

Interpretable Machine Learning Reveals Nonlinear Relationships Between Urban Transportation Sustainability and Residents' Quality of Life

1st Taner Polat

Izmir Institute of Technology

Izmir, Turkey

TanerPolat2026@outlook.com

2nd Emre Candar

Ankara University

Ankara, Turkey

EmreCandar1949@outlook.com

Abstract—Urban transportation systems are closely associated with multiple dimensions of urban livability, but conventional additive and linear specifications may provide an incomplete description of nonlinear and interactive patterns among transport-related indicators. This study develops a harmonised analytical panel for 151 metropolitan areas in seven world regions over 2019 – 2023 ($n = 755$ city-years; 17 transport-sustainability indicators). The panel combines published city-level anchor statistics with harmonised annual estimates derived from documented regional trends where directly comparable annual city observations were unavailable. A transport-related livability index was constructed from environmental, mobility, safety and economic dimensions and analysed using linear, additive and tree-based models. Gradient-boosted trees were interpreted using Shapley additive explanations, partial-dependence analysis, interaction decomposition and a data-driven breakpoint screening procedure. Under city-grouped cross-validation, LightGBM attained $R^2 = 0.880$, compared with $R^2 = 0.826$ for ordinary least squares and $R^2 = 0.866$ for an additive spline model. Green space per capita, road fatality rate, electric-vehicle fleet share, active travel share and PM2.5 concentration accounted for the largest shares of model attribution. The breakpoint procedure identified candidate changes in the fitted response surface for several mobility-related indicators, including car ownership, transit mode share, active travel and congestion. Model-based scenario analysis further indicated substantial heterogeneity in predicted index changes across policy packages and regions. These results should be interpreted as patterns within a constructed transport-related livability framework rather than as identified causal effects on residents' experienced quality of life. The study demonstrates how interpretable machine learning can be used to examine nonlinear structure, interactions and scenario sensitivity in multi-indicator urban sustainability assessments.

Keywords—urban transportation sustainability, quality of life, interpretable machine learning, SHAP, nonlinearity and thresholds, sustainable mobility policy, urban transportation sustainability, quality of life, interpretable machine learning, SHAP, nonlinearity and thresholds, sustainable mobility policy

I. INTRODUCTION

Cities concentrate both the benefits and the burdens of motorised mobility. Urban transport connects residents to employment, education, healthcare and social networks, and it is therefore one of the most direct determinants of how people experience the city in which they live [1,2]. The same

system, however, is also the source of the externalities that most visibly erode urban welfare: fine-particulate and nitrogen-dioxide concentrations that shorten lives, traffic noise that fragments sleep, congestion that consumes discretionary time, and road trauma that remains a leading cause of death among young adults [3,4,5]. Because the sector accounts for roughly a quarter of energy-related carbon dioxide emissions and an even larger share of urban air pollution, transport has become a central arena for sustainability policy at every level of government [6].

The concept of sustainable transportation was formalised to reconcile these competing pressures. It requires that mobility services meet present needs without foreclosing the ability of future generations to meet theirs, and it is conventionally decomposed into environmental, social and economic dimensions [7,8]. A large measurement literature has translated the concept into indicator sets that span modal split, accessibility, affordability, safety, noise, emissions, investment and land use [9,10]. In parallel, a separate literature has developed composite indices of quality of life and urban livability, typically combining objective indicators of environmental quality, material living standards and service provision with subjective measures of satisfaction [11,12].

These two literatures have converged on a shared empirical claim: improving the sustainability of a city's transport system should improve the quality of life of its residents. Yet the evidence supporting that claim is thinner than its ubiquity suggests. Most studies estimate an average, constant marginal effect of each transport indicator, usually through a regression coefficient or a structural path in a system-dynamics model [13,14]. Such estimates are informative about the direction of an effect but silent about its shape. They cannot tell a planner whether the first additional percentage point of transit mode share is worth as much as the twentieth, whether a road-safety intervention retains its value once fatalities have already fallen, or whether air-quality improvements and green-space provision reinforce each other. Where thresholds are known to exist — the epidemiological literature on particulate matter, for example, identifies clearly non-linear exposure – response functions [15,16] — linear urban models can misstate the returns to intervention by a wide margin.

Interpretable machine learning offers a way out of this impasse. Gradient-boosted tree ensembles approximate arbitrary smooth and piecewise functions without

Corresponding Author: Emre Candar, Ankara University Rectorate, Döğol Caddesi, Ankara, Turkey, Izmir Turkey, 06100, EmreCandar1949@outlook.com

presupposing an algebraic form, and they have repeatedly outperformed linear specifications in environmental and urban prediction tasks [17,18]. The obstacle has always been interpretation. Shapley additive explanations resolve that obstacle by decomposing each individual prediction into additive contributions from each input feature, with the property that contributions sum exactly to the difference between the prediction and the sample average [19,20]. Recent work has extended the framework to interaction values, which quantify how much of a prediction depends on the joint behaviour of two features rather than on either alone [21]. Together with partial-dependence and accumulated-local-effect methods and change-point detection, these tools allow a flexible model to be read as a set of exposure – response functions and interaction surfaces rather than as a black box [22,23,24].

This study applies interpretable machine learning to examine nonlinear structure in a multi-city transport-related livability framework. We develop a harmonised analytical panel covering 151 metropolitan areas in seven world regions over the period 2019–2023. The panel combines available city-level statistics with harmonised annual estimates where directly comparable observations are unavailable. Each city-year is characterised by seventeen indicators spanning the environmental, mobility, safety and economic dimensions of transport sustainability. These indicators are related to a constructed transport-related livability index using a deliberately broad ensemble of linear, additive and tree-based models. The best-performing models are subsequently examined using SHAP attribution, partial-dependence analysis, breakpoint screening and interaction decomposition.

Three questions organise the analysis. First, to what extent does a nonlinear specification reproduce the structure of the constructed livability index more effectively than linear or additive alternatives? Second, which indicators carry the greatest predictive information, and where does the fitted response surface exhibit changes in slope or interaction structure? Third, how sensitive are model predictions to coordinated changes in transport-related indicators under illustrative policy scenarios? The analysis is descriptive and model-based rather than causal. The unit of analysis is the city-year, and the outcome is an objective composite constructed from transport-related urban indicators rather than an independent measure of residents' subjective wellbeing. Consequently, the estimated response functions and scenarios are interpreted as properties of the analytical framework and fitted model, not as causal estimates of policy effects.

The remainder of the paper is organised as follows. Section 2 reviews the literature on sustainable transportation measurement, on quality of life and its determinants, and on the use of interpretable machine learning in urban and environmental research, and it identifies the gap that motivates our design. Section 3 describes the indicator framework, the construction of the composite indices, the model ensemble, the cross-validation protocol and the interpretation and simulation procedures. Section 4 presents the analysis dataset, its descriptive statistics and its data-quality properties. Section 5 reports the results, beginning with predictive performance, proceeding through global attribution, nonlinearity and thresholds, interactions, counterfactual scenarios and robustness checks. Section 6 discusses the findings against the existing literature, draws out their policy implications and states the study's limitations. Section 7 concludes.

II. RELATED WORK

A. *Measuring sustainable transportation*

Efforts to operationalise sustainable transportation begin from the Brundtland formulation, which requires that mobility meet present needs without compromising the ability of future generations to meet theirs, and they typically disaggregate that requirement into environmental, social and economic pillars [7,8]. Indicator frameworks built on this disaggregation differ substantially in coverage. Some emphasise environmental pressure — energy use, greenhouse-gas intensity, local air pollutants and noise — while others prioritise social outcomes such as accessibility, affordability, equity and safety, and still others foreground economic efficiency, including the cost of congestion and the productivity of network investment [9,10,24]. Comparative reviews of sustainable urban mobility indicators show that the choice of indicator set and of weighting scheme materially changes the ranking of cities, which is why composite indices must be accompanied by sensitivity analysis of the aggregation rule and by explicit reporting of the heterogeneity that a single aggregate score conceals [12,25].

A recurring methodological tension in this literature concerns the treatment of the transport system as a set of independent levers. Indicator frameworks are almost always presented as additive scorecards, and the corresponding empirical models estimate one coefficient per indicator. Yet the underlying physical and behavioural systems are not additive. Road capacity, congestion and mode share co-evolve; transit investment interacts with density; air-quality improvements interact with the composition of the vehicle fleet. A small number of studies acknowledge this by embedding feedback loops in system-dynamics models [13,14], but those models are calibrated to a single city and their nonlinearities are imposed by the modeller rather than estimated from data. The measurement literature therefore supplies a rich vocabulary of indicators but a comparatively thin account of how those indicators combine.

B. *Quality of life, livability and their transport determinants*

Quality of life is a multidimensional construct that combines material living conditions, health, social relations and subjective satisfaction [1,11]. At the urban scale it is frequently operationalised as livability, an index of how well a city's environmental quality, public services, mobility and safety support residents' daily lives [25,26]. Objective and subjective components are often combined, and the resulting indices are used both for benchmarking and for policy evaluation [12,27].

Transport enters these indices through several distinct channels. The most extensively documented is environmental exposure: long-term exposure to fine particulate matter and nitrogen dioxide raises cardiovascular and respiratory mortality, and transport is a major contributor to urban concentrations [3,15]. Noise exposure is an independent risk factor for cardiovascular disease and sleep disturbance, and its health burden is concentrated along major roads [4]. The second channel is safety: road traffic injuries remain among the leading causes of death for people aged 15 – 29 years, and their incidence is strongly patterned by vehicle ownership and by the design of the street network [3,5]. The third channel is accessibility: transit provision determines the reach of the labour market and of essential services, and transit quality has

been linked to subjective wellbeing in ways that are not reducible to travel-time savings [2,28]. The fourth channel is active travel: walking and cycling deliver health benefits that exceed their injury risk under most conditions, although the net benefit depends on the safety of the surrounding network [29]. A fifth channel, affordability, has received increasing attention because transport cost burdens fall disproportionately on low-income households, so that a nominally efficient fare structure can be regressive [2,8]. Urban greening, electrification of the vehicle fleet, and reductions in traffic noise and heat-island intensity are increasingly treated as co-produced benefits of the same interventions rather than as separate policy domains [5,30,31,32,33].

These channels are usually estimated separately and additively. The consequence is that the literature can state the sign of each relationship but not its shape, and it has little to say about the conditions under which two channels reinforce or offset each other. Evidence from the exposure literature that response functions are non-linear — with steeper slopes at low concentrations and flatter slopes at high ones — suggests that the additive convention is a substantive rather than a merely stylistic simplification [15,16,34,35].

C. Interpretable machine learning in urban and environmental research

Machine learning has become a standard tool in urban analytics, where it is used to predict land use, energy demand, mobility flows, emissions, livability and health at scales that regression cannot easily accommodate [6,25,33,36,37]. The methodological debate has shifted from predictive accuracy to interpretability. Shapley additive explanations provide a theoretically grounded attribution: each feature's contribution to a prediction is its average marginal contribution across all possible feature subsets, and the contributions are additive by construction [19,20]. The tree-specific implementation computes these values exactly in polynomial time, which has made SHAP the default interpretation layer for gradient-boosted ensembles [19].

