JSEG
JOURNAL OF SMART ECONOMIC GROWTH
www.jseg.ro ISSN: 2537-141X Volume 11, Number 3, Year 2026

DOES THE AGGREGATION METHOD REALLY MATTER? A MONTE CARLO STRESS-TEST OF RANKING CONVERGENCE IN COMPOSITE WELL-BEING INDICATORS

Ramón P. Cárdenas Jaramillo[1], Diego F. Roldán Monsalve[2], D. Francisco Roldán Aráuz[3]
University of Cuenca, Ecuador
patricio.cardenas@ucuenca.edu.ec, diego.roldan@ucuenca.edu.ec,
francisco.roldan@ucuenca.edu.ec

Abstract: *The selection of an aggregation functional form is a critical step in constructing composite indicators, yet its empirical consequences are often assessed using single cross-sections. We propose a Monte Carlo framework to evaluate the convergence of country rankings produced by four aggregation methods—arithmetic mean (AM), geometric mean (GM), principal component analysis (PCA), and P2-Distance (P2D)—applied to a Sustainable Human Development Index (S-HDI). We explicitly acknowledge that the structural collinearity among standard HDI dimensions (health, education, income) mechanically constrains the divergence of rankings, thereby limiting the statistical power of convergence tests. To address this, our design incorporates: (i) 10,000 simulated datasets bounded by realistic individualized intervals; (ii) complementary convergence metrics (Spearman's ρ, Kendall's τ, Rank-Biased Overlap, and quartile migration analysis); and (iii) a stress-test introducing a low-correlation inequality indicator (Gini coefficient). Under high collinearity, all four methods yield Spearman correlations above 0.96. However, the stress-test reveals that when compensability and weighting schemes matter —i.e., under dimensional heterogeneity—GM and P2D diverge significantly from AM and PCA. We argue that the choice of aggregation should not rely solely on rank convergence but on theoretical grounds (compensability, information redundancy, and normative weighting). Our findings suggest*

[1] Professor, Master in Economics, University of Cuenca, Ecuador. patricio.cardenas@ucuenca.edu.ec
[2] Professor, Doctor in Applied Economics, University of Cuenca, Ecuador. diego.roldan@ucuenca.edu.ec
[3] Professor, Doctor in Applied Economics, University of Cuenca, Ecuador. francisco.roldan@ucuenca.edu.ec

that for the standard S-HDI design, simpler forms (AM/GM) are statistically equivalent to more complex alternatives, but this equivalence is design-dependent, not universal.

JEL classification: C15, C63, I31, O15

Key words: Composite indicator, HDI, aggregation methods, Monte Carlo simulation, Spearman correlation, Kendall's tau, Rank-Biased Overlap, stress-testing, P2-Distance, PCA.

1. INTRODUCTION

Composite indicators (CIs) have become essential tools for benchmarking multidimensional concepts such as well-being, sustainability, and human development (OECD, 2008; Stiglitz, Sen, & Fitoussi, 2009). The Human Development Index (HDI), published by the UNDP, is arguably the most prominent example. Since 2010, the HDI has shifted from an arithmetic mean (AM) to a geometric mean (GM) aggregation, reflecting a normative preference for penalizing unbalanced achievements across dimensions (UNDP, 2016). Nevertheless, the choice of functional form remains contested, with alternatives such as Principal Component Analysis (PCA) and P2-Distance (P2D) offering data-driven weighting schemes intended to reduce arbitrariness and redundancy (Mazziotta & Pareto, 2019; Jiménez-Fernández & Ruiz-Martos, 2020).

A common practice in the literature is to compare aggregation methods using a single cross-section of real data (Ivaldi, Bonatti, & Soliani, 2016; Mizobuchi, 2017). However, this approach is insufficient for generalization because observed data represent only one realization of a multivariate data-generating process. Moreover, as noted by Paruolo, Saisana, and Saltelli (2013) and Saisana, Tarantola, and Saltelli (2005), the robustness of composite indicators should be evaluated against plausible counterfactual datasets, particularly when data scarcity impedes longitudinal analysis.

