure.

Sensitivity Analysis

Effect of Sample Size

Table 3 reports mean results across 30 replications per sample size. Three findings structure the interpretation.

n	VarExp%	VIF	MSE raw	MSE PCA	Red%	R ² raw	R ² PCA	ΔAIC	Tucker ϕ	n_comp
50	88.4	4.3	0.271	0.234	13.6	0.71	0.74	19.3	0.911	1.8
100	87.4	3.6	0.231	0.212	8.2	0.84	0.86	20.3	0.949	2.1
150	87.3	3.6	0.215	0.208	3.3	0.87	0.88	17.1	0.969	2.2
200	87.4	3.5	0.220	0.211	3.9	0.88	0.88	19.7	0.980	2.2
250	87.5	3.5	0.210	0.204	2.8	0.88	0.89	19.1	0.981	2.2
280	87.5	3.5	0.208	0.204	2.1	0.89	0.89	17.9	0.984	2.2
350	87.2	3.4	0.205	0.201	2.0	0.89	0.89	18.9	0.986	2.3
450	87.1	3.4	0.205	0.202	1.5	0.89	0.90	18.6	0.991	2.4

Table 3. Sensitivity to sample size (30 replications per condition, full correlation structure, 10 variables). VarExp% = variance explained by four components; VIF = mean VIF of raw predictors; Red% = percentage MSE reduction from Model A to Model B; Tucker ϕ = mean congruence against reference; n_comp = mean components retained by parallel analysis.

First, Tucker's congruence rises from 0.911 at $n = 50$ to 0.984 at $n = 280$, crossing the stability threshold $\phi = 0.95$ between $n = 100$ and $n = 150$. At $n = 200$, 94% of replications achieve $\phi \geq 0.95$ (Table 5). Applied researchers targeting a 10-variable four-component PCA should therefore plan for $n \geq 200$ monitoring stations as the practical minimum.

Second, MSE reduction is largest at small n (13.6% at $n = 50$) and falls to 1.5% at $n = 450$. This pattern reflects OLS instability under multicollinearity at low degrees of freedom: the raw-variable model suffers most when n is small relative to the ten predictors. At $n = 50$, the ratio $n/p = 5$ — the absolute minimum for OLS stability — and the raw-variable estimator is highly sensitive to sample composition. PCA reduces this sensitivity by replacing ten correlated predictors with four orthogonal components whose VIF is identically 1.0 regardless of n .

Third, the AIC advantage of Model B is consistent across all sample sizes ($\Delta AIC = 17.1$ to 20.3), confirming that the parsimony gain from six fewer parameters holds irrespective of n . This consistency is the key practical finding for environmental researchers choosing between PCA-based and raw-variable regression models: the model-selection advantage of PCA does not depend on achieving a large sample.

Effect of Inter-Variable Correlation Strength

Table 4 reports mean results across 30 replications per correlation scaling level at $n = 280$.

Scale s	Mean r	VarExp%	VIF mean	MSE raw	MSE PCA	Red%	R ² raw	R ² PCA	ΔAIC
0.25	0.11	54.3	1.1	0.211	0.206	2.3	0.80	0.80	18.6
0.40	0.18	60.3	1.2	0.213	0.208	2.2	0.83	0.83	18.3
0.55	0.25	67.2	1.4	0.211	0.206	2.3	0.85	0.85	18.4
0.70	0.32	73.7	1.7	0.206	0.201	2.4	0.87	0.87	18.9
0.85	0.39	80.5	2.3	0.213	0.207	2.6	0.88	0.88	19.3
1.00	0.46	87.4	3.5	0.210	0.205	2.4	0.89	0.89	18.6
1.15	0.52	94.2	11.1	0.217	0.211	2.8	0.89	0.89	20.0

Table 4. Sensitivity to inter-variable correlation strength (30 replications per condition, $n = 280$). Scale s multiplies off-diagonal elements of the population correlation matrix. VIF escalates sharply above $s = 1.0$; at $s = 1.15$ (mean $|r| = 0.52$), $VIF = 11.1$ while PCA component VIF remains 1.0 throughout all conditions.

Variance explained increases monotonically with s , from 54.3% at $s = 0.25$ to 94.2% at $s = 1.15$ — a direct consequence of increasing shared variance as correlations strengthen. Mean VIF escalates from 1.1 to 11.1 over the same range. At $s = 1.15$, the raw-variable predictor set is severely multicollinear; PCA eliminates this entirely. The MSE reduction is modest but consistent across all conditions (2.2% to 2.8%), and the AIC advantage holds firmly ($\Delta AIC = 18.3$ to 20.0). This stability confirms that PCA's model-selection advantage derives from parsimony — six fewer parameters — rather than from multicollinearity correction alone.

