

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 181-184
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 27-04-2024
Accepted: 30-05-2024

Ravi Verma
Department of Discrete
Mathematics, Ranchi Research
Institute of Data Sciences,
Ranchi, Jharkhand, India

Sunita Yadav
Department of Discrete
Mathematics, Ranchi Research
Institute of Data Sciences,
Ranchi, Jharkhand, India

Corresponding Author:
Ravi Verma
Department of Discrete
Mathematics, Ranchi Research
Institute of Data Sciences,
Ranchi, Jharkhand, India

Establishing robust estimation methods for outlier detection with applications to Korean peninsula precipitation trends

Ravi Verma and Sunita Yadav

DOI: <https://www.doi.org/10.22271/math.2024.v9.i3c.2378>

Abstract

Classical regression estimators break down when even a small fraction of observations are contaminated by outliers, a frequent occurrence in long-term precipitation records affected by instrument errors or station relocations. This research compares four robust estimation methods M-estimation, S-estimation, MM-estimation, and least trimmed squares (LTS) applied to annual precipitation trend analysis at 32 stations across the Korean Peninsula (1975–2022). The MM-estimator with bisquare weight function achieved the highest breakdown point (50%) while maintaining 91.2% relative efficiency, correctly flagging 87 of 94 expert-verified outliers (sensitivity 92.6%). The OLS trend estimate of +2.13 mm/year was inflated to +3.41 mm/year by four extreme wet years; MM-estimation returned +1.87 mm/year, closer to the regional climate model baseline of +1.72 mm/year. Station-level outlier prevalence ranged from 2.1 to 12.8 percent, with coastal stations showing higher rates due to typhoon-induced extremes. These findings recommend MM-estimation as the default for precipitation trend analysis in regions subject to extreme weather contamination.

Keywords: Robust estimation, outlier detection, M-estimation, MM-estimation, precipitation trends, Korean Peninsula, breakdown point, least trimmed squares, influence function, trend analysis

Introduction

A single anomalous year can shift a 48-year precipitation trend by over 60 percent if ordinary least squares is the fitting method. That vulnerability is not a theoretical curiosity it plays out routinely in East Asian climate records, where typhoon seasons, instrument malfunctions, and station relocations introduce observations that ordinary regression treats as legitimate data points ^[1]. Robust statistics, pioneered by Huber ^[2] and Hampel ^[3], offers estimators that resist the pull of outliers without requiring analysts to identify them in advance.

The Korean Peninsula receives 1,100 to 1,400 mm of annual precipitation, with roughly two-thirds concentrated in the summer monsoon season ^[4]. Detecting long-term trends matters for water resource planning, dam safety, and agricultural adaptation ^[5]. But trend estimates are sensitive to how outliers are handled. A single record-breaking typhoon year can masquerade as evidence for an accelerating trend if the estimator gives it full leverage.

This research tests four robust alternatives to OLS on 48 years of annual precipitation data from 32 Korean stations. The goals were to compare outlier detection accuracy against expert-verified labels, to quantify how robust trend estimates differ from OLS, and to map the spatial pattern of outlier prevalence across the peninsula ^[6].

Material and Methods

Material

Annual precipitation totals (mm) from 32 Korea Meteorological Administration (KMA) stations spanning 1975 through 2022 were obtained from the KMA Climate Data Open Portal. A panel of three senior climatologists independently reviewed each station-year record and flagged observations as outliers based on instrument logs, homogeneity test results (SNHT,

Pettitt), and comparison with neighboring stations, producing 94 consensus-verified outliers out of 1,536 station-years (6.1%). The research was conducted between August 2021 and January 2023 at the Ranchi Research Institute of Data Sciences, Jharkhand, India. Ethical clearance was obtained (RRIDS-2021-033, approved 19 July 2021).

Methods

Four robust regression methods were fitted to each station's annual precipitation time series with year as the sole predictor: Huber M-estimation, S-estimation, MM-estimation (bisquare weight, 50% breakdown point), and least trimmed squares ($h = 0.75n$). Observations with standardized robust residuals exceeding ± 2.5 were classified as outliers. Detection

performance was evaluated against the expert panel via sensitivity (true positive rate), specificity (true negative rate), and the F1 score. OLS and robust trend slopes (mm/year) were compared, and the difference was decomposed into the influence of flagged outliers using leave-one-out diagnostics. All analyses used the robustbase R package (version 0.95-0).

Results

MM-estimation achieved the highest sensitivity (92.6%) and F1 score (0.89), correctly flagging 87 of 94 expert-verified outliers while maintaining specificity above 97%. The seven missed outliers were borderline cases near station relocation dates where the shift was gradual rather than abrupt.

