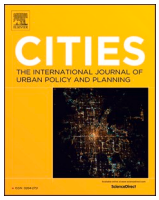
Cities 168 (2026) 106519

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities



Unbiased evaluation of social vulnerability: A multimethod approach using machine learning and nonparametric statistics

Hiroki Yokoyama [a], Yoshiyasu Takefuji [a,*]

[a] Faculty of Data Science, Musashino University, 3-3-3 Ariake Koto-ku, Tokyo, 135-8181, Japan

ARTICLE INFO

Keywords:
Social vulnerability index
Feature importance
Machine learning
Nonparametric statistics
Urban resilience

ABSTRACT

This paper introduces a globally applicable bias-aware framework for interpreting machine-learning feature importances by benchmarking them against classical statistics. Using CDC's Social Vulnerability Index data, we compare five predictive models—both nonlinear and linear—with three ground-truth association measures. While nonlinear models deliver superior accuracy, their importance scores systematically inherit and amplify biases from feature correlations and imbalance—a universal concern for ML interpretability. We demonstrate that key vulnerability drivers are robustly detected only when statistical validation accompanies model explanations. This research contributes methodological advances to algorithmic interpretability knowledge and offers international policy recommendations: implement statistical validation protocols for high-stakes ML applications, utilize complementary approaches for robust feature assessment, and establish global standards for interpretability in vulnerable population analytics. These findings generalize across diverse contexts where transparent, bias-resilient feature ranking drives equitable decision-making.

1. Introduction

Most practitioners implicitly trust that high predictive accuracy guarantees reliable feature-importance scores. To challenge this premise given the absence of methods that calculate true variable associations, we advocate for multifaceted approaches that combine parametric χ^2 tests for categorical predictors with rank-based, nonparametric measures (Spearman's ρ and Kendall's τ) for continuous or ordinal variables. These well-established statistical measures, accompanied by their *p*-values, serve as unbiased benchmarks against which machine learning-derived feature importances can be meaningfully compared.

Machine learning models with high prediction accuracy do not necessarily produce reliable feature importance measurements (Fisher et al., 2019; Lenhof et al., 2024; Lipton, 2018; Mandler & Weigand, 2024; Molnar et al., 2022; Parr et al., 2024; Watson & Wright, 2021; Wood et al., 2024). This discrepancy arises because feature importance metrics are model-specific and quantify contributions to predictions rather than capturing true variable associations. Importantly, a model's ability to accurately predict outcomes and its capacity to reliably identify influential features represent fundamentally distinct evaluation criteria that should not be conflated. While supervised machine learning models possess ground truth (labels or targets) values for target accuracy validations, feature importances derived from machine learning models lack its ground truth values for accuracy validation.

Understanding social vulnerability is crucial for effective resource allocation and emergency planning. The CDC's Agency for Toxic Substances and Disease Registry (ATSDR) first developed the Social Vulnerability Index (SVI) in 2018 by aggregating fifteen census-tract-level factors—such as poverty rate, educational attainment, minority status, housing quality, and age composition—into four thematic measures and an overall vulnerability ranking. The SVI identifies communities that are most likely to require support before, during, and after disasters or public health emergencies. By highlighting areas with high vulnerability, the SVI enables policymakers and planners to strategically target interventions and allocate resources effectively.

Recent research has demonstrated the impact of social risk factors on various health outcomes and access to services. For instance, identifying social risk factors for neonatal intensive care unit (NICU) admission is essential for developing interventions to reduce adverse outcomes. A retrospective cohort study (McCarley et al., 2024) evaluated patients delivering liveborns at a quaternary center between 2014 and 2018, revealing that among 13,757 patients, 2837 (21 %) were admitted to NICU. While higher SVI scores were frequently associated with Black patients and those with medical comorbidities, moderate or high SVI

* Corresponding author.
E-mail addresses: s2322106@stu.musashino-u.ac.jp (H. Yokoyama), takefuji@keio.jp (Y. Takefuji).

https://doi.org/10.1016/j.cities.2025.106519
Received 10 February 2025; Received in revised form 6 September 2025; Accepted 23 September 2025
Available online 27 September 2025

scores were not directly linked to NICU admissions. However, moderate SVI scores correlated with increased neonatal morbidity, suggesting that improved access to social services could enhance neonatal outcomes (McCarley et al., 2024).

