Explainable Machine Learning Reveals Stage-Dependent Predictors of Global PM2.5 Exposure for Environmental Governance

Qifeng Li*, Shuo Wen
Guangdong Police College, Guangzhou, China
**Corresponding Author*

Abstract: We used panel data from the World Bank covering 200 economies from 1990 to 2023. With these data, we built population-weighted PM2.5 exposure models using random forest, XGBoost, LightGBM, and linear regression. The random forest model (R^2 = 0.827) outperformed XGBoost (R^2 = 0.805), LightGBM (R^2 = 0.794), and linear regression (R^2 = 0.419). SHAP analysis showed that population size and GDP per capita made the largest contributions, followed by the share of renewable energy, forest cover, and energy consumption per capita. The marginal contribution of GDP per capita to PM2.5 falls from positive to negative across development stages, which we read as consistent with the environmental Kuznets curve hypothesis. We also observe a sign reversal between the 40%–60% and 60%–80% urbanization bands. Population, renewable energy, and forest cover show stage-specific reversals as well. Our results suggest that economic growth and PM2.5 are not monotonically coupled, and that the link between structural conditions and exposure levels varies across development stages. We believe these findings provide data-driven and interpretable evidence for phased air quality management and sustainable development planning. They also reveal where environment-related public complaints are most likely to accumulate, which gives an evidence base for allocating social governance resources and preventing public order risks.

Keywords: PM2.5; Machine Learning; SHAP; Environmental Kuznets Curve; Sustainable Development; Socioeconomic Correlates

1. Introduction

Ambient air pollution causes millions of premature deaths each year. Together with dietary risks and smoking, it ranks among the major global health risks [1–3]. PM2.5 is one of the pollutants in ambient air pollution. More than 90% of the world's population lives in areas where PM2.5 concentrations exceed the World Health Organization's 5 μg/m³ guideline value [1,4]. High-income countries have achieved substantial emission reductions, yet exposure levels continue to rise in many low- and middle-income countries. No single factor determines exposure levels. Economic activity, energy systems, urbanization, demographic structure, and land use all shape them jointly. Identifying which structural conditions raise or lower exposure levels therefore has practical significance for sustainable development.

A long-standing but still unresolved question is whether economic growth inevitably worsens pollution. A related question is whether growth eventually decouples from pollution once a country crosses a certain development threshold. This is the "Environmental Kuznets Curve" (EKC) hypothesis, first proposed for environmental quality in the early 1990s [5–7]. Early panel data studies found an inverted U-shaped relationship between income and various ambient air pollutants, and later studies extended and systematized the hypothesis [7,8]. Serious disagreement remains, however. Reviews and reanalyses conclude that this empirical regularity is fragile. Results are highly sensitive to sample composition, functional form, and pollutant choice, and the estimated turning points span such a wide range that they are unconvincing [9–13]. For particulate matter, arguably the pollutant with the greatest health impacts, the evidence is especially contradictory. Whether a turning point exists, what shape it takes, and where it lies all remain disputed [9,10,12].

A key methodological difficulty lies behind this impasse. Income, energy consumption, urbanization, population, and industrial structure are highly collinear. Their relationships are also

 Copyright @ STEMM Institute Press

nonlinear and vary with development stage. Traditional parametric panel regression is therefore prone to misattribution. Results may become unstable when the model specification changes, and they are often inconsistent out of sample. The EKC literature itself is full of such warnings: under different functional forms, country samples, or estimation periods, the magnitude and even the sign of coefficients can reverse. We see such fragility as exactly what one would expect when highly collinear independent variables are forced into rigid linear or polynomial models [11,12].

Tree-based model ensembles can capture nonlinear relationships and interactions without a pre-specified functional form. Inspired by this ability, recent studies have increasingly applied machine learning to environmental and air quality prediction [14–16]. These models often achieve high predictive accuracy. Three recurring problems remain, though. First, many studies evaluate models with random or in-sample splits. Such splits can inflate apparent fit, and they cannot establish whether the reported performance truly reflects prediction for unseen years or countries. Second, predictor sets often include contemporaneous or emissions-based variables. Models may then derive the target variable through definitional relationships rather than through the structural predictors that policy actually cares about. Third, even accurate models are often presented as "black boxes". Studies summarize feature importance but do not examine the direction and functional form of each feature's effect. A core economic question therefore goes unanswered: how does the marginal effect of income on pollution change with economic development?

