



How Consistent Are University Sustainability Rankings? Concordance, Consensus, and Sources of Divergence across Three Global Systems

Levent Çinko^{1*} 

¹ Department of Capital Markets, Faculty of Financial Sciences, Marmara University, Istanbul, Türkiye

ARTICLE INFO

Article history:

Received: 10 July 2026

Received in revised form: 20 August 2026

Accepted: 22 August 2026

Available online: 23 August 2026

Keywords:

University sustainability rankings; ranking concordance; ranking divergence; higher education sustainability; sustainability assessment

ABSTRACT

University sustainability rankings have become influential tools for comparing institutions whose environmental, social, governance, and public engagement activities are multidimensional and only partly observable. Their growing use creates an important methodological problem because different ranking systems may assign distinct meanings to sustainability, rely on different data infrastructures, and transform heterogeneous evidence into apparently comparable ordinal positions. This study investigates the extent to which the UI GreenMetric World University Rankings, the Times Higher Education Impact Rankings, and the QS Sustainability Rankings convey a common ordering signal, while also examining the sources and implications of their divergence. The analysis first harmonizes institutional identities and constructs a common comparison base. It then evaluates pairwise concordance and overall agreement, develops a cross-ranking consensus ordering under entropy, CRITIC, and equal weighting, assesses rank stability across the admissible weight space, identifies descriptive divergence profiles, and applies leakage-controlled predictive models to examine whether the components of one ranking explain divergence in another. The results indicate meaningful but incomplete agreement across the systems. The shared signal is strongest among institutions that perform consistently across several dimensions, whereas middle and lower positions are more sensitive to the weighting rule and the composition of the underlying indicators. The consensus analysis therefore offers a structured comparison rather than a claim that the rankings measure a single latent sustainability construct. The divergence profiles further show that institutions can reach similar overall positions through different combinations of environmental, social, governance, research, and engagement-related strengths. Predictive analyses provide limited evidence that the component scores of one systems explain out-of-sample divergence in another. The study concludes that sustainability rankings should be interpreted as complementary information systems rather than interchangeable measures, and that responsible comparison requires explicit attention to methodological design, uncertainty, and the distinct institutional dimensions represented by each ranking.

1. Introduction

University rankings have become influential information intermediaries in higher education. Students and families use them to screen institutions, governments use them to benchmark systems, and university managers use them to communicate institutional performance. By translating

* Corresponding author.

E-mail address: leventicinko@marmara.edu.tr

heterogeneous and partly unobservable attributes into portable ordinal signals, rankings make organizational status and reputation publicly comparable while reducing information asymmetries [1], [2]. Once these measures become consequential, institutions may also respond to the categories and incentives they create [3]. Sustainability rankings extend this informational and reputational role from academic prestige to environmental performance, social contribution, and institutional governance. Among the most visible global instruments are the UI GreenMetric World University Rankings, the Times Higher Education Impact Rankings, and the QS Sustainability Rankings. Although all three operate within a broad sustainability discourse, their participation patterns, indicator coverage, reporting requirements, and assessment logics differ [4], [5].

The central measurement question is whether these instruments can reasonably be treated as interchangeable signals. GreenMetric emphasizes campus operations and management, including energy, climate change, waste, water, transportation, and education and research. The Impact Rankings organize evidence around the Sustainable Development Goals and combines research, stewardship, outreach, and teaching. QS Sustainability gives greater weight to environmental and social impact together with governance and externally sourced indicators. These systems therefore differ in their indicator coverage and assessment logic, even when they address overlapping dimensions of institutional sustainability [4], [5]. Sustainability assessment in higher education is consequently a problem of construct pluralism as well as measurement. A university may be evaluated differently depending on whether the assessment emphasizes campus operations, SDG-linked activities, research, or broader institutional impacts [6], [7].

Existing evidence points to persistent divergence, but it does not yet provide a complete account of its magnitude or structure. Comparative studies have reported weak overlap between academic prestige and environmental performance, differences between GreenMetric and broader university rankings, and limited intersection between elite groups across sustainability ranking systems [7], [9]. Recent studies also indicate that participation and reporting are conditioned by institutional resources, reporting requirements, reputational incentives, and national context; some institutions may therefore be more likely to enter or become visible in these systems than others [10], [11]. Sustainability reporting can also involve selective disclosure or window-dressing risks [12]. These findings motivate a common-sample analysis that distinguishes agreement among the instruments from the consequences of their different measurement choices.

This study addresses that problem through three linked analyses. First, it evaluates common-set concordance among the 2024 GreenMetric, THE Impact, and QS Sustainability rankings using pairwise Spearman correlations and tie-corrected Kendall's W with bootstrap confidence intervals. Second, it constructs a cross-ranking consensus index from normalized rank scores under entropy, CRITIC, and equal weighting. The analysis evaluates uncertainty through cross-scheme comparisons, bootstrap re-estimation, a full Dirichlet weight-simplex sweep, and a tie-capable Plackett–Luce aggregation. Third, it examines the structure of disagreement in a two-dimensional deviation space and evaluates whether ranking components from the other systems have out-of-sample predictive association with divergence. The machine-learning analysis distinguishes a mechanically circular decomposition from a leakage-controlled predictive design.

The empirical questions are therefore as follows. RQ1 asks how strongly the three rankings agree on the universities they have in common. RQ2 asks whether their signals can be combined into a consensus ordering that remains stable under alternative weighting schemes, resampling, and probabilistic rank aggregation. RQ3 asks whether cross-ranking divergence forms stable descriptive profiles and whether components of one ranking provide out-of-sample predictive information about divergence in another ranking after mechanical overlap is removed. These questions connect the

governance literature on consequential rankings with the methodological literature on composite indicators, ordinal aggregation, and predictive validation [1], [3], [13], [15].

The study contributes to four related discussions. First, it advances the literature on ranking governance by showing that sustainability rankings are neither interchangeable signals nor unrelated instruments. Their moderate concordance and patterned divergence indicate that rankings create a plural information environment in which each system makes dimensions of institutional activity visible. This finding refines the information-intermediary perspective because the value of a ranking depends not only on the position it assigns, but also on the measurement architecture through which it selects institutional activities for observation and comparison [1], [3]. Second, it contributes to research on sustainability assessment by showing that cross-ranking synthesis can identify a relatively stable upper ordering without collapsing different measurement systems into a single substantive score. Aggregation can summarize common information, but it cannot resolve construct pluralism or remove uncertainty from the middle of the ordering [6], [10], [13]. Third, it extends comparative rankings research by providing common-set evidence for 113 universities and by identifying divergence profiles that are stable enough to be analytically useful but not sufficiently grounded to be interpreted as institutional types. Fourth, it translates these findings into an applied decision principle. Users should interpret the rankings jointly, report rank intervals or acceptability probabilities where available, and avoid treating a single league-table position as a complete measure of institutional sustainability. Ranking designers should disclose indicator coverage, participation rules, banded ranks, and aggregation procedures more explicitly. The contribution is therefore theoretical and practical as well as methodological, while remaining diagnostic rather than causal.

The results show significant but incomplete agreement. Kendall's W is 0.658, while the consensus ordering is highly similar across weighting schemes and stable at the top of the weight simplex. Three descriptive divergence profiles emerge with adequate bootstrap stability. The leakage-controlled models provide little predictive information for two focal rankings and only a modest, approximately linear association for the Impact target. Taken together, the findings suggest that the rankings share a common component but should not be read as interchangeable measures of one underlying institutional quality. The evidence supports cautious synthesis, especially for high-ranked institutions, while the middle of the ordering should be interpreted as an interval rather than a precise position.

The remainder of the article proceeds as follows. Section 2 reviews rankings as governance and information devices, the commensurability problem, comparative evidence on sustainability rankings, and composite-indicator methodology. Section 3 describes the data and analytical procedures. Section 4 presents the results. Section 5 discusses the implications and limitations, and Section 6 concludes.

2. Literature Review

2.1. Rankings as Information Intermediaries and Governance Instruments

Rankings reduce information costs not simply because they assign ordinal positions, but because they make otherwise heterogeneous and partly unobservable institutional attributes portable across audiences. This function is consequential in higher education, where students, employers, governments, and university managers must compare institutions whose missions, resources, and activities differ. [1] shows that rankings reorganize competition around selected notions of excellence, while [2] emphasize that rankings represent organizational status and reputation in publicly comparable positions and reduce information asymmetries. [3] further demonstrate that

once such measures become consequential, institutions can adapt their behavior to the categories being measured. Rankings are therefore not neutral observations of a pre-existing quality order. They are institutionalized representations built through choices about indicators, populations, weights, aggregation, and presentation. This perspective is central to the present study because the three sustainability systems may provide useful and partially overlapping signals while still embodying different measurement logics. Their moderate concordance is therefore substantively informative. It indicates a shared informational field, while their divergence reveals the limits of treating sustainability as a single commensurable attribute.

