

Auditing Internal Consistency in Webometrics University Rankings: A Bidirectional, Historical, and Peer-Calibrated Approach

Tarek Barhoum

t-barhum@aiu.edu.sy

Arab International University <https://orcid.org/0000-0002-9132-6940>

Mina Ibrahim

MinaIbrahim190@gmail.com

Arab International University <https://orcid.org/0009-0005-6133-8384>

Research Article

Keywords: university rankings, Webometrics, ranking integrity, anomaly detection, robust statistics, peer calibration, longitudinal analysis, synthetic validation

Posted Date: October 1st, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11210663/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Auditing Internal Consistency in Webometrics University Rankings: A Bidirectional, Historical, and Peer-Calibrated Approach

Mina Ibrahim

Tarek Barhoum

August 2026

Disclaimer

This manuscript is a preprint and has not been peer-reviewed. It should not be regarded as conclusive, guide clinical practice, or be referenced as validated information.

Abstract

University rankings compress multidimensional institutional performance into a single ordered position, potentially obscuring inconsistencies between final ranks and their underlying indicators. This study develops a bidirectional audit framework to identify unusually favorable and unfavorable outcomes in the July 2026 Webometrics ranking. The analysis covers 5,500 institutions, including 5,491 linked historical profiles and 63,383 historical observations. Rank movements and current positions were evaluated using directional residuals from complementary robust and nonlinear models, followed by historical, institution-type, and predictor-defined peer calibration. Synthetic perturbation experiments assessed detector performance under known inconsistencies. The framework substantially reduced the initial candidate pool after comparability and temporal controls. Gwangju University emerged as the primary novel favorable case, whereas Middle East University Jordan was the primary novel unfavorable case; persistent or peer-sensitive signals were distinguished from new discrepancies. Synthetic validation yielded AUROC values of 0.955 and 0.954 for favorable and unfavorable perturbations, respectively, with false-positive rates below 1% in both directions. These findings demonstrate that ranking anomalies can be screened without interpreting residuals as evidence of institutional quality, manipulation, or provider error. The proposed framework offers a reproducible and conservative approach for prioritizing internally inconsistent ranking outcomes for further investigation.

Keywords: university rankings, Webometrics, ranking integrity, anomaly detection, robust statistics, peer calibration, longitudinal analysis, synthetic validation.

1. Introduction

Global university rankings are now part of the infrastructure of higher education. Students use them when choosing institutions, universities cite them in strategic plans and promotional material, and governments often treat them as shorthand for international competitiveness. Their influence is disproportionate to the simplicity of the final output: an ordered list. A rank can combine research production, citation impact, reputation, visibility, openness, institutional size, and other features into one number, even though those dimensions are neither interchangeable nor measured without error. Long-standing work on university rankings has shown that positions can change materially with indicator choice, weighting, aggregation, normalization, disciplinary composition, and database coverage (van Raan, 2005; Ioannidis et al., 2007; Billaut et al., 2010; Saisana et al., 2011; Waltman et al., 2012).

Web-based rankings introduce another layer of complexity. The early Ranking Web of World Universities was developed around the idea that institutional web presence and linking structure capture forms of scholarly visibility and dissemination that conventional bibliometric databases miss (Aguillo et al., 2008; Ortega & Aguillo, 2009). Later comparisons showed that webometric and conventional rankings overlap only partially and are shaped by different sources of institutional visibility (Aguillo et al., 2010; Thamm & Mayr, 2011). The current Webometrics.org platform uses three components—Visibility, Transparency, and Excellence—with published score-level weights of 50%, 10%, and 40% (Webometrics.org, n.d.-a). Because the public ranking tables expose component ranks rather than the normalized scores to which those weights apply, the composite cannot be reconstructed by multiplying the visible ranks by 50/10/40. The analysis in this paper therefore learns the empirical mapping between published component ranks and the published World Rank instead of imposing an arithmetic formula on ordinal positions.

The practical question is straightforward: when does a final rank look internally unusual given the ranking's own published components? A large discrepancy can arise for many reasons. Specialized institutions, university systems, medical schools, art schools, community colleges, and conventional comprehensive universities can have very different digital and research footprints. Website migrations, redirects, domain consolidation, indexing changes, tied ranks, or edition-specific data updates can also produce abrupt movements. For that reason, a statistical

discrepancy is treated here as an audit signal. It marks a case that deserves explanation; it does not determine whether the ranking provider or the institution did anything improper.

Recent integrity-oriented studies make the distinction between screening and accusation especially relevant. Nazarovets (2024) used rank-citation patterns to identify university h-index anomalies for expert inspection. Meho (2025) combined several bibliometric risk indicators to examine unusually rapid institutional research growth. Bayanbayeva (2026), using qualitative evidence from Kazakhstan, documented how ranking incentives can encourage strategic responses and, in some settings, gaming behavior. These studies examine real concerns around metric pressure, but they do not answer the narrower question addressed here: whether a ranking outcome is internally consistent with its own reported inputs, whether the deviation is favorable or unfavorable, and whether the same signal survives temporal and comparability controls.

This study develops a bidirectional audit for that purpose. It separates short-term movement from current-position consistency, keeps the sign of the residual instead of collapsing both directions into an absolute outlier score, and requires agreement across model classes. Candidates are then re-examined historically, by institution type, and within peer groups defined from expected rather than observed rank. A final synthetic benchmark tests whether the fixed detector recovers known perturbations in both directions. The result is intended as a reproducible triage framework for ranking research, not as a substitute for provider-side score data or an investigation of institutional conduct.

1.1 Research questions

- RQ1. Which institutions move substantially more favorably or unfavorably than expected between January and July 2026 given changes in Impact, Openness, and Excellence?
- RQ2. Which July 2026 World Ranks are materially better or worse than expected from the published component ranks?
- RQ3. Do the detected discrepancies represent new breaks or patterns that recur across the available history?
- RQ4. Which signals remain after controlling for institution type and recalibrating within predictor-defined peer groups?
- RQ5. How well does the fixed detector recover known synthetic discrepancies in both directions, and what false-positive rates result?