We address these three gaps by modelling global PM2.5 exposure. Our model uses 12 categories of predictors, covering economic, energy, structural, demographic, and land-use factors. We interpret it with SHAP to characterize the nonlinear form of each feature's effect and how that form varies with development stage. We therefore ask two questions. Does an environmental Kuznets curve exist for PM2.5 in the global panel, and if so, where does its turning point lie? And is the association between structural conditions and exposure stable, or does it change with development stage? The study also carries implications for social governance. If we can identify the specific conditions associated with high or rising exposure at different development stages, administrative agencies can better anticipate where public health burdens and environment-related demands are likely to concentrate, and where public order risks may emerge.

2. Materials and Methods

2.1 Data

We used the World Development Indicators dataset from World Bank Open Data [17], and retrieved 16 candidate indicators. We first selected 217 economies that had both geographic -region and income-level classifications. We then dropped economies with less than 20 years of data coverage or with no PM2.5 observations. The final sample comprised 200 economies covering 1990–2023. The outcome variable was population-weighted mean annual exposure to PM2.5 (indicator EN.ATM.PM25.MC.M3, μg/m³). This variable measures average exposure across the population of an economy. It is complete over 1990–2023 for the retained economies (6,800 economy–year observations).

Of the sixteen candidate indicators, one is the outcome itself (PM2.5), twelve are socioeconomic and structural predictors, and three are carbon-emission indicators. The three are CO2 emissions per capita (EN.GHG.CO2.PC.CE.AR5), total CO2 emissions (EN.GHG.CO2.MT.CE.AR5), and CO2 emission intensity (EN.GHG.CO2.RT.GDP.KD). During data exploration, we inspected each indicator's year coverage and completeness. We excluded the three carbon-emission indicators before modelling. They are themselves outputs of the same fossil-fuel combustion process that generates particulate matter, so entering them alongside the PM2.5 outcome would establish a definitionally circular relationship. The retained predictors consequently contain no direct emission or pollution metric (Table 1). Predictor coverage is more variable than outcome coverage. The two energy-scale measures are the sparsest, and Section 2.2 characterizes them.

The twelve retained predictors span four broad domains. Economic scale and activity are captured by GDP per capita, GDP growth, industry value added, agriculture value added, and trade. Energy and structural conditions are captured by renewable energy share, energy use per capita, and energy intensity. Demographic conditions are captured by urbanization rate,

Copyright @ STEMM Institute Press

population size, and population growth. Land use is captured by forest cover. We log-transformed four right-skewed scale variables (GDP per capita, population, energy use per capita, and energy intensity). All variables retain their World Bank indicator codes; Table 1 lists units and codes.

Table 1. Predictor Variables

Predictor	Indicator code	Units	Log-transformed
GDP per capita	NY.GDP.PCAP.KD	constant 2015 USD	Yes
GDP growth	NY.GDP.MKTP.KD.ZG	annual %	No
Industry value added	NV.IND.TOTL.ZS	% of GDP	No
Agriculture value added	NV.AGR.TOTL.ZS	% of GDP	No
Trade	TG.VAL.TOTL.GD.ZS	% of GDP	No
Renewable energy share	EG.FEC.RNEW.ZS	% of total final energy consumption	No
Energy use per capita	EG.USE.PCAP.KG.OE	kg oil equivalent	Yes
Energy intensity	EG.EGY.PRIM.PP.KD	MJ per 2017 PPP USD GDP	Yes
Urbanization rate	SP.URB.TOTL.IN.ZS	% of population	No
Population	SP.POP.TOTL	persons	Yes
Population growth	SP.POP.GROW	annual %	No
Forest cover	AG.LND.FRST.ZS	% of land area	No

2.2 Preprocessing

We pivoted records to a wide economy-by-year matrix. We dropped observations before 1990 and removed economies with fewer than twenty years of coverage. To prevent information from the evaluation period influencing the training data, we split the panel by year before any imputation. We then interpolated missing values within each economy in both directions and filled with a within-block median. We applied this entire imputation procedure separately within the training block (1990–2014) and the test block (2015–2023). One consequence of this choice is that features with structurally sparse coverage are completed from neighbours rather than treated as absent. The outcome itself is complete, so no outcome values are imputed; Section 4 revisits the implications. The two energy-scale variables (energy use per capita and energy intensity) are markedly sparser than the others, with approximately 5,040 and 4,310 observations, respectively. Energy intensity is available only over 2000–2022. Energy use per capita is recorded across the full window but is also far from complete. Section 4 revisits their imputation and interpretation.

