A Reproducible Stability Audit of Multi-Criteria ETF Ranking for FinTech Decision Support

Jiaqi Yao

Business School, University of Dundee, Dundee, United Kingdom
Email: yaojiaqi052505@163.com

Abstract Digital-investment decision support can reduce heterogeneous exchange-traded fund (ETF) risk–return signals to a point ranking, but that order can hide uncertainty from sampling, normalization, criterion choice, market regimes, and the alternative set. We present a retrospective, reproducibility-focused audit of 60 ETFs using public adjusted prices from 2015 through 2025. Six risk–return criteria are evaluated by Entropy-TOPSIS and diagnostic comparators. The Primary case uses 1,000 moving-block bootstrap samples, rank intervals, top-five probabilities, and specification, temporal, weight, criterion, and universe stress tests. Entropy-TOPSIS obtains mean bootstrap Spearman agreement of 0.8902 and a mean 95% rank-interval width of 5.4167 positions. The main grouping audit uses 20 outcome-blind partitions of the same 60-ticker pool: five disjoint 12-ETF panels per partition and 500 draws per panel. Its auxiliary matched worst-gap summary averages 0.3999. In disjoint five-year windows, the same audit yields means of 0.3661 and 0.4319, while 100 early-versus-late full-rank correlations average 0.4026. These results characterize internal regrouping in one USD pool; they neither establish method superiority nor provide a personalized investment recommendation. The contribution is a reproducible audit and reporting protocol around standard ranking methods, not a new ranking algorithm or deployed platform.

Keywords FinTech decision support, exchange-traded funds, multi-criteria decision analysis, rank stability, moving-block bootstrap

1. Introduction

FinTech decision support can use an ETF ranking to organize heterogeneous risk–return information for digital-investment users. Exchange-traded funds offer liquid exposure to equities, bonds, commodities, and real estate, but comparing funds across asset classes requires several incommensurate objectives. Return, volatility, drawdown, reward-to-risk ratios, and tail loss can point in different directions. Earlier ETF studies therefore combine performance and risk measures rather than relying on raw return alone [1, 2]. Multi-criteria decision analysis (MCDA) supplies a transparent way to convert such a criterion matrix into an ordering, and TOPSIS remains attractive because its score has a direct ideal-versus-anti-ideal interpretation.

A point ranking can nevertheless appear more precise than its evidence. ETF criteria are estimated from the same finite return history, while rankings can change with normalization, objective weights, omitted criteria, market regimes, or the set of alternatives. Dependence among criteria may affect weight interpretability or the induced ordering under some constructions [3]; our correlation matrix diagnoses dependence but does not identify a causal bias magnitude. Work on alternative TOPSIS normalizations [4], time-indexed MCDA [5], and explicit reference levels [6] makes the same general point: a ranking is conditional on modeling choices and should be audited accordingly.

This retrospective case study explores grouping sensitivity under one disclosed protocol. Its primary evidence is an outcome-blind grouping rule and a fixed, nonoverlapping calendar-window replication. A secondary five-panel category illustration uses matched block indices and equal budgets. The Matched-Panel Stability Gap (MPSG) is an auxiliary bounded summary defined in Section 3.4; it is not a new ranking rule, and the inferential scope remains the observed histories.

The Primary panel was the initial 12-ETF case. After inspecting its results, we constructed four purposive extensions: Broad assets, Sector, International, and Fixed-alternative. The extensions were not preregistered and may

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

reflect panel-selection bias. Together the panels contain 60 unique ETFs over a common 2015–2025 window using six risk–return criteria and public adjusted prices. Entropy-TOPSIS is the focal standard method; equal-weight and CRITIC-TOPSIS are weighting diagnostics, VIKOR and a weighted-sum model are aggregation diagnostics, and PROMETHEE II is a supplementary implementation check.

This study makes three practical contributions. An outcome-blind partition rule places grouping sensitivity beside sampling, specification, time, and universe checks. Archived price snapshots, full return-to-rank recomputations, structured manifests, hashes, and exact commands support independent auditing. Twenty partitions and two disjoint five-year windows quantify sample-specific grouping dispersion in a 60-ETF pool. The output supports uncertainty-aware reporting in FinTech decision support, not a new ranking method, deployed platform, or investment recommendation.

The rest of the paper reviews the relevant ETF, tail-risk, and MCDA literature, defines the evaluation protocol, reports the empirical results, and discusses their limitations.

2. Related Work

FinTech decision-support systems provide a broad setting for combining financial data with analytical models [7]. In digital investment, evidence on robo-advisory acceptance [8] and on social design effects on investor behavior [9] shows that users encounter ranking outputs through an interface. These studies motivate transparent reporting, but do not test whether an ETF rank remains stable when the analytical pipeline or candidate universe changes.

ETF evaluation spans performance measurement, portfolio construction, and tail-risk analysis. Sharifzadeh and Hojat compare ETFs with matched passive index funds using Sharpe ratios and risk-adjusted returns [1]. A recent study of Asian ETFs combines efficient-frontier analysis with CVaR, STARR, and related reward-to-risk measures [2]. Entropy has also been studied as a financial risk quantity and asset-pricing covariate [10]. These studies motivate a criterion set that contains both central performance and downside outcomes rather than a single mean-variance score.

Tail measures require special care because their observations are sparse and model-sensitive. Sequential portfolio tail-risk estimation emphasizes the uncertainty created by simulation and scenario allocation [11], while work on range Value at Risk formalizes how tail forecasts should be evaluated and compared [12]. Our setting is descriptive rather than predictive, but the lesson carries over: an estimated tail criterion should not be treated as error-free. We therefore resample the entire daily cross-section in blocks and recompute every criterion and rank.