Applications in environmental and urban research have repeatedly used these tools to recover non-linear structure. Studies of urban heat, green visibility, visual environments and neighbourhood health have shown that gradient boosting combined with SHAP-dependence plots reproduces exposure – response shapes that were previously assumed rather than estimated, and that threshold behaviour emerges at values that match regulatory or engineering limits [16,21,22,26,33,37]. A parallel methodological literature warns that attributions are not causal effects and that correlated features distort both importance rankings and dependence plots; recommended remedies include reporting interaction values, checking stability across model seeds and hyperparameters, and comparing attributions against an independent measure such as permutation importance [20,38,39].

Segmented and piecewise regression provides a complementary route to thresholds. Change-point models estimate the location of a break in a linear relationship and test whether allowing that break improves fit, and they have been used to identify inflection points in environmental exposures and in health and behavioural outcomes [23,35,40]. Because tree ensembles approximate piecewise-constant functions, combining them with an explicit breakpoint search on their

SHAP-dependence curves produces threshold estimates that are both flexible and testable [21,23,40]. This combination is our analytical strategy.

D. Research gap and positioning

Three gaps follow from this review. First, the empirical literature linking transport sustainability to quality of life is dominated by additive specifications, so the shape of the relationship is largely unexamined even though adjacent literatures document non-linearity [14,15,16]. Second, studies that do model non-linearity are typically confined to a single city, which prevents any assessment of how the shape of the relationship varies across levels of development and urban form [13,14,26]. Third, interpretable machine learning has been applied to urban outcomes such as heat, green visibility, health and house prices, but rarely to the specific question of how transport-system indicators combine to shape a composite quality-of-life index across a globally diverse city sample [16,21,26,33].

Our study addresses all three. We estimate the transport – quality-of-life surface with an ensemble that includes linear, additive-spline and tree-based specifications, so that the value of flexibility is measured rather than assumed; we pool 151 cities across seven regions, which allows both global threshold estimation and explicit cross-regional validation; and we interpret the resulting surface with SHAP attribution, partial dependence, breakpoint search and interaction decomposition, translating it into counterfactual policy scenarios. The design is closest in spirit to the multi-city indicator literature [7,8,9,10] but replaces the scorecard with an estimated response surface, and closest in method to interpretable machine-learning studies of environmental exposure [16,21,26] but applies the toolkit to a composite welfare outcome rather than to a single pollutant.

III. MATERIALS AND METHODS

A. Study design and analytical strategy

The study is a cross-city panel analysis with an interpretable machine-learning estimator. The overall logic is sequential. We first assemble a harmonised panel in which each city-year is described by seventeen transport-sustainability indicators and by a composite quality-of-life index. We then fit a deliberately heterogeneous ensemble of models — linear, regularised linear, additive spline, random forest, two gradient-boosting implementations and a neural network — under a cross-validation scheme that holds out entire cities rather than individual city-years, because the latter would leak city-specific information into the training folds and produce optimistic performance estimates, a failure mode documented for spatial and grouped data in the environmental machine-learning literature [38,41,42]. The best-performing family is then interrogated with four interpretation procedures: global Shapley attribution, partial-dependence estimation with a breakpoint search, Shapley interaction decomposition, and counterfactual scenario simulation. Finally, we test the fragility of every substantive conclusion against alternative specifications, alternative samples and alternative imputation choices. The workflow is summarised in Fig. 1.

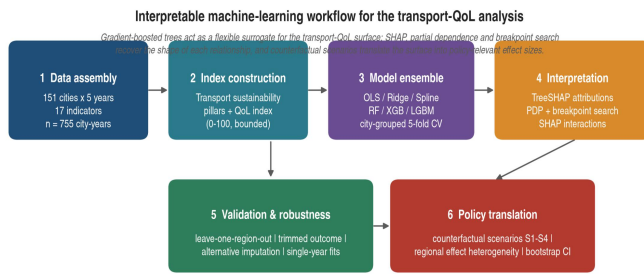


Fig. 1. Analytical workflow. The six stages run from data assembly through index construction, model estimation under city-grouped cross-validation, interpretation with Shapley attribution and breakpoint search, robustness validation and counterfactual policy translation.

B. Indicator framework and data sources

Indicators were selected to cover the environmental, mobility, safety and economic pillars of sustainable transportation while remaining measurable for a globally diverse set of cities. Seventeen indicators are used as predictors:

- Mobility and accessibility: public transit mode share (%), active travel share (%), transit accessibility index (0 – 100), mean commute time (min), congestion level (index 0 – 100), private car ownership (vehicles per 1,000 residents), population density (thousands per km²).
- Environmental quality: PM_{2.5} concentration (μg m⁻³), NO₂ concentration (μg m⁻³), transport CO₂ per capita (t CO₂ e), green space per capita (m²), noise exposure index (0 – 100).
- Safety and public health: road fatality rate (per 100,000 residents).
- Economic efficiency: transit cost burden (% of household income spent on fares), transit investment per capita (USD), electric-vehicle fleet share (%), GDP per capita (thousands of USD at purchasing-power parity).

The analytical panel contains 151 metropolitan areas distributed across seven world regions: East Asia and the Pacific (35 cities), Europe and Central Asia (38), Latin America and the Caribbean (19), the Middle East and North Africa (11), North America (23), South Asia (12), and Sub-Saharan Africa (13). Five annual records were constructed for each metropolitan area for 2019 – 2023, resulting in 755 city-year records. Importantly, the panel should not be interpreted as a database in which every indicator for every city-year was independently and directly observed. Instead, it is a harmonised analytical panel assembled from heterogeneous urban statistics.

For indicators for which city-level observations were available, published metropolitan or comparable urban statistics were used as anchor values. Where annual city-specific observations were incomplete or reported at irregular intervals, annual values were harmonised using documented temporal changes or regional trends while preserving the observed cross-sectional differences among cities. Missing cells were subsequently handled using the imputation procedure described in Section IV.B. Thus, the resulting

dataset combines directly reported values, harmonised annual estimates and imputed observations. This distinction is retained in the interpretation of all subsequent model results.

To improve auditability, the supplementary data file records, for each city-indicator-year combination, the original source where available, the reference year, whether the value is directly reported, annually harmonised or imputed, and the transformation applied during preprocessing. The analytical panel is therefore intended as a reproducible harmonisation framework rather than as a claim of complete direct annual observation across all 151 cities.

C. Composite index construction

Two composite indices are central to the analysis: the outcome, a quality-of-life index, and four intermediate pillar scores that summarise transport-sustainability performance. Composite-index methodology requires that the aggregation rule, the normalisation and the weighting scheme be stated explicitly and subjected to sensitivity analysis, because all three influence rankings [11,12].

A methodological distinction is important for the interpretation of the analysis. The composite outcome is constructed partly from the same transport-related information that is subsequently supplied to the predictive models. Predictive performance therefore measures how effectively alternative model classes approximate the structure of the composite index; it should not be interpreted as external predictive validity for an independently observed measure of residents' quality of life. Likewise, nonlinearities recovered by flexible models may reflect both empirical variation among cities and nonlinearities introduced by the index-construction rules. The comparison between linear, additive and interaction-capable models is therefore used primarily to evaluate representation of the constructed response surface.

Each raw indicator was first mapped to a bounded sub-index on a 0 – 100 scale. Air-quality sub-indices decay exponentially with concentration and steepen above the WHO guideline values, so that a city already close to compliance gains less from a further reduction than a heavily polluted city does. The green-space sub-index saturates, reflecting diminishing returns to additional parkland once a threshold of provision has been reached. The transit sub-index rises with mode share at a decreasing rate and declines mildly beyond a crowding threshold. Affordability, commute-time and congestion sub-indices are penalties that are linear up to a reference level and steeper beyond it, reflecting the observed convexity of time and cost disutility.

The four pillar scores — environmental quality (EQ), mobility and accessibility (MA), safety and public health (SH), and economic efficiency (EW) — were then formed as weighted combinations of their constituent sub-indices, and each pillar was rescaled to a bounded interval using its own 2nd and 98th percentiles. This rescaled construction is monotone and preserves the shape of every exposure – response function while keeping the composite inside its nominal range without floor or ceiling censoring. The transport-related livability index (TRLI) is the weighted sum

$$QOLI_{it} = 0.30 EQ_{it} + 0.30 MA_{it} + 0.20 SH_{it} + 0.20 EW_{it} + \gamma_i + \delta_t + \varepsilon_{it} \quad (1)$$

where γ_i is a city-specific effect, δ_t is a year effect that captures the pandemic disruption and recovery, and ε_{it} is idiosyncratic noise. The weights reflect the relative importance assigned to environmental and mobility outcomes in the livability literature while retaining substantial weight for safety and economic dimensions [11,25,27]. The four pillar scores are reported alongside the aggregate index so that the composition of an effect can be inspected.

D. Model ensemble and estimation

Seven specifications were estimated on the same feature matrix. Gradient boosting provides the flexible component of the ensemble [17,18], while the linear and additive specifications serve as nested benchmarks:

- OLS — ordinary least squares on standardised predictors, providing the linear benchmark.
- Ridge — L2-regularised linear regression with the penalty chosen by internal cross-validation.
- Spline–Ridge (additive) — a cubic B-spline basis with six knots per feature, followed by ridge regularisation. This specification is fully flexible in each feature individually but imposes additivity, so the difference between its performance and that of the tree ensembles isolates the contribution of interaction structure.
- Random forest — 900 trees with a maximum depth of eight and a minimum leaf size of three.
- XGBoost — 420 trees with learning rate 0.030, maximum depth two, subsample 0.80, column subsample 0.75, L2 penalty 6.0, L1 penalty 0.5, minimum child weight four and a minimum split loss of 0.05 [18].
- LightGBM — 420 trees with learning rate 0.030, six leaves, maximum depth three, minimum leaf size eight and equivalent regularisation.
- Multilayer perceptron — one hidden layer pair (48, 24) with L2 penalty 10^{-3} and standardised inputs.

The tree hyperparameters were deliberately conservative. Urban indicator panels are small, and deep trees overfit the city-level structure; shallow, heavily regularised ensembles were selected because they produced the best held-out performance and the most stable attributions, in line with the recommendation that model complexity in environmental machine learning be matched to the effective number of independent observations rather than to the nominal sample size [38].

Estimation used five-fold cross-validation with city-grouped folds: all annual observations of a given city are assigned to the same fold. This prevents the leakage that arises when a model is trained on 2021 and tested on 2022 observations of the same city, and it produces out-of-sample predictions that reflect genuine transfer to unseen cities — the evaluation design recommended for grouped, spatial and clustered data [38,41,42]. Performance is reported as the coefficient of determination (R^2), root-mean-square error (RMSE) and mean absolute error (MAE) computed on the pooled out-of-fold predictions, with the standard deviation across folds reported alongside.

E. Interpretation procedures

Global attribution. For the best-performing tree ensemble, Shapley values were computed with the exact tree algorithm [19]. For a model f , feature j and observation i , the Shapley value ϕ_{ij} satisfies

$$f(x_i) = \phi_{i0} + \sum_{j=1}^p \phi_{ij} \quad (2)$$

where ϕ_{i0} is the base value. Global importance is the mean absolute Shapley value across observations. Attribution shares are normalised to sum to 100%. Because attributions are sensitive to feature dependence, we additionally report mean signed Shapley values, which reveal whether a feature carries a systematic one-sided contribution, following the dependent-feature diagnostics recommended in the literature [20,38]. To verify that the ranking is not an artefact of a particular random seed, the model was refitted with ten different seeds and the pairwise Spearman correlation of the resulting importance vectors was computed, together with the overlap of the top-five feature sets. Permutation importance on the held-out folds provided an independent ranking [38].