The informational role of rankings also has a governance dimension. [3] show that consequential measures can become objects of organizational response, while [2] emphasize that rankings distribute status and provide comparative representations of organizational attributes. Institutions may consequently adapt reporting practices, resource allocation, and strategic communication to the categories that a ranking makes salient. Such reactivity may draw attention to neglected activities, but it can also encourage selective disclosure and investment in indicators that are visible to the ranking system. Sustainability rankings are especially exposed to these tensions because participation is often voluntary and evidence is assembled from heterogeneous institutional reports and external data sources. The resulting visibility should therefore be interpreted as partly shaped by participation and reporting infrastructures rather than as a complete representation of institutional performance. Recent research identifies self-selection, reporting asymmetries, administrative burden, and reputational incentives as important conditions shaping participation in sustainability rankings [10], [12], [16].

The proliferation of sustainability rankings has consequently produced a multiple-signal environment. GreenMetric, THE Impact, and QS Sustainability share a broad concern with institutional sustainability, but they differ in what they observe and how they translate observations into scores. The question is not simply which ranking is correct. It is whether the signals can be interpreted together, which dimensions are shared, and where the measurement systems produce materially different orderings. This is a practical problem for users who may otherwise treat distinct rankings as interchangeable evidence of the same institutional attribute.

2.2. The Commensurability Problem and the Berlin Principles

The comparability of a ranking depends on the relation between its stated purpose, its indicators, its data sources, and the interpretation assigned to its output. The Berlin Principles emphasize transparency of purpose, methodological clarity, fitness for purpose, and caution in the interpretation of league-table positions [17]. These principles are relevant to sustainability rankings because sustainability is inherently multidimensional. Campus operations, research, education, community engagement, social inclusion, and governance may all be relevant, but they are not necessarily substitutable or equally observable. Different ranking systems can therefore produce different positions for legitimate substantive reasons. The empirical task is not to impose agreement as a criterion of validity, but to examine whether disagreement is consistent with distinct measurement logics or reveals instability within a purportedly common construct.

From this perspective, concordance is informative but not decisive. High agreement would indicate a common ordering signal, whereas moderate or low agreement would require closer examination of what each ranking makes visible and what it leaves outside the measurement frame. The present study therefore treats cross-ranking comparison as an external validity check on the interpretation of the ranking family, while avoiding the assumption that a single consensus can

resolve the multidimensionality of sustainability. This distinction motivates the separate reporting of common-set agreement, consensus-ordering uncertainty, and divergence profiles.

2.3. Comparative Studies of University Sustainability Rankings

Comparative research has examined sustainability rankings both against one another and against general academic rankings, but the evidence remains fragmented across pairwise comparisons, selected samples, and different reporting contexts. [7] showed that GreenMetric and the World University Rankings rely on substantially different indicator structures, but their analysis preceded the current three-ranking environment and did not quantify full common-set concordance. [8] found that academic ranking performance and environmental sustainability follow different geographic and institutional patterns. [19] show how intermediary organizations can facilitate the local adoption and implementation of global Sustainable Development Goal frameworks in higher education. [18] compare sustainability practices with academic performance across ranking systems and document a lack of uniformity in the resulting relationships; their analysis also illustrates why university-level controls are important when distinct ranking systems are related empirically. More recent studies have documented variation across sustainability rankings and weighting procedures, as well as persistent decoupling between sustainability participation and academic prestige [9], [20].

A related methodological literature treats sustainability rankings as multi-criteria decision problems. Studies using entropy, CRITIC, TOPSIS, and equal weighting show that alternative aggregation rules can change individual positions even when broad patterns remain similar [20], [23]. These studies provide direct evidence that weighting and normalization decisions can affect ranking positions, but they do not resolve whether the underlying ranking systems agree when they are applied to the same institutions. That distinction motivates the present focus on common-set concordance and cross-ranking synthesis.

Taken together, the reviewed literature supports three conclusions. First, sustainability rankings and general academic rankings capture different institutional hierarchies, and sustainability rankings themselves can produce different participation and performance patterns [7], [9]. Second, the systems differ in their indicator bases, purposes, reporting requirements, and assessment logics, so apparent agreement cannot be assumed to reflect a single underlying construct [4], [6]. Third, weighing and aggregation choices can affect the resulting order even when broad patterns remain similar [13], [15], [20],[22]. What remains less clear is how much the three leading sustainability systems agree across their full common distribution, whether a consensus ordering is stable under broad weighting uncertainty, and whether observed divergence can be described without confusing mechanical decomposition with predictive evidence.

This study addresses two related gaps. The first concerns measurement agreement. Previous studies have relied mainly on pairwise associations, small, selected samples, or overlapping at the top of the ranking. The second concerns synthesis and the structure of divergence. Existing work has not combined common-set concordance, uncertainty-aware consensus construction, probabilistic rank aggregation, stability-validated divergence profiles, and leakage-controlled prediction in one empirical design. The present analysis is intended to fill this methodological gap while remaining explicit that common-set results do not establish external validity for the full ranking populations.

3. Data and Methodology

3.1. Data

The analysis uses the 2024 editions of three international university sustainability rankings, namely the UI GreenMetric World University Rankings, the Times Higher Education Impact Rankings, and the QS Sustainability Rankings. Data were obtained from the providers' official results pages and archived 2024 files [24], [26]. Provider-reported coverage and usable analytical records differ. GreenMetric lists 1,477 participating institutions, and 1,475 unique records were available for analysis. Times Higher Education lists 2,152 universities across 125 countries and territories, and 1,963 records contain usable exact or banded ranks. QS Quacquarelli Symonds lists 1,403 universities across 95 countries, and 1,397 records were usable. These differences reflect record availability and duplicate removal.

The three systems cover overlapping but distinct dimensions of sustainability. UI GreenMetric emphasizes campus operations and sustainability management through setting and infrastructure, energy and climate change, waste, water, transportation, and education and research [26]. The Times Higher Education Impact Rankings assess university contributions to the United Nations Sustainable Development Goals through research, stewardship, outreach, and teaching [25]. QS Sustainability combines environmental impact, social impact, and governance, with governance introduced as a distinct category in the 2024 edition [24]. This distinction defines the scope of the concordance analysis, which compares related ranking signals across systems with different measurement domains [2], [6].

The study uses a cross-sectional design based on the 2024 editions. A common reference year places the three systems within the same measurement period. The QS governance lens appears in the 2024 edition, which limits comparability with earlier editions. Contemporaneous comparability therefore defines the temporal scope of the analysis.

UI GreenMetric and QS report exact ordinal positions. THE reports exact positions for the top 100 institutions and banded positions thereafter. For the 101–200 band, the analysis assigns the midpoint value of 150.5. Midpoint coding retains the order of the reported bands while reducing within-band precision. Sensitivity analysis uses the lower and upper bounds of the THE bands.

Table 1.

Data sources and common sample

Ranking	Edition	Coverage (N)	Rank format	Measurement focus
UI GreenMetric (GM)	2024	1,475	Exact ordinal ranks	Campus operations: setting & infrastructure, energy & climate, waste, water, transportation, education & research
THE Impact Rankings (TI)	2024	1,963	Exact ranks to 100; bands thereafter (midpoints used)	SDG performance: research, stewardship, outreach, teaching
QS Sustainability (QS)	2024	1,397	Exact ordinal ranks	Environmental impact, social impact, governance
Common sample	2024	113		Intersection after name normalization + fuzzy matching (cutoff 94)

Institutional matching proceeded in two stages. Names were normalized by lowercasing, removing diacritics and punctuation, and harmonizing country names and common institutional variants. Candidate pairs were evaluated with fuzzy string matching at a prespecified similarity cutoff of 94. Borderline cases were inspected manually. The threshold prioritizes precision because false-

positive matches can merge distinct institutions and create artificial disagreement. False-negative matches reduce the size of the common set. The procedure identified 113 universities present in all three systems, as summarized in Figure 1.

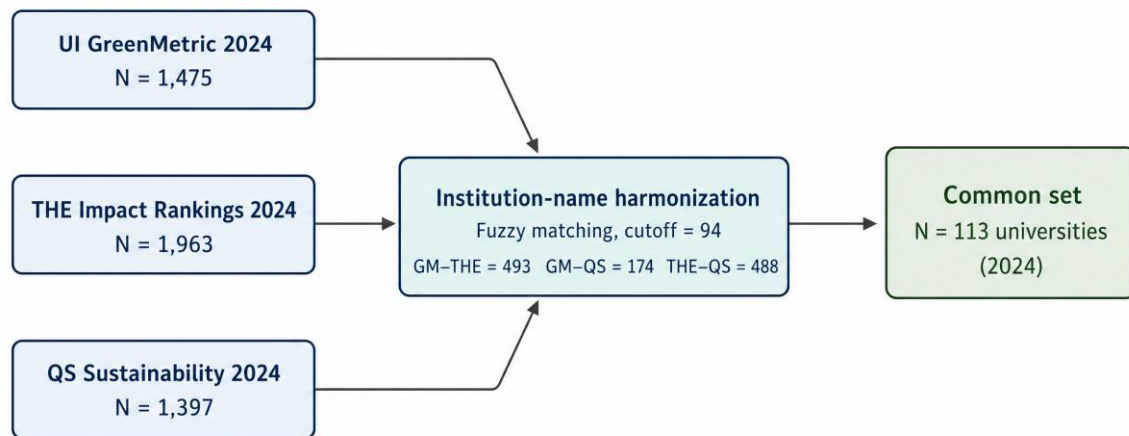


Fig. 1. Construction of the common university sample

The common sample spans Europe, Asia, Latin America, and Africa. It comprises universities that appear in all three systems. Relative to QS Sustainability universities outside the common set, it includes larger shares from the United Kingdom, Taiwan, Japan, and Australia. The noncommon population includes larger shares from the United States and China. The median QS position is 557 in the common set and 766 among universities outside it. The common set is therefore a selected, upward-skewed subset for ranking participants. Participation in sustainability rankings is associated with institutional resources, reporting capacity, and reputational incentives [10], [11].