MCDA research identifies several other sources of instability. Independent component analysis has been combined with TOPSIS to address dependent criteria [3]. Zolfani et al. show that the normalization rule can alter VIKOR and TOPSIS results [4]; a tensorial extension instead incorporates time-series features directly into MCDA [5]. TOPSIS-RAD targets rank reversal by fixing decision-maker reference levels rather than deriving both ideals solely from the observed alternative set [6].

Weight uncertainty has been studied through interpretable TOPSIS views [13], cone-based scalarizations [14], randomized Entropy–CRITIC weights [15], and global sensitivity [16]. They motivate diagnostic perturbations rather than treating one objective-weight vector as immutable preference evidence.

Hybrid subsorting uses two preference orders to resolve weakly differentiated rankings [17], reinforcing that aggregation agreement is an empirical quantity rather than an assumption.

We use CVaR modestly as one historical cost criterion. Our resampling target is the conventional Spearman and Kendall ordering, not Chatterjee's coefficient; this distinction matters because the ordinary bootstrap can be inconsistent for the latter statistic [18]. Moving blocks are used here to preserve serial and cross-sectional dependence, not as a claim that every rank functional admits the same bootstrap theory. Moving-block resampling has established uses for temporally dependent series [19]; our 10-day block is an empirical primary setting, bracketed by 5- and 20-day checks, rather than a claim of theoretically optimal block length.

Gupta et al.'s financial-sector TOPSIS comparison remains a close point-ranking application [20], whereas this audit recomputes full rank distributions. All underlying MCDA equations remain standard.

The bounded search was used to identify component-level antecedents, not to prove a literature gap. In the screened sample, those components occur in different research designs. Sielska's investment-fund study is the closest domain antecedent and compares TOPSIS, PROMETHEE, and weighted-sum stability under weight simulation [21]. Table 1 states the screened nearest-neighbor distinction directly. Agostinelli et al. are a transferability-design antecedent outside finance [22], not an ETF baseline.

Table 1. Nearest screened work. Full recomputation means returns through ranks; matched panels use common resampling indices. N/O denotes not observed in the screened sample.

Study	Domain	Return bootstrap	Full recomputation	Time stress	Matched panels	Artifacts
Sielska [21]	Funds	N/O	N/O	N/O	N/O	N/O
Basilio et al. [15]	Financial MCDA	N/O	N/O	N/O	N/O	N/O
Gupta et al. [20]	Financial sector	N/O	N/O	N/O	N/O	N/O
Agostinelli et al. [22]	Nonfinancial	N/O	N/O	Setups	N/O	N/O
This study	ETFs	Blocks	Yes	Disjoint windows	Yes	Indexed

Rank-reversal work [14, 6] and normalization studies [4] remain complementary antecedents.

The organizing distinction is the unit of evidence: prior studies modify or compare ranking methods, whereas this audit reports matched rank distributions after full-pipeline recomputation. This is a scope distinction, not a priority claim.

To make the antecedent boundary auditable, we ran a bounded metadata search across five sources through 12 July 2026; Supplementary Material Table S3 gives the queries, screening decisions, source failures, and comparison mappings. Table 1 therefore states an observed screened boundary rather than a systematic-review claim.

At the research-objective level, ETF performance studies estimate returns and tail risk, while financial MCDA applications usually seek a point order or a method comparison. This study instead audits stability under outcome-blind grouping, full-pipeline bootstrap, and rolling summaries. The contribution boundary is the one stated in the Introduction; no broader MCDA novelty is claimed here.

3. Stability-Audited Ranking Protocol

Figure 1 summarizes the protocol. A validated adjusted-price panel is converted into six criteria, oriented so that larger normalized values are preferred, and evaluated with six diagnostic MCDA configurations. The full pipeline is then rerun under sampling and specification perturbations. Every perturbation starts from returns or the criterion matrix; no synthetic price path is used in the primary data.

The reusable implementation unit is a stability-audit controller that invokes the standard normalization, entropy weighting, and TOPSIS functions repeatedly under sampling, specification, time, and universe perturbations. It is an experiment and reporting controller, not a seventh ranking method; all underlying MCDA equations remain unchanged.

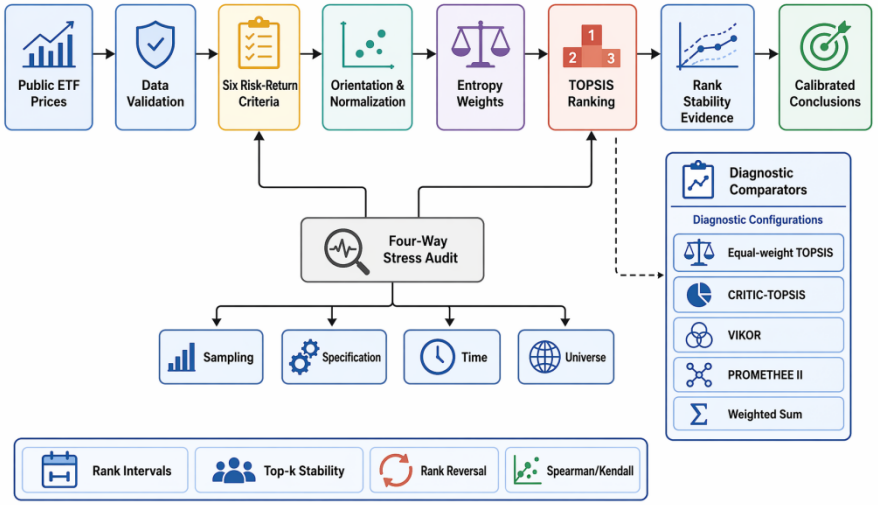

Figure 1. Stability-audit and uncertainty-reporting workflow. Public adjusted prices are validated before criterion construction and diagnostic MCDA evaluation. Sampling, specification, temporal, and universe perturbations produce rank distributions and agreement diagnostics. The workflow surrounds standard Entropy-TOPSIS; it is not a new scoring equation.

