JINAV: Journal of Information and Visualization *ISSN: 2746-1440 (online)*
Vol. 7 No. 1 (2026)
https://doi.org/10.35877/454RI.jinav4864

Developing a Robust Median and MAD-Based Estimator for Casewise and Cellwise Outlier Detection in South Sulawesi Socio-Economic Data

Agung Tri Utomo[a], Abdul Rahman[a,*], & Sharifah Aishah Syed Ali[b]

[a]*Universitas Negeri Makassar, Makassar, Indonesia*
[b]*Faculty of Defence Science & Technology, National Defence University of Malaysia, Malaysia*

Abstract

Socio-economic data analysis at the regional level frequently faces problems of spatial heterogeneity and extreme variability that give rise to outliers. Traditional approaches using casewise deletion often discard valuable information entirely just because of anomalies in a small fraction of attributes. This study proposes the development and application of a robust estimator based on Median and Median Absolute Deviation (MAD) to detect outliers using a cellwise paradigm (data cell-based) and compares it with a casewise approach (observation row-based) on socio-economic datasets of Regencies/Cities in South Sulawesi Province. The analyzed indicators include the Human Development Index (HDI), Regional Gross Domestic Product (RGDP), Percentage of Poor Population, Per Capita Expenditure, and the Open Unemployment Rate (OUR). Experimental results show that the robust estimator with a threshold of 2.24 is able to map cellwise outliers accurately without reducing the dimension of the observation data. Heatmap and Robust Z-Score Scatter Plot visualizations reveal that Makassar City is an extreme outlier in the RGDP indicator, while other anomalies are found in the OUR indicator in the Enrekang, Bulukumba, and Bone regions. Ultimately, this approach proves to be superior in maintaining the integrity of macroeconomic datasets compared to classical methods.

Keywords: robust estimator, median absolute deviation, cellwise outlier, casewise outlier, socio-economic data

Received: 5 January 2026 Revised: 30 March 2026 Published: 30 April 2026

1. Introduction

In public policy formulation, regional macroeconomic and social data play a highly crucial role (Badan Pusat Statistik Provinsi Sulawesi Selatan, 2025). However, high inter-regional disparity, such as the economic dominance of Makassar City in South Sulawesi Province, frequently generates extreme data variations—an anomaly known in statistical studies as an outlier (Ahmar et. al., 2024).

Classical statistical methods, such as mean and variance, have a breakdown point of 0%, meaning that the presence of even a single extreme observation can distort the overall estimation (Huber, 1981). In traditional multivariate data analysis practices, outlier handling is generally conducted through a casewise paradigm, where if a row (observation) is detected as an outlier, the entire row is deleted from the analysis (Rousseeuw & Leroy, 2005). In regional macroeconomic data, removing the data of an entire regency (for instance, deleting all data for Makassar City just because its RGDP is extreme) leads to the loss of valuable information on other indicators that might still be within normal limits, such as the HDI or poverty rates.

Classical approaches that rely on standard deviation around the mean often fail to detect anomalies accurately because the mean itself is highly susceptible to distortion by the presence of outliers. Therefore, the use of absolute deviation around the median (Median Absolute Deviation/MAD) is highly recommended as a robust dispersion parameter (Leys et al., 2013).

Consequently, a new paradigm in outlier handling, namely cellwise outliers, becomes increasingly relevant (Alqallaf et

* Corresponding author.
E-mail address: abdul.rahman@unm.ac.id

al., 2009; Raymaekers & Rousseeuw, 2021). This paradigm assumes that anomalies can occur independently in the cells of a data matrix rather than across the entire row. To overcome the weaknesses of classical methods, this study utilizes robust univariate estimators, specifically the Median as a measure of central tendency and the Median Absolute Deviation (MAD) as a measure of dispersion.

1.1. Research Objectives

The primary objectives of this study are:

(1). To apply a cellwise outlier detection algorithm using a robust standardization score based on Median and MAD on South Sulawesi socio-economic data;
(2). To compare the impact of cellwise versus casewise detection on observation integrity; and
(3). To provide an analytical visualization framework (Heatmap and Scatter Plot) to facilitate interpretation for policymakers.