3.2. Concordance Analysis

Pairwise agreement was measured with Spearman's rank correlation coefficient. Overall agreement was measured with Kendall's coefficient of concordance W . Kendall's W was calculated with a tie correction because THE Impact reports banded ranks beyond the top 100 and midpoint coding produces tied values. Because the common sample contained 113 universities, the chi-square approximation used 112 degrees of freedom. [27]

Uncertainty was quantified with 2,000 percentile-bootstrap resamples over universities. In each resample, the 113 common universities were redrawn with replacement, and all pairwise correlations and W were recomputed. Kendall's W was also recomputed under the lower and upper bounds of the THE bands. Its invariance to common-sample percentile, inverse-normal, and logit transformations was checked because W depends only on rank ordering.

3.3. Cross-ranking Consensus Index Construction and Robustness

The consensus index aggregates the three normalized ranking signals. Its construction follows the standard composite-indicator sequence of defining the measurement object, scaling the components, specifying the weights, aggregating the scores, and testing ordering sensitivity [13],

[14]. The measurement object is cross-ranking agreement. The index summarizes the joint ordering of the three systems while retaining their distinct substantive domains. Before entropy weighting, normalized rank scores were rescaled within the common sample to the interval [1e-6, 1]. This transformation is part of the weighting calculation and supports exact replication. Entropy, CRITIC, and equal weighting were then applied to the same transformed inputs.

Entropy weights were calculated from the dispersion of each normalized ranking score. After each score was converted to its share of the column total, entropy was defined as shown in Equation (1).

$$e_r = -k \sum_i p_{ir} \ln(p_{ir}), \quad k = \frac{1}{\ln(N)} \quad (1)$$

The diversification degree was calculated as shown in Equation (2).

$$d_r = 1 - e_r \quad (2)$$

The entropy weight was obtained from Equation (3). The resulting weights reflect within-sample dispersion in the three ranking signals.

$$w_r = \frac{d_r}{\sum_r d_r} \quad (3)$$

CRITIC weights were calculated using Equation (4), where σ_r is the standard deviation of ranking r and ρ_{rs} is its correlation with ranking s .

$$w_r \propto \sigma_r \sum_s (1 - \rho_{rs}) \quad (4)$$

The procedure combines contrast intensity with information that is not duplicated by the other rankings.

Equal weighting assigned one third to each ranking and provided a transparent benchmark.

The consensus score was calculated as a weighted sum of the three transformed ranking scores. Linear aggregation provides a reproducible summary of the joint ordering signal. Robustness was assessed in four ways. Orderings generated by entropy, CRITIC, and equal weights were compared. Entropy and CRITIC weights were re-estimated in 500 bootstrap samples, and each bootstrap ordering was correlated with the full-sample ordering. Next, 10,000 weight vectors were drawn from a uniform Dirichlet (1,1,1) distribution over the three-component simplex. The resulting rank distributions were summarized with median ranks, 95% rank intervals, and top-10 and top-20 acceptability probabilities. Finally, a tie-capable Plackett–Luce model was fitted to the three rankings, and its worth ordering was compared with the entropy-weighted ordering [28], [30]. These checks quantify sensitivity to aggregation, resampling, and alternative probabilistic ordering within the common sample. [14]

3.4. Divergence profiles in deviation space

Disagreement was represented by pairwise differences among normalized rank scores. The third difference is the linear function of the first two, as shown in Equation (5).

$$d_3 = n_{TI} - n_{QS} = d_2 - d_1 \quad (5)$$

The standardized coordinates were defined as shown in Equation (6).

$$d_1 = n_{GM} - n_{TI}, \quad d_2 = n_{GM} - n_{QS} \quad (6)$$

Universities were clustered in this standardized plane. K-means solutions for $k = 2$ through 6 were compared using average silhouette values and a 200-replication bootstrap adjusted Rand index. A three-cluster solution was selected because it produced the most stable partition and substantively distinct profiles. The clusters summarize observed disagreement patterns in the common sample.

3.5. Machine-learning Attribution of Divergence Through Mechanical Decomposition and Leakage-Controlled Prediction

Cross-ranking divergence was examined with Random Forest regressors implemented in scikit-learn. Random forests were used as a flexible predictive benchmark for nonlinear associations and interactions. The analysis separated model fitting from performance evaluation, handled mechanically overlapping variables, and assessed the stability of out-of-fold predictions. Two designs were estimated. The first provided a mechanical decomposition of the constructed ranking signal. The second estimated leakage-controlled predictive associations [31], [34].

The first design is a mechanical decomposition. The target is the standard deviation of the three normalized ranks. The predictors are the six GreenMetric dimensions and the three QS Sustainability lenses. THE Impact sub-scores were not used because they were not consistently available for the common sample and the relevant years. Because the target is calculated from ranking scores constructed from ranking components, the predictor–target relation contains an accounting component. Permutation importance locates the contribution of component variables within this construction.

The second design estimated leakage-controlled predictive associations. For each focal ranking r , the target was defined as shown in Equation (7).

$$d_r = n_r - \text{mean}(n_s, n_t) \quad (7)$$

All components of the focal ranking were excluded from the predictor set. Thus, d_{QS} was predicted from the six GreenMetric dimensions, d_{GM} from the three QS lenses, and d_{TI} from the combined GreenMetric and QS components. Predictive performance was evaluated with three repetitions of five-fold cross-validation and reported as mean out-of-fold R^2 with its standard deviation. Permutation importance was used to summarize variable contributions within each model. Predictor dependence can influence marginal importance estimates [35]. The resulting models estimate exploratory, cross-sectional associations among ranking components.

4. Results

4.1. Concordance Among the Rankings

Table 2 and Figure 2 show a consistent pattern of positive but uneven pairwise association. The strongest relationship is between THE Impact and QS Sustainability, followed by the relationship between GreenMetric and THE Impact. The GreenMetric and QS Sustainability rankings have the

weakest association. The three confidence intervals remain above zero, and the omnibus Kendall's W test indicates substantial agreement across the complete set of rankings. Thus, the common sample contains a shared ordering component, but the pairwise structure is not uniform across the three systems.

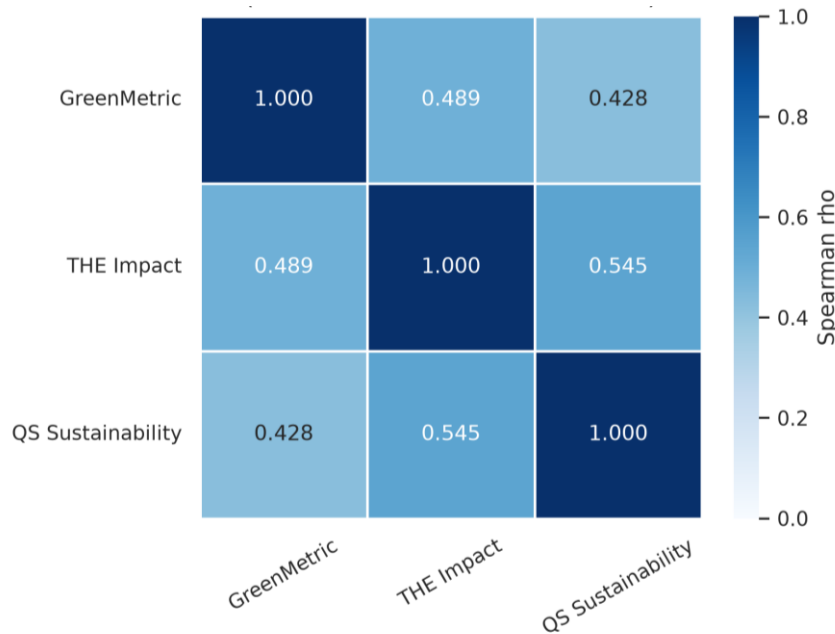


Fig. 2. Pairwise rank correlations among the sustainability rankings

Table 2.

