Explainable AI for Identifying the Environmental Drivers of Public Satisfaction in Urban Spaces

Gregory Zamora
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
gregoryz@unr.edu

Blazej Poletek
Department of Computer Science, University of North Texas, Denton, TX, USA.
poletekblazej@unt.edu

Abstract

Urban public satisfaction is shaped by a dense mixture of physical streetscape qualities, accessibility conditions, ambient exposures, and socially produced activity patterns. Machine-learning models can predict satisfaction from such environmental data with considerable flexibility, but predictive accuracy alone does not identify which features should inform design or policy. This manuscript develops a methodological framework for explainable artificial intelligence in urban satisfaction research that treats environmental drivers as spatially embedded, interacting, and uncertainty-qualified quantities rather than isolated ranking scores. The framework fuses street-view imagery, geotagged social media signals, and tabular environmental indicators; estimates satisfaction with a regularized predictive model; explains predictions through Shapley attribution and interaction analysis; diagnoses spatial dependence and scale sensitivity; and converts attributions into uncertainty-aware decision rules. The central contribution is an integrated analytical specification that connects multimodal urban feature construction, model-agnostic explanation, spatial diagnostics, calibration, and intervention screening in one coherent workflow. The framework also clarifies the difference between predictive attribution and causal effect, explains why tree-based models and conditional attribution are often preferable for tabular environmental features, and identifies failure modes associated with correlated variables, spatial confounding, platform bias, and the modifiable areal unit problem. Practical significance lies in producing driver rankings and interaction profiles that can be audited, spatially inspected, and translated into design priorities without overstating what a predictive model can establish. The approach supports more defensible use of explainable machine learning in urban space assessment and public satisfaction research.

Keywords

explainable artificial intelligence; urban satisfaction; street-level imagery; spatial confounding; Shapley values; environmental drivers

1. Introduction

Urban public satisfaction is a multidimensional outcome reflecting how residents and visitors evaluate the physical, social, and experiential qualities of places [1]. It is not reducible to a single density, greenery, or accessibility measure. Design research has shown that perceived walkability and streetscape character depend on subtle qualities such as enclosure, human scale, transparency, and complexity [2]. At the same time, large-scale audit methods

have made it possible to characterize neighborhood environments from street-level imagery rather than relying only on aggregated land-use inventories [3]. These developments create an opportunity for machine learning to connect fine-grained environmental conditions with satisfaction outcomes at scales that were previously difficult to analyze.

A parallel stream of urban sensing research uses geotagged social media and other digital traces to characterize activity, noise, and place use [4]. Methodological work has also fused geotagged social media with street-view imagery in deep neural networks to model urban streetscapes [5], demonstrating that physical form and socially produced activity signals can be represented jointly. The resulting feature spaces are rich but difficult to interpret. A model may predict satisfaction well while relying on correlated visual features, neighborhood composition, or incidental platform signals that do not correspond to actionable environmental drivers. This creates a central analytical problem: how can urban satisfaction models identify environmental features that are predictively relevant, spatially credible, interaction-aware, and sufficiently stable to support design or policy judgment?

Explainable artificial intelligence offers tools for this problem, but not a complete solution. Shapley-value attribution provides a principled additive decomposition of model output [6], while local surrogate methods illustrate how individual predictions can be explained through simpler approximations [7]. Evaluation frameworks emphasize that explanation quality depends on the intended use and the credibility of the explanatory model [8]. These methods are necessary but insufficient for urban satisfaction research. Environmental features are spatially dependent, measured at multiple scales, and often strongly correlated. An attribution that ignores spatial structure can confuse neighborhood context with site-specific design, and an explanation that ignores interaction can misstate the value of a single feature when its effect depends on adjacent conditions. The knowledge gap is therefore not the absence of explanation algorithms but the absence of an integrated framework that aligns multimodal feature construction, flexible prediction, attribution, spatial diagnosis, and uncertainty communication for urban satisfaction.

The contribution developed here is an explainable modeling framework for identifying environmental drivers of public satisfaction in urban spaces. It specifies the target construct, multimodal feature representation, predictive objective, attribution and interaction layer, spatial diagnostics, calibration logic, and intervention-screening rule. The framework is methodological rather than an empirical evaluation; it provides a defensible analytical architecture for studies that have appropriate satisfaction, imagery, geotagged, and environmental data. Its main argument is that driver identification should be treated as an inference workflow rather than a single post hoc ranking, and that spatial and uncertainty diagnostics belong inside that workflow rather than after it.