2.3 Feature Engineering

We kept feature engineering deliberately minimal. We log-transformed (log1p) four right-skewed scale variables (GDP per capita, population, energy use per capita, and energy intensity) to compress their long upper tails. We entered the remaining predictors, which are ratios or growth rates, untransformed (Table 1). We constructed no interaction terms, polynomial expansions, or manual discretizations as model inputs. The tree ensembles capture nonlinearities and interactions natively, and we applied any binning only post hoc to the SHAP outputs for interpretability (Section 2.6). As an automated feature-selection screen, we fitted a least absolute shrinkage and selection operator (LASSO) model (LassoCV with five-fold cross-validation) to identify predictors with negligible linear contribution [18]. All twelve features survived the screen, so the full predictor set entered every model. Section 2.2 describes missing-value treatment.

2.4 Modeling and Validation

We compared four estimators: linear regression, random forest [14], XGBoost [15], and LightGBM [16]. We performed hyperparameter optimization for the two boosting models under five-fold grouped cross-validation with economies as groups. No economy therefore contributes observations to both a training fold and its paired validation fold, which prevents the optimistic bias that arises when a single country's correlated years leak across the split. The LASSO feature screen used ordinary five-fold cross-validation restricted to the training block. We also restricted the grouped cross-validation used for tuning and the Optuna search to the training block (1990–2014), so that no observation from the 2015–2023 evaluation period informed model selection or hyperparameter choice.

2.5 Hyperparameter Tuning

We tuned LightGBM and XGBoost hyperparameters with Optuna [19], using a tree-structured Parzen estimator sampler (seed 42). The objective was the grouped cross-validation root-mean-squared error, minimized over 50 trials for LightGBM and 30 trials for XGBoost.

 Copyright @ STEMM Institute Press

Table 2 reports the search spaces and winning configurations.

Table 2. Hyperparameter Search Spaces and Winning Configurations for LightGBM (50 Trials) and XGBoost (30 Trials)

Model	Hyperparameter	Search range	Type	Winning value
LightGBM	num_leaves	15–127	integer	118
LightGBM	learning_rate	0.01–0.2 (log scale)	float	0.011
LightGBM	min_child_samples	5–60	integer	53
LightGBM	feature_fraction	0.6–1.0	float	0.61
LightGBM	bagging_fraction	0.6–1.0	float	0.60
XGBoost	max_depth	3–10	integer	5
XGBoost	learning_rate	0.01–0.2 (log scale)	float	0.021
XGBoost	min_child_weight	1–10	integer	1
XGBoost	subsample	0.6–1.0	float	0.84
XGBoost	colsample_bytree	0.6–1.0	float	0.73

2.6 SHAP-Based Analysis

We computed feature contributions with Shapley additive explanations (SHAP) using TreeExplainer on the best-performing model identified in Section 3.1 (the random forest) [20,21]. SHAP values decompose each prediction into an additive sum of feature contributions with a principled game-theoretic grounding [22]. This permits local explanation of individual predictions and, through aggregation, a global account of feature importance and direction. We summarized global feature importance by the mean absolute SHAP value (mean |SHAP|) over the test set.

We recovered the functional form of the income–pollution relationship in two complementary ways. First, we inspected the SHAP dependence of GDP per capita, plotting its contribution against its own value and against urbanization to visualize the sign reversal. Second, we stratified mean SHAP contributions across four urbanization bands (<40%, 40–60%, 60–80%, >80%) as a proxy for development stage, which yielded the stage-dependent contribution profile of every feature. We assessed feature coupling with Spearman rank correlation between Shapley vectors. This quantifies redundancy between feature contributions without requiring an exact interaction decomposition.