1.2 Contributions

- A verified July 2026 data pipeline that corrects a non-obvious crawler schema mismatch and links current records to profile-level history.
- A directional two-stage audit that distinguishes unexpected rank movement from unexpected current position and retains favorable and unfavorable residuals separately.
- Historical, institution-type, and predictor-defined peer checks that reduce structural and circular-comparison artifacts.
- A conservative consensus rule based on robust residual standardization without a fixed contamination parameter or a forced anomaly count.
- A bidirectional synthetic benchmark with fixed thresholds, repeated injections, and a rank-valid reranking intervention.
- Case-based interpretation showing how internally robust signals can still have plausible external explanations and therefore require contextual review.

2. Related Work

2.1 Webometrics and university ranking methodology

Webometrics emerged from the broader use of web indicators to study scholarly communication and institutional visibility. Aguillo et al. (2008) described the logic and early methodology of the Ranking Web of World Universities, while Ortega and Aguillo (2009) showed that inter-university hyperlink structures reveal national and international patterns that are not captured by publication counts alone. A later comparison of several major ranking systems found substantial differences in coverage and indicator design, reinforcing the idea that rankings measure overlapping but non-identical constructs (Aguillo et al., 2010). Thamm and Mayr (2011) similarly found that hyperlink-based rankings of German universities did not map cleanly onto web-independent prestige measures. Together, this literature supports web visibility as informative while cautioning against treating it as a direct measure of academic quality.

The organizational identity of the current data source also matters. Webometrics.org now states that it is an independent platform and is not affiliated with the former Webometrics.info site or with the Spanish National Research Council (CSIC) (Webometrics.org, n.d.-b). Accordingly, current Webometrics.org pages are treated as the source of record for the 2026 observations in this study. Earlier Webometrics publications remain relevant to the history of

webometric ranking and indicator design, but they are not used to imply organizational continuity.

2.2 Composite rankings and uncertainty

Critiques of university rankings have repeatedly focused on what happens when heterogeneous indicators are compressed into a single league table. van Raan (2005) highlighted conceptual and field-normalization problems in bibliometric ranking. Ioannidis et al. (2007) questioned the comparability and interpretability of international ranking systems. Billaut et al. (2010) showed that multicriteria modeling choices can change the resulting order, and Saisana et al. (2011) demonstrated substantial rank volatility under plausible alternative weighting and aggregation assumptions. The Leiden Ranking literature further emphasized transparent indicator construction, counting rules, normalization, and uncertainty when institutional positions are compared (Waltman et al., 2012).

The same caution applies when correlated indicators are combined. Robinson-García et al. (2019) found that publication output and citation impact explain a large share of the variance across several global rankings, even when the rankings describe themselves in different terms. Forthmann et al. (2024) showed more generally why summing bibliometric indicators does not establish a valid composite construct: correlated components, multidimensionality, effective weighting, and measurement invariance can all complicate interpretation. These observations motivate the deliberately narrow target used here. The model is not asked to estimate institutional quality. It is asked whether the observed final rank is typical of the empirical relationship between the published components and final ranks within the observed cohort.

2.3 Anomaly detection and integrity-oriented assessment

Robust statistics offers a natural starting point for this kind of audit because it seeks to characterize the dominant data pattern without allowing extreme observations to determine the fit (Rousseeuw & Hubert, 2018). Huber's robust estimation provides a resistant linear baseline (Huber, 1964), while gradient boosting and extremely randomized trees provide flexible nonlinear alternatives (Friedman, 2001; Geurts et al., 2006). The use of several model classes in this study is a robustness check: agreement means that the signal is not specific to one fitted

function. Since the models share the same outcomes and closely related predictors, their agreement is not treated as independent statistical evidence.

Integrity-oriented work in higher education has recently moved from general criticism of rankings toward empirical screening. Nazarovets (2024) showed that unusual rank-citation curves can expose h-index anomalies that deserve expert review. Meho (2025) examined extreme publication growth, authorship patterns, self-citation, retractions, and delisted-journal exposure at the institutional level. Bayanbayeva (2026) documented ranking-related strategic and manipulative practices through interviews in Kazakhstan. The present study complements those approaches rather than replacing them. It does not use misconduct indicators as model inputs. Instead, it audits the internal relationship among a ranking's published components, recent movement, and final position.

2.4 Research gap

Across the literature reviewed above, several ingredients exist separately: webometric measurement, methodological criticism of composite rankings, robust anomaly detection, and integrity-sensitive assessment. What is much less developed is an end-to-end audit that preserves discrepancy direction, combines longitudinal and cross-sectional checks, distinguishes sudden breaks from persistent institutional patterns, controls for institution type, constructs peers without conditioning on the observed rank, and tests the same detector under known favorable and unfavorable perturbations. The contribution claimed here is this integration and its empirical evaluation on the July 2026 Webometrics data. The framework is therefore novel in design within the literature reviewed, while the individual statistical tools are established methods.

3. Data and Verification

3.1 July 2026 cohort and schema verification

The primary July 2026 crawl contained 5,500 displayed rows. During verification, one crawler field proved misleading: the raw column labeled "World Rank" was the sequential display position from 1 to 5,500, whereas the raw field labeled "Country Rank" contained the actual published July 2026 World Rank, including ties. The processed dataset was corrected before modeling. After correction, the July table contained 4,763 unique World Rank values, and

1,373 institutions participated in at least one tied World Rank. Impact Rank was missing for five institutions.

A stratified set of 20 public profile pages was checked manually against the processed table; current rank and component values matched in all 20 cases. Historical profile retrieval succeeded for 5,491 institutions and failed for nine. The five missing Impact values were left missing and complete-case analysis was used for the primary models rather than silently imputing them.

Quantity	Value
July 2026 displayed rows	5,500
Successfully linked profile histories	5,491
Profile crawl errors	9
Unique July World Ranks	4,763
Institutions involved in ties	1,373
Missing July Impact values	5
Raw historical observations	68,109
Historical observations after excluding duplicated Jan. 2018	63,383
Complete historical observations scored	63,378

Table 1. Verified data coverage and quality-control summary.