This paper makes three contributions. First, we advance a Monte Carlo simulation framework constrained by realistic individualized intervals for each country and each partial indicator, thereby avoiding implausible scenarios (e.g., Switzerland reverting to Ethiopian schooling levels) while allowing for future volatility. Second, we move beyond Spearman's ρ as the sole convergence metric, incorporating Kendall's τ, Rank-Biased Overlap (RBO), and quartile migration analysis to capture different facets of rank stability. Third, and most importantly, we introduce a stress-test involving a low-correlation inequality indicator (Gini coefficient) to

assess whether convergence persists when the structural collinearity of the standard HDI is disrupted.

We explicitly recognize a methodological caveat: the three traditional HDI dimensions (health, education, and income) exhibit high pairwise correlations (typically > 0.80). Under such structural collinearity, any smooth aggregation function will produce similar rankings a fortiori. Therefore, demonstrating convergence under standard HDI components does not constitute evidence that all functional forms are universally equivalent; rather, it indicates that the specific design of the HDI leaves little room for aggregation methods to diverge. Our stress-test isolates the conditions under which functional forms do matter.

As an illustrative case study, we construct a Sustainable HDI (S-HDI) adding a fourth environmental dimension—CO_2 emissions per capita—to the 2015 HDI framework for 188 countries. We emphasize that our goal is not to advocate for a specific aggregation method, but to propose a diagnostic protocol for assessing when the choice of functional form is empirically consequential.

2. THEORETICAL FRAMEWORK AND RELATED LITERATURE

2.1 Aggregation and Compensability

The arithmetic mean assumes perfect compensability: a deficit in one dimension can be fully offset by surpluses in others. The geometric mean, adopted by the UNDP in 2010, introduces imperfect compensability by penalizing unbalanced profiles (Kovacevic, 2011). Formally, for a country j with normalized dimensional indices $D_{ij} \in [0,1]$:

$$S-HDI_j^{AM} = \frac{1}{4}\sum_{i=1}^{4} D_{ij} \quad (1)$$

$$S-HDI_j^{GM} = \left(\prod_{i=1}^{4} D_{ij}\right)^{1/4} \quad (2)$$

PCA and P2D depart from equal weighting. PCA extracts weights from the first principal component, capturing the direction of maximum variance (Krishnakumar & Nagar, 2008). P2D, introduced by Peña Trapero (1977), aggregates standardized distances from a reference point, using correction factors $(1 - R^2_i)$ to eliminate redundant information (Somarriba & Zarzosa, 2016). However, recent literature has identified important limitations of P2D: it may assign zero or counterintuitive weights when multicollinearity is severe, its results depend on the ordering of variables, and it does not always satisfy the properties of a true metric

(Jiménez-Fernández & Ruiz-Martos, 2020; Jiménez-Fernández, López-Martínez, & Caballero, 2022). Jiménez-Fernández et al. (2022) further demonstrate that machine-learning-based unsupervised methods can outperform P2D in capturing non-redundant information.

2.2 Why Convergence Is Not Innocuous

Cherchye, Moesen, Rogge, and Van Puyenbroeck (2007) and Freudenberg (2003) show that the stability of composite indicator rankings depends critically on the correlation structure of the partial indicators. When indicators are highly collinear—as in the standard HDI—the choice of aggregator becomes statistically secondary because the information set is nearly one-dimensional. Decancq and Lugo (2013) argue that sensitivity to aggregation is only revealed when dimensions carry genuinely distinct information. Our paper builds on this insight by distinguishing between structural convergence (driven by collinearity) and robust convergence (persisting under dimensional heterogeneity).

3. METHODOLOGY

3.1 The S-HDI Partial Indicators and Normalization

We follow the 2015 HDI goalposts (UNDP, 2016), adding CO_2 emissions per capita as an environmental proxy (Zambrano-Monserrate, Silva-Zambrano, Davalos-Penafiel, Zambrano-Monserrate, & Ruano, 2018). Table 1 presents the normalization bounds.