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

3.1 Criteria and decision matrix

For ETF i on trading day t, the log return is $r_{i,t}=\log(P_{i,t}/P_{i,t-1})$. Annualized return and volatility are $252\,\bar{r}_i$ and $\sqrt{252}s_i$, respectively. Maximum drawdown is the largest peak-to-trough loss of the normalized wealth path. With a zero risk-free rate, the Sharpe ratio divides annualized return by annualized volatility, and the Sortino ratio uses annualized downside deviation. The 95% CVaR loss is the negative mean simple return below the empirical fifth percentile. Return, Sharpe, and Sortino are benefit criteria; volatility, drawdown, and CVaR are cost criteria.

After reversing cost directions, the primary min–max normalization is

$$z_{ij}=\frac{y_{ij}-\min_i y_{ij}}{\max_i y_{ij}-\min_i y_{ij}},$$

where y_{ij} is the oriented value of ETF i on criterion j. Vector and primary min–max normalizations form a focused sensitivity check: entropy weights are calculated once from the min–max matrix and then held fixed while the shifted-vector matrix is scored. A z-score branch is omitted because its subsequent min–max rescaling is algebraically redundant for nonconstant criteria.

3.2 Entropy weights and TOPSIS score

Let $p_{ij}=z_{ij}/\sum_i z_{ij}$, where m denotes the number of ETFs. Entropy, diversification, and the objective criterion weight are

$$e_j=-\frac{1}{\log m}\sum_i p_{ij}\log p_{ij},\quad d_j=1-e_j,\quad w_j=\frac{d_j}{\sum_\ell d_\ell}.$$

Zero shares contribute zero to $p\log p$. For the weighted matrix $v_{ij}=w_j z_{ij}$, let v_j^+ and v_j^- be the column maximum and minimum. The distances and closeness score are

$$D_i^+=\sqrt{\sum_j (v_{ij}-v_j^+)^2},\quad D_i^-=\sqrt{\sum_j (v_{ij}-v_j^-)^2},$$

$$C_i=\frac{D_i^-}{D_i^+ + D_i^-}.$$

ETFs are ranked by decreasing C_i. Equal weights and CRITIC weights provide two weighting diagnostics. VIKOR and a weighted-sum model provide aggregation diagnostics, not a controlled causal decomposition of method effects.

3.3 Uncertainty and stress tests

The primary sampling analysis draws 1,000 moving-block bootstrap panels [19] with a 10-day block. Ten days is an engineering choice bracketed by the reported 3–30-day sensitivity grid, not an estimated optimum. Each sampled daily return vector retains cross-ETF dependence; prices, criteria, weights, scores, and ranks are rebuilt in every replication. Overlapping noncircular blocks start uniformly from valid indices, are sampled with replacement, concatenated, and truncated to the original return length. The first reconstructed price is one, so every replication preserves length without importing an observed price level. This is a full return-to-rank recomputation, not resampling of final scores. For each method we report mean Spearman and Kendall agreement with the full-sample ordering. ETF-level uncertainty is summarized by median rank, 2.5th and 97.5th percentiles, and top-three and top-five probabilities.

Here b denotes a bootstrap replication. Its agreement with the corresponding full-sample rank vector is summarized by Spearman's ρ_b and Kendall's τ_b. We average these coefficients over the 1,000 replications rather than pooling all ETF–replication pairs. The percentile interval for ETF i is $[Q_{0.025}(R_i),Q_{0.975}(R_i)]$, where R_i is its empirical bootstrap-rank distribution. Thus, method-level reproducibility and ETF-level rank uncertainty remain distinct summaries. ETF rank intervals are reported as bootstrap percentiles; no between-ETF interval-width contrast is used for population inference.

A conclusion is calibrated here only when its wording is conditioned on four observed outputs: percentile rank width, top-set inclusion probability, rank-correlation agreement, and pairwise reversal rate. A point rank is never promoted to a forecast. Wide intervals or weak temporal agreement trigger period-conditional wording even when the full-sample score is well separated.

The calendar-window replication was specified before its outcomes were computed: 2015–2019 is the early window, calendar 2020 is omitted as a gap, and 2021–2025 is the late window. The same 20 seeded symbol partitions are reused, but prices, criteria, weights, scores, ranks, and 500 block draws per panel are recomputed separately. This rule was added after the original full-sample analysis, so it is an independent-window replication rather than a preregistered or prospective test.

3.4 Matched-Panel Stability Gap

Let P denote the number of panels, B the number of matched moving-block draws, and ρ_{pb} Spearman agreement for panel $p\{1,\ldots\ldots P\}$ under draw $b\{1,\ldots\ldots B\}$. We define the replication-level transfer gap and its aggregate as

$$g_b=\max_p \rho_{pb}-\min_p \rho_{pb},\qquad \mathrm{MPSG}=\frac{1}{B}\sum_{b=1}^{B} g_b.$$

Because Spearman correlation lies in [-1,1], $0\leqslant \mathrm{MPSG}\leqslant 2$; zero requires equal panel agreement within every matched draw. Its percentiles describe draw-level gaps, not uncertainty for the mean. Common block indices hold the resampled dates fixed while panels change. The bootstrap remains conditional on observed histories and does not

support population inference.