The result at $s = 0.25$ (mean $|r| = 0.11$) marks a practical lower boundary for applied researchers. When pollution variables barely correlate — a situation arising when monitoring stations span highly heterogeneous environments where different pollution sources dominate in different locations — the four-component solution explains only 54.3% of variance. Components become harder to interpret because each explains only slightly more than random chance. Applied researchers should verify that mean $|r|$ exceeds 0.30 before applying PCA to a multi-pollutant dataset; below this threshold, individual-variable analysis may be more informative.

Component Recovery: Tucker's Congruence

Table 5 summarizes Tucker's congruence across 100 replications per sample size.

n	Mean ϕ	Min ϕ	$\phi \geq 0.95$ (%)	$\phi \geq 0.85$ (%)
50	0.918	0.774	29.0	89.0
100	0.956	0.814	69.0	98.0
150	0.968	0.778	88.0	97.0
200	0.976	0.820	94.0	98.0
280	0.982	0.888	98.0	100.0
350	0.988	0.953	100.0	100.0

Table 5. Tucker's congruence coefficient between recovered and reference ($n = 5,000$) loading matrices, 100 replications per sample size. $\phi \geq 0.95$ = stable and interpretable; $\phi \geq 0.85$ = acceptable with minor instability (Lorenzo-Seva & ten Berge, 2006).

At $n = 50$, only 29% of replications achieve $\phi \geq 0.95$, though 89% achieve $\phi \geq 0.85$ — indicating that solutions are not severely distorted but are insufficiently precise for fine-grained component interpretation. The stability rate reaches 94% at $n = 200$ and 100% at $n = 350$. For environmental monitoring networks where the target is publishing interpretable, replicable component structures — as required for regulatory reporting or inter-city comparisons — $n = 280$ is the recommended minimum, where 98% of solutions are stable. For preliminary exploratory analyses, $n = 100$ is acceptable if the researcher acknowledges 31% probability of a solution deviating from the population structure.

The Dynamic Pollution Signature: Longitudinal Extension

Seasonal Pollution Dynamics in ASEAN Cities

ASEAN cities experience two climatically distinct seasons that produce systematically different pollution profiles. The dry season (November to April) brings reduced rainfall, biomass burning from land clearance in mainland Southeast Asia, and intensified photochemical activity under high solar radiation. PM_{2.5}, PM₁₀, NO₂, and O₃ peak together during this period, producing conditions under which F1 (Vehicular-Industrial Air) and F2 (Photochemical Oxidant Stress) dominate simultaneously. The monsoon season (May to October) brings wet deposition that suppresses airborne particulates but intensifies urban runoff, elevating BOD, TDS, and Fecal Coliform in urban waterways as surface drainage carries accumulated

street contamination into rivers. F3 (Water Contamination) becomes the dominant component for riverside monitoring districts during monsoon onset.

Urban heat intensity follows a seasonal arc tied to solar zenith angle and vegetation phenology: LST and UTCI peak in the pre-monsoon transition months (March to May) when solar radiation is highest and vegetation cover is lowest before rainy-season growth. F4 (Urban Heat and Noise Burden) thus peaks in the hot-dry transition rather than the peak of the dry season. These seasonal dynamics mean that the dominant pollution component for any given district shifts across the calendar year in a pattern that is predictable from physical drivers but variable across districts depending on their urban morphology, proximity to water bodies, and traffic infrastructure.

Formal Signature Definition and Longitudinal Design

The dynamic pollution signature $D = (D(1), \dots, D(T))$ for a district across T seasonal observations is computed as follows. A reference loading matrix is estimated from the full monitoring network at baseline (dry season, year 1). At each subsequent observation t , each district's ten standardized variable values are projected onto this reference loading matrix to produce four component scores ($z_1(t)$, $z_2(t)$, $z_3(t)$, $z_4(t)$) without refitting PCA — a score projection procedure that preserves comparability across time points. The dominant component $D(t) = \operatorname{argmax}_j |z_j(t)|$ identifies which pollution domain most strongly characterizes that district at that time.