Concordance estimates

Measure	Statistic	Value	95% CI
Spearman ρ : GreenMetric – THE Impact	ρ	0.489***	[0.312, 0.639]
Spearman ρ : GreenMetric – QS Sustainability	ρ	0.428***	[0.237, 0.585]
Spearman ρ : THE Impact – QS Sustainability	ρ	0.545***	[0.377, 0.676]
Kendall's W (tie-corrected)	W	0.658***	[0.561, 0.733]
Test of W	$\chi^2(112)$	221.1***	

Note. *** $p < 0.001$. Pairwise coefficients computed on ranks (THE Impact band midpoints); W computed across the three rankings. Confidence intervals are percentile intervals from 2,000 bootstrap resamples over universities.

The pairwise results therefore distinguish two features of the data. All three systems move in the same broad direction across the 113 universities, while the magnitude of association depends on the pair being compared. Figure 2 makes this asymmetry visible through the darker cell for the Impact and QS Sustainability pair and the lighter cell for the GreenMetric and QS Sustainability pair. Table 2 provides the corresponding uncertainty intervals and the tie-corrected omnibus estimate.

The concordance pattern was stable across the sensitivity checks. Alternative matching cutoffs and rank transformations preserved the positive ordering of the pairwise coefficients. Re-estimation under the lower and upper treatments of THE banded ranks also produced only small changes in the pairwise associations and Kendall's W. The bootstrap distribution remained centered near the full-sample estimates, indicating that the observed agreement was not generated by a small number of individual universities.

Taken together, the results establish a common-sample ordering with substantial but differentiated agreement. The Impact and QS Sustainability systems are the closest pair in the data, whereas the GreenMetric and QS Sustainability systems are the least closely aligned. These pairwise differences are retained in the consensus analysis rather than being averaged away at the outset.

4.2. Cross-ranking Consensus Index

Table 3 and Figure 3 compare the consensus ordering under entropy, CRITIC, and equal weighting. Entropy weighting assigns the largest share to QS Sustainability, with smaller shares assigned to THE Impact and GreenMetric. CRITIC weighting produces a more balanced allocation across the three systems, whereas equal weighting assigns one third to each. The resulting rankings remain close despite these different allocations. Cross-scheme Spearman correlations range from 0.926 to 0.996, indicating that the allocation rule changes the relative emphasis more than it changes the overall ordering.

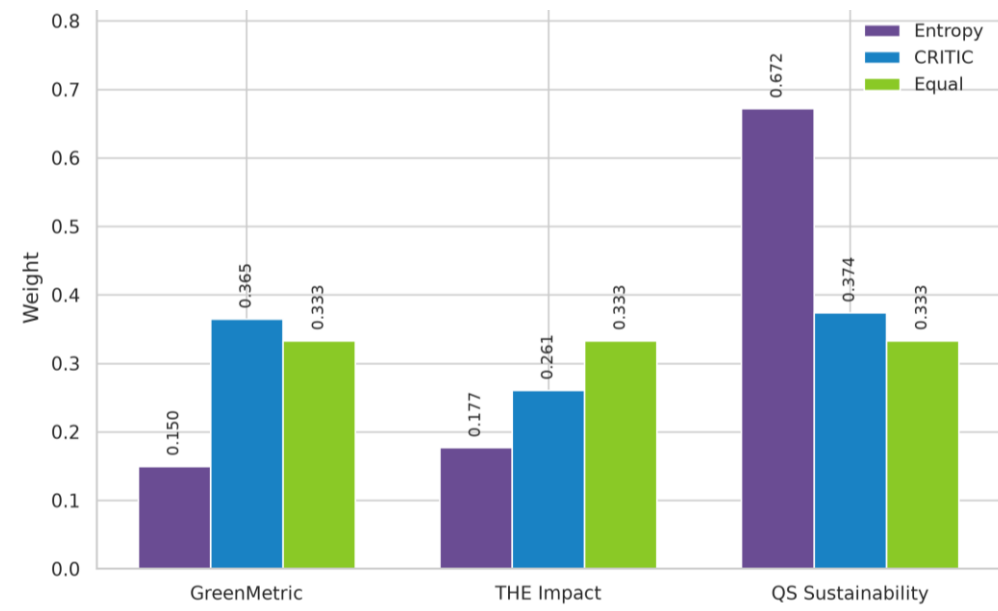


Fig. 3. Consensus rankings under alternative weights

Table 3.

Consensus weights and robustness

Component	Entropy	CRITIC	Equal
GreenMetric (GM)	0.150	0.365	0.333
THE Impact (TI)	0.177	0.261	0.333
QS Sustainability (QS)	0.672	0.374	0.333
Robustness			
Cross-scheme consensus rank correlations (Spearman ρ)	0.926–0.996***		
Bootstrap consensus rank stability (500 resamples, weights re-estimated)	0.998 [0.987, 1.000]		
Weight-simplex analysis (10,000 Dirichlet draws): median simplex rank vs. entropy rank	$\rho = 0.921$		
Plackett–Luce consensus rank vs. entropy-weighted consensus rank	$\rho = 0.883$		

Note. *** $p < 0.001$. Bootstrap entries report mean Spearman ρ between bootstrap and full-sample consensus rankings with 95% confidence interval (the 2.5th and 97.5th percentiles of the bootstrap distribution) in brackets.

The bootstrap analysis reinforces this result. When the weights were re-estimated in 500 resamples, the consensus ranking remained highly correlated with the full-sample ordering. Figure 3 shows the near-parallel movement of the three consensus rank profiles, while Table 3 reports the corresponding rank correlations and bootstrap interval. The similarity is therefore evident both in the point estimates and under resampling.

4.2.1. Weight-simplex stability

Figure 4 extends the comparison beyond the three prespecified weighting schemes by evaluating 10,000 weight vectors drawn from the three-component simplex. The median rank under these draws' correlates strongly with the entropy-weighted rank. Stability is concentrated among the leading institutions. Six universities enter the top ten with a probability of at least 0.90, and twelve enter the top twenty with that level of probability. Eighteen universities enter the top twenty with a probability of at least 0.50. By contrast, the mean width of the 95% rank interval is 32.5 positions across all 113 universities, showing that uncertainty is greater outside the most stable top group.

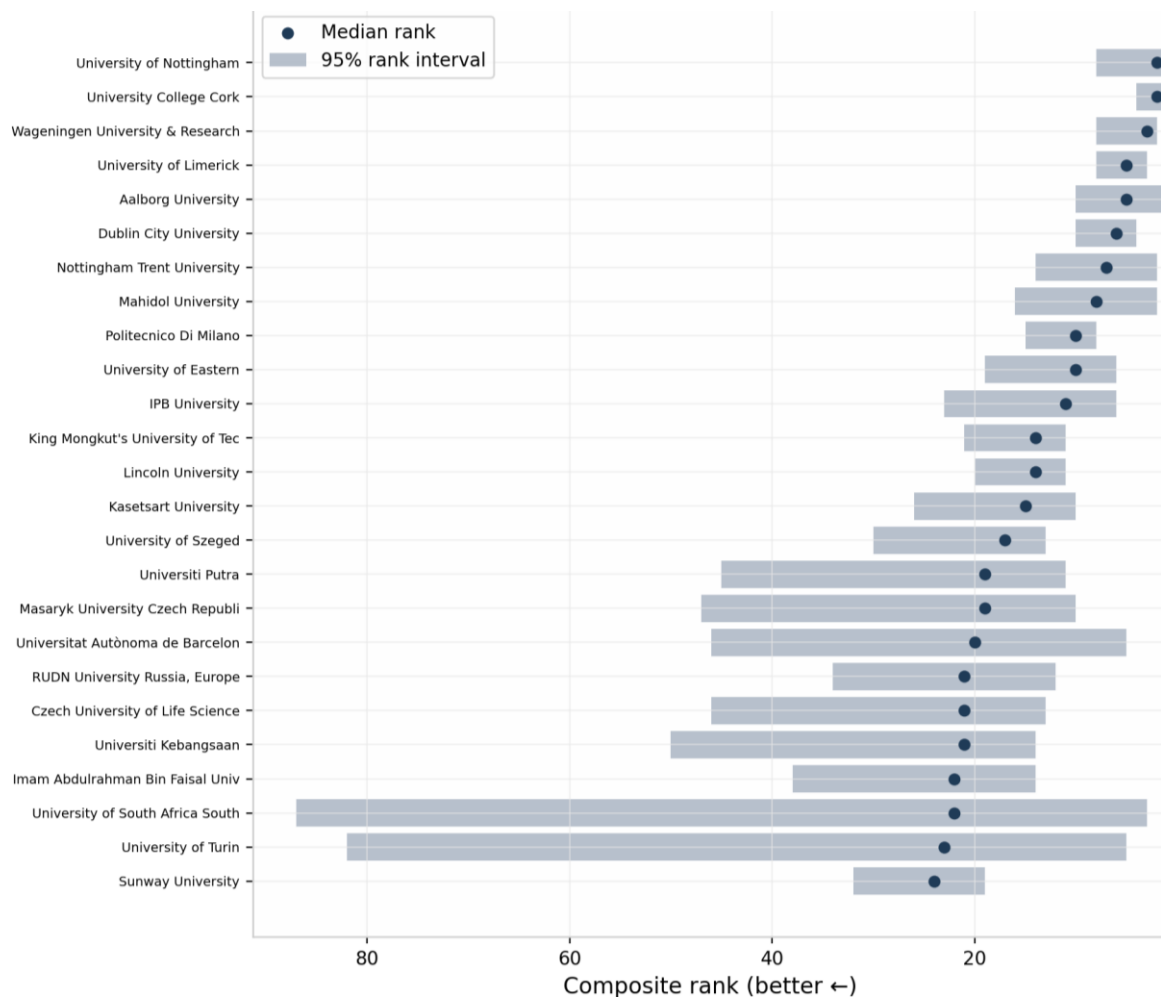


Fig. 4. Rank stability across the weight simplex

Table 4.