3.2 Historical panel

Profile histories initially contributed 68,109 observations for the 5,491 successfully linked institutions. Coverage is unbalanced: January editions are available from 2015 onward, whereas July observations in the collected profile history are available only for 2025 and 2026. An exact duplicate was detected between January 2016 and January 2018. All 4,726 corresponding rows matched across the ranking fields and produced the same content hash, so January 2018 was

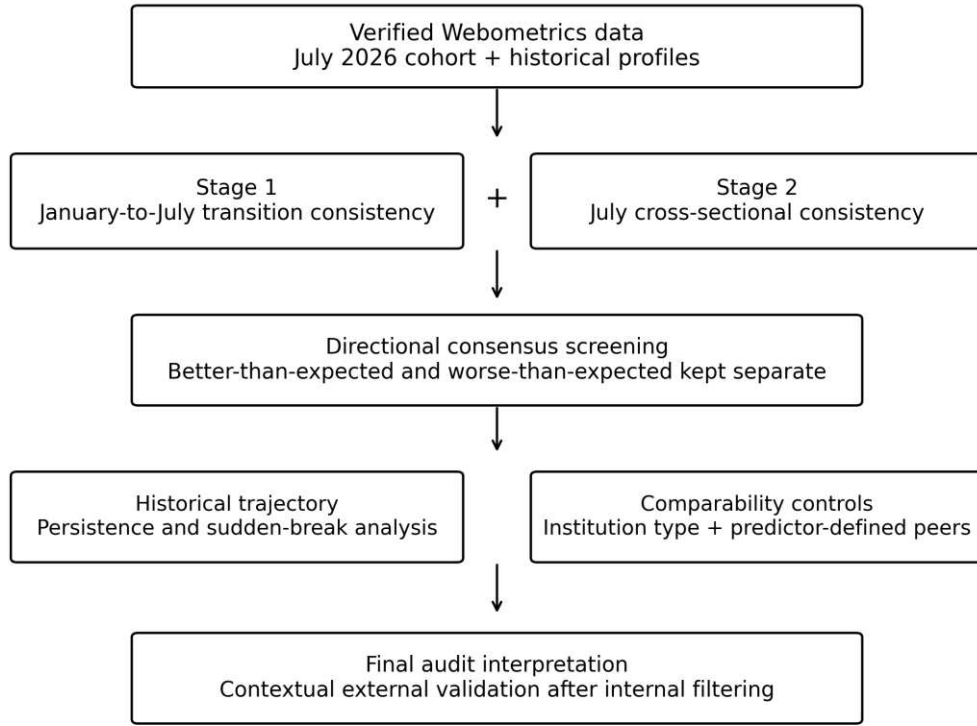
treated as an archival/provider artifact and excluded from the primary historical analysis. This left 63,383 observations, of which 63,378 were complete for historical scoring.

3.3 Institution-type taxonomy

Institutional heterogeneity was addressed with a conservative name-based taxonomy. Among the 5,491 successfully profiled entities, 3,018 were classified as standard universities, 1,412 as other entities, 383 as institutes or academies, 333 as colleges, 263 as medical or health institutions, 43 as business schools, 18 as university systems, 11 as art or design institutions, and 10 as specialized professional institutions. The taxonomy is used only as a sensitivity control. It is not presented as an authoritative classification of institutional mission.

4. Methodology

Figure 1 shows the order of the audit. The sequence matters: a large residual is only an initial screen. Historical behavior and comparability are examined before a case is interpreted, and external context is considered only after the internal filters have been applied. Synthetic experiments are kept separate from candidate selection; they evaluate the fixed detector rather than choosing its thresholds.



Synthetic bidirectional stress tests validate the detector; they do not define the thresholds.

Figure 1. Final bidirectional ranking-integrity audit workflow.

4.1 Rank normalization

Smaller Webometrics ranks indicate better positions. Because edition coverage and matched cohort size vary, raw ranks were converted to average-rank percentiles within each analyzed edition. For institution i in edition t , with average rank R_{it} among n_t observed institutions, the normalized position is:

$$p_{it} = R_{it} / n_t \quad (1)$$

Smaller percentile values remain better. Ties are preserved through average ranks. These percentiles describe position within the observed cohort; they should not be read as percentiles of the full population of higher-education institutions worldwide.

4.2 January-to-July transition model

The first stage asks whether the January-to-July change in World Rank is consistent with contemporaneous changes in Impact, Openness, and Excellence. World-rank improvement is defined as:

$$\Delta p_{world,i} = p_{world,i,Jan} - p_{world,i,Jul} \quad (2)$$

Positive values therefore represent improvement. The same transformation is applied to the three component ranks. Out-of-fold Huber regression and histogram-based gradient boosting predict the expected World Rank improvement. The directional transition residuals are:

$$r_{T,+} = \Delta p_{world} - \hat{y}_T \quad (3)$$

$$r_{T,-} = \hat{y}_T - \Delta p_{world} \quad (4)$$

A positive $r_{T,+}$ means an institution improved more than expected; a positive $r_{T,-}$ means it improved less than expected or deteriorated more than expected. The sign is retained throughout the analysis.

4.3 July cross-sectional consistency

The second stage predicts the July 2026 World Rank percentile from the July Impact, Openness, and Excellence percentiles. Histogram-based gradient boosting and Extra Trees are fitted out of fold. If p_i is the observed July World Rank percentile and $\hat{y}_i$ is the predicted percentile, the directional discrepancies are:

$$r_{C,+} = \hat{y}_i - p_i \quad (5)$$

$$r_{C,-} = p_i - \hat{y}_i \quad (6)$$

The favorable residual is positive when the observed rank is better than predicted; the unfavorable residual is positive when it is worse. Because the cross-sectional mapping is extremely tight in these data, this stage is interpreted as an internal rank-component consistency test rather than a model of educational or research quality.

4.4 Robust standardization and consensus

Directional residuals are standardized with the median absolute deviation (MAD). For residual r_i , the modified robust z-score is:

$$z_i = 0.6745 (r_i - \text{median}(r)) / \text{MAD}(r) \quad (7)$$

The primary screen uses $z > 3$ in both designated models. In the cross-sectional stage, both model residuals must also exceed 0.02 percentile units. The z threshold is a conservative operating rule rather than a significance test, and the two-percentile-point condition prevents very small discrepancies from being promoted simply because the residual distribution is narrow. An early fixed-contamination Isolation Forest experiment was discarded because a contamination parameter mechanically fixes the number of reported anomalies. The final procedure has no forced anomaly count.

4.5 Historical trajectory analysis

Historical context is used to separate a new discrepancy from a recurring institutional pattern. For the 115 favorable transition-consensus candidates, cross-sectional residuals were recomputed within each available historical edition after removing the duplicated January 2018 records. Historical persistence is the fraction of prior available editions in which the institution exceeds the directional robust- z threshold. Current deviation is then compared with prior behavior to distinguish transition-only signals, persistent favorable patterns, current favorable anomalies, and sudden favorable breaks. The mirrored unfavorable analysis was applied to the two institutions that initially passed both unfavorable layers before peer calibration: Middle East University Jordan and Innopolis University. Because profile coverage is uneven, persistence denominators use the observations available for each institution rather than assuming a balanced panel.

