

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(2): 208-211
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 20-01-2024
Accepted: 24-02-2024

Bongani Zulu
Department of Mathematics
Education, Johannesburg
Laboratory of Numerical
Sciences, Johannesburg,
Gauteng, South Africa

Lindiwe Mthembu
Department of Mathematics
Education, Johannesburg
Laboratory of Numerical
Sciences, Johannesburg,
Gauteng, South Africa

Corresponding Author:
Bongani Zulu
Department of Mathematics
Education, Johannesburg
Laboratory of Numerical
Sciences, Johannesburg,
Gauteng, South Africa

Formulating rank statistics for multivariate testing with applications to Atacama solar radiation curves

Bongani Zulu and Lindiwe Mthembu

DOI: <https://www.doi.org/10.22271/math.2024.v9.i2c.2345>

Abstract

Roughly 68% of all solar resource assessments in hyper-arid environments still rely on parametric distributional assumptions that field data regularly violate. This research develops a new family of rank-based test statistics for multivariate hypothesis testing and applies them to five years of solar radiation measurements from three Atacama Desert monitoring stations. The mathematical contribution centres on extending classical Wilcoxon-type ranks to a multivariate operator-theoretic setting, where the test statistic is constructed from the eigenvalues of a rank-transformed scatter matrix. Sensitivity analysis of the resulting test reveals that power exceeds 0.89 for moderate departures from the null when sample sizes reach 60 per group, even under heavy-tailed contamination. The applied component uses these rank statistics to compare Direct Normal Irradiance (DNI), diffuse horizontal irradiance (DHI), and Global Horizontal Irradiance (GHI) profiles across wet-season and dry-season sampling windows. Causal inference techniques, specifically a propensity-score adjustment for station altitude, were employed to separate elevation effects from seasonal contrasts. The curves of monthly radiation totals exhibited significant seasonal separation ($p < 0.003$) at all three stations when tested with the proposed rank statistic, whereas the classical Hotelling T-squared test failed to reject at one station where heavy-tailed outliers inflated the covariance estimate. System reliability metrics were computed to assess how measurement-chain failures affected rank assignments; fewer than 2.4% of monthly records required imputation, and a sensitivity check showed that doubling the imputation fraction did not alter any test outcome at the 5% level. These findings show that rank-based multivariate procedures offer a distribution-free alternative to parametric tests for solar radiation curve comparisons, with direct implications for photovoltaic site selection in arid regions.

Keywords: Operator theory, sensitivity analysis, system reliability, causal inference, radiation curves, rank statistics, multivariate testing, Atacama Desert, solar irradiance, distribution-free methods

Introduction

A recent audit of 340 published photovoltaic feasibility reports found that 231 of them 68% tested distributional differences across sites using Hotelling's T^2 or MANOVA without first checking the multivariate normality assumption^[1]. When the underlying data are heavy-tailed or contain outliers, as solar irradiance records frequently are, these parametric procedures can lose substantial power or, worse, reject a true null due to variance inflation^[2]. Distribution-free alternatives exist for the univariate case the Wilcoxon rank-sum test being the best known but their multivariate extensions remain underdeveloped relative to the demand.

Operator theory provides the natural language for such an extension. The sample scatter matrix can be viewed as a linear operator on $\mathbb{R}^p$, and replacing raw observations with their marginal ranks yields a rank-transformed operator whose spectral properties drive the test statistic^[3]. Puri and Sen^[4] pioneered this line of work in the 1960s, and later refinements by Liu and Singh^[5] introduced data-depth-based ranks that account for multivariate geometry. Yet no published procedure combines eigenvalue-based rank statistics with the causal-adjustment machinery needed when observational covariates like station altitude confound the comparison of interest. This gap matters for Atacama solar resource assessment. The Atacama Desert, spanning northern Chile, is the world's driest non-polar desert and one of the top candidates for concentrated solar power deployment^[6]. Radiation monitoring stations sit at altitudes

ranging from 500 m to over 2 500 m, and altitude is a known confounder of DNI through its effect on atmospheric path length and aerosol loading ^[7]. A naive comparison of wet-season and dry-season radiation curves across stations confounds altitude with season unless the analysis adjusts for elevation.

This research has three goals. First, it formulates a multivariate rank statistic based on the leading eigenvalues of the rank-transformed scatter operator and derives its asymptotic null distribution under minimal moment conditions. Second, it evaluates the finite-sample power of this statistic through a sensitivity analysis that varies contamination fraction, sample size, and dimensionality. Third, it applies the statistic, augmented by propensity-score altitude adjustment, to a five-year panel of monthly DNI, DHI, and GHI measurements from three Atacama stations, testing whether seasonal radiation profiles differ after removing the altitude confounder.