3. Results

3.1 Predictive Performance

The socioeconomic and structural predictors yielded high-fidelity predictions of held-out PM2.5 exposure (Table 3). Linear regression captured only 41.9% of the variance. Tree ensembles performed far better, which reflects pronounced nonlinearity in the predictors of exposure. The random forest achieved the best out-of-sample fit (R^2 = 0.827, RMSE = 6.53 μg/m³, MAE = 4.39 μg/m³), ahead of tuned XGBoost (R^2 = 0.805) and tuned LightGBM (R^2 = 0.794). The fixed-configuration random forest even slightly outperformed both tuned boosting models on the held-out years. All three tree ensembles nevertheless lie within a narrow band (R^2 = 0.794–0.827). The roughly 41-point gap between linear regression and the tree ensembles confirms that the predictors of PM2.5 operate nonlinearly, a central premise of the analysis that follows. Figure 1 compares predicted and observed exposure across the 1,800 held-out observations. The random forest tracks the observed exposure gradient without systematic bias across the range.

3.2 Feature Importance

Population size was the single most influential predictor of PM2.5 exposure, with a mean |SHAP| of 4.85 μg/m³. GDP per capita followed closely (4.40 μg/m³) (Figure 2). Renewable energy share (1.93 μg/m³), forest cover (1.47 μg/m³), and energy use per capita (1.40 μg/m³) formed a second tier. Urbanization rate, energy intensity, trade, population growth, and sectoral shares contributed more modestly. GDP growth was essentially uninformative (0.04 μg/m³).

The overall signed direction of each contribution is also informative (Figure 3). Averaged across all test observations, GDP per capita carried a negative mean contribution (−1.49 μg/m³). We attribute this to the predominance of high-income, low-exposure economies in the sample. Population, by contrast, carried a positive mean contribution (+0.89 μg/m³). These global averages conceal the stage dependence documented in Section 3.3. In our view, the stage-stratified analysis is more informative than a single average. Figure 3 summarizes the full distribution of each feature's contributions as a

Copyright @ STEMM Institute Press

beeswarm plot, along with the mapping from high or low feature values to positive or negative contributions.

Table 3. Out-of-Sample Predictive Performance (Held-Out Years 2015–2023)

Model	R^2	RMSE (μg/m³)	MAE (μg/m³)
Linear regression	0.419	11.98	8.57
Random forest	0.827	6.53	4.39
XGBoost (tuned)	0.805	6.93	5.09
LightGBM (tuned)	0.794	7.13	5.12

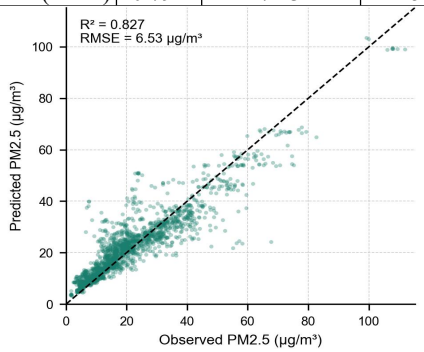


Figure 1. Predicted Versus Observed Population-Weighted PM2.5 Exposure for 1,800 Held-Out Economy–Years (2015–2023), Random Forest; the Dashed Line is the 1:1 Identity

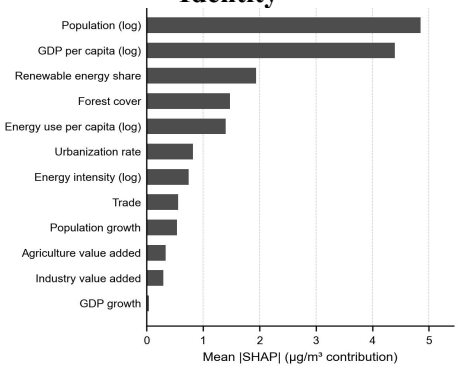


Figure 2. Global Predictor Importance for PM2.5 Exposure on the Held-Out Test Years (2015–2023), Ranked by Mean Absolute SHAP Value (μg/m³)

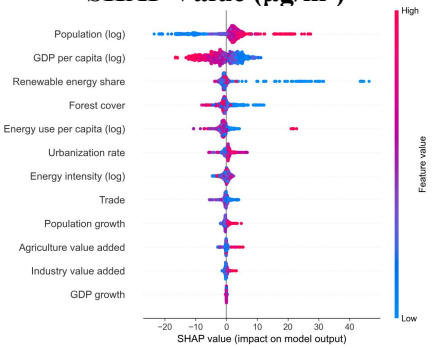


Figure 3. SHAP Beeswarm on the Held-Out Test Years; Each Point is one Economy–Year, Colored by Predictor Value (Red High, Blue Low), Ordered by Mean Absolute SHAP Value as in Figure 2