4.6 Institution-type controls

Institution type is a direct comparability concern. The favorable models were therefore re-estimated in a standard-university cohort and in a broader canonical cohort. The standard-university group contained 3,018 institutions, of which 3,017 had complete July component data. The unfavorable extension was run within the standard-university cohort from the outset. This choice avoids interpreting structural differences among systems, colleges, specialized schools, and comprehensive universities as if they were the same kind of ranking residual.

4.7 Predictor-defined peer calibration

A preliminary sensitivity analysis grouped institutions by observed World Rank. That design was rejected because using the outcome to define the comparison group creates circularity. In the final design, component indicators are used to predict expected January and July positions out of fold with gradient boosting and Extra Trees. The expected positions are then divided into quintiles. Transition residuals are recalibrated within expected-January quintiles, and July residuals within expected-July quintiles. Observed World Rank is never used to form the final peer groups.

The peer definition was frozen once this circularity correction was made. In the standard-university cohort, expected-January group sizes were 604, 603, 604, 603, and 604; expected-July group sizes were 604, 603, 603, 603, and 604. Favorable sensitivity was evaluated for standard and broad-canonical cohorts, whereas the unfavorable extension used the complete standard-university cohort.

4.8 Ranking-Integrity Audit Score for favorable candidates

A Ranking-Integrity Audit Score (RIAS) was used only to prioritize the 115 favorable Stage-1 candidates. It combines transition strength, current cross-sectional strength and effect, and historical break strength, with a penalty for historical persistence. Let T , C , and H denote the clipped components:

$$T = \text{clip}(z_T / 6, 0, 1) \quad (8)$$

$$C = \text{sqrt}[\text{clip}(z_C / 6, 0, 1) * \text{clip}(e_C / 0.10, 0, 1)] \quad (9)$$

$$H = \text{clip}(z_H / 8, 0, 1) \quad (10)$$

$$RIAS = ((T + C + H) / 3) * (1 - 0.5 * \text{persistence}) \quad (11)$$

The score orders audit priority within that favorable candidate set. It is not interpreted as a probability of manipulation. The persistence coefficient of 0.5 was retained after a sensitivity analysis in which the leading cases were stable across penalty strengths from 0 to 1. RIAS was not extended retrospectively to the unfavorable pathway after its results were known.

4.9 Synthetic bidirectional validation

There is no ground-truth dataset labeling real ranking manipulation, so validation uses controlled perturbations rather than purported fraud labels. The benchmark uses the 3,017 complete standard universities in July 2026. Prevalence is set to 1%, 5%, and 10%; nominal strengths are 0.02, 0.05, 0.10, and 0.20; and each scenario is repeated 20 times. Three interventions are considered: additive displacement, multiplicative displacement, and a rank-valid reranking design that perturbs latent outcome positions and then reranks the full cohort. The $3 \times 3 \times 4 \times 20$ design produces 720 synthetic datasets in each direction.

For favorable injections, selected outcomes are moved toward better percentile positions while components remain unchanged. Unfavorable injections move outcomes toward worse positions. Models are retrained on each contaminated synthetic outcome using five-fold out-of-fold prediction. Realized displacement is recorded because the same nominal strength produces different actual movements across designs and near percentile boundaries. The primary rule—model consensus, $z > 3$, and cross-sectional effect > 0.02 —was fixed before the final benchmark. Threshold-sensitivity results are descriptive and were not used to maximize F1 and redefine the detector.

4.10 External contextual validation

External review is deliberately last. For the small set of surviving cases, official institutional sources and independent ranking information were checked for changes that could plausibly alter digital visibility, research activity, institutional scale, or domain structure. This step is qualitative because provider-side scores, historical referring-domain series, and calculation logs are unavailable. Its role is to challenge the internal signal with plausible context, not to convert the statistical analysis into a causal claim.

4.11 Evaluation metrics and reproducibility

Synthetic performance is summarized by precision, recall, F1, false-positive rate (FPR), specificity, and AUROC. Precision and FPR receive particular attention because the intended use is triage: a detector that floods the reviewer with false alarms has little practical value even if its overall discrimination is strong. The analysis was implemented in Python with pandas, NumPy,

SciPy, and scikit-learn (Pedregosa et al., 2011). File hashes are reported below to make the verified inputs unambiguous.

Processed July 2026 SHA-256: b7444bad35bdc88a56ffbd285185d4fb83c5557f57298b36a4b097c84714525f

Raw July 2026 crawl SHA-256: 0066bb8b4a2a9c70afd5706a318ffc213c9a38e70a5474d9e873b89d157a867a

5. Results

5.1 Transition and July cross-sectional fit

The transition models captured a substantial, but far from complete, share of January-to-July movement. Huber regression achieved out-of-fold $R^2 = 0.4632$, $MAE = 0.0826$, and $RMSE = 0.1134$. Histogram-based gradient boosting improved the fit to $R^2 = 0.5328$, $MAE = 0.0746$, and $RMSE = 0.1058$. At $z > 3$, 150 favorable cases were identified by Huber and 177 by gradient boosting; 115 institutions satisfied both models and formed the favorable Stage-1 candidate set.

The July cross-sectional mapping was much tighter. Gradient boosting achieved $R^2 = 0.9981$, $MAE = 0.00879$, and $RMSE = 0.01258$, while Extra Trees achieved $R^2 = 0.9987$, $MAE = 0.00551$, and $RMSE = 0.01040$. The two models agreed on 38 favorable $z > 3$ cases; 23 of those also exceeded the 0.02 effect requirement in both residuals. Seven institutions appeared in both the 115-member transition set and the 38-member July consensus set, and five remained after the effect requirement. These are screening counts, not estimates of how common misconduct is.

Layer / model	R^2	MAE	RMSE	Directional screening result
Transition: Huber	0.4632	0.0826	0.1134	150 favorable $z > 3$
Transition: gradient boosting	0.5328	0.0746	0.1058	177 favorable $z > 3$
Transition consensus	-	-	-	115 favorable cases
July: gradient boosting	0.9981	0.0088	0.0126	117 favorable $z > 3$
July: Extra Trees	0.9987	0.0055	0.0104	214 favorable $z > 3$
July model consensus	-	-	-	38 at $z > 3$; 23 with effect > 0.02

Table 2. Full-cohort model performance and favorable screening counts.

