

PEER REVIEW REPORT

The manuscript titled “*Using Machine Learning Clustering to Build an ESG-Based Taxonomy of Heat Vulnerability*” presents an unsupervised machine-learning framework to construct a cross-national taxonomy of heat vulnerability by integrating the novel indicator HI35 — the average annual days with an apparent temperature exceeding 35°C — with Environmental, Social, and Governance (ESG) variables drawn from the World Bank database. The study covers 193 countries and applies six clustering algorithms (Density-Based, Fuzzy C-Means, Hierarchical, Model-Based, K-Means, and Random Forest) across three ESG pillars (E, S, G) as well as a combined Principal Component Analysis (PCA) and hierarchical Ward linkage strategy, ultimately producing 20 national vulnerability clusters. A central conceptual contribution is the authors' explicit rejection of the standard assumption of local exogeneity for climatic variables, arguing instead that HI35 should be reinterpreted as an endogenous, diagnostically rich indicator of systemic ESG fragility.

However, there are several questions that require resolution. Below, the authors find a list of suggestions and comments per section.

1. The introduction is generally well-structured and provides a broad intellectual context for the research. The authors make a credible case for repositioning HI35 beyond its meteorological origin and for challenging local exogeneity assumptions. The literature review is extensive, touching on corporate ESG performance under heat stress, social vulnerability, governance accountability, digital ESG platforms, and sectoral dynamics. However, several concerns should be addressed:

- The dual labelling of sub-sections — for instance, 'Relaxing the common assumption of local exogeneity' appears twice (as a sub-heading in the Introduction and again as a standalone sub-section) — creates redundancy. The authors should consolidate or clearly differentiate these passages.
- Self-citation density is notably high. Works by Costantiello & Leogrande (2023a, 2023b), Drago et al. (2025a, 2025b), and Leogrande & Costantiello (2023) appear repeatedly across both the introduction and analytical sections. While not disqualifying, the editors may wish to flag potential over-reliance on authors' own prior output.
- Several references in Table 1 (Literature Synthesis) cite works in the future tense (e.g., 'Drago et al., 2025a'; 'Yu et al., 2025c') without confirmed publication details. These should be verified and updated with full bibliographic information.
- The discussion of 'Emerging Technologies and Digital ESG' (Section 2) reads as tangential to the core empirical contribution. Unless computational or digital ESG infrastructure is meaningfully connected to the clustering results, this sub-section should be condensed or removed.

2. The conceptual innovation of reframing HI35 as an endogenous, systemic vulnerability indicator is intellectually stimulating and well-articulated. However, the empirical design creates a significant tension with the conceptual claims:

- The authors state explicitly that their framework is 'deliberately descriptive and associative' and that 'no assertion of pinpointing causality' is made. This is methodologically appropriate for unsupervised clustering, but it sits in tension with the repeated assertion in the introduction and conclusions that HI35 'results from' ESG conditions or that countries' ESG setups 'contribute to its formation'. The language of causality recurs throughout the manuscript while the methods are expressly non-causal. This inconsistency must be resolved — either the causal language should be removed throughout, or a qualifying caveat should be added wherever directional relationships are implied.

- The decision to treat HI35 as the 'dependent variable' in some sub-sections (Section 4) contradicts the unsupervised clustering framework in which no variable is designated as dependent or independent. The authors should clarify whether they are running any parametric regressions or whether all analyses are purely cluster-based.
- The justification for using country-level ESG indicators to characterize vulnerability (rather than, for example, sub-national or population-weighted data) is not explicitly defended. Large federal states (China, USA, India, Brazil) are treated as single observations, which aggregates over enormous intra-national heterogeneity.

3. The data section (Section 3 and Table 2) provides a clear inventory of all variables. The sourcing from the World Bank ESG Portal is appropriate and publicly reproducible. Nonetheless, the following issues need attention:

- Temporal scope is underspecified in the main body. The authors aggregate observations by country across the entire available panel, which spans up to 64 years for some indicators. How aggregation is performed (mean, median, last available year?) is not stated in Section 3 or 4; it only appears obliquely in Section 6. This must be clarified explicitly in the methods section.
- The Under-5 Mortality Rate (MR) variable in the social cluster analysis records a negative minimum value of -55.61 per 1,000 live births, which is physically impossible. The authors acknowledge this anomaly in Appendix B but do not address it analytically. A data cleaning protocol should be described, and observations with impossible values should be excluded or imputed before analysis.
- Missing data are handled using the missForest algorithm (Section 6), which is a reasonable choice. However, the extent of missingness — notably for EIN (2,388 missing of 4,623 total observations, >50%), RDE (2,580 missing), and PUSM (1,839 missing) — is substantial enough to warrant sensitivity analysis. Results derived from imputed majority-missing variables may be unstable.
- The rationale for including Cooling Degree Days (CDD) as an 'environmental' ESG variable rather than as a climate exposure variable is not discussed. CDD is closely correlated with HI35 by construction, introducing potential collinearity in the environmental clustering step.