3.3 A Stage-Dependent Income Gradient Consistent with an Environmental Kuznets Curve

The SHAP dependence of GDP per capita was nonlinear. It declined from positive to negative across the observed income range. When we stratified by development stage, the contribution of GDP per capita declined from strongly positive at low urbanization (+2.80 μg/m³ for <40%) to weakly positive at 40–60% (+0.58 μg/m³). It then turned increasingly negative at higher urbanization (−2.76 μg/m³ for 60–80%; −6.04 μg/m³ for >80%) (Figure 4, Figure 5). We read this systematic sign reversal as consistent with an Environmental Kuznets Curve for PM2.5. At early development, higher income goes with higher exposure. At advanced development, higher income goes with markedly lower exposure. The sign reversal occurs between the 40–60% and 60–80% urbanization bands, near the 60% urbanization level typical of many middle-income economies. In the income dimension, the marginal contribution crosses zero near log GDP 8.3 (about USD 4,100 per capita), the income level at which an inverted-U peak would lie. The four-band gradient falls monotonically, as an inverted U in levels would require.

3.4 Nonlinear Structural Predictors

The stage-dependent profile was not confined to income. Population contributed positively through the low, middle, and upper stages (+1.51, +1.72, and +1.28 μg/m³). It turned negative among the most urbanized economies (−0.98 μg/m³). We read this as evidence that in the most developed economies, a larger aggregate population no longer translates into higher exposure. Presumably their energy and transport systems have been sufficiently decarbonized and regulated. Urbanization contributed negatively at the two lower stages (−0.59 and −0.12 μg/m³) but turned positive at the two upper stages (+0.94 and +0.82 μg/m³). Higher urbanization therefore goes with higher exposure across the upper half of the development spectrum (Figure 5).

Renewable energy share was associated with lower predicted exposure in low- and middle-development stages (−0.49, −0.14, −0.01 μg/m³), which is consistent with a mitigation role. Among the most urbanized economies, however, this association turned strongly positive (+2.16

 Copyright @ STEMM Institute Press

μg/m³). The share averages 17.2% in that band, below the less urbanized bands, so this reversal is not a saturation effect; it reflects the heterogeneous composition of renewable sources (biomass, hydro, wind, and solar). Energy use per capita contributed negatively across all stages (−0.31 to −1.31 μg/m³), while energy intensity was weak across stages (+0.12 to −0.27 μg/m³). These energy measures are collinear with income, so we read their signs as reflecting affluence and the maturity of fossil-fuel energy systems rather than direct exposure effects. Population growth was negligible at the two lower stages (+0.07, −0.04 μg/m³) and more clearly negative at the upper stages (−0.29 and −0.12 μg/m³), another stage-dependent shift beyond income. Forest cover acted as a negative contributor through the low and middle stages (−0.64, −0.47, −0.27 μg/m³) but turned positive among the most urbanized economies (+0.38 μg/m³). This reversal likely reflects the heterogeneous composition of the small set of highly urbanized economies rather than a causal property of forests. Industry value added and trade were weakly negative across most stages. Agriculture value added was modest and sign-mixed, and GDP growth was negligible throughout.

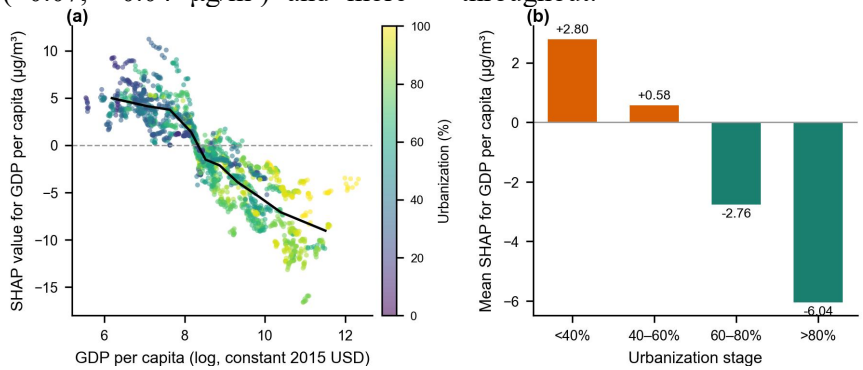


Figure 4. Stage-Dependent Sign Reversal Consistent with an Environmental Kuznets Curve: (a) SHAP Dependence of GDP Per capita, Colored by Urbanization Rate; (b) Mean SHAP of GDP Per capita by Urbanization Stage, Reversing from +2.80 to −6.04