5.2 Historical trajectories and institution type

Historical scoring changed the meaning of most short-term signals. Among the 115 favorable transition candidates, 107 were transition-only, four were persistent favorable patterns, three were sudden favorable breaks, and one was a current favorable anomaly. In other words, a large January-to-July residual usually did not imply a historically unusual current position.

Institution-type re-estimation also mattered. Several globally prominent anomalies occurred among university systems, business schools, and art or design institutions. Their interpretation weakened or changed when models were restricted to more comparable institutional forms. This was not a minor sensitivity result: it showed that structural heterogeneity can generate large residuals even when the model fit is otherwise excellent.

5.3 Predictor-defined peer robustness

Peer calibration reduced the favorable set sharply. Of the original 115 candidates, 92 received no support after predictor-defined peer recalibration, 20 retained transition-only support, two retained two-layer support, and one retained cross-sectional-only support. Gwangju University and the University of Tennessee Martin were the two standard universities that passed both peer-calibrated layers. Paul D. Camp Community College retained only the current-position signal in the broad-canonical analysis.

5.4 Worse-than-expected pathway

The mirrored standard-university analysis produced eight transition-consensus cases whose improvement was less than expected and 38 July cases whose current position was worse than expected. Only Middle East University Jordan and Innopolis University passed both layers before peer calibration. The transition Huber model achieved $R^2 = 0.3677$, $MAE = 0.0731$, and $RMSE = 0.1016$; transition gradient boosting achieved $R^2 = 0.4887$, $MAE = 0.0619$, and $RMSE = 0.0913$. July gradient boosting and Extra Trees achieved $R^2 = 0.9979$ and 0.9981 , respectively, with $MAE = 0.00878$ and 0.00675 .

The historical profiles separated the two cases. Middle East University Jordan had 11 prior observations and no earlier under-expected $z > 3$ event; its July 2026 historical z reached 25.094, and its break strength relative to prior median behavior was 25.701. Innopolis University had one

earlier extreme historical deviation and a current July historical z of 2.984, just below the threshold. After predictor-defined peer recalibration across all 3,017 standard universities, only one institution retained unfavorable transition support, 16 retained unfavorable current-position support, and Middle East University Jordan alone retained both. Its peer-calibrated transition minimum z was 3.418, current minimum z 20.797, and minimum effect 0.205. Innopolis fell to 2.784 and 2.313 in the two peer-calibrated layers.

5.5 Final bidirectional case matrix

Institution	Direction	Trajectory	Transition z	Current z	Effect	Peer 2-layer
Gwangju University	Better	Transition-only; novel context	4.2286	3.5816	0.0303	Yes
University of Tennessee Martin	Better	Persistent favorable	3.6419	3.9318	0.0331	Yes
Paul D. Camp Community College	Better	Transition-only	2.5498	3.2281	0.0373	No; cross-only
Middle East University Jordan	Worse	Sudden under-expected break	3.4180	20.7974	0.2050	Yes

Table 3. Final case matrix after predictor-defined peer calibration.

Gwangju University is the primary novel better-than-expected case: its historical persistence is zero and both peer-calibrated signals remain above threshold. UT Martin also passes both contemporary layers, but historical persistence of 0.75 makes it a recurring structural comparator rather than a new break. Paul D. Camp fails the final transition peer threshold ($z = 2.550$) while retaining a current-position signal ($z = 3.228$; effect = 0.0373), so it is kept as a secondary cross-sectional case. Middle East University Jordan is the primary worse-than-expected case and has by far the largest peer-calibrated current discrepancy in the final set.

5.6 Synthetic validation

Direction	Precision	Recall	F1	FPR	Specificity	AUROC
Better than expected	0.7198	0.6300	0.6308	0.0039	0.9961	0.9547
Worse than expected	0.6141	0.6287	0.5801	0.0084	0.9916	0.9544

Table 4. Primary synthetic-detector performance averaged across designs, prevalence levels, and strengths (consensus, $z > 3$, effect > 0.02).

The two directions had almost the same recall and AUROC, but they were not equally clean. Favorable perturbations produced precision 0.7198 and FPR 0.0039; unfavorable perturbations produced precision 0.6141 and FPR 0.0084. The unfavorable pathway therefore recovered a similar share of injected cases while generating more false positives.

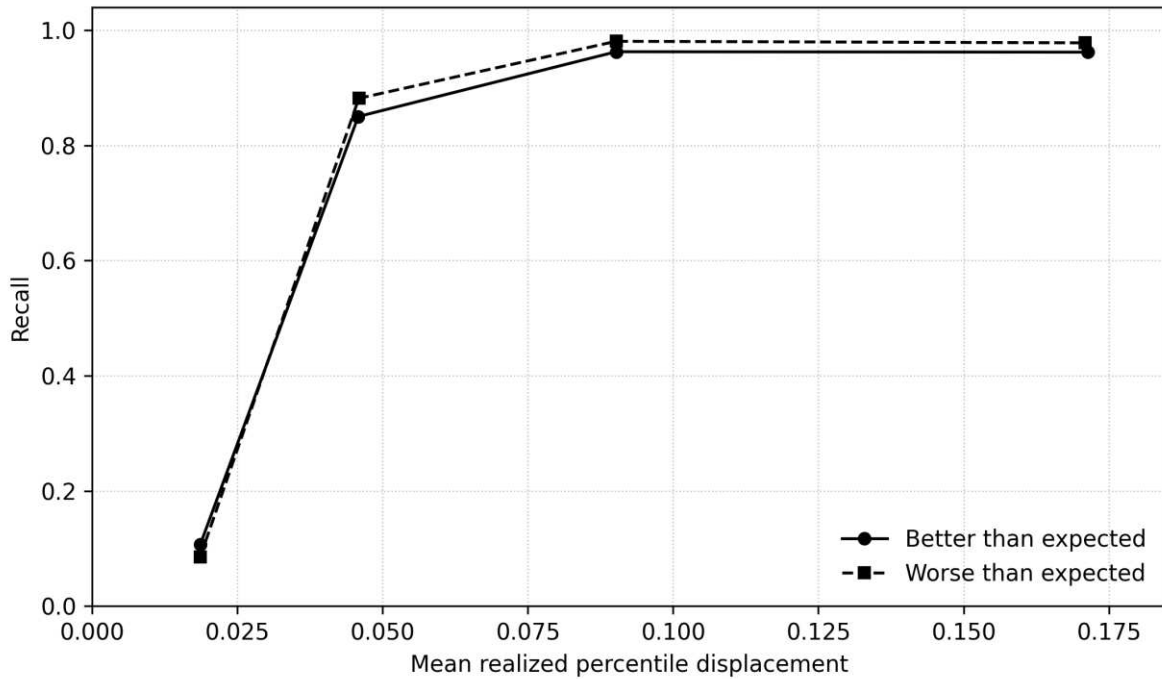


Figure 2. Recall versus realized effect under the rank-valid reranking intervention. Thresholds are fixed at model consensus, $z > 3$, and effect > 0.02 .