A four-season annual design yields signatures of length $T = 4$: dry season, pre-monsoon, monsoon, and post-monsoon. A four-factor structure produces $4^4 = 256$ possible signature types, though in practice ASEAN urban districts cluster into a small number of recurring trajectory types determined by their urban form. Traffic-dominant inner-city districts show (Air, Air, Water, Air) trajectories. Riverside industrial districts show (Air, Air, Water, Water) trajectories. Peri-urban residential districts with low traffic and greenspace show (Heat, Heat, Heat, Heat) or (Heat, Oxidant, Water, Heat) trajectories depending on proximity to industrial corridors.

Policy Applications of the Signature

The dynamic pollution signature translates directly into temporally targeted environmental policy in three ways. First, seasonal intervention timing: a district with $D(\text{dry}) = \text{Air}$ and $D(\text{monsoon}) = \text{Water}$ requires emission control enforcement concentrated in October–November (pre-dry season onset) and pre-monsoon drainage maintenance in March–April — two entirely different interventions targeting different infrastructure in different months. Uniform annual emission standards applied to both seasons simultaneously are simultaneously too weak in peak season and unnecessarily restrictive in the off-season for each domain.

Second, inter-district prioritization: within a city, the signature distribution identifies which districts face the most severe and broadest pollution burden. A district with four-season dominance across multiple components ($D = \text{Air, Water, Heat, Oxidant cycling}$) faces compounding cross-domain burden requiring integrated intervention. A district with single-component dominance across all seasons ($D = \text{Water, Water, Water, Water}$) is a more tractable single-domain problem. City environmental authorities can allocate limited intervention budgets by signature severity rather than by single-pollutant rankings that miss cross-domain compounding.

Third, temporal trend detection: year-on-year shifts in the signature distribution across the city provide the earliest signal of environmental policy effectiveness. If vehicle electrification reduces F1 (Vehicular–Industrial Air) dominance across dry-season observations over five years, the proportion of districts with $D(\text{dry}) = \text{Air}$ falls before statistically significant changes in individual PM_{2.5} or NO₂ time series emerge, because the component integrates co-variation across multiple indicators simultaneously. The signature is a more sensitive policy-tracking instrument than any individual monitored variable.

Discussion

What the Simulation Establishes

Three findings have direct implications for environmental monitoring practice in ASEAN. First, PCA recovers a four-component multi-domain pollution structure reliably at $n \geq 200$ monitoring stations (Tucker $\phi \geq 0.95$ in 94% of replications). ASEAN member states with urban monitoring networks of 200 or more stations — Thailand, Indonesia, Vietnam, the Philippines — can apply PCA with high confidence in component stability. Smaller national networks should treat PCA as preliminary and replicate across multiple seasonal datasets before drawing regulatory conclusions.

Second, the AIC advantage of PCA over raw-variable regression is stable across all seven correlation conditions tested ($\Delta\text{AIC} = 18.3$ to 20.0). This stability implies that the model-selection case for PCA does not depend on achieving high inter-variable

correlation — a reassuring finding for environmental researchers working in heterogeneous urban environments where domain-crossing correlations may be modest. The parsimony gain from six fewer parameters consistently exceeds the cost of approximating the full variable structure with four orthogonal components.

Third, MSE reduction is most pronounced at small n , where OLS multicollinearity damage is greatest. Environmental researchers working with small networks ($n < 100$) stand to gain most from PCA preprocessing in predictive models, precisely because their raw-variable estimators are most unstable. The simulation quantifies this gain precisely: 13.6% MSE reduction at $n = 50$ vs. 1.5% at $n = 450$.

Limitations

Three limitations qualify the simulation findings. First, multivariate normality is assumed throughout. Real environmental monitoring data carries non-normality from extreme pollution events, instrument detection limits producing left-censored distributions, and spatial autocorrelation among geographically proximate stations. Robustness of PCA to non-normality has been examined in air quality contexts (Henry et al., 1984; Paatero & Tapper, 1994), but the specific sensitivity boundaries in Tables 3 and 5 may shift under non-normal distributions. Future work should vary the distributional assumptions.

Second, the simulation assumes temporal independence across replications. Real monitoring networks exhibit spatial autocorrelation — neighboring stations tend to produce correlated observations — that violates the independence assumption of standard PCA. Geographically weighted PCA (GWPCA) extends the standard framework to account for spatial non-stationarity (Harris et al., 2011) and should be considered when monitoring station spacing is small relative to the spatial scale of pollution gradients.