2. Analytical Framework

2.1 Satisfaction as a spatially embedded latent construct

Public satisfaction can be represented as a latent evaluation y_i^* for observation i, where i may denote a geolocated survey response, a household or resident record, or an aggregated spatial unit. The observed outcome y_i may be an ordinal rating, a bounded scale score, or an area-level average. Environmental features are not plausibly independent of neighboring conditions: satisfaction in one place can be influenced by the quality of adjacent streets, nearby destinations, and broader neighborhood context. A useful starting specification is

therefore a spatially lagged latent model in which environmental predictors, neighboring satisfaction, unobserved area effects, and idiosyncratic error operate together.

$$y_i\hat{} * = g(x_i) + \rho \sum_j w_ij\, y_j\hat{} * + u_i + \varepsilon_i$$

Here x_i is the vector of environmental features for observation i, g is a flexible predictive function, ρ captures the strength of spatial dependence in satisfaction, w_ij are pre-specified spatial weights, u_i represents unobserved area or group effects, and ε_i is an observation-level error term. The equation is not a claim that satisfaction is literally generated by a linear spatial process; it is a conceptual target that forces the analyst to distinguish site-specific environmental effects from neighborhood spillovers and omitted contextual effects. Without this distinction, a model may attribute satisfaction to a local design feature when the apparent effect is largely a proxy for neighborhood quality.

2.2 Multimodal environmental feature representation

The environmental feature space should combine physical streetscape information, socially produced activity signals, and tabular environmental indicators. Street-view images can be encoded to represent visible greenness, sky view, building enclosure, sidewalk presence, facade variation, and street furniture. Geotagged social media and other digital traces can represent activity intensity, place visitation, and socially salient locations, although such signals are subject to platform, demographic, and sampling biases. Tabular indicators can include accessibility, noise exposure, air quality, land-use mix, population density, and distance to public amenities. The feature vector can be written as a fusion of modality-specific encoders.

$$x_i = \varphi(I_i;\, \theta_I) \oplus \psi(T_i;\, \theta_T) \oplus z_i$$

In this expression, I_i denotes street-level imagery associated with location i, T_i denotes geotagged text or activity records, z_i denotes tabular environmental variables, φ and ψ are modality-specific encoders parameterized by θ_I and θ_T, and $\oplus$ denotes concatenation or another fixed fusion operation. The encoders should be trained or pre-trained in a way that does not leak satisfaction outcomes into feature construction. This constraint matters because image or text embeddings learned jointly with the satisfaction target can absorb idiosyncratic associations that are not interpretable as environmental drivers. The fused feature space should also preserve interpretable low-dimensional variables, such as visible green fraction or distance to transit, so that attribution can be cross-checked against directly measurable quantities.

2.3 Predictive model and explanation layer

The predictive model g should be flexible enough to capture nonlinearity and interaction but constrained enough to support stable attribution. Tree ensembles are often strong candidates for tabular and fused environmental features [12], and comparative evidence suggests that tree-based models remain competitive on typical tabular data relative to deep networks [13]. Deep encoders remain valuable for imagery and text [10], [11], but their high-dimensional internal representations are less directly interpretable than explicit environmental features. A practical design is therefore a hybrid: deep encoders reduce raw imagery and text to moderately dimensioned, reproducible features, while a tree ensemble or regularized additive model estimates satisfaction from the fused feature set. A taxonomy of interpretability methods helps clarify that this design uses post hoc, model-agnostic explanations for a model whose inputs remain semantically auditable [9].

The predictive function is estimated by minimizing a regularized empirical loss.

$$\min_g \sum_{i=1}^{n} \ell\big(y_i, g(x_i)\big) + \lambda\Omega(g)$$

Here ℓ is a loss function appropriate to the outcome, such as ordinal logistic loss for ratings or squared error for continuous area averages, λ is a nonnegative regularization parameter, and Ω(g) penalizes model complexity. The loss specification should match the measurement process: treating an ordinal satisfaction scale as continuous can distort residual structure, while treating an area average as if it were an individual response can produce overconfident explanations. The regularization term controls overfitting and reduces the chance that attribution reflects idiosyncratic noise rather than stable environmental structure.