Weight-simplex stability

Quantity	Value
Spearman ρ : median simplex rank vs. entropy-weighted consensus rank	0.921
Universities in top 10 with probability ≥ 0.90	6
Universities in top 20 with probability ≥ 0.90	12
Universities in top 20 with probability ≥ 0.50	18
Mean width of 95% rank interval (all $N = 113$)	32.5 positions
Most stable top group ($p_{\text{top}20} = 1.000$)	University College Cork; Nottingham; Wageningen; Limerick; Aalborg; Dublin City

The weight-simplex results show a clear rank-position gradient. The universities with the highest consensus scores remain near the top across a wide range of admissible weights. Middle-ranked universities have wider rank intervals and lower top-position probabilities. University College Cork, the University of Nottingham, Wageningen University and Research, the University of Limerick, Aalborg University, and Dublin City University form the most stable top group in simplex analysis.

4.2.2. Probabilistic rank aggregation

The Plackett–Luce model produced a consensus ordering that was positively associated with the entropy-weighted ordering. The Spearman correlation between the two orderings was 0.883. The agreement is lower than the correlations among the three deterministic weighting schemes, but it remains sufficiently high to show that the leading features of the consensus ranking persist under a probabilistic aggregation model. Table 3 reports this comparison with the other robustness results.

Table 5 shows the composition of the leading part of the consensus ranking. The top ten institutions combine strong positions in at least two of the three systems, although their profiles differ. Nottingham and University College Cork rank highly in GreenMetric and QS Sustainability while occupying a THE Impact band midpoint below the leading exact positions. Aalborg has the strongest THE Impact position among the top six but a lower QS Sustainability position. The University of Eastern Finland, Universitat Autònoma de Barcelona, Politecnico di Milano, and Nottingham Trent University enter the top ten through different combinations of system-specific positions. The table therefore illustrates how the consensus score combines nonidentical ranking profiles into one ordering.

Table 5.

Top universities by consensus rank

Consensus rank	University	GM rank	TI rank	QS rank	Consensus score
1	University of Nottingham	8	150.5	33	0.970
2	University College Cork	4	67.0	88	0.952
3	Wageningen University & Research	1	150.5	75	0.951
4	Dublin City University	11	150.5	143	0.917
5	University of Limerick	19	61.0	175	0.909
6	Aalborg University	42	4.0	183	0.908
7	University of Eastern Finland	33	250.5	164	0.896
8	Universitat Autònoma de Barcelona	162	500.5	114	0.8840
9	Politecnico di Milano	66	150.5	201	0.8835
10	Nottingham Trent University	2	61.0	246	0.876

Note: Consensus score = entropy-weighted sum of normalized rank scores (higher is better). TI banded ranks shown at band midpoints.

4.3. Divergence Profiles in Deviation Space

Figure 5 displays the three clusters in the two-dimensional deviation space, and Table 6 reports the model-selection results. The average silhouette increases from 0.388 for two clusters to 0.405 for three clusters and 0.407 for four clusters. The three-cluster solution has the highest bootstrap adjusted Rand index among the stable candidate solutions, with a mean ARI of 0.853 and a 95% interval from 0.658 to 1.000. The four-cluster solution has a lower mean ARI of 0.745. The three-cluster partition was therefore retained for the profile description because it combines strong resampling stability with a compact representation of the observed deviations.

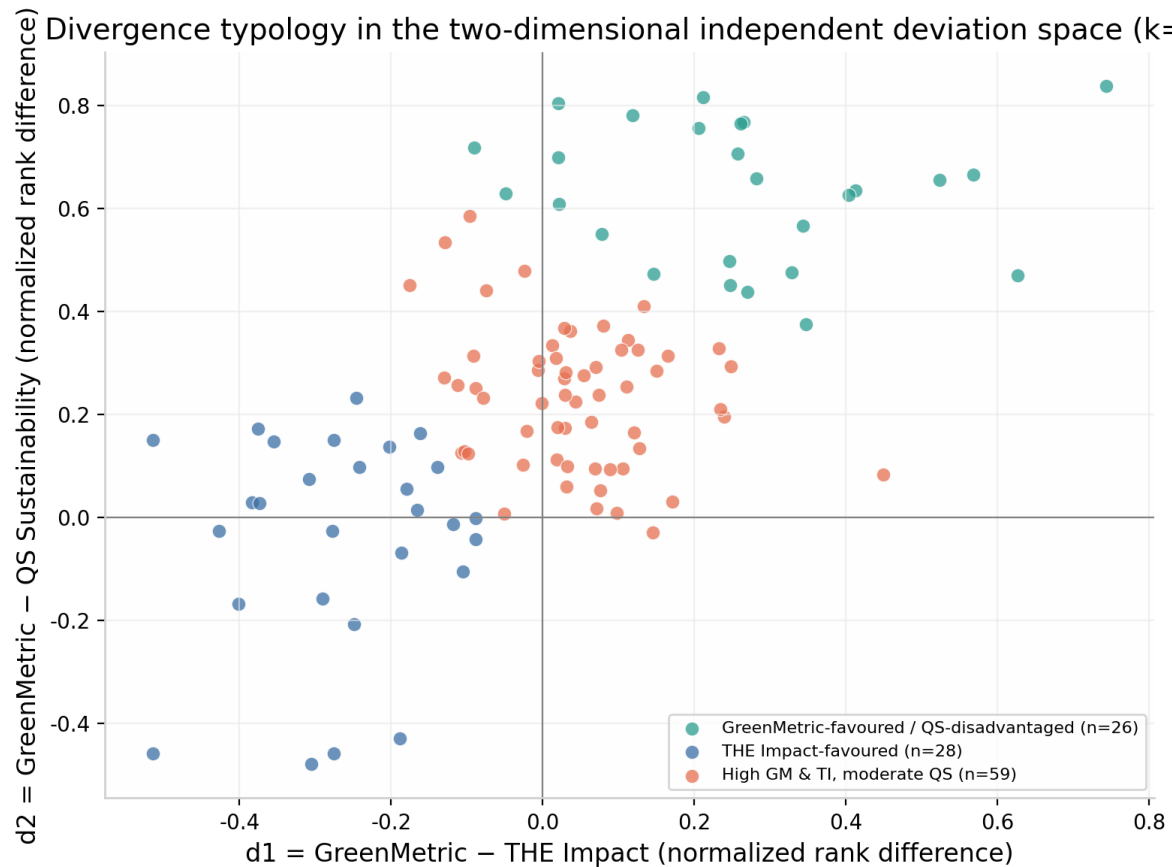


Fig. 5. Divergence profiles in the deviation space

Table 6.

Cluster model selection

k	2	3	4	5	6
Average silhouette	0.388	0.405	0.407	0.400	0.364
Bootstrap ARI (mean)	0.608	0.853	0.745		
Bootstrap ARI 95% interval	[0.148, 1.000]	[0.658, 1.000]	[0.407, 1.000]		

Note: ARI = adjusted Rand index between each bootstrap partition and the full-sample partition; intervals are 2.5th–97.5th percentiles over 200 resamples. Bootstrap ARI was computed for $k = 2-4$, the interpretable range.

Table 7.

Divergence profile characteristics

Cluster	Label	N	Mean GM	Mean TI	Mean QS	GM-TI	GM-QS	TI-QS
0	GreenMetric-favoured / QS-disadvantaged	26	0.865	0.603	0.233	+0.262	+0.632	+0.370
1	THE Impact-favoured	28	0.453	0.719	0.492	-0.266	-0.039	+0.227
2	High GM & TI, moderate QS	59	0.794	0.749	0.562	+0.045	+0.232	+0.187

Note: GM, TI, QS = normalized rank scores in [0, 1] (higher is better); deviations are differences between normalized scores. Deviations are computed from cluster means.