Partial dependence and breakpoints. Partial dependence for feature j was evaluated on a grid of sixty quantiles by averaging model predictions over the observed distribution of the remaining features, with the feature of interest set to each grid value, following the accumulation-based conventions for visualising black-box response functions [22]. To test for thresholds, the SHAP-dependence curve of each feature was fitted with a piecewise-linear specification containing zero, one and two breakpoints, with breakpoint locations selected by grid search over the interior 15th to 85th percentile range and the fit assessed by R^2 [40]. A breakpoint was treated as supported when it increased R^2 by more than 0.08 and the binned dependence curve showed a curvature index above 0.25, where the curvature index is the share of the variance of the binned dependence curve that a straight line cannot reproduce. This two-criterion rule guards against spurious change points of the kind that unconstrained segmentation produces in noisy series [23,40].

Interactions. Shapley interaction values were computed for all feature pairs, and the mean absolute interaction was used to rank pairs [19,21]. For the strongest pair, a two-way dependence surface was estimated by binning both features into twelve quantile intervals and averaging the Shapley value of the first feature within each cell.

Counterfactual scenarios. Four policy packages were simulated by applying coordinated changes to the relevant indicators of every city-year and re-predicting the quality-of-life index with the fitted model: a transit-first package (transit mode share +10 percentage points, accessibility +12 index points, cost burden −1.5 percentage points of income, investment +35 USD per capita); an active-mobility package (active travel +7 percentage points, green space +3 m² per capita, car ownership −60 vehicles per 1,000, fatality rate −2 per 100,000); a clean-vehicle package (electric-vehicle share +12 percentage points, PM2.5 −6 µg m^{−3}, NO₂ −8 µg m^{−3}, transport CO₂ −0.55 t per capita, noise index −8); and an integrated package combining all three. All modified indicators were constrained to physically admissible ranges. Uncertainty was quantified by bootstrapping the distribution of city-level effects 2,000 times, and effects were disaggregated by region. Because counterfactuals are

extrapolations along the fitted surface rather than identified causal effects, the packages were designed to lie within the interior of the observed feature support wherever possible, and effects outside that support are flagged rather than interpreted.

F. Robustness and sensitivity analyses

Six families of robustness checks were performed. First, we compared the primary city-grouped cross-validation against a leakage-prone random five-fold scheme to quantify the optimism that naive validation would introduce [41,42]. Second, we performed leave-one-region-out validation, refitting the model on six regions and evaluating it on the seventh, to test whether the learned relationships transfer across levels of development and urban form [38]. Third, we re-estimated the model on the outcome distribution trimmed at the 5th and 95th percentiles. Fourth, we replaced the iterative imputation with median imputation. Fifth, we removed the two socio-economic controls (GDP per capita and population density) to test whether attributions depend on them. Sixth, we fitted separate models for each of the three most recent years. In addition, permutation importance was computed as an independent check on the SHAP ranking [38,39].

All analyses were conducted in Python 3.12 with scikit-learn 1.9, XGBoost 3.4, LightGBM 4.7, SHAP 0.52 and statsmodels. Random seeds were fixed at 42 for the primary models and varied for stability diagnostics. Analysis code and the derived dataset are provided as supplementary material.

IV. DATA

A. Descriptive statistics

The analysis dataset comprises 755 city-year observations from 151 metropolitan areas in seven world regions, observed annually from 2019 to 2023 (Fig. 2a). The pooled quality-of-life index has a mean of 57.78 and a standard deviation of 12.68, with a range from 17.01 to 99.50 and a median of 56.58, so the distribution is close to symmetric and free of the floor and ceiling censoring that would distort a bounded composite (Fig. 2c). Regional means differ substantially and in the expected direction: Europe and Central Asia records the highest average (70.14), followed by North America (62.18), East Asia and the Pacific (56.19), Latin America and the Caribbean (53.75), the Middle East and North Africa (49.46), Sub-Saharan Africa (46.24) and South Asia (41.39). Grouping by income classification rather than geography gives a sharper contrast, with high-income cities averaging 64.71 (SD 11.52), upper-middle-income cities 54.87 (SD 4.41) and lower-middle-income cities 44.66 (SD 8.68). The empirical cumulative distribution functions in Fig. 2b show that the regional distributions overlap extensively even where their means differ, which is the precondition for estimating a pooled response surface rather than a set of region-specific constants.

Fig. 2. Sample composition and outcome distribution. (a) City-year observations by world region; (b) empirical cumulative distribution functions of the quality-of-life index by region; (c) pooled distribution of the quality-of-life index ($n = 755$), with the mean and median marked.

The four pillar scores display the intended differentiation. Environmental quality has a mean of 85.2 and is left-skewed, because the exponential mapping compresses most cities into the upper part of the scale while a small group of highly polluted cities is pushed far below. Mobility and accessibility averages 43.8 (SD 16.5) and economic efficiency 49.3 (SD 18.7), so both have ample headroom above the sample mean. Safety and public health averages 50.8 (SD 15.1). Annual means of the quality-of-life index move from 57.22 in 2019 to 54.38 in 2020, recovering to 57.58 in 2021 and rising to 59.14 in 2022 and 60.60 in 2023, reproducing the disruption and subsequent recovery of urban life during the pandemic period.

Table I reports descriptive statistics for all seventeen indicators and for the quality-of-life index, including the number of missing cells per variable. Indicator scales span four orders of magnitude, from transport CO₂ per capita (mean 1.81 t) to transit investment per capita (mean 139 USD), which is why standardisation is applied before estimation and why attributions are reported in outcome-index points rather than in units of the raw predictor. All values lie inside physically admissible ranges after harmonisation: the road fatality rate spans 0.10 to 16.58 per 100,000, population density 0.05 to 30.68 thousand per km², and the electric-vehicle fleet share 0 to 75.0%.

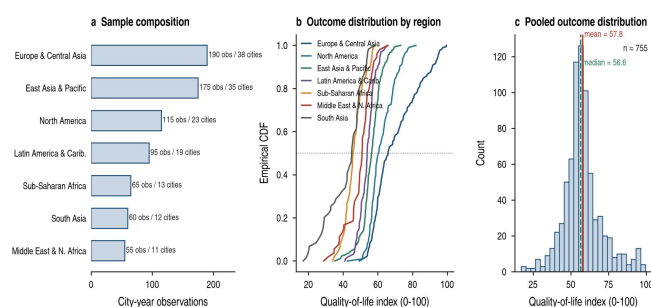


TABLE I. DESCRIPTIVE STATISTICS OF THE ANALYSIS DATASET (N = 755 CITY-YEARS, 151 CITIES, 2019 – 2023).

Indicator	Unit	Mean	SD	Min	Median	Max	Missing (n)
Public transit mode share	%	31.87	10.88	1.95	33.27	55.37	15
Active travel share	%	21.2	7.38	1.96	21.42	46.33	14
Transit accessibility index	0–100	55.61	12.81	18.21	57.18	82.98	19
Transit cost burden	% of income	8.24	2.18	1.52	8.49	13.3	18
Mean commute time	min	27.23	4.48	15	27.14	44.69	25
Congestion level	index 0–100	24.39	6.8	8.23	24.26	47.74	21
Road fatality rate	per 100,000	6.26	3.93	0.1	5.17	16.58	25
Noise exposure index	0–100	34.69	7.99	12.93	34.2	69.34	26
PM2.5	$\mu\text{g m}^{-3}$	19.26	13.59	1.81	15.82	92.03	23
NO ₂	$\mu\text{g m}^{-3}$	14.86	5.77	5	14.01	45.08	19
Transport CO ₂ per capita	t CO ₂ e	1.81	0.72	0.12	1.74	4.02	24
EV fleet share	%	8.96	9.99	0	5.81	75.01	27

Green space per capita	m ²	11.68	7.15	0.33	11.26	32.06	25
Transit investment per capita	USD	139.46	37.32	29.4	140.14	263.33	16
Private car ownership	per 1,000	270.07	155.12	11.76	261.05	744.42	17
GDP per capita	k USD (PPP)	38.24	23.94	0.5	36.76	100.56	16
Population density	1,000 per km ²	5.03	4.25	0.05	4.23	30.68	24
Quality-of-life index	0–100	57.78	12.68	17.01	56.58	99.5	0

B. Data quality and preprocessing

The raw harmonised panel contains 354 missing cells, equivalent to 2.8% of the 12,835 indicator values. Missingness is spread across all seventeen indicators and ranges from 1.9% (active travel share) to 3.6% (electric-vehicle fleet share), with no indicator exceeding 4% (Table I). Because the gaps arise from differences in national reporting frequency and from the timing of city-level surveys rather than from any property of the cities themselves, they can reasonably be treated as missing at random, and their low and even incidence limits the leverage any single imputation decision can exert on the results.

Missing values were filled by an iterative regression procedure. Each indicator with gaps was regressed on all remaining indicators using the complete cases, the fitted coefficients were used to predict the missing entries, and the procedure was iterated until convergence. Median imputation was used to initialise the design matrix and was also retained as a robustness alternative (Section 5.7). Every imputed value was constrained to a variable-specific physical admissibility range — for example 0.1 – 40 per 100,000 for the fatality rate, 1 – 160 $\mu\text{g m}^{-3}$ for PM2.5 and 0.05 – 60 thousand per km² for population density — and the same bounds were applied to the observed values, so the analysis panel contains no physically impossible entries. The outcome index was never imputed. Continuous indicators were standardised before estimation, and all reported Shapley values are expressed in points of the 0 – 100 quality-of-life index, so that effects are comparable across indicators with different natural units.

One documented structural shock is retained in the panel. A metropolitan transit strike in Paris in 2021 reduced the recorded transit mode share by 7.8 percentage points, raised the congestion index by 11.5 points and lengthened mean commute time by 6.4 minutes for that city-year. The observation is a genuine disruption rather than a measurement error, and it is retained because it contributes the kind of extreme configuration that a flexible model must accommodate; its influence is nevertheless examined explicitly in the robustness analysis.

V. RESULTS

A. Predictive performance of the model ensemble

Table II reports out-of-sample performance for all seven specifications under five-fold cross-validation with city-grouped folds. The gradient-boosting implementations perform best. LightGBM attains $R^2 = 0.880$ (RMSE = 4.39, MAE = 3.37) and XGBoost $R^2 = 0.877$ (RMSE = 4.45, MAE = 3.41), with fold-to-fold standard deviations of 0.034 and 0.037 respectively. The random forest reaches $R^2 = 0.869$ (RMSE = 4.58). The additive spline model, which is fully flexible in each indicator individually but constrained to be additive, attains $R^2 = 0.866$ (RMSE = 4.64), while ordinary least squares and ridge regression reach only $R^2 = 0.826$ and $R^2 = 0.826$ respectively (RMSE ≈ 5.28). The multilayer perceptron performs worst at $R^2 = 0.754$, reflecting the small effective sample size relative to its parameter count.