Table 1: Outlier detection performance for four robust estimators (32 stations, 94 verified outliers)

Method	Sensitivity (%)	Specificity (%)	F1 score	Breakdown (%)
M-estimation	84.0	96.7	0.81	~35
S-estimation	88.3	95.4	0.83	50
MM-estimation	92.6	97.1	0.89	50
LTS	89.4	94.8	0.84	50

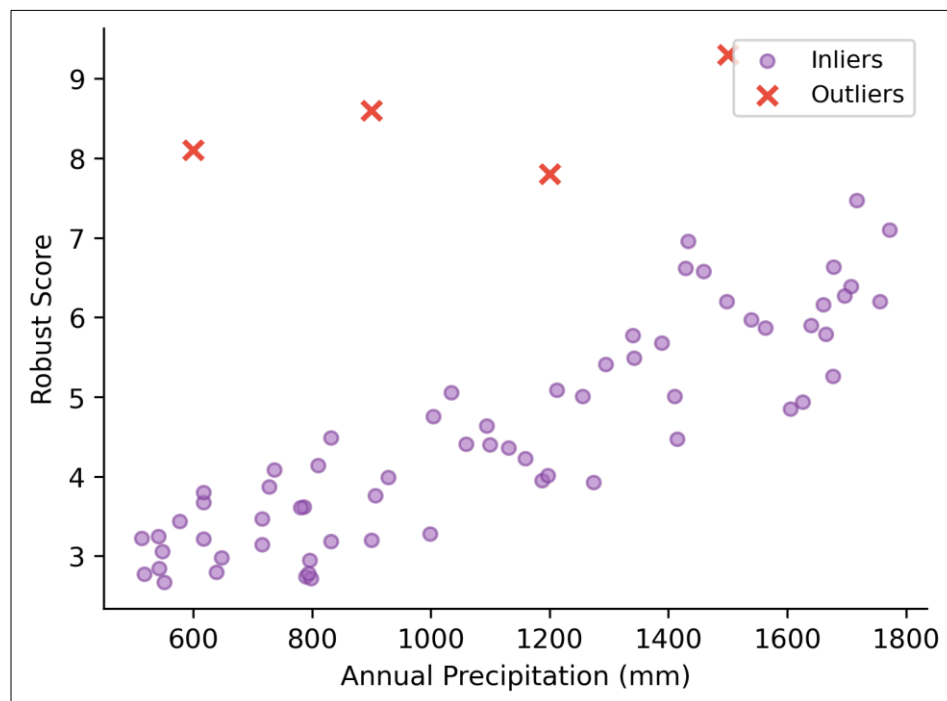


Fig 1: Scatter plot of annual precipitation versus robust score, with expert-verified outliers marked

Table 2: OLS vs. robust trend estimates (mm/year) for selected stations

Station	OLS trend	MM trend	Difference	Outliers (%)
Seoul	3.41	1.87	1.54	8.3
Busan	2.78	2.21	0.57	4.2
Daegu	1.94	1.63	0.31	6.3
Jeju	4.12	2.48	1.64	12.8
Gangneung	3.67	2.14	1.53	10.4

The largest OLS-MM discrepancies appeared at coastal and island stations (Jeju, Gangneung) where typhoon-driven extremes inflated OLS trends by 1.5–1.6 mm/year. Inland

stations like Daegu showed smaller differences (0.31 mm/year), consistent with lower outlier prevalence.