3. Methodological Framework

3.1 Attribution and interaction analysis

Driver identification requires a decomposition of model output into feature contributions. Shapley values provide an additive attribution with an efficiency property: the contributions sum to the difference between the model prediction and a baseline value. For a model g and feature set F, the Shapley value for feature k at prediction x is defined as a weighted average of marginal contributions over all subsets of the remaining features.

$$\varphi_k(x) \;=\; \sum_\{S \subseteq F \setminus \{k\}\} |S|!\,(|F| - |S| - 1)!/|F|!\;\; (g(x_\{S \cup \{k\}\}) - g(x_S))$$

In this equation, S is a subset of features that excludes k, x_S denotes the prediction input in which only features in S are set to their observed values while the remaining features follow a specified reference distribution, and φ_k(x) is the attributed contribution of feature k. The reference distribution is a substantive choice. Interventional Shapley values, which break feature dependence, answer what happens when a feature is changed while others are held to a reference. Conditional Shapley values, which respect feature correlation, answer how a feature contributes within the observed dependence structure. For urban design questions, interventional attribution is often more interpretable for hypothetical changes, but it can assign credit to features that are not manipulable. Conditional attribution can be more faithful to the observed landscape but can split credit among correlated variables such as greenery, sidewalk presence, and neighborhood income.

A single-feature attribution can also obscure environmental interaction. The value of a bench, shade tree, or sidewalk may depend on adjacent traffic, enclosure, or activity. Pairwise interaction attribution can be expressed as an expected second difference over subsets.

$$\Phi_\{k,l\}(x) \;=\; E_S[g(x_\{S \cup \{k,l\}\}) - g(x_\{S \cup \{k\}\}) - g(x_\{S \cup \{l\}\}) + g(x_S)]$$

Here Φ_{k,l}(x) captures the incremental joint contribution of features k and l beyond their separate contributions, and E_S denotes an expectation over feature subsets under the chosen reference distribution. Positive interaction means that the two features reinforce one another; negative interaction means that one feature diminishes the value of the other. This is directly relevant to urban spaces because design features rarely operate in isolation. A greenery effect without controlling for enclosure or traffic exposure may be an average over heterogeneous places and may not transfer to a new site.

3.2 Spatial diagnostics and scale sensitivity

Attribution maps must be examined for spatial dependence and scale artifacts. If positive residuals cluster geographically, the model is missing a spatially structured process, and feature attributions may be contaminated by neighborhood context. Local indicators of spatial association provide a diagnostic for residual clustering and local spatial outliers [14]. For a standardized residual or satisfaction value z_i, with m_2 equal to the average of squared deviations, a local Moran-type statistic can be written as follows.

$$I_i = \left(\frac{z_i}{m_2}\right)\sum_j w_i j z_j$$

In this expression, I_i measures the local association between location i and its neighbors, w_ij are spatial weights, and z_i is the value being diagnosed. A cluster of high explanatory residuals in one district may indicate that the model has absorbed a neighborhood-level effect into site-level features. Spatial nonstationarity is a related concern: the association between environmental features and satisfaction may vary across the city. Geographically weighted regression and related local models provide a conceptual reference for allowing relationships to vary over space [15]. The point is not to replace machine learning with a global regression but to use spatial diagnostics to decide whether attribution should be reported globally, by district, or through spatially varying summaries.

Scale is equally consequential. Environmental variables measured at the block, buffer, tract, or city level can produce different attributions, a manifestation of the modifiable areal unit problem [16]. An explanation that ranks greenery as a dominant driver at a coarse buffer may not hold at the street-segment scale. The modeling workflow should therefore specify the unit of analysis in advance and repeat key attributions across defensible spatial scales. Agreement across scales strengthens a driver claim; disagreement is itself informative and should be reported rather than hidden by a single preferred geography.

3.3 Calibration and uncertainty

Attribution without uncertainty can create false confidence. A feature may rank highly in a global summary but have wide local variation or unstable sign across resamples. Conformal prediction offers a distribution-free way to attach uncertainty to model output under exchangeability assumptions. For a fitted model g and a calibration sample, a marginal prediction interval can be formed from absolute residuals.

$$C_\alpha(x) = \{y : |y - g(x)| \leq q_\{1-\alpha\}\}$$

Here C_α(x) is the prediction set at coverage level 1−α, and q_{1−α} is the empirical quantile of calibration residuals. Spatial data complicate the exchangeability assumption: nearby observations are not independent, so calibration should be performed with spatial blocking or leave-one-region-out schemes when the goal is transfer to new areas. For driver identification, uncertainty should be attached not only to predictions but also to attributions. Attribution intervals can be generated through bootstrap resampling of observations, spatial blocks, or model seeds. A driver ranking is more credible when the sign and relative magnitude of contributions remain stable across these perturbations.