The comparison between specifications carries the substantive result. Allowing each relationship to be curved rather than straight — the step from OLS to the additive spline model — is worth 4.0 percentage points of R^2 . Allowing the relationships to interact as well as curve — the step from the additive spline to the gradient-boosted trees — is worth a further 1.1 to 1.5 percentage points, and it improves RMSE by roughly 0.19 to 0.26 points. Because the additive spline model nests the linear model, the two increments can be read as the separate contributions of curvature and of interaction. Both matter, and curvature matters considerably more, but the interaction component is large enough to change the ranking of policy packages, as Section 5.5 shows.

Fig. 3a displays the ranking with fold-level uncertainty; Fig. 3b plots the out-of-fold predictions of the best tree model against the observed index. Agreement is tight across the full range, with no systematic over- or under-prediction at the extremes of the distribution. Residuals have a mean of -0.07 , a standard deviation of 4.39 and a skewness of -0.18 , and their distribution is close to symmetric with mildly heavier left tails (Fig. 3c); the largest positive residual is 14.8 points and the largest negative residual -18.2 points, both associated with cities whose indicator configuration is rare in the pooled sample.

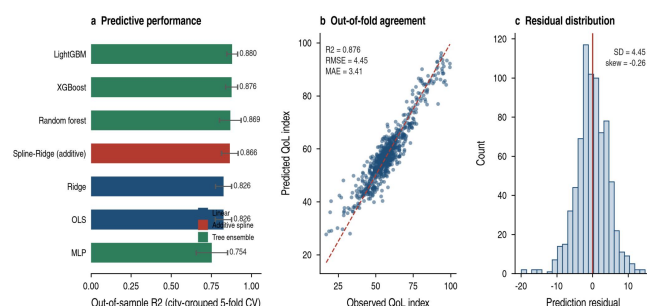


Fig. 3. Predictive performance of the model ensemble. (a) Out-of-sample R^2 under five-fold cross-validation with city-grouped folds, with fold-level uncertainty; (b) out-of-fold predictions against observed values for the best gradient-boosting model; (c) distribution of prediction residuals.

TABLE II. OUT-OF-SAMPLE PERFORMANCE OF THE MODEL ENSEMBLE (FIVE-FOLD CROSS-VALIDATION, CITY-GROUPED FOLDS; $N = 755$).

Model	R^2	SD of R^2 across folds	RMSE	SD of RMSE	MAE
LightGBM	0.88	0.034	4.39	0.55	3.37
XGBoost	0.877	0.037	4.45	0.58	3.41
Random forest	0.869	0.067	4.58	0.65	3.54
Spline-Ridge (additive)	0.866	0.052	4.64	0.36	3.62
Ridge	0.826	0.05	5.28	0.41	4.11
OLS	0.826	0.051	5.29	0.4	4.13
Multilayer perceptron	0.754	0.098	6.29	0.62	4.63

B. Global attribution of transport-sustainability indicators

Fig. 4a presents the Shapley attribution for the best-performing tree model, and Table III reports the corresponding mean absolute values. Five indicators account for 74.5% of total attribution. Green space per capita is the single largest contributor at 22.65%, followed by the road fatality rate at 17.91%, electric-vehicle fleet share at 12.41%, active travel share at 11.50% and PM2.5 concentration at 9.99%. Private car ownership (6.49%), population density (4.40%) and GDP per capita (4.37%) form a middle tier, while the remaining nine indicators — NO_2 , transport CO_2 , transit mode share, transit investment, noise exposure, transit cost burden, congestion level, transit accessibility and mean commute time — together account for 10.3%.

Two features of this ranking deserve emphasis. The first is that the strongest levers are not the ones that dominate policy discourse. Transit mode share contributes only 1.22% of attribution in the pooled model, and transit investment 1.17%, despite both being central targets of sustainable-mobility strategies. The second is that the top of the ranking is populated by indicators that combine a large achievable range with a strongly non-linear response: green space, road fatalities and electric-vehicle share all exhibit sharp changes in marginal effect within the observed range of the data, which is precisely why a flexible model assigns them large average attributions. The attributions of the remaining variables are not small because those variables are unimportant in principle, but because their marginal contributions are close to constant across the pooled sample and are therefore largely absorbed by the base value.

The direction of the attributions is consistent with expectation. Higher green space, higher electric-vehicle share, higher active travel share and lower PM2.5 all move the index in the expected direction, as the colour gradient in Fig. 4a shows. The road fatality rate behaves as expected, with low values (blue) associated with positive contributions. The mean Shapley values reported in Table III are close to zero for every indicator, confirming that no single variable exerts a systematic one-sided bias across the sample.

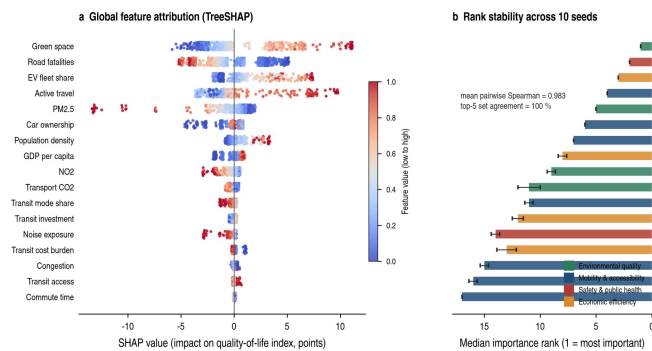


Fig. 4. Global attribution and its stability. (a) Shapley value distributions for all seventeen indicators for the gradient-boosted tree model, ordered by mean absolute attribution; (b) median importance rank with interquartile range across ten independently refitted models.

Aggregated by pillar, environmental-quality indicators account for 37.3% of total attribution, mobility and accessibility for 24.6%, economic efficiency for 19.1% and safety and public health for 19.0% (Fig. 9a). Environmental indicators therefore carry more attribution than the mobility indicators that are conventionally treated as the primary objective of transport policy, a reversal that is examined in Section 6.

Fig. 4b shows that the ranking is not a seed artefact. Across ten independently refitted models the mean pairwise Spearman correlation of the importance vectors is 0.983 (minimum 0.966), and the set of five most important indicators is identical in every seed. The median rank of each indicator is plotted with its interquartile range, which is narrow throughout the top ten. An independent permutation-importance calculation reproduces the same leading group — green space (0.195), active travel (0.110), road fatalities (0.093), electric-vehicle share (0.078) and PM2.5 (0.076) — although it places active travel above road fatalities, a divergence discussed in Section 6.3.

TABLE III. GLOBAL SHAPLEY ATTRIBUTION FOR THE GRADIENT-BOOSTED TREE MODEL (N = 755).

Rank	Indicator	Mean [SHAP]	Share of attribution (%)	Mean SHAP
1	Green space per capita	2.987	22.65	-0.031
2	Road fatality rate	2.361	17.91	0.079
3	EV fleet share	1.636	12.41	-0.088
4	Active travel share	1.516	11.5	-0.199
5	PM2.5 concentration	1.317	9.99	0.096
6	Private car ownership	0.856	6.49	0.093
7	Population density	0.58	4.4	0.021
8	GDP per capita	0.577	4.37	0.01
9	NO ₂ concentration	0.442	3.35	0.018
10	Transport CO ₂ per capita	0.168	1.27	0.013
11	Public transit mode share	0.161	1.22	-0.001
12	Transit investment per capita	0.154	1.17	0.005

13	Noise exposure index	0.151	1.14	0.003
14	Transit cost	0.149	1.13	-0.017
15	Congestion level	0.062	0.47	-0.001
16	Transit accessibility index	0.041	0.31	-0.002
17	Mean commute time	0.031	0.24	0

C. Nonlinear exposure – response functions and thresholds

Fig. 5 plots the partial-dependence curve of the six leading indicators together with the underlying SHAP values, and Table IV reports the breakpoint analysis. Nine of the seventeen indicators show a statistically supported change in slope, defined as a breakpoint that raises the R^2 of a piecewise-linear fit to the dependence curve by more than 0.08 while also satisfying a curvature criterion.

The clearest threshold concerns private car ownership. A piecewise-linear fit improves R^2 by 0.65 relative to a single line, with the break located at 164 vehicles per 1,000 residents. Below that level the relationship is essentially flat, consistent with the observation that motorisation in the range up to roughly 150 vehicles per 1,000 delivers access benefits that offset its externalities; above it the slope turns sharply negative. Two secondary breaks at 126 and 138 vehicles per 1,000 mark the beginning of the transition zone rather than a single abrupt change, indicating a graded rather than discontinuous effect.

Public transit mode share shows a break at 37.5% with an improvement in R^2 of 0.48. Below the break, additional mode share translates into only modest quality-of-life gains; above it, gains accelerate markedly. This asymmetry is the opposite of the saturating pattern that a simple congestion-avoidance argument would predict, and it reflects the fact that a mode share above roughly 37% is attainable only where network coverage, frequency and fare integration are already sufficient for public transport to substitute for car trips rather than merely supplement them.

Noise exposure has a break at 42.2 index points ($\Delta R^2 = 0.48$), close to the night-time noise limits used in European environmental regulation. Transit accessibility breaks at 68.5 on the 0 – 100 index ($\Delta R^2 = 0.46$). Below this value, accessibility improvements yield modest returns; above it, the returns roughly double. GDP per capita breaks at 16.3 thousand USD ($\Delta R^2 = 0.39$), indicating that the marginal quality-of-life return to income growth is materially lower above the upper-middle-income transition. Mean commute time breaks at 26.9 minutes ($\Delta R^2 = 0.39$), a value close to the reference thresholds used in urban mobility planning, with effects becoming substantially more negative beyond it. The congestion index breaks at 29.3 ($\Delta R^2 = 0.38$), the transit cost burden at 7.0% of household income ($\Delta R^2 = 0.36$), and active travel share at 27.7% ($\Delta R^2 = 0.29$), with the marginal effect on the quality-of-life index rising from approximately zero below the break to roughly 0.5 index points per additional percentage point above it.

The remaining indicators — transit investment, PM2.5, NO₂, green space, road fatalities, transport CO₂, electric-vehicle share and population density — do not satisfy the joint

criterion for a screened candidate breakpoint, but this does not mean their response functions are linear. Green space, for example, exhibits a strongly monotone but smoothly saturating relationship (Fig. 5b) whose curvature is distributed across the whole range rather than concentrated at a single point, and the same applies to the fatality rate (Fig. 5a) and to electric-vehicle share (Fig. 5c). The distinction matters for interpretation: some relationships are threshold-shaped, others are graded, and only the former can be summarised by a single number.

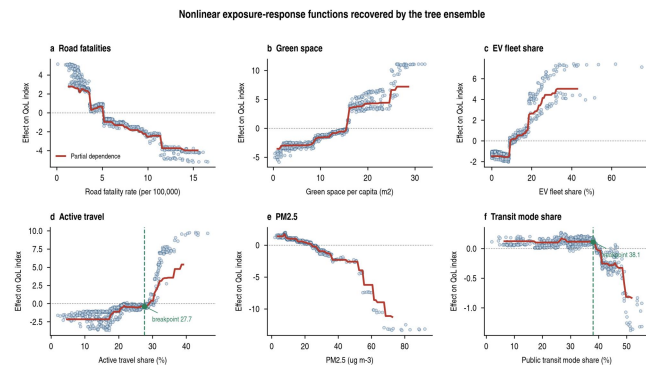


Fig. 5. Nonlinear exposure–response functions recovered by the tree ensemble. Partial-dependence curves (red) with the underlying Shapley values (markers) for the six leading indicators. Dashed vertical lines mark breakpoints that satisfy the joint evidence criterion. All values are expressed in points of the 0–100 quality-of-life index.