Its matched interpretation differs from marginal summaries. For three matched draws (rows) across three panels (columns), consider

$$A=\begin{bmatrix}.2 & .2 & .2\\ .5 & .5 & .5\\ .8 & .8 & .8\end{bmatrix}, B=\begin{bmatrix}.2 & .5 & .8\\ .5 & .8 & .2\\ .8 & .2 & .5\end{bmatrix}.$$

Both arrays have identical panel-wise marginal distributions, means, variances, and Friedman statistic zero, yet MPSG is 0 versus 0.6. Its only special role is the worst simultaneous separation on the Spearman scale. It is extreme-sensitive and does not explain panel differences; smoother standard-deviation and pairwise-gap summaries are reported alongside it.

Specification checks use block lengths of 5 and 20 days, two fixed-entropy-weight normalization variants, 1,000 weight perturbations, 24 one-at-a-time weight changes, six leave-one-criterion-out variants, and two grouped ablations. Temporal checks cover three stages, four reported regimes, and eight rolling three-year windows. Alternative-set stress removes each ETF and each asset class in turn. Rank reversal, top-five Jaccard overlap, Spearman correlation, and Kendall correlation are computed against the full specification.

Random weights are drawn on the nonnegative unit simplex from a Dirichlet distribution with concentration $100w+1$, where w is the full-sample entropy vector. The 24 deterministic checks multiply one of six weights by 1 ± 0.1 or 1 ± 0.2 (Tex translation failed) and renormalize the vector. Random draws probe joint changes around the entropy solution; deterministic changes isolate each criterion and direction.

The comparator design is diagnostic rather than strictly matched. Equal-weight and CRITIC variants share min–max preprocessing and the TOPSIS distance rule. The weighted sum shares the oriented min–max matrix and entropy weights. VIKOR necessarily uses its defining best–worst normalization and compromise coefficient $v=0.5$, so its differences cannot be assigned to aggregation alone.

PROMETHEE II is an auxiliary implementation check and follows the complete net-flow ordering [23] with a linear zero-indifference preference on the [0,1] matrix, $P_j(a,b)=\max(z_{aj}-b_{aj},0)$; equivalently $q=0$ and $p=1$ for every criterion. Since $\max(x,0)-\max(-x,0)=x$, the resulting net flow is affine in the weighted-sum model; the two outputs are therefore an implementation-consistency check, not independent comparator evidence. This equivalence is the study's direct derivation under the displayed preference function. All fixed values are disclosed engineering choices, not universal defaults. Accordingly, the results are diagnostic reference outputs rather than a fairness-controlled method benchmark.

4. Empirical Evaluation

4.1 Data and implementation

The initial universe contains SPY, QQQ, IWM, EFA, EEM, TLT, IEF, LQD, HYG, GLD, VNQ, and DBC. Yahoo Finance's public chart endpoint supplied adjusted daily prices. The common panel spans 2 January 2015 through 31 December 2025, with 2,766 price rows and 2,765 return rows, no missing cells, and no duplicate dates. Raw JSON responses, request parameters, timestamps, validation results, and SHA-256 hashes are preserved. Supplementary Material Table S1 maps each main figure and table to an artifact, hash, and minimal command; Table S2 lists panel membership and Table S3 records search and screening evidence. The archive contains acquisition records for all 60 tickers and every stress-test row. All reported metrics were computed in remote, reproducible Python environments; run-specific package versions and the seed are recorded in the anonymous supplement. The supplement contains acquisition and validation records, ranking matrices, bootstrap samples, stress tables, panel definitions, and search decisions. These are anonymous supplementary routes inside the submitted archive, not public web addresses. Table 2 centralizes the input rule, replication budget, and stable reviewer route for each core experiment.

Table 2. Minimum reproducibility map inside the anonymous Supplementary Material.

Experiment	Input and rule	Core parameters	Stable reviewer route
Primary bootstrap	12 aligned ETFs; return-to-price reconstruction	1,000 draws; 10-day blocks	Table S1, command R1
Grouping audit	60 symbols; seeded permutation into panels	500 draws per generated panel	Table S1, command R2
Calendar replication	same groups; disjoint 2015–19/2021–25 windows	500 draws per panel-window	Table S1, command R3

4.2 Point ranking and objective weights

Table 3 reports the full-sample Entropy-TOPSIS ordering. HYG, LQD, and IEF occupy the first three positions, while EEM and VNQ appear at the bottom. These are historical, criterion-conditional positions rather than recommendations.

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

Table 3. Full-sample Entropy-TOPSIS ranking.

Rank	ETF	Score	Rank	ETF	Score
1	HYG	0.7202	7	EFA	0.4116
2	LQD	0.6565	8	TLT	0.3341
3	IEF	0.6466	9	DBC	0.2851
4	GLD	0.6387	10	IWM	0.2564
5	SPY	0.4880	11	EEM	0.2499
6	QQQ	0.4353	12	VNQ	0.2105

Figure 2 compares Entropy-TOPSIS with four diagnostic outputs: CRITIC- and equal-weight TOPSIS, VIKOR, and weighted sum. PROMETHEE II appears only in the supplement as an implementation check. The first three TOPSIS columns are matched weighting diagnostics; VIKOR and weighted sum are unmatched aggregation diagnostics, not benchmark evidence. Figure 3 helps explain why the ordering favors several defensive exposures: the entropy scheme assigns its two nearly equal largest weights to CVaR and annualized volatility, followed by annualized return.