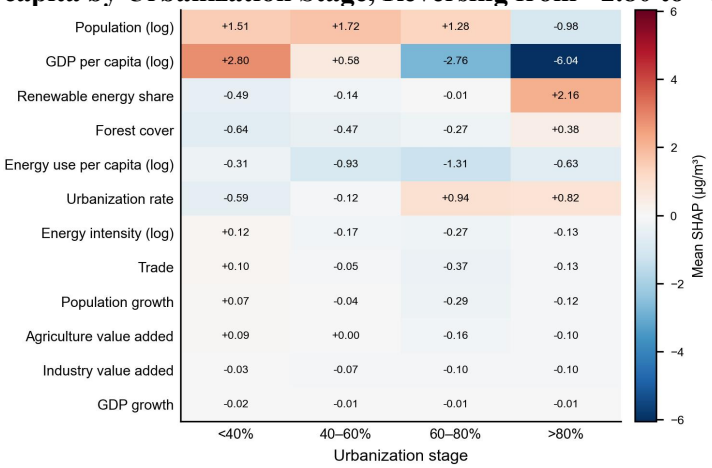


Figure 5. Mean SHAP Contribution (μg/m³) of Each Predictor by Urbanization Stage

3.5 Feature Coupling

To assess whether the contribution decomposition was dominated by a few redundant dimensions, we computed Spearman rank correlations between Shapley vectors across the test set (Figure 6). Coupling between feature contributions was mild to moderate. The strongest associations involved GDP per capita with renewable energy share ($\rho \approx 0.43$), with population growth ($\rho \approx 0.42$), and with population size ($\rho \approx 0.41$). All remaining pairwise correlations were below 0.40. The only notable negative coupling was between GDP per capita and urbanization rate ($\rho \approx -0.34$). These correlations reflect the shared development axis running through income, demography, energy, and sectoral structure. We therefore interpret the contribution profile of GDP per capita alongside its correlated development indicators rather than in isolation. No pair approaches a near-duplicate coupling ($\rho > 0.9$), so the stage-dependent profiles retain feature-specific information rather than collapsing onto a single redundant

dimension.

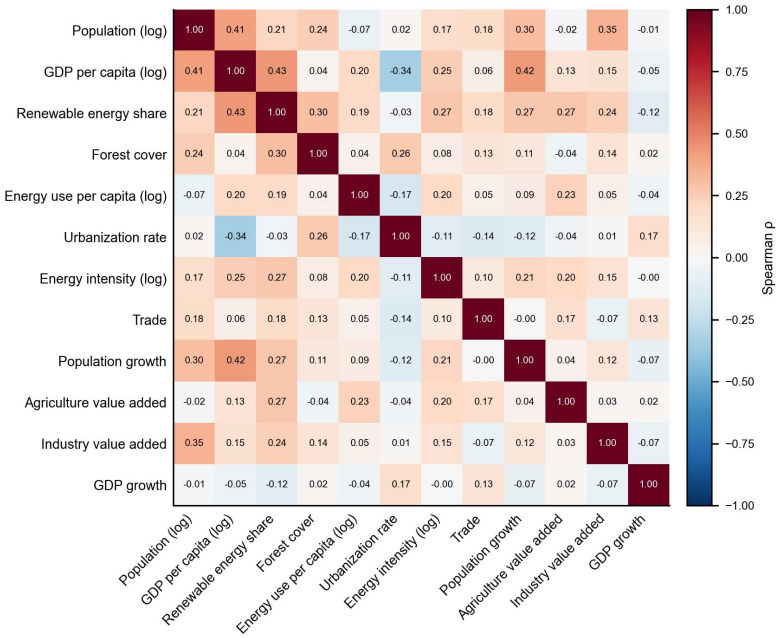


Figure 6. Spearman Rank Correlations between the Shapley Contribution Vectors of the Twelve Predictors on the Held-Out Test Years, Ordered by Global Importance