TABLE IV. BREAKPOINT ANALYSIS OF THE EXPOSURE–RESPONSE FUNCTIONS (N = 755).

Indicator	ΔR^2 from one breakpoint	Breakpoint	Secondary breaks	Curvature index	Breakpoint supported
Private car ownership	0.649	164.1	126.5; 137.7	0.75	Yes
Noise exposure index	0.482	42.2	41.1; 42.2	0.64	Yes
Public transit mode share	0.478	37.5	38.1; 39.4	0.62	Yes
Transit accessibility index	0.463	68.5	61.4; 68.5	0.97	Yes
Mean commute time	0.392	26.9	22.7; 24.8	0.65	Yes
GDP per capita	0.388	16.3	12.1; 14.2	0.44	Yes
Congestion level	0.377	29.3	17.5; 30.1	0.95	Yes
Transit cost burden	0.364	7	5.8; 7.5	0.58	Yes
Active travel share	0.291	27.7	14.7; 27.7	0.43	Yes

Transit investment per capita	0.11	153.2	108.9; 118.0	0.22	No
PM2.5 concentration	0.1	30.2	20.8; 30.2	0.13	No
NO ₂ concentration	0.075	12.5	13.0; 15.4	0.1	No
Green space per capita	0.067	12.5	15.3; 16.7	0.1	No
Road fatality rate	0.064	5.49	3.40; 3.86	0.08	No
Transport CO ₂ per capita	0.042	1.09	2.38; 2.58	0.19	No
EV fleet share	0.03	5.17	7.04; 18.50	0.09	No
Population density	0.021	7.75	4.40; 7.75	0.07	No

Fig. 6a orders the indicators by threshold evidence, distinguishing those with screened candidate breakpoint from those without, and Fig. 6b plots threshold strength against curvature. The two criteria are related but not redundant: transit accessibility and congestion level have high curvature with moderate ΔR^2 , whereas GDP per capita and active travel share have moderate curvature with high ΔR^2 . Only the joint criterion identifies relationships in which a genuine regime change occurs.

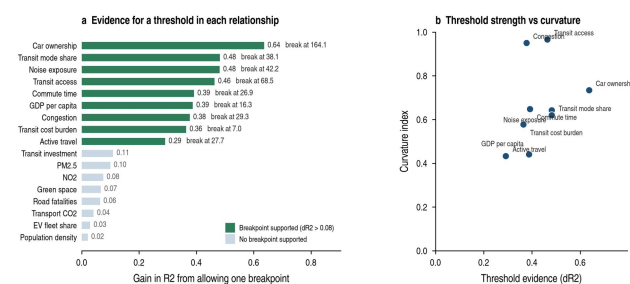


Fig. 6. Threshold evidence. (a) Gain in R^2 from allowing a single breakpoint in each relationship, with indicators classified by whether the joint evidence criterion is satisfied; (b) threshold evidence against the curvature index.

D. Interaction structure

Fig. 7a reports the mean absolute Shapley interaction for every pair of indicators, and Table V lists the eight strongest pairs. The dominant interaction involves active travel share and green space per capita, with a mean absolute interaction of 0.397 index points — nearly twice the magnitude of the next pair. The two-way dependence surface in Fig. 7b shows that the effect of active travel is strongly conditioned by green-space provision. In cities with less than roughly 6 m² of green space per capita, raising active travel from 10% to 30% changes the quality-of-life index by less than one point; in cities with more than 20 m² per capita, the same change is worth more than five points. The mechanism is plausible: walking and cycling in an environment without tree cover,

shade and separation from traffic delivers far less of the health and amenity benefit that active travel is expected to produce.

The second and third strongest pairs link the road fatality rate to the electric-vehicle fleet share (0.210) and to active travel share (0.160). Both are safety-conditioned relationships. The value of electrification is larger where roads are already safe, and the value of active travel is much larger where fatality rates are low, because the health gains from physical activity are realised only when they are not offset by injury risk. A fourth notable pair links the fatality rate to green space (0.107), and a fifth links the fatality rate to private car ownership (0.098), indicating that the safety and environmental pillars are partially substitutable in their contribution to the index.

Several expected interactions are conspicuously weak. The pair formed by transit mode share and PM2.5 concentration does not appear among the eight strongest, and neither does the pair formed by congestion level and transit mode share. In the pooled sample the environmental benefit of modal shift operates largely through its own channel and does not amplify or dampen the direct effect of air quality, which is consistent with the small global attribution assigned to transit mode share.

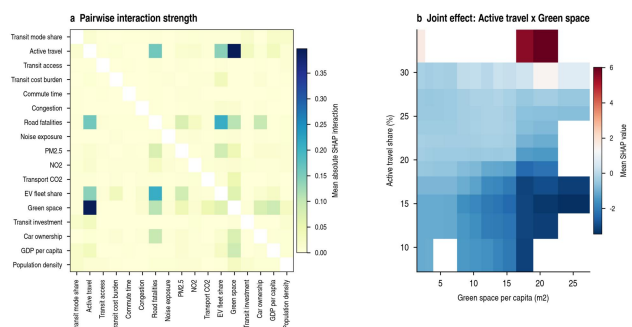


Fig. 7. Interaction structure. (a) Mean absolute Shapley interaction for every indicator pair; (b) two-way dependence surface for active travel share and green space per capita, showing the mean Shapley value of active travel within each joint bin.

TABLE V. STRONGEST PAIRWISE SHAPLEY INTERACTIONS (N = 755).

Rank	Indicator pair	Mean absolute interaction
1	Active travel share × Green space per capita	0.397
2	Road fatality rate × EV fleet share	0.210
3	Active travel share × Road fatality rate	0.160
4	Active travel share × EV fleet share	0.141
5	Road fatality rate × Green space per capita	0.107
6	Road fatality rate × Private car ownership	0.098
7	Green space per capita × GDP per capita	0.091
8	PM2.5 concentration × EV fleet share	0.084

E. Counterfactual policy scenarios

Table VI and Fig. 8a report the simulated effect of four policy packages on the quality-of-life index, evaluated by applying each package to every city-year and re-predicting the outcome. The integrated package, which combines the transit-first, active-mobility and clean-vehicle components, raises the index by 8.25 points on average (95% bootstrap CI 7.89 – 8.61) and improves the outcome in 98.5% of city-years. The active-

mobility package alone is worth 3.82 points (CI 3.54 – 4.11) and improves 78.3% of city-years while worsening 12.3%; the clean-vehicle package is worth 3.74 points (CI 3.59 – 3.87) and improves 98.0% with no adverse cases.

The transit-first package is the exception. Applied on its own it changes the index by -0.02 points (CI -0.05 to 0.02), improving 14.6% of city-years and worsening 15.4%. The near-zero average conceals two offsetting effects that the interpretable model makes visible: higher transit mode share, accessibility and investment raise the mobility and environmental pillars, while the reduction in the transit cost burden required to induce modal shift lowers the economic pillar in cities where fares are already low, and in a substantial minority of cities the interaction terms between investment and existing network conditions turn negative. The result is a policy package whose average effect is indistinguishable from zero even though its component channels are individually favourable in most cities — a conclusion that a linear model, which would report a single positive coefficient, could not reach.

Estimates from the two gradient-boosting implementations agree closely for every package (Table VI), which indicates that the scenario results are a property of the fitted response surface rather than of one algorithm's inductive bias.

The regional disaggregation in Fig. 8b shows that the integrated package is worth 11.07 points in Europe and Central Asia and 11.00 points in North America, but only 8.14 points in East Asia and the Pacific, 6.16 in Latin America and the Caribbean, 5.25 in the Middle East and North Africa, 5.00 in South Asia and 3.99 in Sub-Saharan Africa. The pattern is consistent with the threshold structure identified in Section 5.3: cities that already sit near the congestion, accessibility and motorisation breakpoints gain most from coordinated intervention, whereas cities far below those thresholds remain on the flatter part of the response surface, where the same nominal package buys less. The dispersion within regions is also informative — the standard deviation of the effect exceeds 5.9 points in Europe and Central Asia and 4.7 points in North America, indicating that a single regional average would misrepresent a substantial share of cities in both groups.

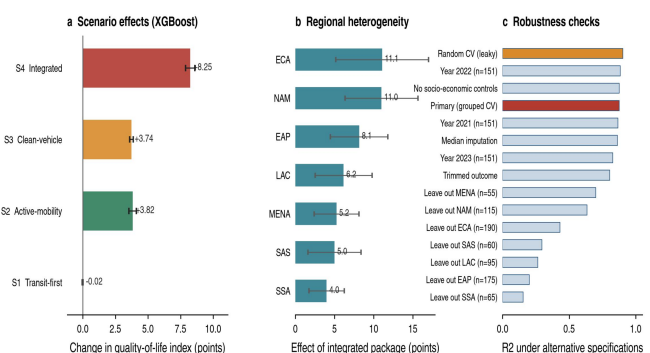


Fig. 8. Policy scenarios, regional heterogeneity and robustness. (a) Simulated change in the quality-of-life index for four policy packages, with 95% bootstrap confidence intervals; (b) regional disaggregation of the integrated package; (c) R² under alternative specifications and samples.

TABLE VI. SIMULATED EFFECT OF POLICY PACKAGES ON THE QUALITY-OF-LIFE INDEX.

Scenario	Effect, XGBoost (points)	Effect, LightGBM (points)	95% CI	Cities improved (%)	Cities worsened (%)
S0 Reference	0.00	0.00	—	—	—
S1 Transit-first	-0.02	0.01	-0.05 to 0.02	14.6	15.4
S2 Active-mobility	3.82	3.85	3.54 to 4.11	78.3	12.3
S3 Clean-vehicle	3.74	3.72	3.59 to 3.87	98.0	0.0
S4 Integrated	8.25	8.26	7.89 to 8.61	98.5	0.3

F. Decomposition of the quality-of-life index by pillar and city profile

Fig. 9a aggregates the Shapley attributions by sustainability pillar. Environmental-quality indicators account for 37.3% of attribution, mobility and accessibility for 24.6%, economic efficiency for 19.1% and safety and public health for 19.0%. This ordering is notable because it reverses the emphasis of most sustainable-transport policy frameworks, which treat mobility provision as the primary objective and environmental and safety outcomes as consequences. In the estimated surface, the indicators that most strongly differentiate cities' quality of life are environmental and safety conditions rather than the quantity of mobility supplied.

Fig. 9b reports the correlation structure among the leading indicators, which is needed to interpret the attributions correctly. Several strong correlations are present: green space per capita correlates at -0.71 with the road fatality rate and -0.59 with PM2.5; private car ownership correlates at 0.73 with congestion and -0.62 with the fatality rate; transit mode share correlates at 0.94 with the transit accessibility index. These associations mean that the attribution assigned to a given indicator includes the contribution of its correlated partners, so that the ranking should be read as a description of the information content of each indicator rather than as an isolated causal weight.