Table 7 describes the three profiles in normalized score space. Cluster 0 contains 26 universities and has the highest mean GreenMetric score, 0.865, but the lowest mean QS Sustainability score, 0.233. Its GreenMetric to QS deviation is 0.632, the largest gap in the table. Cluster 1 contains 28 universities and has the highest means THE Impact score, 0.719. It's THE Impact score exceeds its GreenMetric score by 0.266 and its QS Sustainability score by 0.227. Cluster 2 is the largest group, with 59 universities. It combines high mean GreenMetric and THE Impact scores, 0.794 and 0.749, with a lower but still positive mean QS Sustainability score of 0.562. The corresponding deviations are smaller than those observed in Cluster 0.

The cluster profiles are asymmetric across the three ranking pairs. The largest separation is the GreenMetric to QS Sustainability gap in Cluster 0. Cluster 1 is distinguished primarily by its THE Impact position, whereas Cluster 2 represents the largest group and has the smallest within-profile separation between GreenMetric and THE Impact. No cluster of comparable size is centered on a QS Sustainability advantage over both other systems. Figure 5 shows these differences as distinct spatial concentrations rather than as a single continuous cloud.

The clustering results organize the common sample into three empirically distinct divergence profiles. Their different sizes, centroid locations, and resampling stability indicate that disagreement varies systematically across universities in the common set. The next analysis evaluates whether the underlying component scores can predict these cross-ranking deviations out of sample.

4.4. Predictive Attribution of Ranking Divergence

The predictive analysis separates a mechanical decomposition from a leakage-controlled specification. The naive model uses components that enter the constructed target and therefore describes the internal composition of the ranking signal. The leakage-controlled models exclude the focal ranking's own components and evaluate prediction using components from the other ranking systems. Table 8 presents the mechanical permutation importance, whereas Table 9 and Figure 6 present the out-of-fold predictive results.

4.4.1. Mechanical attribution

The naive Random Forest achieved a five-fold cross-validated R^2 of 0.40. Its permutation-importance profile is highly concentrated. QS Environmental Impact has an importance of 0.959, substantially larger than QS Social Impact at 0.229 and GreenMetric Transportation at 0.210. The remaining six predictors have importances below 0.060. Table 8 therefore shows that the naive model is dominated by a small number of components that are mechanically close to the constructed target.

Table 8.

Mechanical feature attribution

Rank	Feature	Importance	SD
1	QS: Environmental Impact	0.959	0.111
2	QS: Social Impact	0.229	0.023
3	GM: Transportation	0.210	0.032
4	GM: Setting & Infrastructure	0.059	0.008
5	QS: Governance	0.040	0.006
6	GM: Energy & Climate Change	0.035	0.003
7	GM: Education & Research	0.031	0.006
8	GM: Waste	0.030	0.006
9	GM: Water	0.026	0.003

Note: Importance = mean decrease in predictive performance after permuting the feature, averaged over repeated permutations.

The importance of ranking in Table 8 is sharply ordered. QS Environmental Impact is the leading feature by a wide margin. QS Social Impact and GreenMetric Transportation form a second tier, while GreenMetric Setting and Infrastructure, QS Governance, and the remaining GreenMetric dimensions have much smaller importances. The standard deviations are also small relative to the leading importance, indicating that the ordering of the principal features is consistent across the repeated permutations.

4.4.2. Leakage-controlled attribution

The leakage-controlled results in Table 9 and Figure 6 differ across the three focal targets. The out-of-fold R^2 is close to zero for QS divergence and slightly negative for GreenMetric divergence. THE Impact divergence has a positive mean out-of-fold R^2 of 0.131 with a standard deviation of 0.028. Within that model, Waste has the largest permutation importance, followed by Energy and Climate Change. The QS-divergence model gives its largest within-model importance to Education and Research, Waste, and Transportation, but its overall predictive performance remains near zero. The GreenMetric-divergence model assigns the largest importances to Social Impact, Environmental Impact, and Governance, while its out-of-fold performance is also near zero.

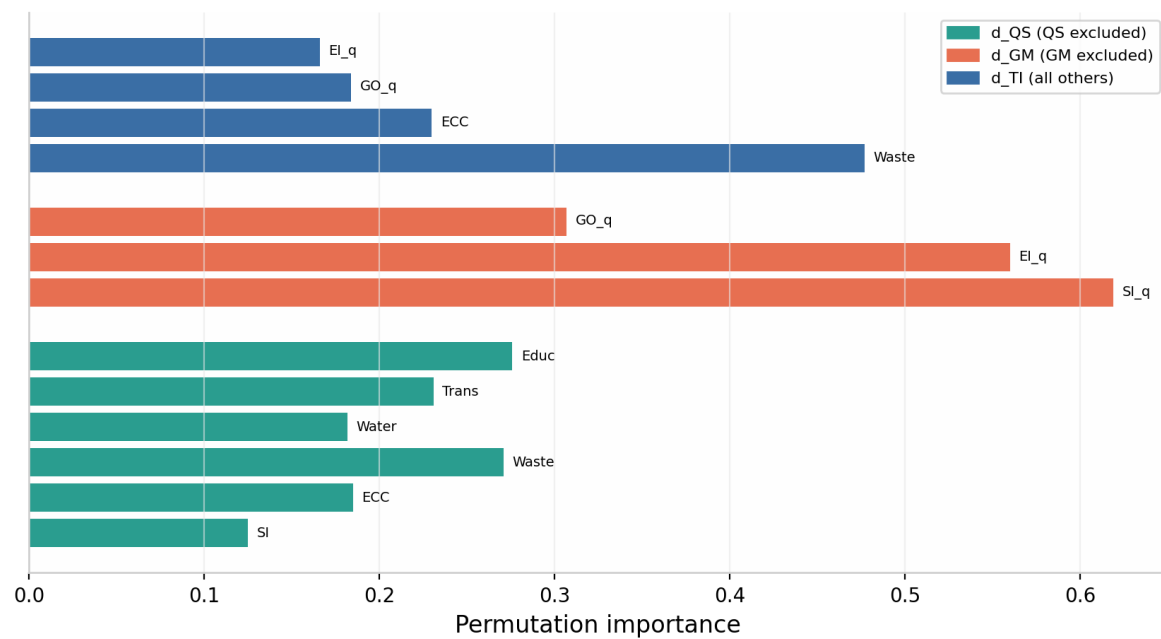


Fig. 6. Leakage-controlled predictive attribution

Table 9.

Leakage-controlled predictive performance

Target	Predictor set	OOF R^2 (mean $\pm$ SD)	Most important predictors
d_QS = nQS – mean(nGM, nTI)	GreenMetric dimensions only	0.003 $\pm$ 0.037	Education & Research (0.276); Waste (0.271); Transportation (0.231)
d_GM = nGM – mean(nTI, nQS)	QS lenses only	–0.007	Social Impact (0.619); Environmental Impact (0.560); Governance (0.307)
d_TI = nTI – mean(nGM, nQS)	GreenMetric + QS components	0.131 $\pm$ 0.028	Waste (0.477); Energy & Climate Change (0.230)

Note: OOF R^2 = out-of-fold coefficient of determination averaged over 3 repetitions of 5-fold cross-validation. A negative OOF R^2 indicates performance worse than predicting the target mean. Permutation importances are within-model descriptive summaries, reported only where informative.

Table 10 provides the elastic-net benchmark and the permutation null comparison. For QS and GreenMetric divergence, both Random Forest and elastic-net R^2 values are close to zero or negative. For THE Impact divergence, elastic-net produces an out-of-fold R^2 of 0.252 compared with 0.127 for Random Forest. The observed values exceed the 95th percentile of the corresponding permutation null distributions for all three targets, with permutation p values of 0.02. The size of the predictive association nevertheless differs substantially across targets, with the clearest out-of-fold performance observed for THE Impact divergence.

Table 10.

Elastic-net benchmarks and permutation null test

Target	Random Forest OOF R^2	Elastic-net OOF R^2	Permutation null 95th percentile	Permutation p
d_QS	–0.002	–0.011	–0.058	0.02
d_GM	–0.006	–0.014	–0.067	0.02
d_TI	0.127	0.252	–0.010	0.02

Note: Benchmark values are computed with the replication scripts at a single fixed random seed (42); random-forest OOF R^2 therefore differs marginally from the repeated-CV averages in Table 9, without changing any substantive conclusion. The permutation null is the distribution of OOF R^2 over 50 target-shuffled refits.

Across the predictive specifications, the results show limited and target-specific out-of-fold performance. QS Sustainability and GreenMetric divergence remain close to the mean-prediction benchmark, while THE Impact divergence is predicted more successfully by both Random Forest and elastic-net. The predictive findings complement the cluster analysis by showing that the observed divergence profiles are empirically structured, but their predictability varies across the focal ranking system.

5. Discussion

5.1. Shared signals and distinct measurement regimes

The empirical pattern has two sides that belong together. The three rankings are positively associated on the common sample, and Kendall's W indicates a substantial shared ordering. At the same time, the pairwise relationships remain moderate rather than close. The systems therefore recognize a common institutional orientation toward sustainability without assigning the same meaning to it. This result fits the view of higher-education sustainability as a multidimensional field without one universally accepted assessment objective [6]. The important point is that partial agreement is not an incidental imperfection in the data. It is evidence that the rankings capture overlapping but distinct representations of institutional sustainability.