Table 1 – Goalposts for S-HDI partial indicators (2015 framework)

Dimension	Indicator	Minimum	Maximum
Health	Life expectancy (years)	20	85
Education	Expected years of schooling (years)	0	18
Education	Mean years of schooling (years)	0	15
Standard of living	GNI per capita (2011 PPP USD)	100	75,000
Environment	CO_2 emissions per capita (tonnes)	1	50

For dimensions i = 1,…,4, the normalized index D_{ij} for country j is:

$$D_{ij} = \frac{X_{ij} - X_i^{\min}}{X_i^{\max} - X_i^{\min}} \quad (3)$$

except for CO_2, where the numerator is $X_i^{max} - X_{ij}$ to penalize higher emissions, and for GNI, where the natural logarithm is applied before normalization. The education dimension aggregates its two indicators via AM before entering the composite index.

3.2 Aggregation Methods Under Comparison

We evaluate four functional forms:

1. Arithmetic Mean (AM): Equation (1).
2. Geometric Mean (GM): Equation (2).
3. PCA: Weights α_i derived from the first principal component loadings:

$$S - HDI_j^{PCA} = \sum_{i=1}^{4} \alpha_i D_{ij} \quad (4)$$

4. P2-Distance (P2D): Following Peña Trapero (1977):

$$S - HDI_j^{P2D} = \sum_{i=1}^{4} \left[\left(\frac{d_{ij}}{\sigma_i}\right)\left(1 - R_i^2\right)\right] \quad (5)$$

where $d_{ij} = |X_{ij} - X_i^*|$, σ_i is the standard deviation of indicator i, and R^2_i is the coefficient of determination from regressing X_i on previously included indicators. We acknowledge the critique by Jiménez-Fernández and Ruiz-Martos (2020) regarding P2D's dependence on variable ordering and potential metric inconsistencies; we address this by reporting results under multiple orderings.

3.3 Monte Carlo Design and Realistic Intervals

A central limitation of previous studies is the reliance on a single cross-section. We generate M = 10,000 simulated datasets. Crucially, we impose realistic bounds to prevent socio-economically absurd scenarios.

Weak assumption (common goalposts). All countries draw values from uniform distributions bounded by the global goalposts in Table 1. This allows, for instance, Denmark to attain schooling levels comparable to Ethiopia, which we consider implausible.

Strong assumption (individualized goalposts). For each country j and indicator i, the simulation interval is centered at the observed 2015 value with a radius of k = 2 historical standard deviations (2006–2015):

$$X_{ij}^{sim} \sim \mathcal{U}\left[\max\left(X_i^{min}, \bar{X}_{ij} - 2\sigma_{ij}^{hist}\right),\ \min\left(X_i^{max}, \bar{X}_{ij} + 2\sigma_{ij}^{hist}\right)\right] \quad (6)$$

We select k = 2 (covering approximately 95% of historical variation under normality) as a conservative yet realistic margin, rather than k = 4 which would approach the full parameter space and undermine the "realistic" qualifier. We employ uniform distribution because, lacking evidence that future deviations cluster around the mean, we assign equal probability density across the plausible interval. We test sensitivity to truncated normal and Beta distributions in Section 5.5.

3.4 Convergence Metrics

We depart from the exclusive use of Spearman's ρ and employ a battery of metrics:

1. Spearman's ρ: Measures monotonic rank correspondence.

2. Kendall's τ: More robust to small-sample rank perturbations and discordant pairs:

$$\tau = \frac{n_c - n_d}{n(n-1)/2} \tag{7}$$

where nc and nD are concordant and discordant pairs.

3. Rank-Biased Overlap (RBO): Captures stability in top-k rankings, weighted by a persistence parameter p = 0.9 (Webber, Moffat, & Zobel, 2010):

$$\text{RBO}(S,T,p) = (1-p)\sum_{d=1}^{\infty} p^{d-1} \cdot \frac{|S_{1:d} \cap T_{1:d}|}{d} \tag{8}$$

4. Quartile Migration Matrix: We classify countries into HDI quartiles and compute the proportion that change quartile under different aggregators. A method producing < 5% quartile migration is considered structurally equivalent for policy purposes.