4. Discussion

Our central finding is that economic growth and PM2.5 exposure are not monotonically coupled. Across the global panel of 200 economies, the marginal contribution of GDP per capita shifted from positive to negative as urbanization increased, crossing zero near log GDP 8.3 (about USD 4,100 per capita). This sign reversal is consistent with an inverted-U relationship of the kind the Environmental Kuznets Curve describes, with the stage transition located between the 40–60% and 60–80% urbanization bands, near the 60% urbanization level. Our qualitative conclusion agrees with the classic inverted-U hypothesis [5–8]. We recover it here under a temporal out-of-sample design, however: the pattern was reproduced from held-out years that the model had never seen. Much of the EKC literature reports estimates that are fragile to specification and sample [11,12]. The nonlinear, interaction-capable estimator we use reaches the same qualitative conclusion from out-of-sample prediction of unseen years.

GDP per capita and population size were the two main influencing factors, and both are closely related to energy demand, transport volume, and land-use change. Their contributions also vary with development stage in instructive ways. Across the low, middle, and high development stages, population size contributes positively to predicted exposure. In the most urbanized economies its sign reverses. We take this to mean that exposure intensity for a given population size differs depending on an economy's existing energy and transport systems. In the low and middle development stages, the share of renewable energy is negatively associated with predicted exposure, a pattern consistent with the energy-structure dimension of the growth–exposure relationship. At the highest urbanization level, this association turns positive; the share averages 17.2% there, below the less urbanized bands, and the indicator aggregates biomass, hydro, wind, and solar.

The sign-reversing behaviour is not confined to income and population. Renewable energy share and forest cover also reversed the direction of their association across urbanization stages. The association between structural conditions and exposure therefore behaves as a stage-dependent profile rather than a stable constant. This carries a direct implication for modelling strategy. A single linear coefficient for forest cover or urbanization, estimated over the pooled panel, would average across opposing signs. It would yield a weak or null estimate even where strong, interpretable stage-dependent associations exist. The stage-stratified SHAP profile therefore retains information that pooled parametric models systematically discard. In our view this offers one methodological account of why pooled estimates of the income–pollution

 Copyright @ STEMM Institute Press

relationship have proven so fragile [9–12].

Several limitations bound these inferences, and they fall into two groups. The first group concerns data coverage and preparation. The energy-scale variables (energy use per capita and energy intensity) have the sparsest coverage in the panel. The negative contributions we recover for energy use per capita likely reflect its collinearity with income and affluence rather than a direct physical pathway. We completed missing predictor values by interpolation and median filling within each split. This may smooth genuine temporal variation and attenuate within-country dynamics. National means, moreover, mask within-country heterogeneity: a country-level GDP per capita does not separate the many subnational pathways by which income acts.

The second group concerns interpretation and design. The sign-reversing contributions of forest cover and renewable energy at high urbanization in particular may reflect compositional confounding among a small set of economies rather than a mechanistic relationship. Urbanization bands serve as a proxy for development stage, so the transition is located on the urbanization axis rather than directly on income. The exclusion of emission variables prevents circularity, but it also omits the direct physical pathway and confines the analysis to distal socioeconomic and structural predictors.

Notwithstanding these caveats, the results carry a practical message for air-quality and sustainable-development policy. For social governance in particular, the stage-dependent profile indicates where exposure-related health burdens and environment-related public complaints are most likely to accumulate. That information is relevant to administrative resource allocation, urban grid management, and the prevention of public-order risks. Because the association between income and exposure changes sign with development stage, growth-oriented objectives and exposure-reduction objectives are not uniformly in tension. The tension is sharpest in middle-income economies within the middle urbanization bands. Advanced economies, by contrast, display an association consistent with growth having been decoupled from exposure. Exposure mitigation accordingly needs to engage urbanization and demographic trajectories rather than income growth alone. Our patterns further generate a hypothesis: interventions are best calibrated to development stage, for example by prioritizing the energy mix in economies in the upper-middle urbanization range.