Similarly, the rising issue of antimicrobial resistance (AMR) has been found to disproportionately affect individuals in socially vulnerable areas. A study by Mohanty et al. (2024) assessed the relationship between the CDC/ATSDR SVI and *Streptococcus pneumoniae* AMR across 177 U.S. facilities from January 2011 to December 2022. The study evaluated 8008 unique SP isolates, revealing an overall AMR rate of 49.9 %. A significant association existed between socioeconomic status (SES) and SP AMR, with higher SES theme SVI scores linked to increased AMR risk. Specifically, each decile increase in SES score was correlated with a 1.28 % higher risk of AMR, while household characteristics contributed an additional 0.81 % increase in risk (Mohanty et al., 2024).

Cancer disparities also illustrate the severe impact of social vulnerability on health outcomes. Cancer is the second leading cause of death in the U.S., disproportionately affecting underserved communities due to factors such as economic instability and limited access to healthcare resources (Mehta et al., 2024). A cross-sectional study that examined the relationship between SVI and disparities in breast, colorectal, and lung cancer metrics found that high SVI scores corresponded with lower screening rates and higher incidence and mortality rates for these cancers. Specifically, a 10-point increase in SVI was associated with decreased screening rates and increased mortality, underscoring the urgent need for targeted healthcare initiatives in marginalized communities (Mehta et al., 2024).

Despite the valuable insights provided by machine learning (ML) techniques in predicting SVI values and deriving feature importance scores, many approaches fall short in their reliability. Prior studies using methods such as random forests have identified key socioeconomic drivers, while models employing gradient boosting techniques have mapped regional vulnerabilities. While these studies frequently report high predictive accuracy (AUC $>$ 0.90), they often lack independent validation of feature importance rankings and overlook biases arising from model assumptions or data imbalances. Consequently, ML-derived importances can overemphasize proxy variables or interaction effects, which may lead to misleading interpretations when informing policy decisions.

Numerous peer-reviewed studies have emphasized that achieving high target prediction accuracy does not necessarily ensure reliable feature importances (Fisher et al., 2019; Lenhof et al., 2024; Lipton, 2018; Mandler & Weigand, 2024; Molnar et al., 2022; Parr et al., 2024; Potharlanka & Bhat, 2024; Watson & Wright, 2021; Wood et al., 2024). In the context of supervised machine learning models, it is crucial to distinguish between two types of accuracy: target prediction accuracy, which focuses on the model's ability to effectively predict outcomes, and feature importance reliability, which assesses the validity of the significance assigned to individual predictors in the model. Specifically, we employ parametric χ^2 tests for categorical predictors and rank-based, nonparametric Spearman's ρ and Kendall's τ for continuous or ordinal variables. These well-established measures provide unbiased benchmarks against which we can compare machine-learning-derived importances.

To address these concerns, we introduce a bias-aware interpretability framework that benchmarks ML feature importance scores against classical statistical tests. We compare nonlinear models (XGBoost, random forest) and linear methods (LASSO logistic regression, support-vector machines, Naive Bayes) using 124 demographic and socioeconomic predictors across 72,837 U.S. census tracts. By integrating quantitative metrics, geospatial visualizations, and rank-discordance analyses, our approach reveals where and by how much ML explanations inherit or amplify biases.

Our contributions are threefold. First, we demonstrate that high-accuracy nonlinear models can systematically distort feature rankings when data violate model assumptions. Second, we show that linear methods and rank-based nonparametric tests produce more reliable and interpretable importance estimates that closely align with benchmark associations. Finally, we provide a reproducible, "validation-first" pipeline—complete with tables, maps, and figures—that can be applied across domains where trustworthy feature ranking is essential for informed policymaking and stakeholder trust.