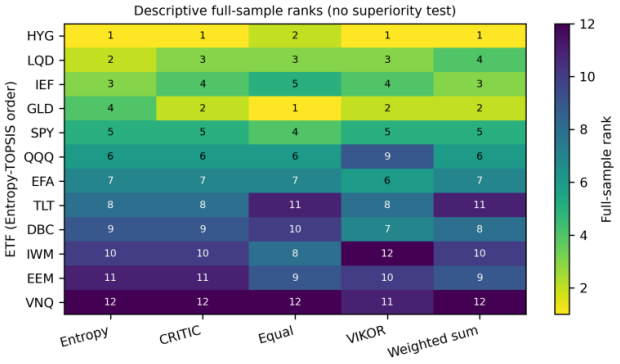

Figure 2. Primary-panel descriptive result: full-sample rank matrix for Entropy-TOPSIS and four diagnostic comparators; it is not a superiority test.

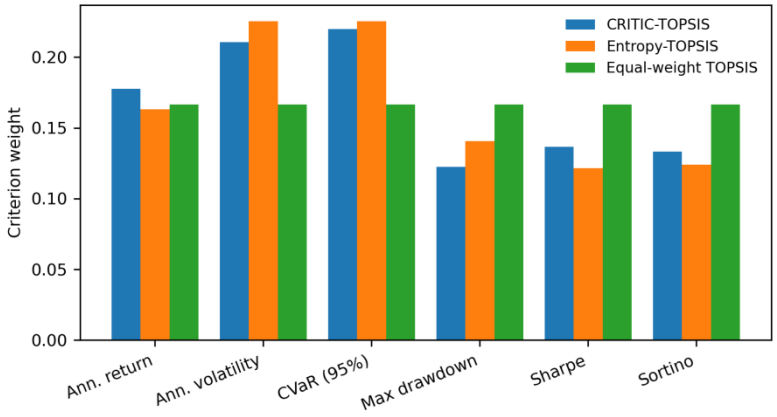

Figure 3. Diagnostic weights for annual return, annual volatility, maximum drawdown, Sharpe, Sortino, and 95% CVaR under Entropy, equal, and CRITIC schemes.

4.3 Bootstrap stability and method comparison

Table 4 reports five diagnostic outputs over 1,000 primary bootstrap replications. Mean Spearman agreement is 0.8902 for Entropy-TOPSIS, 0.8843 for CRITIC-TOPSIS, and 0.8224 for equal-weight TOPSIS. The paired Entropy-minus-CRITIC mean is 0.0059, with a bootstrap percentile interval of $[-0.1680, 0.1259]$. Every paired Entropy-minus-comparator interval crosses zero. The table therefore supplies within-panel descriptive robustness, not evidence that one method is superior. Across seeds 20260712–20260714, Entropy mean agreement is 0.8868–0.8902, mean interval width is 5.4167–5.5000, and the Entropy-minus-CRITIC mean is 0.0043–0.0059; every paired interval includes zero. The temporal and universe variants enumerate fixed slices and deletions and therefore do not consume a random seed.

Table 4. Primary-panel agreement distributions and paired contrasts. Parentheses give MC SE; brackets in the middle column are empirical percentiles.

Method	Mean ρ (MC SE)	Empirical 2.5–97.5% ρ	Empirical 2.5–97.5% Δ
Entropy-TOPSIS	.8902 (.0023)	[.6993,.9860]	reference
CRITIC-TOPSIS	.8843 (.0022)	[.7273,.9720]	[-.1680,.1259]
VIKOR	.8705 (.0043)	[.4754,.9860]	[-.1049,.2939]
Weighted sum	.8527 (.0027)	[.6502,.9720]	[-.0769,.1608]
Equal-weight TOPSIS	.8224 (.0029)	[.6082,.9512]	[-.0909,.2587]

Figure 4 is the principal ETF-level uncertainty evidence. HYG has median rank 1, a 95% interval of [1,5], and top-five probability 0.985. LQD and IEF have top-five probabilities of 0.952 and 0.970. Across all ETFs, the mean interval width is 5.4167 positions, which cautions against reading adjacent point ranks as cleanly separated.

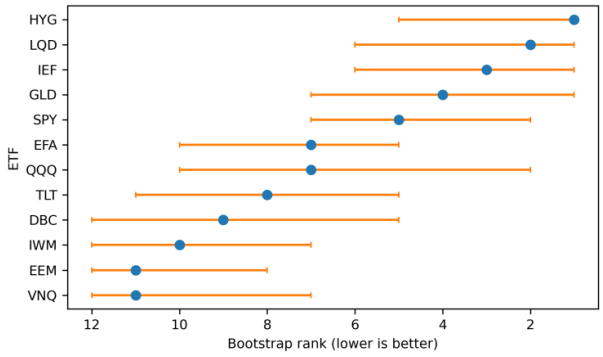

Figure 4. ETF-level uncertainty evidence: Entropy-TOPSIS median ranks and 95% bootstrap percentile intervals; lower rank is better.

4.4 Specification, time, and universe stress

The fixed-entropy-weight normalization sensitivity leaves the point ordering unchanged: entropy weights are calculated once from the primary min–max matrix, then held fixed while shifted-vector normalization is scored. The resulting Spearman agreement is 1.0000, the rank-reversal rate is 0, and top-five Jaccard overlap is 1.0000. The z-score branch is omitted because a later min–max rescaling would be algebraically redundant. Six criterion deletions retain mean Spearman agreement of 0.9650 (minimum 0.9091), and two grouped ablations retain 0.9441 (minimum 0.9231). One-at-a-time weight changes are also mild, with mean agreement 0.9948 and minimum 0.9790. Changing the bootstrap block from 10 days to 5 or 20 days lowers mean agreement only to 0.8778 and 0.8874.