5. Conclusions

Using an explainable machine-learning framework and a global panel of 200 economies, we showed that purely socioeconomic and structural predictors account for approximately 83% of the variance in held-out PM2.5 exposure across the test years. GDP per capita and population size were the two leading contributors, closely matched and well ahead of the rest. Renewable energy share, forest cover, and energy use per capita followed, and we recovered their contributions with interpretable directions and functional shapes. These contributions were not uniform. Their strength varied with development stage, and for income even their sign did. The result is a distinctly stage-dependent predictor structure. We provide evidence consistent with an Environmental Kuznets Curve for PM2.5, with the sign reversal located between the 40–60% and 60–80% urbanization bands, near the 60% urbanization level. We also find a mitigation role for renewable energy at earlier development stages. The stage-dependent, sign-reversing profile of income, population, renewable energy, and land use suggests that the associations of structural conditions with exposure vary with development stage. Pooled parametric estimates discard exactly this information. We believe these findings offer a data-driven, explainable basis for aligning environmental governance and public-service planning with the specific development conditions of each economy. They also illustrate the value of leakage-free evaluation and Shapley-based explanation for governance-oriented environmental analysis more broadly.

Acknowledgment

This paper was funded by the "Golden Eagle Scholar" Team Project of Guangdong Police College (2026JY01), the "Young Eagle Program" Project of Guangdong Police College (2026QY12), the "Young Eagle Program" Project of Guangdong Police College (2025FY22), and 2024 Guangdong Province Undergraduate Teaching Quality and Teaching Reform Project (No. 879).

Copyright @ STEMM Institute Press

References

[1 World Health Organization. WHO Global Air Quality Guidelines: Particulate Matter (PM2.5 and PM10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide; World Health Organization: Geneva, Switzerland, 2021.

[2] Cohen, A.J.; Brauer, M.; Burnett, R.; Anderson, H.R.; Frostad, J.; Estep, K.; et al. 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. Lancet 2017, 389, 1907–1918.

[3] Burnett, R.; Chen, H.; Szyszkowicz, M.; Fann, N.; Hubbell, B.; Pope, C.A., III; et al. Global Estimates of Mortality Associated with Long-Term Exposure to Outdoor Fine Particulate Matter. Proc. Natl. Acad. Sci. USA 2018, 115, 9592–9597.

[4] Shaddick, G.; Thomas, M.L.; Mudu, P.; Ruggeri, G.; Gumy, S. Half the World's Population Are Exposed to Increasing Air Pollution. npj Clim. Atmos. Sci. 2020, 3, 23.

[5] Grossman, G.M.; Krueger, A.B. Economic Growth and the Environment. Q. J. Econ. 1995, 110, 353–377.

[6] Shafik, N. Economic Development and Environmental Quality: An Econometric Analysis. Oxf. Econ. Pap. 1994, 46, 757–773.

[7] Selden, T.M.; Song, D. Environmental Quality and Development: Is There a Kuznets Curve for Air Pollution Emissions? J. Environ. Econ. Manag. 1994, 27, 147–162.

[8] Holtz-Eakin, D.; Selden, T.M. Stoking the Fires? CO2 Emissions and Economic Growth. J. Public Econ. 1995, 57, 85–101.

[9] Stern, D.I. The Rise and Fall of the Environmental Kuznets Curve. World Dev. 2004, 32, 1419–1439.

[10]Dinda, S. Environmental Kuznets Curve Hypothesis: A Survey. Ecol. Econ. 2004, 49, 431–455.

[11]Harbaugh, W.T.; Levinson, A.; Wilson, D.M. Reexamining the Empirical Evidence for an Environmental Kuznets Curve. Rev. Econ. Stat. 2002, 84, 541–551.

[12]Carson, R.T. The Environmental Kuznets Curve: Seeking Empirical Regularity and Theoretical Structure. Rev. Environ. Econ. Policy 2010, 4, 3–23.

[13]Kaika, D.; Zervas, E. The Environmental Kuznets Curve (EKC) Theory—Part A: Concept, Causes and the CO2 Emissions Case. Energy Policy 2013, 62, 1392–1402.

[14]Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32.

[15]Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

[16]Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 3149–3157.

[17]World Bank. World Development Indicators; World Bank: Washington, DC, USA, 2025.

[18]Tibshirani, R. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B 1996, 58, 267–288.

[19]Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2623–2631.

[20]Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.

[21]Lundberg, S.M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J.M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; Lee, S.-I. From Local Explanations to Global Understanding with Explainable AI for Trees. Nat. Mach. Intell. 2020, 2, 56–67.

[22]Štrumbelj, E.; Kononenko, I. Explaining Prediction Models and Individual Predictions with Feature Contributions. Knowl. Inf. Syst. 2014, 41, 647–665.

 Copyright @ STEMM Institute Press