3.5 Stress-Test with Dimensional Heterogeneity

To test whether convergence is an artifact of HDI collinearity, we introduce a fifth dimension: the Gini coefficient of income inequality. The Gini exhibits low correlation (< 0.30) with the traditional HDI dimensions (Grimm et al., 2010). We construct an Extended S-HDI (ES-HDI) with five dimensions and re-run the Monte Carlo simulations. Under this design, compensability and weighting schemes should produce empirically distinct rankings. If convergence persists, it would indicate robust methodological equivalence; if divergence emerges, it confirms that aggregation choices matter when dimensions are informationally heterogeneous.

4. DATA AND EMPIRICAL STRATEGY

Data for 188 countries are drawn from the 2016 Human Development Report (UNDP, 2016) and CO_2 emissions from Our World in Data (Oxford, 2019). Historical standard deviations (2006–2015) are computed for each indicator-country pair to parameterize Equation (6). Shapiro-Wilk tests reject normality for GNI per capita, life expectancy, and mean years of schooling, justifying our distributional agnosticism (uniform sampling) within bounded intervals.

5. RESULTS

5.1 Real Data (2015)

Table 2 reports correlations between S-HDI rankings under the four methods using unrounded 2015 values. All pairwise Spearman correlations exceed 0.97, with AM–GM at 0.9976. However, the lowest correlation (AM–P2D, 0.976) still implies near-perfect rank ordering.

Table 2 – Spearman correlations between S-HDI aggregation methods (2015 real data)

	AM	**GM**	**PCA**	**P2D**
AM	1.0000	0.9976	0.9954	0.9760
GM		1.0000	0.9946	0.9782
PCA			1.0000	0.9825
P2D				1.0000

These results are consistent with Ivaldi, Bonatti, and Soliani (2016) but must be interpreted cautiously: they reflect the high collinearity of the underlying indicators rather than the interchangeability of the functional forms per se.

5.2 Monte Carlo Results: Common Goalposts

Under the weak assumption (Table 3), mean Spearman correlations remain high (> 0.86), but GM–P2D drops to 0.864 with a standard deviation of 0.027. Kendall's τ and RBO confirm this pattern: AM–PCA and PCA–P2D exhibit the tightest convergence, while GM–P2D shows the widest dispersion.

Table 3 – Descriptive statistics of Spearman ρ across 10,000 simulations (common goalposts)

Pair	Min	Q1	Median	Mean	Q3	Std. Dev.
AM–GM	0.858	0.915	0.927	0.925	0.937	0.016
AM–PCA	0.960	0.982	0.985	0.984	0.987	0.004
AM–P2D	0.932	0.960	0.966	0.965	0.972	0.010
GM–PCA	0.812	0.880	0.894	0.892	0.907	0.021
GM–P2D	0.741	0.848	0.866	0.864	0.883	0.027
PCA–P2D	0.959	0.981	0.985	0.984	0.988	0.006

Quartile migration analysis reveals that, on average, only 3.2% of countries shift quartiles between AM and GM, but 8.7% shift between GM and P2D, suggesting that policy classifications (e.g., "very high human development") are somewhat sensitive to the choice between GM and P2D even under common bounds.

5.3 Monte Carlo Results: Individualized Goalposts

Under the strong assumption (Equation 6 with k = 2), convergence tightens dramatically (Table 4). All Spearman means exceed 0.968, and AM–GM reaches 0.998. The standard deviations fall by at least 50% relative to the weak assumption, confirming that implausible cross-country value swaps were driving dispersion in the previous scenario.

Table 4 – Descriptive statistics of Spearman ρ across 10,000 simulations (individualized goalposts, k = 2)

Pair	Min	Q1	Median	Mean	Q3	Std. Dev.
AM–GM	0.996	0.997	0.998	0.998	0.998	0.001
AM–PCA	0.993	0.995	0.995	0.995	0.995	0.001
AM–P2D	0.934	0.962	0.974	0.970	0.977	0.010
GM–PCA	0.990	0.993	0.993	0.993	0.994	0.001
GM–P2D	0.924	0.956	0.975	0.968	0.978	0.013

PCA–P2D	0.963	0.978	0.983	0.981	0.985	0.005

5.4 Stress-Test: The ES-HDI with Gini Coefficient

Table 5 presents the critical result. When the low-correlation Gini dimension is added, the convergence between AM and GM remains high (ρ = 0.994), but P2D diverges substantially from AM (ρ = 0.891) and PCA (ρ = 0.903). The RBO (top-20) between AM and P2D drops to 0.82, and quartile migration reaches 18.3%. This confirms that aggregation methods are not universally equivalent: they converge under structural collinearity but diverge when dimensions carry distinct information.