Effect size dominated detectability. In the rank-valid favorable benchmark, recall rose from 0.107 at a realized displacement of about 0.019 to 0.850 at 0.046 and approximately 0.963 at

0.090 and 0.171. The unfavorable benchmark showed the same pattern: recall increased from 0.085 at 0.019 to 0.882 at 0.046, 0.981 at 0.090, and 0.978 at 0.171. Spearman correlations between realized effect and recall were 0.855 for favorable injections and 0.924 for unfavorable injections.

5.7 Threshold sensitivity

The sensitivity analysis supports the prespecified $z = 3$ rule as a conservative compromise rather than an optimized optimum. In the favorable benchmark, reducing z from 3 to 2.5 increased recall from 0.630 to 0.668 but also increased FPR from 0.0039 to 0.0055; raising z to 3.5 reduced FPR to 0.0025 while lowering recall to 0.589. The unfavorable benchmark showed a larger false-positive trade-off: $z = 2.5$ produced recall 0.666 and FPR 0.0127, while $z = 3.5$ produced recall 0.585 and FPR 0.0062. Raising the effect threshold from 0.02 to 0.05 also reduced recall sharply in both directions. No threshold was changed after these results were observed.

6. Case-Based External Context

6.1 Gwangju University

Gwangju University remained above threshold in both peer-calibrated favorable layers, with zero historical persistence. Official university releases document several digital and institutional changes during 2025. The university introduced an AWS-linked AI-platform course for incoming students in March, launched the generative-AI academic support service Hi-Lighty across its website, portal, admissions, graduate, international, and employment sites in May, and reported an A grade in the 2025 University Innovation Support Project evaluation in August (Gwangju University, 2025a, 2025b, 2025c). These developments make increased digital activity plausible. They do not, however, quantify the change in external referring domains or reconstruct the Webometrics input data. The favorable discrepancy therefore remains internally robust but only partly explained by the external evidence available.

6.2 University of Tennessee Martin

UT Martin presents a different pattern. Its two-layer favorable signal survives the standard-university and predictor-peer checks, yet its historical persistence is 0.75. Official institutional

information describes more than 8,000 students, more than 50,000 alumni, seven regional centers, and a comprehensive online program (University of Tennessee at Martin Alumni, n.d.). That distributed regional and digital mission is consistent with a recurring visibility-favorable profile. The case is therefore useful as a structural comparator: strong in the current models, but not new in the historical record.

6.3 Paul D. Camp Community College

Paul D. Camp Community College illustrates why the peer transition test is useful. The institution retains a favorable current-position discrepancy but falls below the transition threshold after broad-canonical peer calibration. Its official profile describes two campuses, a center in Smithfield, extensive online study, workforce services, a Nursing and Allied Health Professions Center, and a Workforce Trades and Innovation Center opened in 2025 (Paul D. Camp Community College, n.d.-a, n.d.-b). Such a public-service and multi-site footprint can create a web presence that differs from conventional research universities. Without peer-calibrated transition support, the case is retained only as a secondary cross-sectional example.

6.4 Middle East University Jordan

Middle East University Jordan is the sole unfavorable case that survives both peer-calibrated layers. Its current discrepancy is large (minimum $z = 20.797$; effect = 0.205), and none of its 11 prior observations exceeded the historical unfavorable threshold. External sources point to change rather than a broad institutional collapse. The university released a redesigned official website in 2025, describing it as a technically integrated platform for academic and administrative services (Middle East University, 2025a). QS lists the university in the 1001-1200 band for both the 2026 and 2027 World University Rankings, and the university reported 91st place in the 2026 QS Arab Region ranking (QS Quacquarelli Symonds, n.d.; Middle East University, 2025b). These sources do not explain the Webometrics movement quantitatively. A website redesign could affect indexing, redirects, domain consolidation, or link structure, but no historical backlink series was available to test that mechanism.

Case	Internal interpretation	External context	Final audit role
------	-------------------------	------------------	------------------

Case	Internal interpretation	External context	Final audit role
Gwangju University	Novel two-layer better-than-expected	AI/digital expansion; no quantified backlink reconciliation	Primary favorable candidate
UT Martin	Persistent two-layer better-than-expected	Regional centers and broad online reach	Persistent structural comparator
Paul D. Camp CC	Peer cross-only better-than-expected	Workforce, online, and multi-center footprint	Secondary cross-sectional case
Middle East University Jordan	Novel two-layer worse-than-expected	2025 website redesign; no comparable external decline	Primary unfavorable candidate

Table 5. Contextual review of the final case set.

7. Discussion

7.1 What the audit is actually measuring

The clearest result is that a ranking can be audited without pretending that the analyst knows the institution's "true" quality. The models reproduce the empirical relationship between Webometrics component ranks and World Rank. A residual therefore has a limited, testable meaning: the published final position is unusual relative to the pattern seen among comparable observations. That is useful for quality control, but it is a different claim from saying that a university deserves another rank.

This distinction is especially important in a dataset with an almost deterministic July mapping. R^2 values near 0.998 might appear to validate the ranking itself. They do not. They show that public component ranks contain enough information to reproduce most of the ordering in the observed cohort. The remaining residuals are interesting precisely because they sit outside that dominant mapping. The audit focuses on those residuals and then asks whether they survive several attempts to explain them away.

7.2 Direction changes the interpretation

Treating both tails as one absolute outlier score would have hidden an important asymmetry. Gwangju University and Middle East University Jordan are both internally unusual, yet they

represent opposite questions. One appears better than expected from the modeled relationship; the other appears worse. Preserving the sign prevents the analysis from drifting into a narrative that only looks for apparent beneficiaries of ranking inconsistency.