The common-set design makes this interpretation empirically precise. The comparison concerns the same 113 universities in all three systems and uses the full observed overlap rather than an elite subset. The associations consequently describe variation within one matched population. They are not generated by comparing separate provider populations, yet they still fall short of the equivalence that would be expected if the rankings measured the same institutional attribute. The result establishes a strong basis for discussing both convergence and divergence within one institutional field.

The architecture of the three systems helps account for this result. UI GreenMetric measures sustainability through campus operations and management domains such as infrastructure, energy, waste, water, transport, education and research [26]. THE Impact organizes evidence around institutional contributions to the Sustainable Development Goals, including research, stewardship, outreach and teaching [25]. QS Sustainability combines environmental impact, social impact and governance dimensions [24]. The rankings thus address related institutional concerns, but they do so through different measurement questions and different evidentiary expectations.

Earlier comparative studies have documented differences in indicator structures, participation arrangements and institutional coverage across sustainability rankings [7], [8]. The present analysis advances that literature by showing the consequence of those design differences in the positions assigned to the same universities. It also develops the information-intermediary perspective on rankings, where complex institutional activities are condensed into signals that acquire reputational and governance significance [1], [3]. The theoretical contribution is a distinction between common ordering and construct equivalence. Moderate concordance is not an ambiguous middle ground. It is an empirical signature of partial construct overlap, and the structure of the divergence indicates which dimensions remain specific to each ranking regime.

5.2. Consensus aggregation and ranking uncertainty

The consensus analysis establishes that the broad ordering is not a consequence of one arbitrary weighting rule. Entropy weighting gives greater influence to QS Sustainability because its normalized scores are more dispersed in the common sample. CRITIC weighting distributes influence more evenly because it considers both contrast and redundancy, while equal weighting provides a transparent benchmark. Despite these different allocation principles, the resulting consensus ordering remains closely aligned. The main ordering is therefore supported by convergent evidence across alternative aggregation logics.

This finding confirms the composite-indicator literature's emphasis on explicit weighting and aggregation choices, but it adds an important qualification to the usual sensitivity argument. Weighting matters, yet its influence is not uniform across the ranking. In these data, alternative schemes modify the allocation of positions more than they alter the broad structure of the ordering. The effect is concentrated among universities with relatively close scores, especially away from the top. Robustness is therefore a property of portions of the ordering rather than a single attribute of the ranking as a whole [13], [14].

The weight-simplex analysis makes this uneven uncertainty visible. Universities near the top retain high inclusion probabilities across a wide range of admissible weight combinations, whereas middle positions have wider rank intervals and lower inclusion probabilities. The distinction between a stable position and a stable ordering is consequently important. A university may remain within a broad upper group while its exact rank changes across plausible weighting schemes. This result gives empirical content to the methodological argument that rank uncertainty is an inherent feature of composite measures, not merely a technical inconvenience [15], [22].

The Plackett–Luce aggregation supports the same broad structure through a probabilistic treatment of the source rankings. Probabilistic aggregation preserves information about ordering uncertainty that is lost when every observed rank is treated as a fixed-point estimate [30]. In the present analysis, its agreement with the compensatory consensus ordering strengthens the evidence for a shared upper structure, while the remaining dispersion across positions identifies where the synthesis is less determinate. The methodological contribution is therefore more than a robustness check. It provides a transparent way to report a consensus ordering together with the empirical conditions under which that ordering is stable.

5.3. Divergence, prediction and the limits of attribution

The deviation-space analysis shows that disagreement is organized in the common sample. The clusters separate universities according to the ranking system in which their relative position is stronger or weaker, and their stability under bootstrap resampling indicates that these profiles are not produced by a single accidental partition. Comparative research has established that sustainability rankings differ in indicator structures and institutional coverage [7]. The present results add a positional account of that divergence. They show how the same university can be relatively advantaged in one system and relatively disadvantaged in another, even when all three rankings are addressing sustainability.

The predictive analysis identifies a different boundary. After leakage control, the source components provide limited out-of-fold information about rank deviations. Feature importance remains informative about associations within the fitted models, but it does not identify a mechanism that produces a ranking position. This distinction is consequential in small samples, where correlated predictors, model selection and target construction can make in-sample importance appear stronger

than out-of-sample performance warrants [33]–[34]. The analysis therefore separates a visible attribution pattern in the observed data from evidence that the pattern will generalize.

The contrast between clustering and prediction is not a weakness of the analytical design. It reveals two distinct properties of ranking divergence. A pattern can be recurrent enough to describe and classify the observed common sample while remaining too contingent for reliable prediction in new observations. The clusters identify structured forms of relative advantage and disadvantage. The predictive models test whether the available ranking components contain transferable information about those forms. The stronger clustering result and the weaker predictive result jointly define the empirical boundary of what the data supports.

Taken together, the analyses establish a multi-level framework for ranking comparison. Pairwise concordance measures shared ordering. Consensus analysis evaluates how that ordering responds to aggregation. Deviation-space clustering identifies the empirical forms of divergence. Leakage-controlled prediction examines whether those forms generalize beyond the fitted sample. The value of this framework lies in connecting these levels without collapsing them into single statistics. It shows not only whether rankings agree, but also where their agreement is stable, how their disagreement is organized and how far the observed structure can support prediction.

5.4. Theoretical and practical contributions

At the theoretical level, the study establishes that ranking concordance is best understood as partial construct overlap rather than as a binary condition of agreement or disagreement. A common ordering and systematic divergence can coexist. This matters for the interpretation of rankings because it separates the existence of a shared institutional signal from the assumption that the signal represents one underlying construct in the same way across providers. The evidence shows that rankings can be sufficiently related to support comparison while remaining theoretically distinct in the dimensions of sustainability they make visible. This formulation turns disagreement into an object of explanation rather than a residual problem to be removed.

At the methodological level, the study provides an integrated framework for moving from association to uncertainty, divergence and generalization. The common-set design establishes comparability. The consensus procedures test the stability of the resulting ordering. The weight-simplex analysis locates uncertainty across the ranking rather than summarizing it with one global statement. Deviation-space clustering identifies recurring profiles, and leakage-controlled prediction tests whether those profiles transfer beyond the observed sample. The contribution is cumulative. Each stage answers a different question, and together they create a more demanding evaluation of ranking concordance than pairwise correlation alone.

For ranking users, the findings support a more discriminating use of published positions. The consensus results identify an upper ordering that remains comparatively stable across alternative weights, while the simplex analysis shows greater sensitivity in the middle of the distribution. This distinction has direct consequences for benchmarking, strategic comparison and institutional reputation. A clear separation between universities near the top is more informative when it persists across aggregation rules. Small differences among middle-ranked universities require greater caution because their positions are more dependent on weighting and measurement design. Users can therefore distinguish robust comparative information from rank precision that the underlying data do not support.

For ranking designers, the results identify the information that must accompany a published position if the ranking is to be analytically useful. Providers should report the eligible population,

participation conditions, indicator domains, missing-data rules, treatment of banded ranks and aggregation procedure in a single accessible methodological record. They should also disclose how coverage and reporting conditions enter the final position. Such reporting would make the source of cross-ranking divergence observable rather than leaving users to infer it from the final ordinal list. It would also allow independent researchers to evaluate whether a change in rank reflects a change in institutional evidence, a change in the participating population or a change in the aggregation rule.

For universities and public decision-makers, the findings favour domain-specific monitoring over a single institutional target. Improvements in campus operations, SDG reporting, governance evidence or research visibility need not produce parallel movement across all three rankings because those systems expose different dimensions of institutional activity. Internal assessment can therefore track the underlying domains separately and use external rankings as contextual signals. This approach preserves the informational value of rankings while preventing one published position from becoming a substitute for a broader assessment of institutional sustainability.

5.5. Limitations and future research

The evidence is conditional on the 113 universities appearing in all three 2024 editions. This common set is analytically useful, but it excludes institutions that participate in only one or two systems and may overrepresent universities with stronger reporting capacity or international visibility. Cross-sectional design also prevents a distinction between persistent institutional characteristics and edition-specific conditions.

The published rankings impose further measurement constraints. THE Impact reports exact positions for the upper part of the table and bands thereafter. Midpoint conversion provides a reproducible numerical representation of those bands, but it does not recover the unobserved position within an interval. Name matching can also leave residual uncertainty, and the analysis conditions on the observed common set rather than modelling participation and coverage as additional sources of uncertainty. The clustering solution is descriptive, even though its resampling stability is adequate.

Future research should combine several comparable editions with larger or differently defined common samples. Longitudinal designs could separate persistent institutional differences from changes in provider methodology. Indicator-level data could identify which reporting practices are associated with recurring divergence. Prediction could also be strengthened by adding institution-level covariates and by defining a clear exposure and outcome before applying heterogeneous-effect methods. Such designs would provide a stronger basis for evaluating conditional associations than the present cross-sectional ranking data allows.