Temporal diagnostics are reported descriptively, without an across-family population test. Three broad stages have mean Spearman agreement 0.8392, eight rolling windows have 0.8278, and the four market regimes average 0.7727 with a minimum of 0.6294. By contrast, leave-one-ETF and leave-one-asset-class tests average 0.9735 and 0.9818. The heat map in Figure 5 displays every audit family's mean diagnostics. The pandemic shock/recovery regime is the visible low point (ρ=0.6294), while removing TLT is the strongest single-ETF universe perturbation (ρ=0.8273).

Supplement S4 reports every variant row.Table 5 is the main robustness summary. The rows are not treated as exchangeable samples, so no formal across-family significance claim is made.

The diagnostics answer different questions: rank intervals localize ETF uncertainty, top-set probabilities address membership, correlations summarize global agreement, and reversal rates count changed pairwise decisions. They are an author-selected reporting set, not a necessary or sufficient set.

Finally, the supplementary criterion-correlation matrix is the dependence diagnostic. Because Sharpe and Sortino ratios share return and downside information with other criteria, the matrix helps readers assess how much apparently separate objectives overlap. Sharpe and Sortino form the visually strongest positive cluster, whereas cost criteria also move together. The entropy weights must therefore be read as marginal dispersion in this dependent matrix, not as six independent measures of economic importance.

Table 5. Primary-panel robustness summary against the full Entropy-TOPSIS order.

Family	Variants	Mean ρ	Minimum ρ
Fixed-weight normalization	2	1.0000	1.0000
Criterion deletion	8	.9598	.9091
Single-weight change	24	.9948	.9790
Temporal windows	15	.8154	.6294
ETF/class deletion	23	.9775	.8273

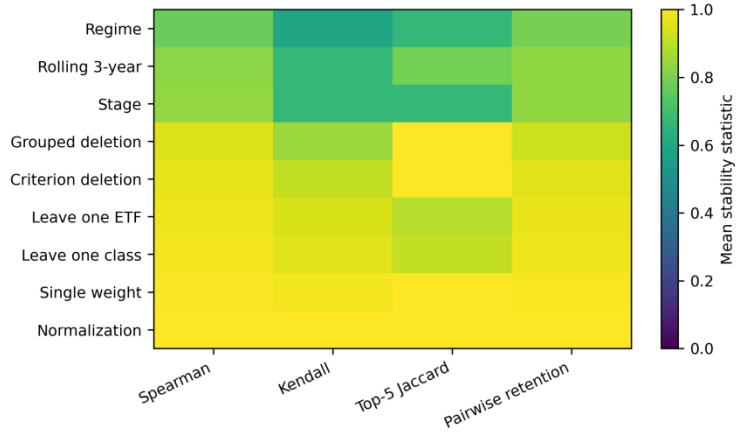

Figure 5. Ablation and robustness evidence relative to the full Entropy-TOPSIS order. Each cell is the family mean of the named agreement or pair-retention statistic; color increases from low to high stability.

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

Under the linear zero-threshold preference used here, PROMETHEE II and the weighted-sum model are rank-equivalent in all 1,000 replications; they are not independent evidence. Supplement S4 reports method standard deviations and percentiles. Entropy- and CRITIC-TOPSIS remain a paired near tie, so the evidence concerns audit behavior rather than weighting superiority.

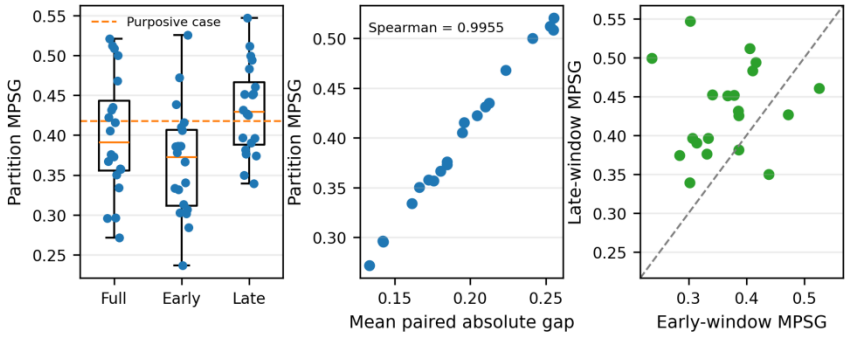

Figure 6. Main result: outcome-blind grouping sensitivity within the observed 60-ETF pool. Each of 20 partitions contains five disjoint 12-ETF panels with 500 draws per panel. Left: full-period and disjoint-window MPSG distributions; the dashed line is the purposive case. Center: full-period MPSG versus mean pairwise gap. Right: paired early- and late-window MPSG.

4.5 Extended sensitivity and exploratory grouping audit

The public 13-week Treasury yield is the daily annual-percent adjusted close for ticker `^IRX`, aligned to Primary dates over 2015–2025 and averaged arithmetically (1.9643%); the endpoint is https://query2.finance.yahoo.com/v8/finance/chart/%5EIRX. Substituting it for zero changes the point ordering mildly (ρ=0.9790). CVaR tail probabilities from 2.5% through 10% leave the point order unchanged. Table 6 reports matched bootstrap summaries; the replication-level empirical intervals span zero and do not support an Entropy-versus-CRITIC advantage. A 3–30-day block grid yields Entropy mean agreement from 0.8813 to 0.8906, treating 10 days as a disclosed engineering setting rather than an estimated optimum. Changing VIKOR's compromise coefficient from 0.5 to 0.25 or 0.75 retains point-rank agreement of 0.9720 and 0.9790, respectively.

Table 6. Primary-panel sensitivity. $\bar{\rho}_B$ is mean bootstrap Spearman agreement, MC SE is simulation standard error, and Δ is the replication-level Entropy-minus-CRITIC empirical interval.