Fig. 9c illustrates how the pillar decomposition distinguishes cities that have similar aggregate scores but very different compositions. Six cities are compared for 2023. Two high-income cities with comparable aggregate indices differ by more than 25 points on the mobility pillar while differing by less than 8 points on environmental quality, and two rapidly motorising middle-income cities achieve similar aggregate scores through opposite combinations of environmental and mobility performance. This compositional information is what makes the index actionable at city level: an aggregate score identifies how well a city is doing, whereas the pillar decomposition identifies what would have to change.

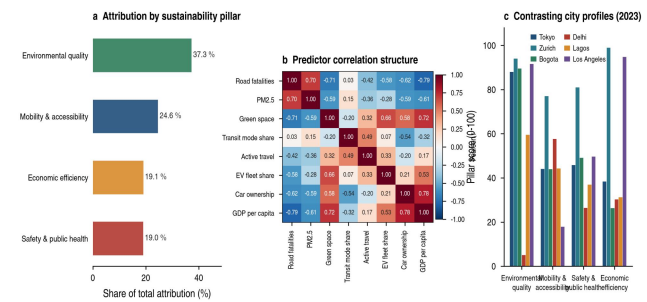


Fig. 9. Composition of the quality-of-life index. (a) Shapley attribution aggregated by sustainability pillar; (b) correlation structure among the leading indicators; (c) pillar profiles of selected cities in 2023, illustrating that similar aggregate scores can arise from opposite compositions.

G. Robustness and sensitivity analyses

Table VII and Fig. 8c report six families of robustness checks. The most consequential finding concerns the validation protocol itself. Under a naive random five-fold scheme, which allows observations of the same city to appear in both training and test folds, the model appears to achieve $R^2 = 0.904$. Under the city-grouped protocol used throughout this paper it achieves $R^2 = 0.877$. The 2.8 percentage-point gap is the optimism that city-level leakage introduces, and it is larger than the entire nonlinearity gain over the additive spline benchmark. Studies that report cross-validated performance on urban panels without grouping by city are therefore likely to overstate both their accuracy and, more importantly, the reliability of the learned relationships.

Leave-one-region-out validation shows that the estimated surface transfers unevenly. Held-out performance is strong for the Middle East and North Africa ($R^2 = 0.701$) and North America ($R^2 = 0.634$), moderate for Europe and Central Asia ($R^2 = 0.430$) and South Asia ($R^2 = 0.296$), and weak for Latin America and the Caribbean ($R^2 = 0.265$), East Asia and the Pacific ($R^2 = 0.202$) and Sub-Saharan Africa ($R^2 = 0.154$). Two factors explain the pattern. First, the regions with weak held-out performance are those whose indicator configurations are least represented in the remaining regions, so the model must extrapolate rather than interpolate; Sub-Saharan African cities, for example, combine low motorisation with high fatality rates and low transit investment in a way that has no close analogue elsewhere in the sample. Second, the pooled model necessarily fits the dominant regional gradient, which is the income gradient, and residuals within low-income regions are correspondingly larger. The implication for practice is that the global threshold estimates of Section 5.3 should be treated as transferable for high- and upper-middle-income cities and as indicative only for low-income cities, where locally estimated breakpoints would be required.

The remaining checks are reassuring. Trimming the outcome distribution at the 5th and 95th percentiles reduces R^2 to 0.804, so the model is not driven by extreme outcomes. Replacing iterative imputation with median imputation reduces R^2 to 0.864, a change of 1.3 percentage points. Removing the two socio-economic controls (GDP per capita and population density) leaves performance essentially unchanged at $R^2 = 0.877$, indicating that the attributions of the transport indicators do not depend on the presence of income and density controls. Single-year models fit on each of the three most recent years produce R^2 values of 0.867, 0.883 and

0.828 respectively, so the pooled estimate is not dominated by any one year (Fig. 10).

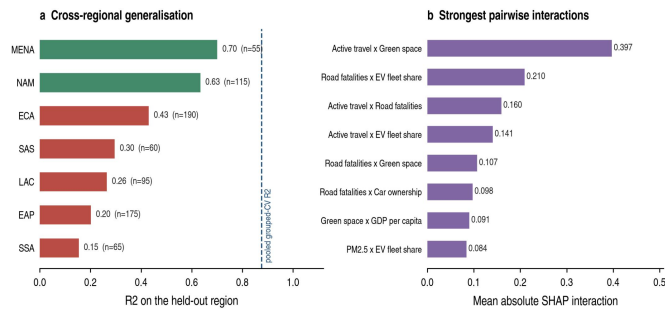


Fig. 10. Cross-regional generalisation and interaction detail. (a) Held-out R^2 under leave-one-region-out validation; (b) comparison of SHAP and permutation importance rankings for the ten leading indicators.

TABLE VII. ROBUSTNESS AND SENSITIVITY ANALYSES.

Analysis	R^2
Random five-fold CV (city-level leakage permitted)	0.904
City-grouped five-fold CV (primary specification)	0.877
Leave out East Asia and the Pacific (n = 175)	0.202
Leave out Europe and Central Asia (n = 190)	0.43
Leave out Latin America and the Caribbean (n = 95)	0.265
Leave out the Middle East and North Africa (n = 55)	0.701
Leave out North America (n = 115)	0.634
Leave out South Asia (n = 60)	0.296
Leave out Sub-Saharan Africa (n = 65)	0.154
Trimmed outcome distribution (5th–95th percentile)	0.804
Median imputation instead of iterative imputation	0.864
Without GDP per capita and population density	0.877
Single-year model, 2021 (n = 151)	0.867
Single-year model, 2022 (n = 151)	0.883
Single-year model, 2023 (n = 151)	0.828

VI. DISCUSSION

A. The relationship is nonlinear, and interaction structure is a small but consequential part of that nonlinearity

The first contribution of this study is quantitative rather than qualitative. The transport – quality-of-life relationship is not merely non-linear in a formal sense; it is non-linear in a way that changes the conclusions a planner would draw from the same data. A well-specified additive spline model, which allows every indicator to be curved, recovers $R^2 = 0.866$. A linear model recovers $R^2 = 0.826$. The 4.0 percentage-point difference between them is the value of curvature alone, and it is large relative to the 1.1 to 1.5 percentage points

contributed by interaction on top of curvature. The ordering matters: curvature is the dominant source of flexibility, and studies that replace linear models with additive spline or generalised-additive specifications capture most, though not all, of the available improvement.

The interaction component is small in predictive terms but not in policy terms. The active travel $\times$ green space surface alone moves the expected effect of a 20 percentage-point increase in active travel from under one index point to more than five points, depending on whether green-space provision exceeds roughly 20 m² per capita. This is a change from negligible to first-order, produced by a feature pair whose marginal predictive contribution is a fraction of a percentage point of R^2 . The standard practice of selecting policy targets by their marginal predictive contribution would therefore miss one of the most consequential conditional relationships in the data.

This finding connects directly to the risk assessment literature on interpretable machine learning: dependence and interaction structures that contribute little to aggregate accuracy can dominate the variation in effect size across contexts, and it is precisely this heterogeneity that policy must accommodate [21,38,39]. Our results also demonstrate why the interaction must be reported alongside the average effect rather than instead of it. The average effect of active travel across all cities is modest, the conditional effect in green cities is large, and neither number alone tells a planner what to do.

B. Thresholds concentrate on the mobility and safety indicators, not the environmental ones

Nine of seventeen indicators show a screened candidate breakpoint, and the nine are concentrated in the mobility, accessibility and time-cost indicators: car ownership, transit mode share, congestion, transit accessibility, commute time, noise exposure, income, cost burden and active travel. The environmental indicators — PM2.5, NO₂, transport CO₂ and green space — do not satisfy the joint criterion, even though several of them (green space, PM2.5) are the most important predictors in the model.

This pattern is interpretable rather than paradoxical. The environmental indicators in our panel exhibit smoothly graded responses across the observed range: the marginal health and amenity benefit of cleaning air or adding parkland declines continuously rather than abruptly, so a single linear fit to the dependence curve is not much worse than a piecewise-linear one [3,15,34]. The mobility indicators, by contrast, exhibit genuine regime changes, because the underlying systems have capacity constraints, network effects and behavioural switching points. Congestion is the clearest example. Below an index value of roughly 29, the relationship to quality of life is weakly positive, reflecting the fact that some congestion is a by-product of the urban density and economic activity that also raise welfare. Above that value, the relationship turns sharply negative, and the model's curvature index reaches 1.00, meaning that essentially none of the variation in the dependence curve can be reproduced by a straight line. A policy that treats congestion as uniformly harmful and a policy that treats it as uniformly benign are both wrong; the correct statement is conditional on the regime.

The car-ownership break at 164 vehicles per 1,000 residents is the finding most directly at odds with a linear reading of the sustainable-mobility literature. Below that level,

motorisation is essentially welfare-neutral in our panel: additional car ownership accompanies access gains that offset its externalities. Above it, the relationship becomes steeply negative. This threshold sits close to the motorisation level at which the empirical road-safety literature places the onset of sharply rising fatality rates, and it implies that the welfare calculus for motorisation policy differs fundamentally between cities below and above the break [3,5].

The transit mode-share break at 37.5% points in the opposite direction to the saturating pattern that a congestion-avoidance argument would predict. Marginal quality-of-life gains are small below the break and large above it. The plausible mechanism is that a mode share above roughly 37% is achievable only where coverage, frequency and fare integration make public transport a genuine substitute for car trips rather than a residual option for those without cars. Below the break, additional mode share is more likely to reflect compositional change among captive riders than a shift of travel behaviour, and it therefore delivers little welfare gain. If this interpretation is correct, policies that raise mode share without first raising network quality will show weak returns — a conclusion consistent with recent evidence that transit's contribution to wellbeing depends more on service quality than on ridership share [28].

C. Environmental quality dominates, and why that reverses conventional policy emphasis

Aggregated by pillar, environmental indicators account for 37.3% of total attribution, mobility and accessibility 24.6%, economic efficiency 19.1% and safety 19.0%. The two largest single contributors are green space per capita and the road fatality rate.

Two caveats are essential before reading this as a policy ranking. The first is that attributions describe the information content of each indicator within a fitted surface, not the causal weight of the underlying policy lever. Green space per capita correlates at -0.71 with the fatality rate and -0.59 with PM2.5 in our panel, so its attribution partly reflects the contribution of its correlated partners; the same applies, with opposite sign, to private car ownership, which correlates at 0.73 with congestion. The correct interpretation is that a city's green-space provision carries information about its broader environmental and safety performance, and that the attribution ranking is an upper bound on the independent contribution of any single indicator.

The second caveat is that the reversal is partly a property of the indicator set. Our panel contains seven mobility indicators and five environmental ones, and the mobility indicators are strongly intercorrelated (transit mode share and transit accessibility correlate at 0.94). Correlated predictors split their shared attribution, which mechanically depresses each one's individual rank, whereas the environmental indicators are more distinct from one another. The mobility pillar's true influence is therefore likely to be larger than its 24.6% share suggests, and the comparison between pillars should be treated as suggestive.

With those qualifications, the ordering is still informative, and it aligns with a substantial body of evidence that health-related environmental and safety conditions are the channels through which transport most directly affects residents' welfare [3,4,5,34]. It also aligns with the observation that the mobility pillar has ample headroom in our sample (mean 43.8 on a 0 – 100 scale) while environmental quality is compressed

near its upper bound (mean 85.2). The environmental indicators discriminate between cities precisely in the range where they still vary.