6. Conclusion

This study compared the 2024 UI GreenMetric, THE Impact and QS Sustainability rankings for 113 universities appearing in all three systems. The rankings share a moderate common ordering, but their agreement is incomplete because each system makes different institutional activities visible. The consensus analysis shows that the broad ordering is stable across alternative weighting rules and the full weight simplex, while middle positions remain more sensitive than positions near the top. The divergence analysis identifies recurring relative-position profiles, whereas leakage-controlled prediction provides limited out-of-sample information about their generalization. The study therefore contributes a common-set framework for ranking concordance, an uncertainty-aware

consensus procedure and a clear separation between descriptive divergence and predictive attribution. The rankings are most informative when their shared signal and their system-specific dimensions are considered together.

Data Availability Statement

The analysis used publicly available 2024 ranking data published by the UI GreenMetric World University Rankings, the Times Higher Education Impact Rankings, and the QS World University Rankings Sustainability. The processed common-sample data, institution-matching documentation, and analysis code are not publicly deposited. They may be requested from the corresponding author for legitimate scholarly purposes on reasonable request, subject to the access terms of the original data providers.

Author Contributions

Conceptualization, methodology, formal analysis, data curation, writing of the original draft, review and editing were carried out by L. Çinko. The author has read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Conflicts of Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

- [1] Hazelkorn, E. (2015). *Rankings and the reshaping of higher education: The battle for world-class excellence* (2nd ed.). Palgrave Macmillan. <https://doi.org/10.1057/9781137446671>
- [2] Rindova, V. P., Martins, L. L., Srinivas, S. B., & Chandler, D. (2018). The good, the bad, and the ugly of organizational rankings: A multidisciplinary review of the literature and directions for future research. *Journal of Management*, 44(6), 2175-2208. <https://doi.org/10.1177/0149206317741962>
- [3] Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American journal of sociology*, 113(1), 1-40. <https://doi.org/10.1086/517897>
- [4] Leal Filho, W., Viera Trevisan, L., Wahaj, Z., Déda Araújo Nunes, D., Ruy Portela de Vasconcelos, C., Aparecida Dibbern, T., ... & Pimenta Dinis, M. A. (2024). University rankings and sustainable development: the state of the art. *International Journal of Sustainability in Higher Education*, 26(7):1615–1636. <https://doi.org/10.1108/IJSHE-11-2023-0530>
- [5] Urbano, V. M., Arena, M., Azzone, G., & Mayeres, M. (2025). Sustainable development in higher education: An in-depth analysis of Times Higher Education Impact Rankings. *Journal of Cleaner Production*, 501, 145302. <https://doi.org/10.1016/j.jclepro.2025.145302>
- [6] Caeiro, S., Sandoval Hamón, L. A., Martins, R., & Bayas Aldaz, C. E. (2020). Sustainability assessment and benchmarking in higher education institutions—A critical reflection. *Sustainability*, 12(2), 543. <https://doi.org/10.3390/su12020543>
- [7] Galleli, B., Teles, N. E. B., Santos, J. A. R. D., Freitas-Martins, M. S., & Hourneaux Junior, F. (2022). Sustainability university rankings: a comparative analysis of UI green metric and the times higher education world university rankings. *International Journal of Sustainability in Higher Education*, 23(2), 404-425. <https://doi.org/10.1108/IJSHE-12-2020-0475>
- [8] Muñoz-Suárez, M., Guadalajara, N., & Osca, J. M. (2020). A comparative analysis between global university rankings and environmental sustainability of universities. *Sustainability*, 12(14), 5759. <https://doi.org/10.3390/su12145759>

- [9] Wilhelm, E. M. D. S., Ferreira, C. L., Cantorani, J. R. H., & Pilatti, L. A. (2026). Sustainability Without Prestige? The Rise of UI GreenMetric and the Persistent Decoupling from Global University Rankings (2010–2025). *Sustainability*, 18(7), 3555. <https://doi.org/10.3390/su18073555>
- [10] Alberti, C., Civera, A., Lehmann, E. E., Meoli, M., Otto, J., & Paleari, S. (2025). University sustainability rankings: a critical literature review on the UI GreenMetric ranking system. *The Journal of Technology Transfer*, 50(6), 2752-2801. <https://doi.org/10.1007/s10961-025-10223-9>
- [11] Buckner, E., & Zhang, Y. (2026). Sustainability rankings in higher education: 'The right thing to do' or the pursuit of global recognition?. *Higher Education*, 91(5), 1775-1791. <https://doi.org/10.1007/s10734-025-01495-z>
- [12] Lakhno, M. (2024). Green or green-washed? Examining sustainability reporting in higher education. *Higher Education Quarterly*, 78(3), 1099-1122. <https://doi.org/10.1111/hequ.12513>
- [13] Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2019). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social indicators research*, 141(1), 61-94. <https://doi.org/10.1007/s11205-017-1832-9>
- [14] OECD/European Union/EC-JRC (2008), *Handbook on Constructing Composite Indicators: Methodology and User Guide*, OECD Publishing, Paris, <https://doi.org/10.1787/9789264043466-en>.
- [15] Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168(2), 307-323. <https://doi.org/10.1111/j.1467-985X.2005.00350.x>
- [16] Prior, D. D., Seshadrinath, S. M., Zhang, M., & McCormack, M. (2025). Measuring sustainable development goals (SDGs) in higher education through semantic matching. *Studies in Higher Education*, 50(7), 1556-1569. <https://doi.org/10.1080/03075079.2024.2386625>.
- [17] International Ranking Expert Group. (2006). *Berlin principles on ranking of higher education institutions*. Institute for Higher Education Policy. <https://www.ihep.org/publication/berlin-principles-on-ranking-of-higher-education-institutions/>
- [18] Irungu, R. W., & Liu, Z. (2024). Exploring the nexus between university sustainability practices and academic performance: An empirical analysis of the QS sustainability ranking and four world university rankings. *Plos one*, 19(10), e0306286. <https://doi.org/10.1371/journal.pone.0306286>
- [19] Lakhno, M., & Ortiz-Gervasi, L. (2024). Intermediary organizations and their role in advancing the SDGs in higher education. *Policy Reviews in Higher Education*, 8(1), 56-84. <https://doi.org/10.1080/23322969.2024.2323735>
- [20] Abacioğlu, S., Ayan, B., & Pamucar, D. (2025). The race to sustainability: Decoding green university rankings through a comparative analysis (2018–2022). *Innovative Higher Education*, 50(1), 241-275. <https://doi.org/10.1007/s10755-024-09734-4>
- [21] Akyol Özcan, K. (2023). Sustainability ranking of Turkish universities with different weighting approaches and the TOPSIS method. *Sustainability*, 15(16), 12234. <https://doi.org/10.3390/su151612234>
- [22] Kelemen, A., Szabó, Z. K., Bozóki, S., Szádóczi, Z., & Hartvig, Á. D. (2024). A sensitivity analysis of composite indicators: Min/max thresholds. *Environmental and Sustainability Indicators*, 23, 100453. <https://doi.org/10.1016/j.indic.2024.100453>
- [23] Latifis, K., Sidiropoulos, A., & Evangelidis, G. (2026). International university rankings: composite indicators and methodological approaches. *Frontiers in Research Metrics and Analytics*, 11, 1828850. <https://doi.org/10.3389/frma.2026.1828850>
- [24] QS Quacquarelli Symonds. (2024). "QS World University Rankings: Sustainability 2024." <https://www.topuniversities.com/sustainability-rankings/2024>.
- [25] Times Higher Education. 2024. "Impact Rankings (2024): Results announced." <https://www.timeshighereducation.com/rankings/impact/overall/2024>.
- [26] UI GreenMetric World University Rankings. (2024). "Overall rankings 2024." <https://uigreenmetric.com/rankings/university/overall-rankings-2024>.
- [27] Kendall, M. G., and Babington Smith, B. (1939). "The problem of m rankings." *The Annals of Mathematical Statistics* 10(3):275–287. <https://doi.org/10.1214/aoms/1177732186>.
- [28] Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.
- [29] Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2), 193-202. <https://doi.org/10.2307/2346567>.
- [30] Turner, H. L., van Etten, J., Firth, D., & Kosmidis, I. (2020). Modelling rankings in R: the PlackettLuce package. *Computational Statistics*, 35(3), 1027-1057. <https://doi.org/10.1007/s00180-020-00959-3>
- [31] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>

- [32] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12, 2825-2830.
- [33] Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11), e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- [34] Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11, 2079-2107. <https://jmlr.org/papers/v11/cawley10a.html>.
- [35] Chamma, A., Engemann, D. A., & Thirion, B. (2023). Statistically valid variable importance assessment through conditional permutations. *Advances in neural information processing systems*, 36, 67662-67685. <https://doi.org/10.52202/075280-2959>
- [36] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics* 47(2):1148–1178. <https://doi.org/10.1214/18-AOS1709>.
- [37] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228-1242. <https://doi.org/10.1080/01621459.2017.1319839>