Setting	Point ρ	$\bar{\rho}_B$	MC SE	Empirical 2.5–97.5% Δ
Public short rate	.9790	.8643	.0037	[-.1818,.0876]
CVaR 97.5%	1.0000	.8875	.0031	[-.1748,.1189]
CVaR 90%	1.0000	.8908	.0030	[-.1715,.1365]

Four nonoverlapping retrospective extension panels were defined after the Primary analysis and recorded before their own public-price requests. Together with Primary, they form five nonoverlapping 12-ETF panels (60 unique tickers): Primary, Broad assets, Sector, International, and Fixed-alternative. Every panel has the same 2,766 price dates, criterion definitions, 10-day blocks, and full return-to-rank recomputation. The transfer test uses 500 replications for every panel with the same seed and, because date lengths coincide, matched block indices. Each row of the resulting 500×5 matrix is one common block-index draw and each column one panel; these rows enter the Friedman and label-permutation diagnostics. They are conditional bootstrap replications, not independent market histories. Metrics were computed in a remote reproducible Python environment with seed 20260712.

Supplementary Table S4 reports the exploratory purposive-panel case. Mean agreement ranges from 0.6456 to 0.9120, every paired Entropy-minus-CRITIC interval contains zero, and three seeds yield MPSG from 0.4176 to 0.4454. Broad assets and Sector are weaker than the other three panels, but the retrospective construction prevents population or superiority claims. Across 20 outcome-blind random repartitions of the same 60 tickers into five groups of 12, 500 draws per generated panel give mean MPSG 0.3999 (MC SE 0.0168; partition-level 2.5–97.5 percentiles [0.2830,0.5167]). MPSG correlates 0.9970 with within-draw standard deviation and 0.9955 with mean pairwise absolute gap. Of 100 generated panels, 97 Entropy–CRITIC replication-level empirical intervals include zero and three do not. These 100 unadjusted intervals are descriptive, not evidence of weighting superiority. This reduces reliance on the hand-built grouping, but it is not external-market validation. The purposive-case MPSG

(0.4176) lies at the 60th empirical percentile of the 20 partitions (z=0.23), so it is an illustrative category grouping rather than an exceptional result.

The disjoint-window replication preserves the conclusion without implying a constant effect. Early-window MPSG averages 0.3661 (MC SE 0.0154), whereas the late-window mean is 0.4319 (MC SE 0.0127). Their paired difference is 0.0658, with an unadjusted partition-level empirical 2.5–97.5% interval of $[-0.0773, 0.2541]$. Across the same 100 generated panels, early-versus-late full-rank Spearman agreement averages 0.4026; its 2.5th, median, and 97.5th percentiles are -0.1121, 0.4720, and 0.7566. Thus grouping dispersion appears in both windows, but the particular ETF orders are period-dependent. Figure 6 displays both the replicated distributions and their paired partition values.

5. Limitations

The five panels are purposive rather than probability samples. The four extensions were selected after inspection of the Primary result, so category coverage and apparent heterogeneity may reflect retrospective selection. USD-denominated adjusted prices do not capture taxes, trading costs, bid–ask spreads, investor-specific currency exposure, or fund closures. Results may change with a different risk-free rate, tail probability, criterion set, or sample window. The added short-rate, tail-probability, block-grid, window-length, and multi-panel checks reduce but do not remove that dependence. The bootstrap quantifies historical sampling uncertainty; it does not simulate future markets. These boundaries rule out causal, predictive, and portfolio-optimality claims.

6. Conclusion

TThis audit replaces a single Entropy-TOPSIS ETF order with rank distributions under sampling, specification, time, universe, and grouping changes. In the outcome-blind grouping experiment, 20 five-panel partitions with 500 draws per generated panel produce mean MPSG 0.3999 (MC SE 0.0168) and partition percentiles $[0.2830, 0.5167]$. The disjoint early and late windows reproduce grouping dispersion, with mean MPSG 0.3661 and 0.4319; their paired difference interval crosses zero, while early-versus-late full-rank agreement averages 0.4026. These results concern regrouping within the observed 60-ETF pool only.

The five purposive panels are a secondary category illustration. Their mean agreement ranges from 0.6456 to 0.9120, and stability is mixed: Primary, International, and Fixed-alternative retain higher agreement than Broad assets and Sector. Three of 100 generated-panel Entropy–CRITIC intervals exclude zero. These unadjusted replication-level empirical intervals are descriptive and do not establish weighting superiority.

For FinTech decision support, the audit suggests displaying a point rank with its interval and top-k inclusion probability (where k is the displayed candidate-set size), flagging period-sensitive ETFs, and rerunning when the candidate universe changes. These are untested design implications, not platform outcomes, trading rules, personalized advice, or evidence of a deployed platform.

The contribution is a reproducible reporting protocol around established methods; MPSG is auxiliary. The four extensions were retrospective, and the USD-denominated pool omits taxes, trading costs, fund closures, and investor-specific constraints. Broader currencies, independent markets, and prospectively fixed future windows remain necessary before predictive, portfolio-optimality, or universal claims can be considered.

Reference

[1] Sharifzadeh, Mohammad, and Simin Hojat. 2011. “An Analytical Performance Comparison of Exchange Traded Funds with Index Funds: 2002-2010.” arXiv Preprint arXiv:1111.0389.

[2] Divelgama, Bhathiya, Nancy Asare Nyarko, Naa Sackley Dromo Aryee, Abootaleb Shirvani, and Svetlozar T. Rachev. 2025. “Performance and Risk Analytics of Asian Exchange-Traded Funds.” arXiv Preprint arXiv:2511.12476.