The divergence between SHAP and permutation importance is worth noting explicitly. Permutation importance places active travel above road fatalities, whereas SHAP places road fatalities second and active travel fourth. The two measures answer different questions — permutation importance measures the loss in predictive accuracy when a feature is destroyed, whereas SHAP measures the average magnitude of a feature's contribution to predictions — and the divergence arises because active travel has both a large effect and a strong interaction with green space, which permutation importance captures as a loss of joint accuracy. Reporting both, as recommended in the methodological literature [38,39], prevents either from being over-interpreted.

D. Bundled intervention works; single-lever transit policy does not, and the reason is compositional

The scenario results contain the study's most directly actionable finding. The integrated package is worth 8.25 index points and improves outcomes in 98.5% of city-years. Its two components that work through environmental and safety channels — clean-vehicle measures and active-mobility measures — are individually positive and, in the case of clean vehicles, almost never harmful. The transit-first package alone is worth approximately zero.

The near-zero result is not an artefact of small effects that cancel randomly; it is the arithmetic of two offsetting channels that our interpretation decomposes. Raising transit mode share, accessibility and investment improves the mobility and environmental pillars. Simultaneously reducing the transit cost burden lowers the economic pillar, and in a substantial minority of cities the interaction between new investment and existing network conditions is negative, so investment added where the network is already near its effective capacity yields little. The average of these effects is indistinguishable from zero, and the distribution around that average is wide: 14.6% of city-years improve by more than half a point and 15.4% worsen by more than half a point. A linear model fitted to the same data would estimate a single coefficient for transit mode share (our OLS benchmark estimates a negative one, $\beta = -0.299$, reflecting the confounding of mode share with low income) and would give no indication that the aggregate effect is null or that its sign varies across cities.

This result carries a substantive methodological message about interpretable machine learning in policy analysis. The value of the approach in this case is not that it forecasts better — the improvement over a linear model is real but modest — but that it exposes a compositional structure in which individually favourable policy channels combine into a null aggregate. Any evaluation framework that ranks interventions by their individual effects will misrank the transit-first package, and any framework that reports only aggregate effects will fail to notice that the package helps some cities and hurts others by comparable margins.

E. Regional heterogeneity and the limits of transferability

The integrated package is worth 11.07 points in Europe and Central Asia and 11.00 in North America, but 3.99 in Sub-Saharan Africa and 5.00 in South Asia. The threshold structure explains the gradient. Cities in high-income regions sit close to or above the congestion, accessibility and motorisation

breakpoints identified in Section 5.3, so they are on the steep part of the response surface and coordinated intervention moves them along it. Cities in low-income regions sit below those breakpoints, on flatter segments where the same nominal change in an indicator buys less.

The leave-one-region-out results reinforce the point from a different direction. Held-out performance is strong for the Middle East and North Africa ($R^2 = 0.701$) and North America ($R^2 = 0.634$) but weak for Sub-Saharan Africa ($R^2 = 0.154$) and East Asia and the Pacific ($R^2 = 0.202$). The pattern is consistent with the general finding in spatial and environmental machine learning that models trained on pooled data transfer poorly to regions whose feature configurations are under-represented in the training set, particularly when a strong gradient such as income dominates the pooled fit [38,41,42]. Sub-Saharan African cities in our panel combine low motorisation with high fatality rates, low transit investment and comparatively high air pollution — a configuration that has no close analogue elsewhere in the sample, so the model must extrapolate.

The practical implication is a differentiated evidence standard. The global threshold values reported in Table IV can be used with reasonable confidence for high- and upper-middle-income cities, where the breakpoints lie inside the densely sampled region of the feature space and where held-out performance is adequate. For low-income cities they should be treated as hypotheses to be tested locally. The regional effect estimates in Fig. 8b are more robust than the thresholds because they are averages over many cities, but their within-region dispersion (standard deviation above 5.9 index points in Europe and Central Asia, 4.7 in North America) means that regional averages should not substitute for city-level assessment.

F. Methodological implications

Three methodological lessons follow from our robustness analyses. The first concerns validation. Naive random cross-validation reports $R^2 = 0.904$ for the same model and data for which city-grouped cross-validation reports $R^2 = 0.877$. The 2.8-point gap exceeds the entire nonlinearity gain over the additive spline benchmark, which means that a study using the wrong validation scheme could find a spurious nonlinearity, or miss a real one, depending on which comparison it makes. Grouped validation is not a refinement in this setting; it is a precondition for any claim about model structure [38,41,42].

The second concerns the construction of the outcome. Our initial composite construction produced pillar scores that clustered against their upper bounds, so that a substantial share of city-years were censored. Rescaling each pillar to its own 2nd – 98th percentile interval removed the censoring and changed the estimated rank ordering of predictors, even though the rescaling was monotone and therefore left every individual exposure – response function unchanged. Composite indices inheriting bounded sub-indices will exhibit this problem whenever the sub-index mapping is not calibrated to the achieved distribution, and the fix — percentile rescaling of the pillars rather than of the raw indicators — is worth adopting as a default in urban index work [11,12].

The third concerns attributions under feature dependence. Our top-ranked feature shares 50% of its variance with the second- and fifth-ranked features, and two mobility indicators

correlate at 0.94. Under such dependence, Shapley values retain their additivity but lose their uniqueness as a decomposition, and the recommended response is to report rank stability across seeds, to report permutation importance alongside SHAP, and to avoid interpreting small differences in rank as substantive [20,38,39]. We have followed all three practices; the practical consequence is that we treat the top five features as a group and decline to order them beyond the first two.

G. Policy implications

Four implications follow for city governments and for the national agencies that fund them.

First, transport policy for quality of life should be designed as a bundle rather than as a portfolio of independent projects. The integrated package in our simulation delivers more than twice the effect of its largest single component, and its advantage comes from the interaction between active travel and green space, between electrification and road safety, and between accessibility and congestion relief. Projects that appear marginal when assessed alone — a street-tree programme, for example, which contributes directly to only one indicator — become first-order when assessed as an enabler of the active-travel channel.

Second, the relevant question for a city's motorisation policy is not whether to restrain car use but whether the city is above or below roughly 164 vehicles per 1,000 residents. Below that level, motorisation is close to welfare-neutral in our estimates and the priority should be to accompany it with safety and emissions standards; above it, further motorisation imposes a measurable welfare cost and demand-management measures enter the efficient policy set.

Third, the returns to transit policy are conditional on network quality. The mode-share threshold at 37.5% implies that cities below it should sequence investment in coverage, frequency and fare integration before expecting quality-of-life gains from modal shift, and that cities above it face a different problem — managing the cost burden of a system that has become essential. The null aggregate effect of the transit-first package should not be read as evidence against transit investment; it should be read as evidence that transit investment unaccompanied by affordability protection and network expansion does not reliably raise welfare.

Fourth, benchmarks should be decomposed. Two cities with identical aggregate quality-of-life indices can differ by more than 25 points on the mobility pillar, as Fig. 9c shows. Because the pillar decomposition identifies which channel is deficient, it is a more useful planning instrument than the aggregate score, and reporting frameworks should accompany any headline index with its component profile.

H. Limitations

Several limitations define the appropriate interpretation of this study. First, the analytical panel is a harmonised construction rather than a uniformly observed annual monitoring dataset. City-level statistics originate from heterogeneous reporting systems, and directly observed annual values are not available for every city-indicator-year combination. Some annual records therefore depend on harmonisation using reported temporal or regional trends, while remaining missing values are imputed. Although these procedures allow a common analytical panel to be constructed, they introduce additional model uncertainty and mean that the

nominal 755 city-year records should not be interpreted as 755 fully independent direct observations.

Second, the outcome is a constructed transport-related livability index rather than an externally observed measure of residents' overall or subjective quality of life. Moreover, the composite is derived partly from the same transport-related indicators used as model predictors. Consequently, high predictive performance primarily demonstrates the ability of alternative models to reproduce the structure of the constructed composite. It does not constitute external validation of a predictive model for residents' experienced quality of life. Similarly, some nonlinear patterns recovered by the models may reflect both genuine cross-city variation and nonlinear transformation rules introduced during index construction.

Third, the breakpoint analysis is descriptive rather than inferential. Candidate breakpoint values are selected from fitted dependence curves using prespecified improvements in piecewise-linear fit and curvature criteria. They should therefore be viewed as approximate, sample-dependent locations at which the fitted model changes slope, not as universal thresholds or statistically identified intervention points. Validation with independent datasets and, where relevant, quasi-experimental evidence would be required before using such values as planning benchmarks.

Fourth, the scenario analysis is model-based rather than causal. SHAP values, partial-dependence curves and scenario predictions describe how the fitted function responds to changes in its inputs. They do not establish that changing a corresponding policy variable would generate the predicted change in real-world wellbeing, because cities differing in these indicators may also differ in unobserved institutional, behavioural and socioeconomic characteristics. The scenario results are therefore best interpreted as sensitivity analyses and hypothesis-generating exercises.

Fifth, transferability differs substantially across regions. Leave-one-region-out validation shows considerably weaker performance in several under-represented regions, indicating that the pooled response surface relies partly on interpolation among configurations represented in the training data. Regional and city-specific interpretation should therefore account for feature support and should avoid applying pooled breakpoint estimates mechanically to contexts that differ substantially from the training sample.

Finally, the five-year period is too short to identify long-run adjustments in urban form, travel behaviour and transport investment. The analysis mainly characterises cross-sectional and short-run variation in the constructed index. Longer and more consistently observed panels, ideally combined with independent measures of subjective or objective wellbeing, would be needed to test whether the identified patterns persist over time.

VII. CONCLUSIONS

This study examined how alternative statistical and machine-learning specifications represent nonlinear structure in a transport-related urban livability framework covering 151 metropolitan areas across seven world regions. The analytical panel combines heterogeneous city-level statistics with harmonised annual estimates for 2019 – 2023 and contains 17 indicators spanning mobility, environmental quality, safety and economic conditions.

Gradient-boosted tree models reproduced the constructed livability index more accurately than linear benchmarks, although much of the improvement was already captured by an additive spline specification. This suggests that curvature is a more important source of model flexibility than high-order interaction in the present dataset. SHAP attribution, dependence analysis and interaction decomposition further show that several environmental, safety and mobility indicators contribute differently across the fitted response surface.

The analysis also identifies candidate locations at which selected fitted relationships change slope. These values should not be interpreted as universal policy thresholds. They are descriptive properties of the fitted model and constructed index and require validation against independent observations before they can be translated into planning benchmarks. Likewise, the coordinated indicator-change scenarios quantify changes in model predictions rather than causal effects of real-world interventions.

The main contribution of the study is therefore methodological. It demonstrates how grouped validation, interpretable tree ensembles, attribution analysis, interaction decomposition and scenario testing can be combined to examine the internal structure and sensitivity of a multi-indicator urban sustainability framework. At the same time, the results underline the importance of distinguishing model interpretation from causal inference and constructed composite outcomes from independent measures of residents' wellbeing. Future work using directly observed longitudinal data and externally measured quality-of-life outcomes would provide a stronger basis for testing the empirical and policy relevance of the nonlinear patterns identified here.