3.4 Decision rules for intervention screening

The final step converts explanations into a screening rule rather than a deterministic prescription. Let a denote a candidate intervention, such as adding shade, reducing traffic exposure, improving sidewalk continuity, or increasing access to a public space. The expected utility of intervention a at location i can be compared with its cost.

$$a^{\wedge} * (i) = argmax_a\, E[U(a)|x_i] - c(a)$$

The term E[U(a)|x_i] is the expected satisfaction benefit conditional on the environmental profile, c(a) is intervention cost, and a^(i) is the screened intervention for location i. This expression should be interpreted as a decision aid, not as an estimated causal effect. Its inputs may come from model predictions, design heuristics, or a separate causal design. The value of the rule is that it forces explicit treatment of cost, uncertainty, and local context. It also prevents a common misreading of feature attribution: a high Shapley value for greenery does not by itself imply that planting trees at a given location will produce a proportional satisfaction gain.

4. Technical Discussion

4.1 Model choice and interpretability trade-offs

The proposed framework favors models with strong tabular performance and stable post hoc attribution. Tree ensembles are attractive because they handle nonlinearity and interaction, require limited feature scaling, and support exact or near-exact Shapley computation for tree structures [6], [12]. Deep networks remain appropriate for raw image and text encoders, where representation learning is essential [10], [11], but the explanation target should be the fused environmental features and their interactions, not only low-level image pixels. This choice reflects a trade-off between predictive capacity and auditability. A highly flexible end-to-end model may achieve lower predictive error, yet its attributions can be harder to communicate and more sensitive to spurious correlations. The framework does not assume that one model class is universally superior; it requires that model selection be justified against the intended inferential use and evaluated with spatial and resampling diagnostics.

4.2 Confounding, spatial dependence, and attribution stability

The central interpretive limitation is that predictive attribution is not causal identification. A feature can receive high attribution because it is a proxy for an unobserved process, because it is correlated with another driver, or because the model has learned a neighborhood-level pattern that is not transferable. Spatial dependence intensifies this risk: a local feature may correlate with district-level affluence, service quality, or historical investment, and the model may assign to the feature what belongs to the broader context. Spatial diagnostics help detect residual clustering, but they do not eliminate confounding. Stronger causal claims require quasi-experimental or design-based variation, such as policy changes, matched comparisons, or instrumental variation. The framework therefore distinguishes three levels of evidence: predictive importance, spatially robust attribution, and causal effect. Only the third supports claims that changing a feature will change satisfaction.

Attribution stability also depends on correlated features and reference distributions. When two variables are nearly collinear, Shapley values can divide credit arbitrarily unless the analyst specifies a conditional or grouped-feature attribution. Grouping conceptually related variables, such as all greenery measures or all accessibility measures, can produce more interpretable driver families. Multiverse analysis is useful here because it makes the consequences of reasonable analytical choices visible rather than burying them in a single specification [17]. Reporting rankings across multiple spatial scales, model classes, and reference distributions is more honest than presenting one definitive explanation. Model complexity and regularization choices should likewise be examined through bias-variance reasoning and out-of-sample assessment [18]. A driver that survives these variations is a

stronger candidate for policy attention than one that appears only under a single favorable configuration.

4.3 Boundary conditions and validation logic

The framework has clear boundary conditions. It assumes that environmental features can be measured with acceptable validity at the chosen spatial unit and that satisfaction outcomes can be linked to those units without severe misalignment. Geo-referencing error, image capture date, seasonal variation, and platform bias can all weaken this link. Street-view imagery may overrepresent visible frontage and underrepresent interior courtyards, night-time conditions, or ephemeral uses. Geotagged social media may overrepresent certain demographics and activities. Tabular environmental variables may be modeled at different temporal resolutions from the satisfaction outcome. These are not minor measurement details; they determine whether an attribution is interpretable as an environmental driver.