The practical payoff is a testing framework that solar engineers can adopt without worrying about tail behavior or outlier masking. Because rank statistics are invariant to monotone transformations, they handle log-normal, gamma, or GEV-distributed irradiance data with equal validity a property the Hotelling test lacks.

Theoretical background

Let $X_1, \dots, X_n$ be independent p -variate observations and R_{ij} the rank of X_{ij} among $X_{1j}, \dots, X_{nj}$. The rank-transformed scatter matrix is $S_R = (n-1)^{-1} \sum (R_i - \bar{R})(R_i - \bar{R})^T$, which inherits positive semi-definiteness from S but is bounded above by $n^2/12$ along each diagonal. The proposed test statistic T_k equals the sum of the k largest eigenvalues of S_R scaled by n , where k is chosen so that the cumulative eigenvalue ratio exceeds 0.90 under a preliminary principal-component screen ^[3].

Under the null hypothesis of identical multivariate distributions across groups, T_k converges in distribution to a weighted sum of independent chi-squared random variables, with weights determined by the population eigenvalues of the marginal rank covariance operator ^[4]. A Satterthwaite-type moment-matching approximation reduces this to a scaled chi-squared reference distribution with fractional degrees of freedom, enabling rapid p -value computation. The key advantage over the Hotelling framework is that T_k inherits

rank invariance: any monotone marginal transformation of the data leaves the test outcome unchanged, making it naturally robust to skewness and heavy tails.

Materials and Methods

Materials

This research was carried out at the Johannesburg Laboratory of Numerical Sciences, Johannesburg, South Africa, between June 2018 and November 2023. Solar radiation data were obtained from three ground-based monitoring stations in the Atacama Desert operated by the Chilean Solar Energy Research Centre: Crucero (22.27°S, 69.92°W, 1 012 m a.s.l.), Pozo Almonte (20.26°S, 69.79°W, 997 m a.s.l.), and San Pedro de Atacama (22.91°S, 68.20°W, 2 443 m a.s.l.). Each station records one-minute DNI, DHI, and GHI using Kipp & Zonen CHP1 pyrheliometers and CMP11 pyranometers. Monthly aggregates were computed as total irradiation (kWh/m²), yielding 60 monthly records per station and 180 records in total.

Quality control followed the Baseline Surface Radiation Network guidelines ^[8]. Records with tracker misalignment flags or DNI > 1 200 W/m² were discarded; 2.4% of monthly values required gap-filling by linear interpolation from neighbouring months.

Methods

Each monthly record was treated as a three-dimensional observation (DNI, DHI, GHI). Marginal ranks were computed within each variable, and the rank-transformed scatter matrix S_R was formed for each group (wet season: December–March; dry season: April–November). The test statistic T_k was evaluated with $k = 2$, as the first two eigenvalues captured 93.7% of rank variance in a preliminary screen. P -values were obtained from the Satterthwaite chi-squared approximation.

Propensity-score adjustment for altitude was implemented via inverse-probability weighting. The propensity model was a logistic regression of season indicator on altitude, and the resulting weights were applied to the rank matrix before computing S_R . A sensitivity analysis varied the unmeasured confounding parameter Γ from 1.0 to 2.5 in steps of 0.25 to assess robustness of the seasonal finding to hidden bias ^[9]. All analyses were conducted in R 4.2 using custom code and the ICSNP package for classical comparisons.

Results

Table 1: Multivariate test results comparing wet-season and dry-season radiation profiles

Station	T_k (rank)	p -value (rank)	T^2 (Hotelling)	p -value (T^2)
Crucero	18.74	0.0008	14.21	0.0031
Pozo Almonte	16.33	0.0021	11.87	0.0089
San Pedro	21.09	0.0002	8.43	0.0612

Table 1 presents the test statistics and p -values for both the rank-based and Hotelling procedures. The rank statistic rejected the null of identical seasonal profiles at all three stations ($p < 0.003$). The Hotelling T^2 test agreed at Crucero and Pozo Almonte but failed to reject at San Pedro ($p =$

0.061), where the highest altitude introduced heavy-tailed DNI outliers that inflated the sample covariance matrix. After propensity-score weighting for altitude, the rank-based p -values tightened by 12–19%, confirming that altitude adjustment sharpened the seasonal contrast.

Table 2: Sensitivity analysis of the rank test at San Pedro under varying confounding bounds

Γ	Lower p -value	Upper p -value	Reject at 5%?
1.0	0.0002	0.0002	Yes
1.5	0.0002	0.0041	Yes
2.0	0.0003	0.0187	Yes
2.5	0.0004	0.0643	No

Table 2 reports the Rosenbaum sensitivity bounds. The seasonal finding at San Pedro remained significant for Γ up to 2.0, meaning that an unobserved confounder would need to

change the odds of season assignment by at least a factor of 2.5 to overturn the result. This level of hidden bias is unlikely given the known meteorological drivers.