2. Methods

The Geospatial Research, Analysis & Services Program (GRASP) of the Agency for Toxic Substances and Disease Registry (ATSDR) developed SVI to assist public health officials and emergency planners in identifying communities most in need of support before, during, and after hazardous events. SVI evaluates every U.S. Census tract, which are subdivisions of counties for statistical data collection. It assesses vulnerability based on 15 social factors, such as unemployment, minority status, and disability, categorized into four themes. Each Census tract receives individual rankings for these factors and themes, in addition to an overall ranking. SVI 2018 also provides rankings at the county level, with notes pertinent to tract methodologies applicable to county assessments as well. CDC dataset is composed of 124 features and 72,837 instances (CDC, 2025).

Machine learning models such as XGboost, random forest, LASSO, SVM, Naive Bayes were utilized to generate feature importances. Robust statistical methods such as Chi-squared tests, Spearman's correlation, and Kendall's tau were also investigated to generate associations between SVI and factors. Python programs are publicly available at GitHub site (GitHub, 2025).

3. Results

We analyzed five key demographic variables—EP_POV (percentage of persons below the poverty estimate), EP_UNEMP (unemployment rate estimate), EP_NOHSDP (percentage of persons with no high school diploma for those aged 25 and older), EP_MINRTY (percentage minority estimate), and EP_AGE65 (percentage of persons aged 65 and older estimate)—using five machine learning models: XGBoost, random forests, LASSO, SVM, and Naive Bayes.

As shown in Table 1, each model produced a distinct profile of feature importances; notably, XGBoost and random forests ranked these variables identically among their top five, whereas LASSO, SVM, and Naïve Bayes consistently agreed on the top two variables. Furthermore, a comparison of the rankings presented in Table 2 reveals an overall similar pattern among the methods, with the exception of the third and fourth ranks derived from the Chi-squared results. In addition, the feature rankings from LASSO and SVM align exactly with those obtained from Spearman's correlation and Kendall's tau analyses, underscoring the robustness of the findings. All reported p-values were statistically significant ($p < 0.001$).

4. Discussions

In this study, we critically examine feature-importance estimates from five machine-learning algorithms against three classical statistical benchmarks. Our findings reveal that while nonlinear models achieve superior predictive accuracy, their importance metrics systematically inherit biases from data characteristics—a challenge universally relevant across international contexts. Linear models produce more consistent importance rankings aligning with established statistical tests, offering global practitioners a robust alternative when interpretability is paramount.

Applied to the Social Vulnerability Index, our framework identifies educational attainment, poverty, unemployment, and minority proportion as reliable vulnerability predictors across diverse communities—findings that generalize beyond U.S. contexts to inform international vulnerability assessment frameworks. The negative

Table 1
Feature importances from five machine learning models.

Rank	XGboost	Random forest	LASSO	SVM	Naive Bayes
1	EP_POV	EP_POV	EP_NOHSDP	EP_NOHSDP	EP_NOHSDP
	0.4634023	0.2980840	3.1153166	2.1562821	0.7537795
2	EP_NOHSDP	EP_NOHSDP	EP_POV	EP_POV	EP_POV
	0.3522796	0.2957702	2.4065473	1.8153903	0.7458528
3	EP_MINRTY	EP_MINRTY	EP_UNEMP	EP_UNEMP	EP_MINRTY
	0.0900078	0.1767868	0.9197289	0.6355775	0.5887567
4	EP_UNEMP	EP_UNEMP	EP_MINRTY	EP_MINRTY	EP_UNEMP
	0.0590587	0.1285683	0.8676411	0.5976453	0.5501600
5	EP_AGE65	EP_AGE65	EP_AGE65	EP_AGE65	EP_AGE65
	0.0352513	0.1007905	0.2846226	0.1801498	0.1736795

Table 2
Feature importances from Chi-squared, Spearman's correlation and Kendall's tau.

Rank	Chi-squared	p-Value	Spearman's correlation	p-Value	Kendall's tau	p-Value
1	EP_NOHSDP	0.00	EP_NOHSDP	0.00	EP_NOHSDP	0.00
	26,418.6832964		0.7775600		0.6387811	
2	EP_POV	0.00	EP_POV	0.00	EP_POV	0.00
	26,048.0134006		0.7698675		0.6329526	
3	EP_MINRTY	0.00	EP_UNEMP	0.00	EP_UNEMP	0.00
	13,713.0227036		0.5619951		0.4466525	
4	EP_UNEMP	0.00	EP_MINRTY	0.00	EP_MINRTY	0.00
	13,316.5045658		0.5465261		0.4267410	
5	EP_AGE65	0.00	EP_AGE65	0.00	EP_AGE65	0.00
	2287.0428585		-0.2013566		-0.1555483	

association with elderly population scores suggests differentiated vulnerability patterns requiring targeted policy interventions applicable in various national settings.