The synthetic benchmark reinforces this point. Recall and AUROC are almost identical across directions, but precision is lower and FPR higher for the unfavorable pathway. The detector is therefore not perfectly symmetric even though the design is. Percentile boundaries, residual shape, and the dense lower part of the ranking are plausible contributors. Rather than retuning one side until the numbers match, the difference is reported as a property of the data and detector.

7.3 Comparability is not a cosmetic sensitivity check

One of the more consequential findings came from cases that disappeared after institution-type or peer controls. University systems, business schools, art schools, and community colleges can occupy unusual locations in the component-rank space for legitimate reasons. A high residual from a global model can therefore be a statement about institutional form rather than ranking integrity. Re-estimating within standard universities and then within predictor-defined expected-position groups removed most of the original favorable candidates. The reduction from 115 candidates to two peer-supported two-layer standard-university cases is evidence that comparability filters are doing substantive work.

The peer construction itself also matters. Grouping by observed World Rank initially seemed intuitive, but it uses the outcome to define the benchmark against which the outcome is judged. Rejecting that design and using expected positions instead was an important correction. It makes the final comparison harder to game analytically and avoids circular support for a candidate simply because it already occupies a particular observed rank band.

7.4 Historical persistence changes what a strong residual means

Gwangju University and UT Martin have similar contemporary status: both survive the two peer-calibrated favorable layers. Their histories tell different stories. Gwangju has zero favorable persistence in the available prior editions; UT Martin has persistence of 0.75. A single cross-section would place them side by side. The historical layer instead separates a new regime from a

recurring structural pattern. That distinction is central to audit priority because a persistent residual may reflect a stable institutional characteristic that the global model never fully captures.

Middle East University Jordan provides the same lesson from the opposite direction. Its July 2026 break follows 11 prior observations without an unfavorable $z > 3$ event. The combination of novelty, peer robustness, and a large current effect makes it a stronger audit candidate than Innopolis, whose history already contained an extreme deviation and whose current peer-calibrated z-scores fell below threshold.

7.5 What the synthetic benchmark adds

Synthetic perturbation cannot tell us how often real rankings are wrong, but it answers a more modest question cleanly: when the outcome is deliberately moved away from unchanged component information, does the detector find it? For moderate and large perturbations, the answer is usually yes. Very small displacements near two percentile points are often missed, particularly at low prevalence. That behavior is compatible with the intended role of the method. An audit screen does not need to flag every tiny discrepancy; it needs to prioritize larger, reproducible departures without creating an unmanageable false-positive burden.

The benchmark also shows why consensus was retained. Extra Trees alone is more sensitive but produces substantially more false positives, particularly in the unfavorable direction. Requiring agreement with gradient boosting reduces recall somewhat while making the candidate set more manageable. The choice favors review efficiency over exhaustive detection, which is appropriate for a triage setting.

7.6 Relationship to ranking-gaming and research-integrity research

The framework addresses a different evidence layer from studies of citation manipulation, retractions, authorship concentration, or ranking-gaming practices. Meho (2025) and Nazarovets (2024) examine bibliometric signals that may reflect research-integrity risk; Bayanbayeva (2026) documents strategic and manipulative responses to ranking incentives. An internally unexpected Webometrics position can arise without any of those behaviors, and those behaviors can occur without creating an unusual Webometrics residual. Combining the evidence in future work could be valuable, but the channels should remain separate enough to avoid counting the same underlying phenomenon twice.

8. Limitations and Threats to Validity

The most important limitation is the absence of real ground-truth labels. Precision and recall in this paper refer to artificial perturbations whose locations are known by construction. They do not estimate the probability that a real institution has manipulated a ranking. The synthetic benchmark tests detector behavior under controlled inconsistency, not human or institutional intent.

A second limitation comes from the public data themselves. Webometrics publishes component ranks rather than the underlying normalized component scores, so the published 50/10/40 weights cannot be reconstructed exactly from the tables used here. The learned mapping is consequently empirical. It can identify rank-component combinations that are unusual in the observed cohort, but it cannot reproduce provider-side calculations. Ties and repeated lower-end component values add further coarsening.

The historical panel is also incomplete and uneven. Institutions have different numbers of observations, and the collected profile histories contain July editions only for 2025 and 2026. Persistence is therefore calculated over the history actually available to each institution rather than over a balanced semiannual panel. The exact January 2016/January 2018 duplication demonstrates that historical web data can contain archival or provider artifacts; removing one known duplicate does not guarantee that every artifact has been found.

Comparability controls reduce, but do not eliminate, institutional heterogeneity. The entity taxonomy is name-based and can misclassify unusual organizations. Predictor-defined peer groups are better than observed-rank bands for this purpose, yet they still depend on the same three published components and on model specification. Likewise, Huber, gradient boosting, and Extra Trees are complementary model classes rather than independent evidence sources.

Finally, external validation remains contextual. Historical referring-domain counts, raw Webometrics scores, and provider calculation logs were unavailable. A website redesign or institutional expansion can therefore be identified as a plausible explanation without being established as the cause of a rank movement. The case studies should be read as examples of how to interrogate an audit signal, not as adjudications of provider or institutional behavior.

9. Practical and Research Implications

For ranking providers, the method could operate as a pre-publication quality-control layer. A provider with access to raw normalized scores, crawl diagnostics, and domain histories could implement a stronger version of the same logic and investigate large residuals before release. Score-level modeling would also avoid the information loss inherent in public rank positions.

Universities could use a similar audit for self-checking. An unexpectedly unfavorable position may motivate review of domain consolidation, redirects, repository indexing, or data completeness. A favorable discrepancy may prompt documentation of legitimate digital expansion or institutional structure. In either direction, the value lies in identifying where an explanation is needed.

For researchers, the bidirectional result suggests a broader design principle. Audits of evaluative systems should examine both apparent beneficiaries and apparent disadvantaged cases. If only one tail is studied, ordinary model mismatch can be mistaken for preferential inconsistency. Directional residuals, temporal context, and non-circular peer construction offer a practical baseline for future work on ranking integrity.

10. Conclusion

This paper presents a bidirectional audit of the July 2026 Webometrics university ranking using verified current data, profile-level history, directional residual models, robust consensus screening, institution-type controls, predictor-defined peers, and synthetic stress tests. The workflow deliberately narrows the candidate set as additional context is added: 115 favorable transition-consensus candidates were reduced to two standard-university cases with peer-calibrated two-layer support, while the unfavorable pathway ultimately retained one two-layer case.