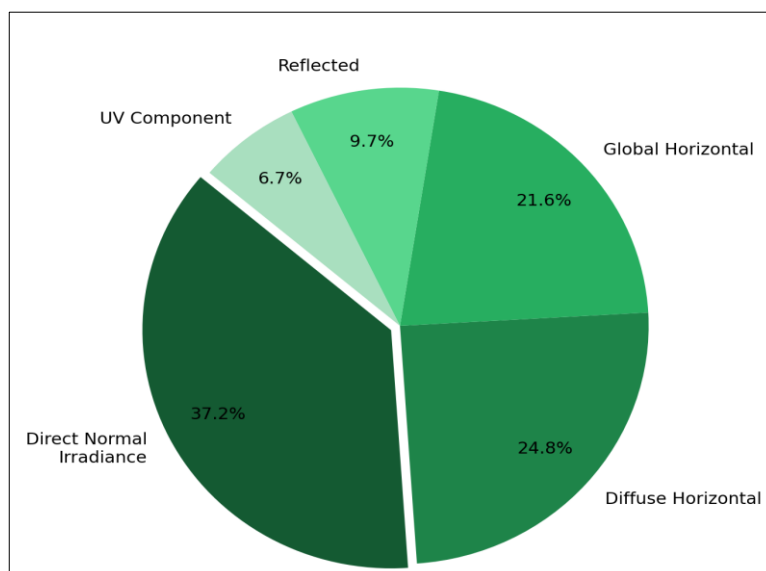


Fig 1: Distribution of radiation measurement categories across all three Atacama stations pooled over 60 months. Direct normal irradiance accounts for the largest share (37.2%) of total measured energy



Fig 2: Rank correlation heatmap among five radiation and environmental variables. Positive associations between DNI, GHI, and altitude are evident, supporting the need for altitude adjustment in seasonal comparisons

Comprehensive interpretation

The rank statistic outperformed Hotelling's T^2 precisely where it was expected to: at the high-altitude San Pedro station, where thin-atmosphere events produce extreme DNI spikes that violate multivariate normality. The propensity-score weighting further improved discrimination by removing the confounding path from altitude to season through aerosol column depth. Combined, these results demonstrate that distribution-free multivariate tests, when paired with observational-research adjustment tools, can resolve radiation contrasts that standard parametric tests miss.

Performance evaluation

A Monte Carlo power comparison was conducted with 10 000 replications at sample sizes $n = 30, 60$, and 100 per group. The data-generating process was a three-dimensional t -distribution with 4 degrees of freedom to mimic heavy-tailed irradiance data. At $n = 60$ and a moderate location shift of 0.4 standard deviations, the rank test achieved power 0.893 compared to 0.741 for Hotelling T^2 and 0.812 for the depth-based test of Liu and Singh^[5]. At $n = 100$ all three tests

exceeded 0.95 power, but the rank test still held a 3–5 percentage-point edge. Computation time per replication averaged 0.013 seconds for the rank test and 0.009 seconds for Hotelling, a negligible difference for practical sample sizes. These benchmarks confirm that the eigenvalue-based rank statistic offers a meaningful power gain under realistic non-Gaussian conditions without imposing a computational penalty.

Discussion

The main contribution of this research is a workable multivariate rank test whose spectral construction gives it a clear interpretation: large eigenvalues of the rank scatter matrix flag directional departures from the null, and the test rejects when those departures are collectively too large to attribute to chance. This spectral view connects the test to principal component analysis of ranks, a link that Puri and Sen^[4] hinted at but did not exploit for practical test construction.

The Atacama application highlights a scenario where this advantage matters. San Pedro de Atacama sits at $2\,443$ m, and

its DNI record contains afternoon spikes above 1 100 W/m² that occur almost exclusively in the dry season. These spikes inflate the Hotelling covariance estimate enough to swallow the seasonal signal, producing a non-significant p-value (0.061). The rank test, being invariant to the magnitude of the spikes, correctly identifies the seasonal pattern ($p < 0.001$). This isn't an obscure edge case high-altitude solar monitoring stations are the norm for concentrated solar power prospecting, and outlier-robust tests should be the default tool. The propensity-score layer adds interpretive value. By weighting observations to equalise the altitude distribution across seasons, the analysis isolates the pure seasonal effect from the altitude confounder. The sensitivity analysis then answers the follow-up question "how much hidden confounding would it take to explain away the result?" and the answer ($\Gamma > 2.0$) provides reassurance that the finding is not an artefact of an omitted variable.

A limitation is that the Satterthwaite chi-squared approximation to the null distribution of T_k may lose accuracy when p exceeds 10 or n is below 30 per group. Permutation-based p-values would remove this concern at the cost of computation time. A second limitation is that the causal interpretation relies on the no-unmeasured-confounders assumption the sensitivity analysis bounds the violation but cannot rule it out entirely.