2. Literature Review

2.1. Definition and Characteristics of Outliers

Hawkins (1980) defines an outlier as an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism. In the context of spatial and socio-economic data, outliers do not always imply data errors; rather, they frequently represent a population structure that is inherently highly unequal (Sembiring et al, 2020).

Conventional outlier detection often falls prey to two primary issues: the masking effect (where an outlier fails to be detected due to the presence of another adjacent outlier) and the swamping effect (where a normal observation is mistakenly identified as an outlier) (Barnett & Lewis, 1994). Leys et al (2013) emphasize that in outlier detection procedures, the use of the mean and standard deviation possesses a fundamental weakness because they have a breakdown point of 0%. Instead, they recommend using the absolute deviation around the median, as it is empirically proven to be more effective in overcoming masking and swamping effects in data with extreme variability.

2.2. Robust Estimator: Median and MAD

To avoid masking and swamping, methods based on robust estimation are proposed. Hampel (1974) introduces the concept of the breakdown point, which is the maximum proportion of outliers in a sample that an estimator can tolerate before it yields nonsensical results. The Median possesses the highest breakdown point, which approaches 50%.

As a companion to the Median for measuring dispersion, the Median Absolute Deviation (MAD) serves as a robust alternative to the standard deviation. MAD is calculated based on the absolute distance of each data point to the median of its distribution:

$$MAD(X) = 1.4826 \cdot median_i\left(\left|x_i - median_j(X)\right|\right) \quad (1)$$

The constant c (typically $c = 1.4826$) is utilized so that the MAD value is asymptotically equivalent to the standard deviation (σ) under the assumption of an underlying normal distribution (Rousseeuw & Croux, 1993). The use of MAD as an outlier-resistant dispersion estimator is proven to enhance detection performance, yielding more true positives and fewer false negatives (Voloh et al., 2020).

2.3. Casewise vs. Cellwise Outliers Paradigm

(1). Casewise Outliers: This paradigm assumes that the i-th row in the data matrix is a contaminated vector. The conventional Mahalanobis distance or the Robust Mahalanobis Distance such as the Minimum Covariance Determinant (MCD)—is frequently used to detect these outliers (Rousseeuw & Van Driessen, 1999).

(2). Cellwise Outliers: Formally introduced by Alqallaf et al (2009), this paradigm highlights that in the era of high-dimensional datasets, the probability of at least one cell being corrupted within a single row is extremely high. If e is the probability of a cell being contaminated, the probability of an observation row being entirely clean is $(1 - e)^p$. In large dimensions p, almost all observations can be flagged as casewise outliers, which causes the

casewise deletion method to fail completely.

2.4. Related Research

Research regarding cellwise outliers is gaining widespread attention. Raymaekers and Rousseeuw (2021) developed the Cellwise Robust M-regression method. In Indonesia, a study by Yulianto et al. (2019) confirms that the use of robust estimators is highly crucial when analyzing regional economic index data in Indonesia, which features archipelagic characteristics with high inequality among its provinces.

3. Research Methods

3.1. Data Source and Characteristics

This study utilizes secondary data obtained from the Central Bureau of Statistics (BPS) of South Sulawesi Province. The units of observation consist of 24 Regencies/Cities in South Sulawesi. The independent variables (indicators) analyzed include:

(a). X_1: Human Development Index (HDI)
(b). X_2: Regional Gross Domestic Product (RGDP) in Billions of Rupiah
(c). X_3: Percentage of Poor Population (%)
(d). X_4: Adjusted Per Capita Expenditure (Thousands of Rupiah)
(e). X_5: Open Unemployment Rate (OUR) in August (%)

3.2. Cellwise Detection Algorithm Framework

The analytical procedure is divided into several mathematical stages computed using a statistical programming language.