[3] Pelegrina, Guilherme Dean, Leonardo Tomazeli Duarte, and Joao Marcos Travassos Romano. 2019. “Application of Independent Component Analysis and TOPSIS to Deal with Dependent Criteria in Multicriteria Decision Problems.” Expert Systems with Applications 122: 262–80. https://doi.org/10.1016/j.eswa.2019.01.008.

[4] Zolfani, Sarfaraz, Morteza Yazdani, Dragan Pamucar, and Pascale Zarate. 2020. “A VIKOR and TOPSIS Focused Reanalysis of the MADM Methods Based on Logarithmic Normalization.” Facta Universitatis, Series: Mechanical Engineering 18 (3): 341–55. https://doi.org/10.22190/FUME191129016Z.

[5] Campello, Betania S. C., Leonardo T. Duarte, and Joao M. T. Romano. 2020. “A Study of the Multicriteria Decision Analysis Based on the Time-Series Features and a TOPSIS Method Proposal for a Tensorial Approach.” arXiv Preprint arXiv:2010.11720.

[6] Costa, Leonardo Fernandes, Helder Gomes Costa, Diogo Lima, and Brunno Rodrigues. 2026. “TOPSIS-RAD: Ranking According to Desires.” arXiv Preprint arXiv:2606.07253.

[7] Pahsa, Alper. 2024. “Financial Technology Decision Support Systems.” Journal of Electrical Systems and Information Technology 11 (1). https://doi.org/10.1186/s43067-023-00130-0.

[8] Seiler, Volker, and Katharina Maria Fanenbruck. 2021. “Acceptance of Digital Investment Solutions: The Case of Robo Advisory in Germany.” Research in International Business and Finance 58: 101490. https://doi.org/10.1016/j.ribaf.2021.101490.

[9] Back, Camila, Stefan Morana, and Martin Spann. 2023.

ISSN 3106-5783 (Print) ISSN 3106-7042 (Online)

"When Do Robo-Advisors Make Us Better Investors? The Impact of Social Design Elements on Investor Behavior." Journal of Behavioral and Experimental Economics 103: 101984. https://doi.org/10.1016/j.socec.2023.101984.

[10] Ormos, Mihaly, and David Zibriczky. 2014. "Entropy-Based Financial Asset Pricing." PLoS ONE 9 (12): e115742. https://doi.org/10.1371/journal.pone.0115742.

[11] Risk, Jimmy, and Michael Ludkovski. 2018. "Sequential Design and Spatial Modeling for Portfolio Tail Risk Measurement." SIAM Journal on Financial Mathematics 9 (4): 1137–74. https://doi.org/10.1137/17M1158380.

[12] Fissler, Tobias, and Johanna F. Ziegel. 2021. "On the Elicitability of Range Value at Risk." Statistics & Risk Modeling 38 (1–2): 25–46. https://doi.org/10.1515/strm-2020-0037.

[13] Susmaga, Robert, Izabela Szczech, and Dariusz Brzezinski. 2024. "Towards Explainable TOPSIS: Visual Insights into the Effects of Weights and Aggregations on Rankings." Applied Soft Computing 153: 111279. https://doi.org/10.1016/j.asoc.2024.111279.

[14] Hamel, Andreas H, and Daniel Kostner. 2023. "Multi-Weight Ranking for Multi-Criteria Decision Making." arXiv Preprint arXiv:2312.03006.

[15] Basilio, Marcio Pereira, Valdecy Pereira, Fatih Yigit, Büşra Ayan, and Seda Abacıoğlu. 2025. "Iterative Random Weight EC-TOPSIS Method for Ranking Companies on Social Media." arXiv Preprint arXiv:2504.04169.

[16] Conejero, J. M., J. C. Preciado, A. E. Prieto, M. C. Bas, and V. J. Bolos. 2024. "Applying Data Driven Decision Making to Rank Vocational and Educational Training Programs with TOPSIS." arXiv Preprint arXiv:2411.00017.

[17] Zheng, Yong, and David Xuejun Wang. 2023. "Hybrid Multi-Criteria Preference Ranking by Subsorting." arXiv Preprint arXiv:2306.11233.

[18] Lin, Zhexiao, and Fang Han. 2023. "On the Failure of the Bootstrap for Chatterjee's Rank Correlation." arXiv Preprint arXiv:2303.14088.

[19] Pilavakis, Dimitrios, Efstathios Paparoditis, and Theofanis Sapatinas. 2019. "Moving Block and Tapered Block Bootstrap for Functional Time Series with an Application to the k-Sample Mean Problem." Bernoulli 25 (4B): 3496–526. https://doi.org/10.3150/18-BEJ1099.

[20] Gupta, Dev, Akanksha Parikh, and Tapan Kumar Datta. 2021. "A Multi-Criteria Decision-Making Approach to Rank the Sectoral Stock Indices of National Stock Exchange of India Based on Their Performances." National Accounting Review 3 (3): 272–92. https://doi.org/10.3934/NAR.2021014.

[21] Sielska, Agata. 2010. "Stability of Multicriteria Ranking: A Comparison." Multiple Criteria Decision Making 5: 229–46. https://mcdm.ue.katowice.pl/index.php?id=Siel09&page=publications.

[22] Agostinelli, Andrea, Michal Pándy, Jasper Uijlings, Thomas Mensink, and Vittorio Ferrari. 2022. "How Stable Are Transferability Metrics Evaluations?" European Conference on Computer Vision, 303–21. https://doi.org/10.1007/978-3-031-20050-2_18.

[23] Goswami, Shankha Shubhra. 2020. "Outranking Methods: Promethee i and Promethee II." Foundations of Management 12: 93–110. https://doi.org/10.2478/fman-2020-0008.