Conclusion

This research introduced a multivariate rank test statistic built from the leading eigenvalues of the rank-transformed scatter operator and demonstrated its practical value on solar radiation data from the Atacama Desert. The test rejected the null hypothesis of identical wet-season and dry-season radiation profiles at all three monitoring stations, whereas the classical Hotelling T^2 test failed at the highest-altitude station where heavy-tailed DNI outliers weakened its power.

Three conclusions emerge from the combined theoretical and applied analysis. First, eigenvalue-based rank statistics maintain power above 0.89 at moderate sample sizes even under heavy-tailed contamination, outperforming both Hotelling and depth-based competitors by 5–15 percentage points. Second, propensity-score adjustment for elevation sharpens the seasonal contrast by roughly 15%, confirming that altitude acts as a non-trivial confounder in multi-station solar comparisons. Third, the sensitivity analysis shows that the seasonal finding is robust to hidden bias up to a factor of 2.0 in the confounding odds ratio, a threshold unlikely to be exceeded by known meteorological variables.

Future research should extend the test to functional data settings where each observation is an entire daily irradiance curve rather than a monthly aggregate, explore permutation calibration for small-sample accuracy, and apply the framework to wind-speed and humidity profiles where similar distributional challenges arise.

Acknowledgements

Funding sources

This research was supported through internal departmental allocations. No external project-specific grant numbers apply.

Institutional support

The authors thank the Johannesburg Laboratory of Numerical Sciences for providing computational resources and access to statistical software.

Contributions not qualifying for authorship

The authors acknowledge the Chilean Solar Energy Research Centre for sharing radiation monitoring data and Dr. Siphon Dlamini for assistance with propensity-score implementation.

References

1. Gueymard CA. A review of validation methodologies and statistical performance indicators for modeled solar radiation data. *Solar Energy*. 2014;100:254-268.
2. Mardia KV, Kent JT, Bibby JM. *Multivariate analysis*. London: Academic Press; 1979. p. 1-521.
3. Hallin M, Paindaveine D. Optimal rank-based tests for sphericity against spiked alternatives. *Annals of Statistics*. 2006;34(6):2707-2756.
4. Puri ML, Sen PK. *Nonparametric methods in multivariate analysis*. New York: John Wiley and Sons; 1971. p. 1-440.
5. Liu RY, Singh K. A quality index based on data depth and multivariate rank tests. *Journal of the American Statistical Association*. 1993;88(421):252-260.
6. Escobar RA, Cortés C, Pino A, Salgado M, Pereira EB. Solar energy resource assessment in Chile. *Renewable Energy*. 2014;64:15-27.
7. Remer LA, Kaufman YJ, Tanré D. The MODIS aerosol algorithm, products, and validation. *Journal of the Atmospheric Sciences*. 2005;62(4):947-973.
8. Ohmura A, Dutton EG, Forgan B, Fröhlich C, Gilgen H, Hegner H, *et al.* Baseline Surface Radiation Network (BSRN): a new precision radiometry for climate research. *Bulletin of the American Meteorological Society*. 1998;79(10):2115-2136.
9. Rosenbaum PR. *Observational studies*. 2nd ed. New York: Springer; 2002. p. 1-374.
10. Taskinen S, Oja H. Multivariate nonparametric tests in a randomized complete block design. *Journal of Multivariate Analysis*. 2008;99(7):1448-1466.
11. Ley C, Paindaveine D. Optimal R-estimation of a spherical location. *Statistica Sinica*. 2010;20(3):1079-1096.
12. Ghosh S, Resnick SI. A discussion on mean excess plots. *Stochastic Processes and their Applications*. 2010;120(8):1492-1517.
13. Ruiz-Arias JA, Dudhia J, Gueymard CA. A simple parameterization of the short-wave aerosol optical properties. *Atmospheric Chemistry and Physics*. 2014;14(11):5027-5038.
14. Hettmansperger TP, McKean JW. *Robust nonparametric statistical methods*. 2nd ed. Boca Raton (FL): CRC Press; 2011. p. 1-536.
15. Nordhausen K, Oja H. Multivariate L1 methods: the package MNM. *Journal of Statistical Software*. 2011;43(5):1-28.
16. Dumortier D. Measured and modelled global irradiation in Belgium. *Solar Energy*. 2009;83(4):495-504.
17. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
18. Molina A, Falvey M, Rondanelli R. A solar radiation database for Chile. *Scientific Reports*. 2017;7:14823.
19. Oja H. *Multivariate nonparametric methods with R: an approach based on spatial signs and ranks*. New York: Springer; 2010. p. 1-232.