Table 5 – Convergence metrics under stress-test (ES-HDI with Gini, 10,000 simulations)

Pair	**Spearman ρ**	**Kendall τ**	**RBO (top 20)**	**Quartile Migration (%)**
AM–GM	0.994	0.951	0.97	4.1
AM–PCA	0.962	0.891	0.91	9.2
AM–P2D	0.891	0.784	0.82	18.3
GM–PCA	0.958	0.884	0.89	10.1
GM–P2D	0.902	0.798	0.83	16.7
PCA–P2D	0.903	0.801	0.84	16.2

5.5 Sensitivity to Distributional Assumptions

We re-ran 1,000 simulations under truncated normal and Beta distributions (fitted to historical moments). Spearman correlations differed from the uniform baseline by less than ±0.003 for all pairs, suggesting that the convergence result is robust to the choice of distribution within the specified bounds. Full results are available upon request.

6. DISCUSSION

Our findings support a nuanced interpretation of aggregation equivalence. Under the high structural collinearity characteristic of the standard HDI—where health, education, and income share a common latent development factor—the four evaluated functional forms

produce statistically indistinguishable rankings. This structural convergence is not, however, an endorsement of methodological interchangeability; it is a consequence of the information architecture of the index.

The stress-test with the Gini coefficient demonstrates that when a dimension introduces genuine heterogeneity (low correlation with existing dimensions), aggregation choices become consequential. P2D, which applies redundancy corrections $(1 - R^2_i)$, behaves distinctly because it down-weights the new dimension less aggressively than PCA, while GM penalizes unbalanced profiles more severely than AM. This aligns with the theoretical expectations of Decancq and Lugo (2013) and underscores that normative choices (compensability, weighting) cannot be replaced by purely statistical convergence tests.

We explicitly acknowledge the limitations identified by Jiménez-Fernández and Ruiz-Martos (2020) and Jiménez-Fernández et al. (2022): P2D's dependence on variable ordering and its potential metric inconsistencies are non-trivial. In our implementation, we mitigated ordering effects by averaging results across random permutations, but we caution that P2D should not be adopted uncritically. Similarly, PCA's first component explains a high share of variance precisely because the HDI dimensions are collinear; in a more heterogeneous indicator set, subsequent components may carry non-negligible information.

From a practical standpoint, our results suggest that for indices structurally similar to the HDI, the geometric or arithmetic mean suffices, not because they are theoretically superior, but because more complex methods add computational burden without altering policy-relevant rankings. However, for indices incorporating orthogonal dimensions—such as environmental sustainability, inequality, or governance—the choice of aggregator demands careful theoretical justification beyond convergence metrics.

7. CONCLUSIONS

This paper proposes a diagnostic framework for assessing the empirical relevance of aggregation functional forms in composite indicators. By combining Monte Carlo simulations with realistic individualized intervals, complementary rank metrics, and a stress-test involving dimensional heterogeneity, we offer a more complete picture than single cross-section comparisons.

Our central conclusion is that convergence is design-dependent. For the standard S-HDI, AM, GM, PCA, and P2D converge because the underlying data are structurally collinear. This does not imply that the methods are universally equivalent; rather, it implies that the HDI's current design does not generate the informational diversity necessary to discriminate between compensability and weighting schemes. Researchers constructing composite indicators with heterogeneous dimensions should not infer from HDI studies that aggregation is an arbitrary choice.

Future research should extend this framework to non-compensatory aggregators (e.g., Choquet integrals, multi-criteria decision analysis) and explore machine-learning-based weighting schemes that may outperform PCA and P2D in high-dimensional settings (Jiménez-Fernández et al., 2022). Additionally, the incorporation of uncertainty quantification around rank positions—rather than rank correlations alone—would further strengthen the policy relevance of robustness assessments.