4. The methodological framework is ambitious and multi-layered. The use of six competing clustering algorithms with a multi-criteria selection procedure (Section 5) is a genuine strength and represents a more rigorous approach than single-algorithm studies. However, there are important weaknesses:

- Algorithm selection across the three ESG pillars is inconsistent: Model-Based clustering is selected for both the Environmental and Governance dimensions, whereas Random Forest proximity clustering is selected for the Social dimension. The decision rule for this asymmetry is not made explicit. The authors state that structural balance (cluster size distribution) was used as a tiebreaker, but they do not provide a pre-registered or reproducible decision protocol. This ex post selection risks circularity.
- The PCA step in Section 6 is incompletely described. The number of principal components retained, the proportion of variance explained, loadings for each component, and the criterion used for component retention (e.g., eigenvalue > 1, scree plot, parallel analysis) are not reported. This information is essential for reproducibility and for interpreting the final cluster assignments.
- The choice of 20 clusters in the hierarchical taxonomy (Section 6) is determined via the gap statistic, but the gap statistic results themselves are not shown. A figure displaying the gap statistic as a function of the number of clusters should be included, along with the rationale for selecting $k = 20$ over other local optima.

- The dendrogram from the hierarchical clustering (Ward's method) is not presented. Its inclusion would substantially aid interpretation of the cluster structure and provide evidence of separation between groups.
- The description 'pure, unintelligent machine learning methodology' in the abstract is imprecise and potentially misleading. The authors presumably mean 'unsupervised' (as opposed to 'supervised'). The term 'unintelligent' has no standard methodological meaning and should be replaced.
- For each clustering analysis, standard validation measures (silhouette plots, within-cluster sum of squares curves) should be reported visually, not only as tabular normalized scores.

5. The results are presented in considerable detail across Sections 5–7. The cluster mean tables (Tables 6, 10, 14) are informative. However, the narrative interpretation of results requires improvement:

- Cluster 17 (Section 6) groups China and the United States together, described as reflecting 'large economies with comparable exposure and adaptive capacity'. This is an unusual finding that warrants deeper scrutiny. The two countries differ dramatically across almost every ESG dimension. The authors should investigate whether this co-clustering is robust across different numbers of clusters or is an artefact of the Ward/PCA pipeline.
- The discussion of Cluster 6 — which records the highest HI35 with strong regulatory quality but weak accountability and legal rights — would benefit from naming which countries are included in this cluster. Azerbaijan alone is listed for Cluster 6 in Table 15, which raises questions about whether any substantive cross-national generalizations can be drawn from a single-country cluster.
- Several single-country clusters (Azerbaijan, Bahrain, Barbados, Belgium, Belize, Benin, Cambodia, China + USA) indicate that the 20-cluster solution may over-partition the data. The authors should discuss whether a more parsimonious solution (e.g., $k = 10\text{--}12$) would yield substantively similar taxonomy patterns with greater statistical stability.
- Results are interpreted largely in terms of the cluster means of standardized variables. Unstandardized or back-transformed values (e.g., actual average HI35 days, actual mortality rates) would help readers assess real-world significance.
- The taxonomy table (Table 16) is a useful synthesis, but the 'Effect on HI35' column uses signs ('+', '-', 'mixed') without confidence intervals or test statistics. In an associative, non-causal framework, this notation risks being interpreted as directional causal claims.

6. Regarding figures and tables, the authors must address the following questions:

- Figures A1–A5 and B1–B5 and C1–C5 are all placed in appendices and are described only in text within the appendix sections. Key distributional plots (e.g., Figure A1) and the scatterplot matrices should be referenced directly in the main analysis sections where they are most relevant.
- Table 15 (country-cluster assignments) contains a formatting error: Cuba appears in both Cluster 4 and Cluster 13. This duplication must be corrected.
- Several figures (A2, B2, C2, C5) are dense scatterplot matrices that are difficult to read at journal resolution. The authors should either replace these with selective pair-plot panels or increase figure resolution and label clarity.
- Appendix D (national HI35 time series) is valuable but its link to the main analytical narrative is weak. The authors mention that it 'complements the main analytical outline' but do not integrate the temporal trends into any clustering or inference.

The English requires substantial revision. While the core ideas are discernible, the prose is often convoluted, repetitive, and at times ungrammatical. Specific issues include:

- Extremely long and syntactically complex sentences that impede comprehension (e.g., the paragraph beginning 'Such analysis explicitly diverges...' in Section 4).
- Repetitive phrasing across sections — for example, the argument about relaxing local exogeneity is restated at least four times across Sections 1, 2, 4, and 9.
- Idiosyncratic word choices and apparent machine-translation artifacts: e.g., 'hedonic stress' (line 1003), 'extremurry' (line 136), 'favorite son metrics' (line 1082), 'decrement of GNI' used ambiguously.
- Several section titles are grammatically non-parallel or redundant (compare 'Relaxing the common assumption of local exogeneity' in Section 2 with 'Relaxig the assumption of local exogeneity' [sic] in Section 4).

The authors should engage a professional academic English editing service before resubmission.

The references are generally appropriate and relevant to the research topic. However, several forward-dated references (2025) lack confirmed publication details and should be verified. Moreover, four of the five listed authors contribute co-authored works that are cited ten or more times in the manuscript. While the cited works may be relevant, the density of self-citation — particularly given that several cited papers appear to be preprints or working papers rather than peer-reviewed publications — warrants editorial scrutiny.

In conclusion, with respect to the necessary adjustment to be made, as suggested above, the overall recommendation is to reconsider after a major revision.