(a). Stage 1: Robust Parameter Estimation per Variable For each variable column , the central tendency estimation (Median, $\tilde{\mu}_j$) and scale estimation (MAD, $\tilde{\sigma}_j$) are calculated as follows:

$$\tilde{\mu}_j = median_{1\le i\le n}(x_{ij}) \tag{2}$$

$$\tilde{\sigma}_j = 1.4826 \cdot median_{1\le i\le n}(|x_{ij} - \tilde{\mu}_j|) \tag{3}$$

(b). Stage 2: Robust Standardization (Robust Z-Score) The data standardization stage in this study does not employ the classical Z-Score, but rather uses a Robust Z-Score based on the Median and MAD. This choice of estimator follows the recommendation of Leys et al. (2013) to guarantee that the detection threshold does not shift due to the extreme outlier values themselves.

Unlike the classical Z-score standardization ($Z = (x - \mu)/\sigma$), this study applies a robust variant:

$$Z_{ij} = \frac{x_{ij} - \tilde{\mu}_j}{\tilde{\sigma}_j} \tag{4}$$

This method ensures that the mean and variance values are not distorted by the presence of extreme outliers (Rousseeuw & den Bossche, 2018).

(c). Stage 3: Threshold Selection and Cellwise Flagging Once the values are obtained, a thresholding test is conducted. The critical threshold value used is . An indicator function is applied to flag a data cell as an outlier:

$$I_{ij} = \begin{cases} 1, & jika\ |Z_{ij}| > 2.24\ (Cellwise\ Outlier) \\ 0, & jika\ |Z_{ij}| > 2.24\ (Inlier\ or\ Normal\ Data) \end{cases} \tag{5}$$

(d). Stage 4: Diagnostic Visualization The data is visualized into an intensity matrix (Heatmap) where the rows represent the Regencies/Cities and the columns represent the Variables (X_j). Outlier tagging is displayed by

applying a black border/frame around the cell. The second approach uses a Scatter Plot to validate the position of cells lying outside the boundary band of $k = \pm 2.24$.

4. Results and Discussion

4.1. Robust Parameter Estimation

The application of classical versus robust estimators yields significant differences. In the RGDP indicator, the presence of Makassar City pulls the classical mean substantially upward, causing the classical variance to inflate artificially (swamping). By utilizing the Median and MAD, the baseline of normalcy for RGDP in South Sulawesi is successfully stabilized based on the conditions of the majority of regencies, rather than the distortion caused by a single large city. The data from this study can be seen in the Table 1.

Table1. RGDP Indicator of Sulawesi Selatan

Regency/City	HDI	GRDP (Billions)	Poverty Rate(%)	Per Capita Expenditure	Open Unemployment Rate (Aug)
Kepulauan Selayar	71.98	8340.51	10.79	10235	2.05
Bulukumba	74.43	20224.35	6.71	11807	2.23
Bantaeng	72.2	13451.36	8.26	12719	2.57
Jeneponto	69.45	13057.09	11.82	10158	2.47
Takalar	72.06	13428.09	7.75	11679	3.84
Gowa	73.71	30748.66	6.85	10700	3.91
Sinjai	71.81	15974.77	7.82	10665	1.52
Maros	74.04	28624	9.32	12209	4.34
Pangkajene dan Kepulauan	73.87	34624.22	12.41	12643	3.99
Barru	74.51	9891.13	8.31	12058	6.42
Bone	70.81	50015.24	9.58	10084	2.28
Soppeng	72.76	16099.18	6.9	10547	3.33
Wajo	73.98	25954.22	6.47	13608	2.31
Sidenreng Rappang	74.81	19614.67	5.02	13209	3.02
Pinrang	75.43	26653.4	8.55	13010	3.12
Enrekang	75.83	10467.33	11.25	12738	1.15
Luwu	73.86	22952.23	11.7	13867	4.14
Tana Toraja	71.94	10158.53	10.79	8319	3.98
Luwu Utara	74.04	19688.66	11.24	12866	2.39
Luwu Timur	76.44	30391.94	6.55	13867	4.58
Toraja Utara	72.31	13257.86	10.73	9292	2.44
Kota Makassar	85.23	243062.7	4.97	18368	9.71
Kota Parepare	80.97	9972.18	5.27	14928	5.23
Kota Palopo	81.25	11083.92	7.35	14369	7.64