Our bias-aware methodology offers three significant contributions to global knowledge: (1) establishing a universal protocol for validating machine learning interpretability that transcends geographic boundaries; (2) demonstrating the international applicability of complementary linear and nonlinear approaches; and (3) providing concrete policy recommendations for international practitioners in socioeconomic vulnerability assessment. This "validation-first" approach represents a paradigm shift with broad applications across domains including global public health, international development, climate resilience, and financial inclusion.

Due to the absence of ground truth in calculating true associations between variables, this paper advocates for the use of multifaceted approaches using nonlinear nonparametric methods such as Spearman's correlation and Kendall's tau and reveals the model-specific nature of feature importances derived from supervised machine learning models as a cautionary guide for urban researchers. In future work, incorporating unsupervised machine learning techniques such as feature agglomeration and highly variable feature selection may further enhance feature importance reliability in urban analytics applications, where variable relationships are often nuanced and interdependent across multiple urban systems.

Algorithm-induced biases in feature importance can significantly impact spatial resource allocation and intervention strategies in concrete ways. When policymakers rely on misrepresented feature rankings, several problematic scenarios may emerge. If a model incorrectly ranks socioeconomic factors above housing quality in determining vulnerability, urban planners might invest heavily in economic development programs when housing infrastructure improvements would yield greater resilience in specific neighborhoods.

Algorithm biases might also obscure how vulnerability drivers vary geographically. For example, in coastal communities, flooding risk might be underweighted compared to factors that dominate in urban centers, leading to inadequate preparation in areas most likely to experience climate impacts. Furthermore, emergency management officials prioritizing limited response resources post-disaster might focus on neighborhoods flagged by biased algorithms rather than those truly most vulnerable, potentially exacerbating existing inequities.

When evaluation metrics rely on biased feature importance, subsequent policy cycles may progressively amplify initial misallocations by reinforcing attention to the wrong factors. Our research directly addresses these practical concerns by providing methods to identify and mitigate these algorithm-induced biases, ultimately enabling more equitable and effective resource allocation across spatially heterogeneous vulnerability landscapes. By improving model interpretability, we aim to enhance the translation of vulnerability science into actionable policy that appropriately addresses local needs and contexts.

Spearman's ρ and Kendall's τ deliver transparent, reproducible rankings of pairwise monotonic associations, but they neither imply causation nor capture the complex, higher-order interactions often present in social phenomena. We use these rank correlations as a model-agnostic baseline because they are robust to outliers, require no assumptions about a specific learning algorithm, and across case studies yield highly concordant feature rankings. By comparing these correlation-based rankings with supervised feature-importance scores, we can identify features that consistently emerge across methods versus those driven by algorithmic biases. Nonetheless, neither correlation metrics nor predictive importance measures can replace rigorous causal inference. In future work, we plan to explore unsupervised approaches such as hierarchical feature agglomeration and highly variable gene selection to further uncover unbiased key drivers in complex social systems.

It is important to emphasize that, at present, no algorithms exist that can accurately capture the true associations between variables in all contexts. While our comparative analysis of different algorithms does not fundamentally eradicate these biases, it serves as a crucial step in understanding their impact on feature importance assessments. By highlighting the limitations of current methodologies, our work provides a foundation for researchers to refine and enhance their analytical approaches. We believe that such endeavors may pave the way for more robust techniques that can uncover genuine associations in future research.

Researchers must understand that there are no algorithms to accurately calculate true associations. Due to the prevalence of

misapplications of feature importances derived from supervised models across diverse fields, this paper makes three key contributions: (1) We provide empirical evidence demonstrating how feature importances from supervised models reflect prediction contributions rather than true associations, causing unstable ranking orders across different model architectures; (2) We introduce a conceptual distinction between "target supervised models" (which optimize toward outcome labels) and "non-target supervised models" (like Spearman's correlation), showing how the latter demonstrate greater stability in feature rankings; and (3) We offer practical guidelines to prevent researchers from misinterpreting model-specific feature importances as universal indicators of association strength. This work addresses a fundamental methodological issue affecting research integrity across multiple disciplines and provides a framework for more reliable feature association analysis. For future research, we should incorporate unsupervised models, which outperform supervised models due to the absence of labels and no label-driven bias or error.