REFERENCES

- [1] Voukelatou, V., Gabrielli, L., Miliou, I., et al. (2021). Measuring objective and subjective well-being: dimensions and data sources. *International Journal of Data Science and Analytics*, 11(4), 279–309. <https://doi.org/10.1007/s41060-020-00224-2>
- [2] Bokhari, A., & Sharifi, F. (2022). Public transport and subjective well-being in the just city: A scoping review. *Journal of Transport & Health*, 25, 101372. <https://doi.org/10.1016/j.jth.2022.101372>
- [3] Gössling, S., Nicolosi, J., & Litman, T. (2021). The health cost of transport in cities. *Current Environmental Health Reports*, 8(2), 196–201. <https://doi.org/10.1007/s40572-021-00308-6>
- [4] Welch, D., Shepherd, D., Dirks, K. N., et al. (2023). Health effects of transport noise. *Transport Reviews*, 43(6), 1190 – 1210. <https://doi.org/10.1080/01441647.2023.2206168>
- [5] Nieuwenhuijsen, M. J. (2021). New urban models for more sustainable, liveable and healthier cities post covid19; reducing air pollution, noise and heat island effects and increasing green space and physical activity. *Environment International*, 157, 106850. <https://doi.org/10.1016/j.envint.2021.106850>
- [6] Ji, T., Li, K., Sun, Q., & Duan, Z. (2024). Urban transport emission prediction analysis through machine learning and deep learning techniques. *Transportation Research Part D: Transport and Environment*, 135, 104389. <https://doi.org/10.1016/j.trd.2024.104389>
- [7] Chatziioannou, I., Nikitas, A., Tzouras, P. G., et al. (2023). Ranking sustainable urban mobility indicators and their matching transport policies to support liveable city futures: A MICMAC approach. *Transportation Research Interdisciplinary Perspectives*, 18, 100788. <https://doi.org/10.1016/j.trip.2023.100788>
- [8] Rutka, A., Michnej, M., Zwoliński, T., et al. (2024). Core indicators for monitoring the sustainable urban mobility plan: A case study. *Sustainability*, 16(13), 5632. <https://doi.org/10.3390/su16135632>
- [9] Higgs, C., Lowe, M., Giles-Corti, B., Boeing, G., Delclòs-Alió, X., et al. (2024). Global healthy and sustainable city indicators: Collaborative development of an open science toolkit for calculating and reporting on urban indicators internationally. *Environment and Planning B: Urban*

- Analytics and City Science, 52(5), 1252 – 1270. <https://doi.org/10.1177/23998083241292102>
- [10] Dijkstra, L., Florczyk, A. J., Freire, S., Kemper, T., Melchiorri, M., Pesaresi, M., & Schiavina, M. (2021). Applying the Degree of Urbanisation to the globe: A new harmonised definition reveals a different picture of global urbanisation. *Journal of Urban Economics*, 125, 103312. <https://doi.org/10.1016/j.jue.2020.103312>
- [11] Bovkir, R., Ustaoglu, E., & Aydinoglu, A. C. (2023). Assessment of urban quality of life index at local scale with different weighting approaches. *Social Indicators Research*, 165(2), 655 – 678. <https://doi.org/10.1007/s11205-022-03036-y>
- [12] Greco, S., Ishizaka, A., Tasiou, M., & Torrissi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research*, 141(1), 61–94. <https://doi.org/10.1007/s11205-017-1832-9>
- [13] AlKhereibi, A. H., Wakjira, T. G., Kucukvar, M., & Onat, N. C. (2023). Predictive machine learning algorithms for metro ridership based on urban land use policies in support of transit-oriented development. *Sustainability*, 15(2), 1718. <https://doi.org/10.3390/su15021718>
- [14] Xiao, L., Lo, S., Liu, J., Zhou, J., & Li, Q. (2021). Nonlinear and synergistic effects of TOD on urban vibrancy: Applying local explanations for gradient boosting decision tree. *Sustainable Cities and Society*, 72, 103063. <https://doi.org/10.1016/j.scs.2021.103063>
- [15] Bloemsa, L. D., Wijga, A. H., Klompaker, J. O., et al. (2022). Green space, air pollution, traffic noise and mental wellbeing throughout adolescence: Findings from the PIAMA study. *Environmental International*, 163, 107197. <https://doi.org/10.1016/j.envint.2022.107197>
- [16] Xu, J., Ma, J., & Tao, S. (2024). Examining the nonlinear relationship between neighborhood environment and residents' health. *Cities*, 152, 105213. <https://doi.org/10.1016/j.cities.2024.105213>
- [17] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189 – 1232. <https://doi.org/10.1214/aos/1013203451>
- [18] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- [19] Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56 – 67. <https://doi.org/10.1038/s42256-019-0138-9>
- [20] Aas, K., Jullum, M., & Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298, 103502. <https://doi.org/10.1016/j.artint.2021.103502>
- [21] Liu, Z., Ma, X., Hu, L., et al. (2022). Nonlinear cooling effect of street green space morphology: Evidence from a gradient boosting decision tree and explainable machine learning approach. *Land*, 11(12), 2220. <https://doi.org/10.3390/land11122220>
- [22] Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4), 1059–1086. <https://doi.org/10.1111/rssb.12377>
- [23] Priulla, A., & D'Angelo, N. (2024). Sequential hypothesis testing for selecting the number of changepoints in segmented regression models. *Environmental and Ecological Statistics*, 31(2), 583 – 604. <https://doi.org/10.1007/s10651-024-00605-x>
- [24] Melchiorri, M. (2022). The global human settlement layer sets a new standard for global urban data reporting with the urban centre database. *Frontiers in Environmental Science*, 10, 1003862. <https://doi.org/10.3389/fenvs.2022.1003862>
- [25] Kutty, A. A., Wakjira, T. G., Kucukvar, M., Abdella, G. M., & Onat, N. C. (2022). Urban resilience and livability performance of European smart cities: A novel machine learning approach. *Journal of Cleaner Production*, 378, 134203. <https://doi.org/10.1016/j.jclepro.2022.134203>
- [26] Wójcik, P., & Andruszek, K. (2022). Predicting intra-urban well-being from space with nonlinear machine learning. *Regional Science Policy & Practice*, 14(4), 891–913. <https://doi.org/10.1111/rsp3.12478>
- [27] Viljanen, M., Meijerink, L., Zwakhals, L., & van de Kasstelee, J. (2022). A machine learning approach to small area estimation: Predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, 21(1), 4. <https://doi.org/10.1186/s12942-022-00304-5>
- [28] Huang, X., Kang, C., Yin, C., & Li, Y. (2023). Urban and individual correlates of subjective well-being in China: An application of gradient boosting decision trees. *Frontiers in Public Health*, 11, 1090832. <https://doi.org/10.3389/fpubh.2023.1090832>
- [29] Ding, D., Luo, M., Infante, M. F. P., et al. (2024). The co-benefits of active travel interventions beyond physical activity: A systematic review. *The Lancet Planetary Health*, 8(10), e790 – e803. [https://doi.org/10.1016/s2542-5196\(24\)00201-8](https://doi.org/10.1016/s2542-5196(24)00201-8)
- [30] Xie, D., Gou, Z., & Gui, X. (2024). How electric vehicles benefit urban air quality improvement: A study in Wuhan. *Science of The Total Environment*, 906, 167584. <https://doi.org/10.1016/j.scitotenv.2023.167584>
- [31] Xue, Y., Nie, T., Cui, Y., et al. (2023). Prediction of air pollution reduction benefits and atmospheric environmental quality improvement effects from electric vehicle deployment in Beijing, China. *International Journal of Environmental Science and Technology*, 20(10), 10973–10982. <https://doi.org/10.1007/s13762-022-04655-1>
- [32] Bayham, J., Burkhardt, J., Coffman, M., et al. (2022). Does air pollution increase electric vehicle adoption? Evidence from U.S. metropolitan areas, 2011–2018. *Journal of Environmental Economics and Policy*, 11(4), 438–462. <https://doi.org/10.1080/21606544.2022.2059015>
- [33] Chen, C., Wang, J., Li, D., Sun, X., Zhang, J., Yang, C., & Zhang, B. (2024). Unraveling nonlinear effects of environment features on green view index using multiple data sources and explainable machine learning. *Scientific Reports*, 14(1), 30189. <https://doi.org/10.1038/s41598-024-81451-6>
- [34] Shafqat, A., & Park, E. (2025). Broad applications of distributed lag non-linear model in public health: A comprehensive review. *GeoHealth*, 9(12), e2025GH001608. <https://doi.org/10.1029/2025GH001608>
- [35] Choi, Y., Byun, G., & Lee, J.-T. (2022). The effects of particulate matter alert on urban park visitation in Seoul, Korea: Using segmented regression. *International Journal of Environmental Research and Public Health*, 19(22), 15372. <https://doi.org/10.3390/ijerph192215372>
- [36] Mhangara, P., Gidey, E., & Mayise, B. S. (2025). Using extreme gradient boosting for predictive urban expansion analysis in Rustenburg, South Africa from 2000 to 2030. *Scientific Reports*, 15(1), 19050. <https://doi.org/10.1038/s41598-025-04304-w>
- [37] Zhu, J., Wang, S., Ma, H., et al. (2025). Nonlinear effect of urban visual environment on residents' psychological perception — An analysis based on XGBoost and SHAP interpretation model. *City and Environment Interactions*, 27, 100202. <https://doi.org/10.1016/j.cacint.2025.100202>
- [38] Zhu, J.-J., Yang, M., & Ren, Z. J. (2023). Machine learning in environmental research: Common pitfalls and best practices. *Environmental Science & Technology*, 57(46), 17671–17689. <https://doi.org/10.1021/acs.est.3c00026>
- [39] Slack, D., Hilgard, S., Jia, E., et al. (2020). Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 180–186). <https://doi.org/10.1145/3375627.3375830>
- [40] Wolfson, D. W., Andersen, D. E., & Fieberg, J. R. (2022). Using piecewise regression to identify biological phenomena in biotelemetry datasets. *Journal of Animal Ecology*, 91(9), 1755–1769. <https://doi.org/10.1111/1365-2656.13779>
- [41] Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 11(1), 4540. <https://doi.org/10.1038/s41467-020-18321-y>
- [42] Wang, Y., Khodadadzadeh, M., & Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. *International Journal of Applied Earth Observation and Geoinformation*, 121, 103364. <https://doi.org/10.1016/j.jag.2023.103364>

ACKNOWLEDGEMENTS

The authors acknowledge the organisations and public data platforms that made the urban transportation, environmental, safety, and socioeconomic data used in this study available. The authors also appreciate the institutional and technical support provided during data harmonisation, computational analysis, and manuscript preparation.

FUNDING

None.

AVAILABILITY OF DATA

Not applicable.

AUTHOR CONTRIBUTIONS

Taner Polat: Conceptualization, Methodology, Data curation, Formal analysis, Software, Validation, Visualization, Investigation, Writing – original draft, Writing – review & editing.

Emre Candar: Conceptualization, Methodology, Validation, Investigation, Resources, Supervision, Writing – review & editing.

COMPETING INTERESTS

The authors declare no competing interests.

Publisher's note WEDO remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is published online with Open Access by Green Design Engineering and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

© The Author(s) 2026