4.2. Cellwise Detection Analysis Based on the Heatmap

Based on Figure 1 (Cellwise Outlier Detection Heatmap), several critical findings can be detailed:

(a). Economic Dominance (Upper Outlier): In the "RGDP (Billions of Rupiah)" column, the observation cell for Makassar City is marked with a highly intense red color and a black border. The $|Z_{ij}|$value for this cell far exceeds 2.24, identifying it as an extreme cellwise outlier. This fact aligns with demographic and economic realities, where Makassar, as the provincial capital, serves as the primary center of monetary circulation.

(b). Anomalies in Social Indicators: Interesting observations occur in Bone, Bantaeng, Barru, Enrekang, and Bulukumba regencies. These regencies show indications of outliers in the HDI and August OUR indicators.

(c). Preservation of Information: Under a traditional approach, if Bone is detected to have an anomaly in RGDP and OUR, the entire data row for Bone would be discarded (casewise deletion). However, the Heatmap confirms that the Percentage of Poor Population and Per Capita Expenditure indicators for these regencies do not exhibit outlier status at all (represented by neutral-colored cells).

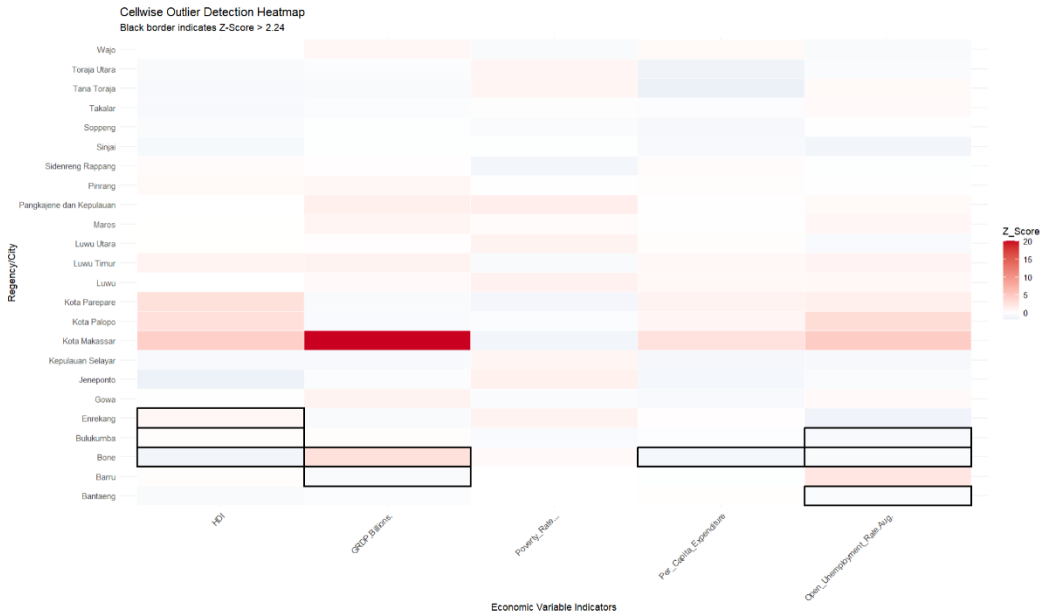

Figure 1. Cellwise Outlier Detection Heatmap

4.3. Robust Score Scatter Plot Analysis

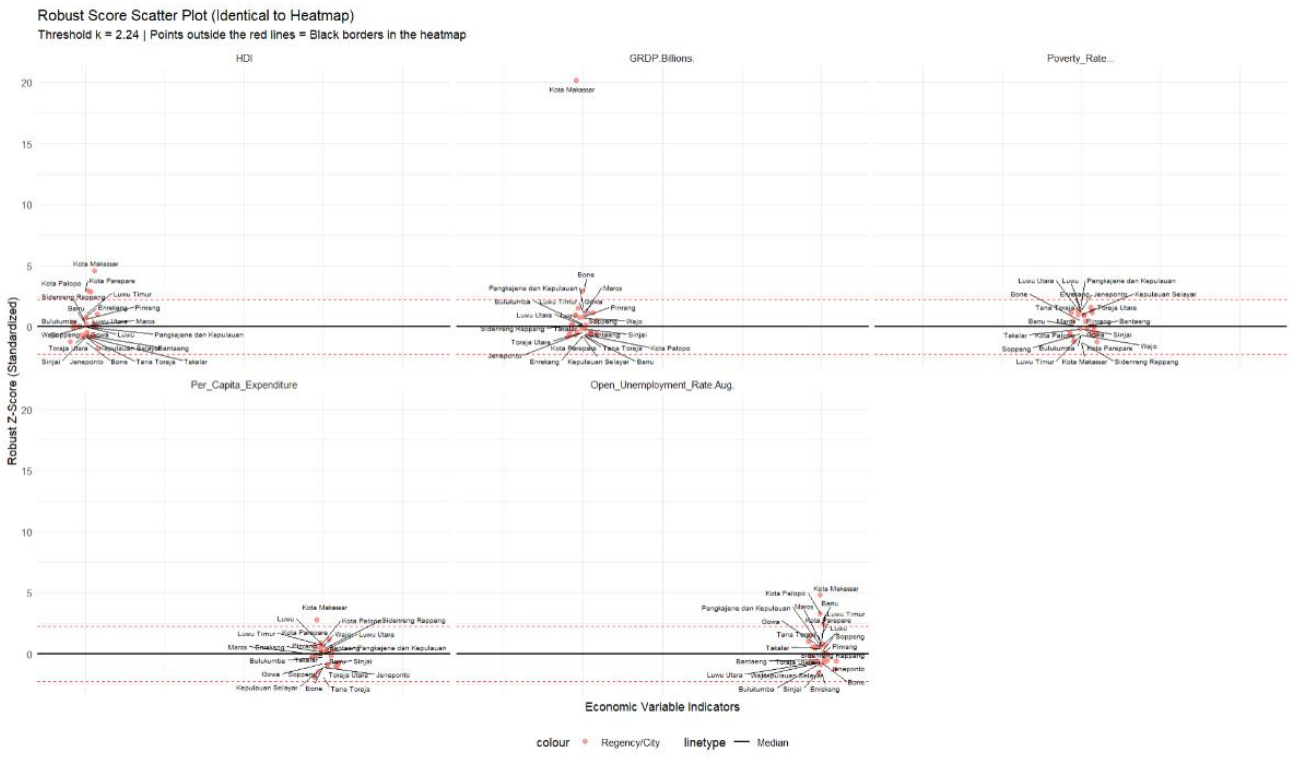

Figure 2. Robust Score Scatter Plot

Figure 2 (Robust Score Scatter Plot) provides a geometric perspective of the Heatmap results:

(a). The Y-axis on the scatter plot represents the Robust Z-Score value (Z_{ij}). The X-axis separates each economic indicator variable.

(b). Two horizontal red dashed lines serve as the visual representation of the thresholds $= -2.24$ and $k = 2.24$. The median point is represented by a solid black line at $Z = 0$.

(c). Distribution of RGDP: The point representing Makassar City sits exceptionally high, approaching a Z-Score value of 20 (nearly 20 times the MAD from the median). This proves the extreme nature of that specific cell.

(d). Open Unemployment Rate (OUR): The distribution of points is more evenly spread and fluctuates around the zero line. However, Makassar City and Tana Toraja Regency lie above the critical upper threshold, indicating an unemployment rate that is significantly higher or possesses unique labor characteristics compared to other agrarian regencies.