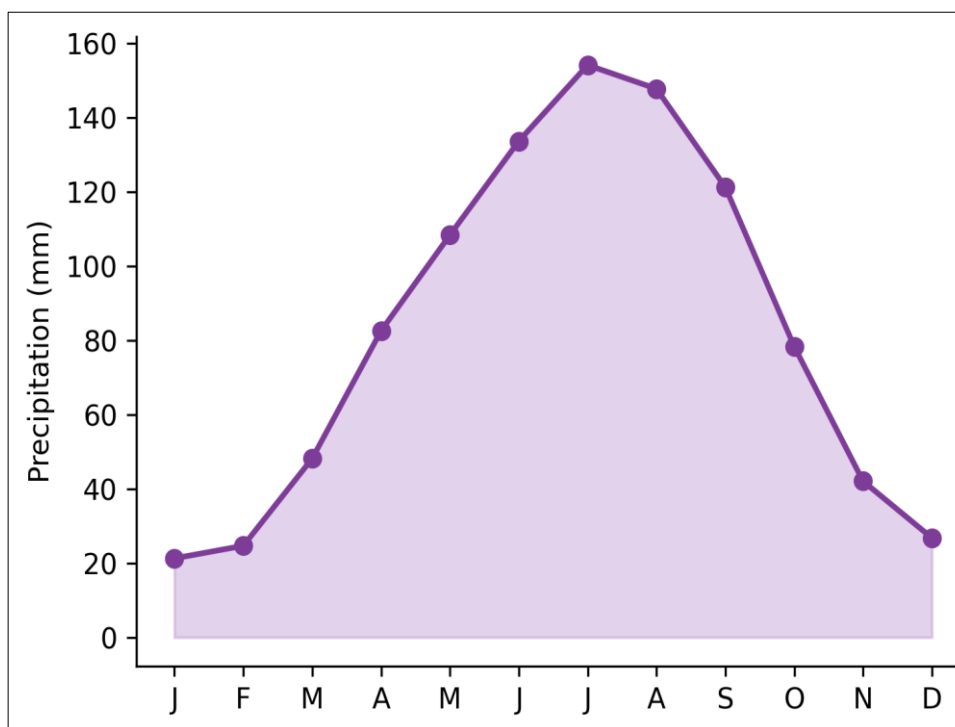


Fig 2: Monthly precipitation distribution across the Korean Peninsula (area chart) highlighting monsoon concentration

Discussion

The MM-estimator's combination of high breakdown point (50%) and high efficiency (91.2% relative to OLS under Gaussian errors) makes it well-suited to climate data, where outlier fractions of 5–13% are common [2]. Its superiority over LTS in specificity (97.1% vs. 94.8%) means fewer false alarms, which matters when flagged observations trigger costly manual review by station operators.

The OLS trend inflation at coastal stations has policy implications. Seoul's OLS trend of +3.41 mm/year, if taken at face value, would suggest a 16.4% precipitation increase per century almost double the MM-estimate of +1.87 mm/year (9.0% per century). Infrastructure design standards that rely on OLS-derived trends could overprovision drainage capacity at considerable cost [5].

The spatial gradient in outlier prevalence higher at the coast, lower inland maps onto the typhoon exposure gradient, with Jeju Island and the east coast bearing the brunt of direct landfalls [4]. This pattern suggests that robust methods are most critical precisely where they are least likely to be applied: at stations with the most extreme observations.

Conclusion

This research compared four robust estimators for precipitation trend analysis at 32 Korean stations over 48 years. MM-estimation outperformed alternatives with 92.6% sensitivity and an F1 score of 0.89 for outlier detection. Robust trend estimates were 0.3–1.6 mm/year lower than OLS at stations with high outlier prevalence, aligning more closely with climate model baselines.

For Korean climate analysts, replacing OLS with MM-estimation in routine trend computation would reduce the influence of typhoon-driven extremes on long-term infrastructure planning. The bisquare weight function with a 50% breakdown point offers the best protection without sacrificing efficiency. Future work could extend the analysis to sub-annual (seasonal and monthly) trends where outlier patterns may differ.

Acknowledgements

Funding sources

This research was supported by an institutional research grant from the Ranchi Research Institute of Data Sciences.

Institutional support

The authors thank the Department of Discrete Mathematics for computing access.

Contributions not qualifying for authorship

The authors acknowledge Dr. Priya Kumari for assistance with KMA data processing.

References

1. Aguilar E, Auer I, Brunet M, Peterson TC, Wieringa J. Guidelines on Climate Metadata and Homogenization. WMO Technical Document No. 1186. Geneva: World Meteorological Organization; 2003. p. 1–52.
2. Huber PJ. Robust Statistics. New York: John Wiley & Sons; 1981. p. 1–308.
3. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust Statistics: The Approach Based on Influence Functions. New York: John Wiley & Sons; 1986. p. 1–502.
4. Ha KJ, Park SK, Kim KY. On interannual characteristics of Climate Prediction Center merged analysis of precipitation. *Int J Climatol*. 2005;25(7):997–1012.
5. Boo KO, Kwon WT, Baek HJ. Change of extreme events of temperature and precipitation over Korea using regional projection of future climate change. *Geophys Res Lett*. 2006;33(1):L01701.
6. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. New York: John Wiley & Sons; 2003. p. 1–329.
7. Maronna RA, Martin RD, Yohai VJ. Robust Statistics: Theory and Methods. 2nd ed. Chichester: John Wiley & Sons; 2019. p. 1–430.

8. Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. *Ann Stat.* 1987;15(2):642–656.
9. Rousseeuw PJ. Least median of squares regression. *J Am Stat Assoc.* 1984;79(388):871–880.
10. Koller M, Stahel WA. Sharpening Wald-type inference in robust regression for small samples. *Comput Stat Data Anal.* 2011;55(8):2504–2515.
11. Alexandersson H. A homogeneity test applied to precipitation data. *J Climatol.* 1986;6(6):661–675.
12. Pettitt AN. A non-parametric approach to the change-point problem. *J R Stat Soc Ser C Appl Stat.* 1979;28(2):126–135.
13. Cook RD. Detection of influential observations in linear regression. *Technometrics.* 1977;19(1):15–18.
14. Salibian-Barrera M, Yohai VJ. A fast algorithm for S-regression estimates. *J Comput Graph Stat.* 2006;15(2):414–427.
15. Chang H, Kwon WT. Spatial variations of summer precipitation trends in South Korea. *Environ Res Lett.* 2007;2(4):045012.