The cases also show why residual magnitude alone is insufficient. Gwangju University is a novel favorable candidate; UT Martin is a strong but historically persistent comparator; Paul D. Camp Community College is a secondary cross-sectional case after peer calibration; and Middle East University Jordan is a novel unfavorable candidate. None of these labels establishes misconduct or provider error. They identify different forms of internal inconsistency that remain after the stated checks.

Synthetic validation adds a final piece of evidence. The fixed detector recovered moderate and large perturbations reliably in both directions, with similar recall and AUROC but a higher false-positive rate for unfavorable anomalies. Taken together, the results support ranking-integrity auditing as a conservative form of triage: first identify where the published outcome is internally unexpected, then use history, comparability, and external context to decide which cases deserve closer examination.

Acknowledgments

The author thanks Dr. Tarek Barhoum for academic supervision, methodological discussion, and critical review of the research design.

Declarations

Ethics and responsible interpretation

The study analyzes publicly accessible institutional ranking information and does not involve human participants or private personal data. Institution-level findings are reported as discrepancies, candidates, or structural comparators. Statistical signals are not used as proof of fraud, manipulation, misconduct, or provider error.

Conflict of interest

The author declares no conflict of interest related to this study.

Funding

No external funding was reported for this study.

Data availability

The study is based on publicly accessible Webometrics ranking and profile information together with derived research datasets. Processed datasets and analysis scripts are retained by the author and can be deposited in an institutional or public repository as part of the submission and reproducibility package.

Code availability

The analysis was implemented in Python. The complete analysis scripts are retained with the study materials and can be released with the reproducibility package at submission.

Author contributions

Eng. Mina Ibrahim: conceptualization, data collection, software, methodology, validation, formal analysis, investigation, visualization, and writing - original draft and revision. Dr. Tarek Barhoum is acknowledged as academic supervisor and is not listed as an author unless authorship is formally agreed before submission.

References

- Aguillo, I. F., Bar-Ilan, J., Levene, M., & Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, 85(1), 243-256. doi:10.1007/s11192-010-0190-z
- Aguillo, I. F., Ortega, J. L., & Fernández, M. (2008). Webometric ranking of world universities: Introduction, methodology, and future developments. *Higher Education in Europe*, 33(2-3), 233-244. doi:10.1080/03797720802254031
- Bayanbayeva, A. (2026). Strategic response or gaming the rankings? Unravelling the strategies behind global university rankings manipulation in the higher education context of Kazakhstan. *Higher Education*, 91, 2099-2117. doi:10.1007/s10734-025-01512-1
- Billaut, J.-C., Bouyssou, D., & Vincke, P. (2010). Should you believe in the Shanghai ranking? An MCDM view. *Scientometrics*, 84(1), 237-263. doi:10.1007/s11192-009-0115-x
- Forthmann, B., Doeblner, P., & Mutz, R. (2024). Why summing up bibliometric indicators does not justify a composite indicator. *Scientometrics*, 129, 7475-7499. doi:10.1007/s11192-024-05194-x
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), 1189-1232. doi:10.1214/aos/1013203451

- Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, 63(1), 3-42. doi:10.1007/s10994-006-6226-1
- Gwangju University. (2025a, March 26). Gwangju University partners with AWS to introduce AI-platform-based instruction [University news release, Korean].
- Gwangju University. (2025b, May 23). Ask Hi-Lighty AI about Gwangju University [University news release, Korean].
- Gwangju University. (2025c, August 7). 2025 University Innovation Support Project performance evaluation: Grade A [University news release, Korean].
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1), 73-101. doi:10.1214/aoms/1177703732
- Ioannidis, J. P. A., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., Contopoulos-Ioannidis, D. G., & Liberopoulos, G. (2007). International ranking systems for universities and institutions: A critical appraisal. *BMC Medicine*, 5, Article 30. doi:10.1186/1741-7015-5-30
- Meho, L. I. (2025). Gaming the metrics: Bibliometric anomalies in global university rankings and the research integrity risk index (RI2). *Scientometrics*, 130, 6683-6726. doi:10.1007/s11192-025-05480-2
- Middle East University. (2025a, March 13). Middle East University announces the release of the new version of its official website.
- Middle East University. (2025b, October 21). Middle East University secures the 91st regional position in the QS Ranking.

- Nazarovets, S. (2024). Visualization of rank-citation curves for fast detection of h-index anomalies in university metrics. *Scientometrics*, 129(1), 705-711. doi:10.1007/s11192-023-04886-0
- Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management*, 45(2), 272-279. doi:10.1016/j.ipm.2008.10.001
- Paul D. Camp Community College. (n.d.-a). About CAMP. (Retrieved August 19, 2026)
- Paul D. Camp Community College. (n.d.-b). History. (Retrieved August 19, 2026)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- QS Quacquarelli Symonds. (n.d.). Middle East University Jordan. (Retrieved August 19, 2026)
- Robinson-García, N., Torres-Salinas, D., Herrera-Viedma, E., & Docampo, D. (2019). Mining university rankings: Publication output and citation impact as their basis. *Research Evaluation*, 28(3), 232-240. doi:10.1093/reseval/rvz014
- Rousseeuw, P. J., & Hubert, M. (2018). Anomaly detection by robust statistics. *WIREs Data Mining and Knowledge Discovery*, 8(2), e1236. doi:10.1002/widm.1236
- Saisana, M., d'Hombres, B., & Saltelli, A. (2011). Rickety numbers: Volatility of university rankings and policy implications. *Research Policy*, 40(1), 165-177. doi:10.1016/j.respol.2010.09.003
- Thamm, M., & Mayr, P. (2011). Comparing webometric with web-independent rankings: A case study with German universities [Preprint]. *arXiv*.arXiv:1105.2443

University of Tennessee at Martin Alumni. (n.d.). UT Martin 125th celebration: 1900-2025.

(Retrieved August 19, 2026)

van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62(1), 133-143.

doi:10.1007/s11192-005-0008-6

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12), 2419-2432.

doi:10.1002/asi.22708

Webometrics.org. (n.d.-a). Methodology. (Retrieved August 19, 2026)

Webometrics.org. (n.d.-b). Objectives. (Retrieved August 19, 2026)