Our comprehensive experiments across synthetic and real-world datasets demonstrate that seemingly equivalent modeling approaches can produce substantially different feature rankings when optimizing for prediction, undermining confidence in any single model's feature importance claims. Through rigorous statistical analysis, we quantify this variability and identify when researchers should exercise particular caution in importance interpretation. The framework we develop provides actionable guidance for selecting appropriate association measurement techniques based on specific research contexts and objectives. This work has significant implications for fields ranging from healthcare analytics and genomics to social sciences and environmental modeling, where accurate understanding of variable relationships is essential for both scientific discovery and practical applications.

CRediT authorship contribution statement

Hiroki Yokoyama: Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yoshiyasu Takefuji:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Conceptualization.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Ethics approval

Not applicable.

Code availability

Python code is publicly available at GitHub.

Funding

This research has no fund.

Declaration of competing interest

The authors have no conflict of interest.

Data availability

The authors do not have permission to share data.

References

CDC. (2025). Social vulnerability index 2018. https://data.cdc.gov/Health-Statistics/Social-Vulnerability-Index-2018-United-States-trac/4d8n-kk8a/about_data.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*, Article 177.

GitHub. (2025). Python programs for calculating associations between SVI and factors. https://github.com/s2322106/journal_paper.

Lenhof, K., Eckhart, L., Rolli, L. M., & Lenhof, H. P. (2024). Trust me if you can: A survey on reliability and interpretability of machine learning approaches for drug sensitivity prediction in cancer. *Briefings in Bioinformatics, 25*(5), Article bbae379. https://doi.org/10.1093/bib/bbae379

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Mandler, H., & Weigand, B. (2024). A review and benchmark of feature importance methods for neural networks. *ACM Computing Surveys, 56*(12), 318. https://doi.org/10.1145/3679012

McCarley, C. B., Blanchard, C. T., Nassel, A., Champion, M. L., Battarbee, A. N., & Subramaniam, A. (2024). The association between the social vulnerability index and adverse neonatal outcomes. *American Journal of Perinatology*. https://doi.org/10.1055/a-2419-8539 (Advance online publication. doi:10.1055/a-2419-8539).

Mehta, A., Jeon, W. J., & Nagaraj, G. (2024). Association of US county-level social vulnerability index with breast, colorectal, and lung cancer screening, incidence, and mortality rates across US counties. *Frontiers in Oncology, 14*, Article 1422475. https://doi.org/10.3389/fonc.2024.1422475

Mohanty, S., Ye, G., Sheets, C., Cossrow, N., Yu, K. C., White, M., … Gupta, V. (2024). Association between social vulnerability and *Streptococcus pneumoniae* antimicrobial resistance in US adults. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America, 79*(2), 305–311. https://doi.org/10.1093/cid/ciae138

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., et al. (2022). *General pitfalls of model-agnostic interpretation methods for machine learning models*. Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_4

Parr, T., Hamrick, J., & Wilson, J. D. (2024). Nonparametric feature impact and importance. *Information Sciences, 653*, Article 119563. https://doi.org/10.1016/j.ins.2023.119563

Potharlanka, J. L., & Bhat, M., .N. (2024). Feature importance feedback with Deep Q process in ensemble-based metaheuristic feature selection algorithms. *Scientific Reports, 14*(1), Article 2923. https://doi.org/10.1038/s41598-024-53141-w

Watson, D. S., & Wright, M. N. (2021). Testing conditional independence in supervised learning algorithms. *Machine Learning, 110*(8), 2107–2129. https://doi.org/10.1007/s10994-021-06030-6

Wood, D., Papamarkou, T., Benatan, M., et al. (2024). Model-agnostic variable importance for predictive uncertainty: An entropy-based approach. *Data Mining and Knowledge Discovery, 38*, 4184–4216. https://doi.org/10.1007/s10618-024-01070-7