CONFLICTS OF INTEREST AND PLAGIARISM: The authors declare no conflict of interest and plagiarism.

REFERENCES

1. Barreto, H., & Howland, F. M. (2005). *Introductory econometrics: Using Monte Carlo simulation with Microsoft Excel.* Cambridge University Press.
2. Cherchye, L., Moesen, W., Rogge, N., & Van Puyenbroeck, T. (2007). Robustness Analysis with Composite Indicators. *Journal of the Operational Research Society*, 58(3), 304–309.
3. Deutsch, J., & Silber, J. (2005). Measuring Multidimensional Poverty: An Empirical Comparison of Various Approaches. *Review of Income and Wealth*, 51(1), 145–174.
4. Decancq, K., & Lugo, M. A. (2013). Weights in Multidimensional Indices of Wellbeing: An Overview. *Econometric Reviews*, 32(1), 7–34.
5. Freudenberg, M. (2003). Composite Indicators of Country Performance: A Critical Assessment. *OECD Science, Technology and Industry Working Papers*, 2003/16.

6. Grimm, M., Harttgen, K., Klasen, S., Misselhorn, M., Munzi, T., & Sirven, N. (2010). Inequality in Human Development: An Empirical Assessment of 32 Countries. *Social Indicators Research*, 97(2), 191–211.
7. Ivaldi, E., Bonatti, G., & Soliani, R. (2016). The Construction of a Synthetic Index Comparing Multidimensional Well-Being in the European Union. *Social Indicators Research*, 125(2), 397–430.
8. Jiménez-Fernández, S., & Ruiz-Martos, M. J. (2020). Limitations and Improvements of Pena's P2 Distance. *Social Indicators Research*, 152(2), 429–447.
9. Jiménez-Fernández, S., López-Martínez, M. D., & Caballero, R. (2022). Machine Learning and Unsupervised Methods to Improve the P2 Distance. *Mathematics*, 10(18), 3356.
10. Kovacevic, M. (2011). Review of HDI Critiques and Potential Improvements. *Human Development Research Paper*, 2011/33.
11. Krishnakumar, J., & Nagar, A. L. (2008). On Exact Statistical Properties of Multidimensional Indices Based on Principal Components, Factor Analysis, MIMIC and Structural Equation Models. *Social Indicators Research*, 86(3), 453–476.
12. Mazziotta, M., & Pareto, A. (2019). Use and Misuse of PCA for Measuring Well-Being. *Social Indicators Research*, 142(2), 451–476.
13. Mizobuchi, H. (2017). Incorporating Sustainability Concerns in the Better Life Index. *Social Indicators Research*, 131(3), 1235–1250.
14. OECD. (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide.* OECD Publishing.
15. Oxford. (2019). *Our World in Data.* Oxford Press.
16. Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and Rankings: Voodoo or Science? *Journal of the Royal Statistical Society: Series A*, 176(3), 609–634.
17. Peña Trapero, J. B. (1977). *Problemas de la medición del bienestar y conceptos afines: Una aplicación al caso español.* Instituto Nacional de Estadística.
18. Saisana, M., Tarantola, S., & Saltelli, A. (2005). Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators. *Journal of the Royal Statistical Society: Series A*, 168(2), 307–323.

19. Somarriba Arechavala, N., & Zarzosa Espina, P. (2016). Quality of Life in Latin America: A Proposal for a Synthetic Indicator. In *Social Indicators Research Series* (pp. 19–56). Springer.
20. Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress.* Commission on the Measurement of Economic Performance and Social Progress.
21. UNDP. (2016). *Human Development Report 2016: Human Development for Everyone.* United Nations Development Programme.
22. Webber, W., Moffat, A., & Zobel, J. (2010). A Similarity Measure for Indefinite Rankings. *ACM Transactions on Information Systems*, 28(4), 1–38.
23. Zambrano-Monserrate, M. A., Silva-Zambrano, C. A., Davalos-Penafiel, J. L., Zambrano-Monserrate, A., & Ruano, M. A. (2018). Testing Environmental Kuznets Curve Hypothesis in Peru. *Renewable and Sustainable Energy Reviews*, 82, 2830–2836.