Validation should therefore be spatial and multimodal. Predictive performance should be assessed with spatial block cross-validation, not only random splits, because random splits overstate performance when nearby observations are correlated. Attribution stability should be checked across model seeds, feature encodings, and spatial scales. Explanation plausibility should be evaluated against domain knowledge and, where possible, against independent audits or direct environmental measurements. The framework's decision rule should be applied only after these checks, and its outputs should be reported with uncertainty and spatial context. This validation logic aligns explanation quality with the intended planning use rather than treating explanation as a decorative post-processing step.

5. Implications for Urban Design and Policy

The framework has several implications for urban design and policy analysis. First, it encourages driver identification to be organized around interpretable feature families rather than isolated variables. Designers can examine whether satisfaction is associated primarily with visible greenness, enclosure, traffic exposure, accessibility, or activity intensity, and then inspect interactions among those families. Second, spatial diagnostics make it possible to distinguish citywide tendencies from district-specific relationships. A feature that is beneficial on average may be neutral or negative in a particular context, and a spatially explicit explanation can reveal that heterogeneity. Third, uncertainty-aware attributions support more defensible prioritization. A planning team can compare candidate interventions not only by predicted benefit but also by the stability of the underlying evidence.

The framework also cautions against overreliance on global feature rankings. Public satisfaction is an outcome of place experience, and place experience is produced by combinations of features. A global ranking may hide the fact that a feature matters only when another condition is present, or that its effect is concentrated in a subset of locations. Interaction analysis and local attribution maps are therefore not optional refinements; they are part of the substantive interpretation. When communicated carefully, these outputs can support design review, participatory planning, and environmental audit by showing where a model's conclusions are strong, where they are uncertain, and where additional data or causal evidence are needed.

6. Conclusion

This manuscript has developed a methodological framework for explainable artificial intelligence in identifying environmental drivers of public satisfaction in urban spaces. The

framework treats satisfaction as a spatially embedded latent construct, represents urban environments through fused street-view, geotagged, and tabular features, estimates satisfaction with a regularized predictive model, decomposes predictions through Shapley attribution and interaction analysis, diagnoses spatial dependence and scale sensitivity, attaches uncertainty through calibration, and converts explanations into a transparent intervention-screening rule. Its central contribution is the integration of these components into a single inferential workflow rather than a collection of disconnected algorithms.

The framework's practical value lies in making driver identification more auditable. It clarifies that predictive attribution is not causal effect, that spatial confounding and measurement bias can distort rankings, and that stable explanations require spatial validation, grouped-feature reasoning, and sensitivity to model and scale choices. These constraints do not diminish the usefulness of explainable machine learning; they specify the conditions under which its outputs can responsibly inform urban design and policy. Used with appropriate data and cautious interpretation, the framework can help researchers and practitioners distinguish robust environmental drivers from artifacts of correlated features, spatial context, or platform bias, and thereby support more defensible decisions about the design of urban public spaces.

References

1. Marans, R. W., & Stimson, R. J. (2011). Investigating quality of urban life: Theory, methods, and empirical research. Springer.

2. Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. Journal of Urban Design, 14(1), 65–84.

3. Rundle, A. G., Bader, M. D. M., Richards, C. A., Neckerman, K. M., & Teitler, J. O. (2011). Using Google Street View to audit neighborhood environments. American Journal of Preventive Medicine, 40(1), 94–100.

4. Zheng, Y., Liu, T., Wang, Y., Zhu, Y., Liu, Y., & Chang, E. (2014). Diagnosing New York City's noises with ubiquitous data. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 715–725). ACM.

5. Aman, J., Matisziw, T. C., Kim, J. B., & Luo, D. (2022, April). Sensing the city: Leveraging geotagged social media posts and street view imagery to model urban streetscapes using deep neural networks. In Proceedings of the International Conference on Computer-Aided Architectural Design Research in Asia (pp. 595-604). Association for Computer-Aided Architectural Design Research in Asia.

6. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates.

7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.

8. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv:1702.08608). arXiv.

9. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Independently published.

10. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

12. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

13. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Advances in Neural Information Processing Systems 35. Curran Associates.

14. Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93–115.

15. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically weighted regression: The analysis of spatially varying relationships. Wiley.

16. Openshaw, S. (1984). The modifiable areal unit problem. Geo Books.

17. Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives on Psychological Science, 11(5), 702–712.

18. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.