Third, the four-factor structure is fixed by design. Real ASEAN urban environments may support three, five, or six meaningful pollution dimensions depending on the mix of industrial, traffic, and agricultural emission sources. The sensitivity boundaries in Tables 3 and 5 apply specifically to the four-factor case; researchers targeting a different number of factors should re-run the sensitivity analysis for their specific loading structure.

Conclusion

This simulation study demonstrated PCA's application to a ten-variable multi-domain urban pollution dataset with a known four-factor structure representing Vehicular-Industrial Air Pollution, Photochemical Oxidant Stress, Water Contamination, and Urban Heat and Noise Burden, quantified component recovery sensitivity across sample size and correlation conditions, and introduced the dynamic pollution signature for longitudinal environmental monitoring.

The principal findings are four. PCA recovers the true four-component structure reliably at $n \geq 200$ (Tucker $\phi \geq 0.95$ in 94% of replications) and fails in 71% of replications at $n = 50$. The AIC advantage of PCA over raw-variable regression is stable across all correlation conditions ($\Delta\text{AIC} = 18.3$ to 20.0), confirming that parsimony — six fewer parameters — is the dominant driver of the model-selection advantage. MSE reduction is largest at small n (13.6% at $n = 50$), providing the strongest quantitative case for PCA preprocessing in small monitoring networks. Applied researchers should not use PCA when mean inter-variable correlation falls below $|r| \approx 0.30$, as components lose interpretability even when computation remains feasible.

The dynamic pollution signature extends these cross-sectional findings into a policy-relevant longitudinal framework. Each district's characteristic seasonal sequence of dominant pollution components is individually informative, directly translatable into temporally targeted intervention priorities, and detects policy effectiveness before individual pollutant time series show significant change. The four components established here — Air, Oxidant, Water, Heat-Noise — constitute the vocabulary of the signature. Longitudinal deployment across ASEAN monitoring networks is the direct next application.

The simulation code is available from the author on request.

The dynamic pollution signature concept draws on an analogy from integrative environmental assessment practice, in which each location presents a shifting pattern of dominant stressors across seasons rather than a fixed pollution profile, and the assessor tracks the current leading stressor to determine the appropriate intervention. The parallel to longitudinal PCA is direct: the signature captures the temporal unfolding of multi-domain environmental burden in a form that is both individually specific and policy-actionable.

References

1. Bañaga, P. A., Braun, R. A., Betito, G., Cruz, M. T., Orbe, C. M., Cambaliza, M. O., & Holzapfel, A. M. (2021). Evidence of long-range transport contributing to the 2019 biomass burning haze episode in Metro Manila, Philippines. *Environmental Research Letters*, 16(1), 014010.

2. Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.
3. Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., ... & Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution. *The Lancet*, 389(10082), 1907–1918.
4. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.
5. Harris, P., Brunsdon, C., & Charlton, M. (2011). Geographically weighted principal components analysis. *International Journal of Geographical Information Science*, 25(10), 1717–1736.
6. Henry, R. C., Cayouette, M., & Savoie, A. (1984). Factor analysis of environmental receptor models. *Atmospheric Environment*, 18(6), 1507–1515.
7. Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2), 179–185.
8. IQAir. (2023). 2022 World Air Quality Report. IQAir.
9. Jiang, Y., Guo, Y., Li, X., & Zhao, Q. (2018). Multivariate statistical analysis of water quality in a Chinese river watershed. *Environmental Monitoring and Assessment*, 190(5), 286.
10. Lorenzo-Seva, U., & ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, 2(2), 57–64.
11. Paatero, P., & Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2), 111–126.
12. Pongpiachan, S., Tipmanee, D., Deelaman, W., Maneechot, L., Uthaiattanacheep, N., & Sompongchaiyakul, P. (2019). Seasonal and spatial variations of carcinogenic risk of PM_{2.5}-bound polycyclic aromatic hydrocarbons in Bangkok, Thailand. *Atmospheric Pollution Research*, 10(4), 1227–1236.
13. Roth, M. (2007). Review of urban climate research in (sub)tropical regions. *International Journal of Climatology*, 27(14), 1859–1873.
14. Santoso, M., Lestiani, D. D., Mukhtar, R., Hamonangan, E., Syafrul, H., Markwitz, A., & Hopke, P. K. (2021). Preliminary study of the sources of ambient air pollution in Serpong, Indonesia. *Journal of the Air & Waste Management Association*, 61(2), 218–225.
15. Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
16. WHO. (2021). WHO global air quality guidelines: Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization.
17. Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99(3), 432–442.



2026 by the authors; Journal of Global Social Transformation (JGST). This is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).