4.4. Comparative Advantages of Cellwise over Casewise in Spatial Modeling

If this study applies casewise detection (for instance, using Mahalanobis Distance), Makassar City, Tana Toraja, and Bone will certainly be eliminated from the matrix before proceeding to the regression or advanced analysis stages. This constitutes a fatal analytical loss of information. Although Makassar's RGDP deviates, its Per Capita Expenditure and HDI data may still reside within a regional linear structure that is highly viable for study.

The cellwise method successfully isolates “dirty cells” from “clean cells” within a “contaminated row.” For both the BPS (Central Bureau of Statistics) and BAPPEDA (Regional Development Planning Agency), this cell-based granular information enables far more precise policy interventions (e.g., specific interventions to tackle unemployment in Toraja without generalizing its macroeconomic issues).

5. Conclusion

This study successfully proves that the use of robust univariate estimators (Median and MAD) is highly effective in accurately identifying outliers in regional-scale data characterized by high spatial variability.

(1). The cellwise detection approach proves to be far more efficient in maintaining the data integrity of South Sulawesi's socio-economic data compared to the casewise method.

(2). Based on the filtering results using the Robust Z-Score with a critical threshold 2.24, outliers do not occur across the entire structure of a region; instead, they remain isolated within specific indicators: Makassar City in the RGDP variable, and several regions such as Tana Toraja, Bone, and Bantaeng in the OUR and HDI variables.

For future methodological development, it is suggested to combine this univariate cellwise detection method with bivariate or multivariate detection that accounts for correlations among variables, such as utilizing the Detecting Deviating Cells (DDC) algorithm. Additionally, data imputation techniques (such as KNN or robust regression) can be further investigated to fill the detected “dirty cells,” ensuring that the entire dataset matrix is ready for machine learning or applied econometrics without removing any observations.

References

Ahmar, A.S., Meliyana, S.M., Rahman, A., & Rusli, (2024). The Accuracy Analysis of Loan Interest Rate Forecasting Using Double Exponential Smoothing Methods. *International Journal of Data Science*, 5(2), 75–79.

Alqallaf, F., Van Aelst, S., Yohai, V. J., & Zamar, R. H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics*, 37(1), 311–331.

Badan Pusat Statistik Provinsi Sulawesi Selatan. (2025). *Provinsi Sulawesi Selatan Dalam Angka 2025*. Badan Pusat Statistik Provinsi Sulawesi Selatan. https://sulsel.bps.go.id/id/publication/2025/02/28/a7beacd81481497cecb44084/provinsi-sulawesi-selatan-dalam-angka-2025.html

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data* (3rd ed.). John Wiley & Sons.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346), 383–393.

Hawkins, D. M. (1980). *Identification of Outliers*. Chapman and Hall.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & Sons.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4), 764–766.

Raymaekers, J., & Rousseeuw, P. J. (2021). Fast robust correlation for high-dimensional data. *Technometrics*, 63(2), 184–198.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424), 1273–1283.

Rousseeuw, P. J., & den Bossche, W. (2018). Detecting Deviating Data Cells. *Technometrics*, 60(2), 135–145.

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust Regression and Outlier Detection*. John Wiley & Sons.

Rousseeuw, P. J., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212–223.

Sembiring, F., & others. (2020). Penerapan Data Mining untuk Analisis Tingkat Kemiskinan di Indonesia. *Jurnal SIFO Mikroskil*, 21(1), 45–56.

Voloh, B., Watson, M. R., Konig, S., & Womelsdorf, T. (2020). MAD saccade: statistically robust saccade threshold estimation via the median absolute deviation. *Journal of Eye Movement Research*, 12. https://doi.org/10.16910/jemr.12.8.3

Yulianto, A., Soelistyo, A., & Handayani, D. (2019). Analisis Robust Principal Component Analysis (RPCA) pada Data Ekonomi Regional. *Jurnal Statistika Nusantara*, 5(2), 112–120.