FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 8, April, 2026, pp 64 – 71
DOI: https://doi.org/10.33003/fjs-2026-1008-4885

IDENTIFYING GLOBAL UNDER-FIVE MORTALITY HOTSPOTS BASED ON SPATIAL OUTLIER DETECTION AND ROBUST REGRESSION ESTIMATORS: POLICY INSIGHTS

[1,2]Ishaq Abdullahi Baba, *[3]Mohammed Bappah Mohammed [3], [1]Rabiu Mohammed Madaki

[1]Department of Statistics, Abubakar Tafawa Balewa University Bauchi, Bauchi State, Nigeria
[2]Mathematics Unit, Department of Art and Science, University of the Gambia, Faraba Banta, West Coast Region, Gambia.
[3]Mathematics and Statistics Department, Federal University of Kashere, 234, Gombe State, Nigeria.

ABSTRACT

Under-five mortality remains a critical public health problem driven by multiple socioeconomic, health, and environmental factors. The severity of this problem varies across regions, depending on the underlying contextual characteristics. Understanding the spatial distribution and temporal pattern of under-five mortality can provide actionable insights for targeted interventions in vulnerable areas and help identify key sources of variation in mortality. In this study, we apply multivariate outlier detection and classification methods to identify countries at high risk of under-five mortality, using the robust multivariate CovMCD and cellMCD techniques. We exploit the concepts of regular, cellwise, and rowwise outliers to classify the risk levels of different countries. The analysis uses 2023 global under-five mortality data for 199 countries obtained from the United Nations Inter-agency Group for Child Mortality Estimation and evaluates robust regression methods for modelling global under-five mortality (U5MR). Our results show that spatial maps derived from CovMCD and cellMCD identify Africa as the most vulnerable region in terms of under-five mortality risk. Furthermore, we apply four regression estimators: ordinary least squares (OLS), least absolute deviations (LAD), least trimmed squares (LTS), and MM to compare their performance in the presence of cellwise and rowwise contamination. Both the real-data analysis and simulation study indicate that the MM estimator outperforms its competitors in terms of average coefficient RMSE and RMSPE across different outlier scenarios. Overall, our findings demonstrate that conventional OLS models can be misleading in the presence of multivariate outliers and reinforce the value of robust methods for policy-relevant child mortality modelling.

Keywords: Outlier Detection, Robust Regression, Mortality Rate, Estimators

INTRODUCTION

Child mortality remains one of the most critical indicators of population health and socio-economic development worldwide. Despite substantial global progress over the past decades, under-five mortality rates remain high in many low and middle income countries, especially in Sub-Saharan Africa and South Asia (Tagoe et al., 2020; Wang, 2003). Understanding the determinants of child mortality is a central in public health research, as empirical evidence derived from statistical models directly informs policy maker about stragtegic resource allocation, and interventions planning (Isiko et al., 2025). Regression analysis has been widely adopted to quantify the relationship between child mortality and explanatory variables such as maternal education, household income, access to healthcare, sanitation, and nutrition (Van Malderen et al., 2019; Musa et al., 2020). Among regression estimation techniques, the Ordinary Least Squares (OLS) estimator has historically been the most widely used due to its computational simplicity and optimal properties under the classical assumptions (Stocker, 2008; Zdaniuk, 2024). However, this estimator is well known to produce corrupted estimates where classical assumptions such as normality, homoscedasticity, and absence of influential outliers are violated (Wilcox, 2012). Empirical evidence shows that child mortality data rarely satisfy these assumptions (Silva, 2012; Kaombe and Manda, 2023). Mortality indicators are often derived from complex survey designs, indirect demographic methods, and model-based estimates that combine multiple data sources of varying quality. Consequently, such datasets frequently exhibit outliers, skewness, heteroscedasticity, and leverage points, which can severely distort OLS estimates and lead to misleading inferences (Kefale et al., 2025). This problem commonly appears in scientific, engineering, and public health datasets (Currit, 2002; Shatz, 2024). For example, the vulnerability of OLS is problematic in health policy research, where regression coefficient estimates directly influence decisions such as resource allocation and intervention prioritization (Ugrinowitsch et al., 2004). In addition misidentified or biased risk factors can lead to ineffective policies and misdirected public health investments (Kiviet and Phillips, 1996; Horrace and Oaxaca, 2006). In response to these challenges, the statistical literature has increasingly emphasized the use of robust regression estimators that are less sensitive to departures from classical assumptions. The Least Absolute Deviations (LAD) estimator minimizes the sum of absolute residuals and is known to be more robust to extreme observations than OLS (Dodge, 1997; Khan et al., 2021).

The Least Trimmed Squares (LTS) estimator further enhances robustness by fitting the regression model to a subset of the data after trimming observations with the largest residuals, thereby achieving a high breakdown point (Rousseeuw and Van Driessen, 2006). The MM estimator, which combines a high-breakdown initial estimator with an efficient M-estimation step, has been shown to offer an attractive balance between robustness and statistical efficiency (Khotimah et al., 2020; Lakshmi and Sajesh, 2023). Recent studies have demonstrated that robust estimators often outperform OLS in the presence of contaminated data, both in simulation settings and applied research contexts (Fitrianto and Xin, 2022). In health and demographic studies, where extreme observations may arise from conflict-affected regions, weak health systems, or data reporting errors, robust methods are particularly relevant (Van Malderen et al., 2019). Beyond classical statistical outliers, child mortality data often exhibit spatial outliers,

which arise when mortality levels in a particular region deviate markedly from those of its surrounding locations. Recent advances in spatial statistics emphasize that ignoring spatial dependence when detecting outliers can lead to incorrect identification and biased parameter estimation (Yildirim and Mert Kantar, 2020; Cai et al., 2024). Robust Mahalanobis-distance-based approaches, using estimators such as the Minimum Covariance Determinant (MCD), have been shown to effectively identify anomalous spatial units while maintaining resistance to masking and swamping effects (Gimàez et al., 2012; Rousseeuw and Bossche, 2018; Hu et al., 2024). Although robust regression estimators such as LAD, LTS, and MM have been shown to perform better under data contamination, their application in child mortality modeling remains limited and fragmented in recent empirical literature. Furthermore, most existing studies either focus on simulated datasets or apply robust methods without systematic comparison against OLS using real demographic data. Consequently, there is insufficient empirical evidence to guide researchers on the most appropriate estimator for analyzing child mortality data characterized by outliers, skewness, and heteroscedasticity. The primary aim of this paper is to propose a detection and classification algorithm based on minimum covariance determinant for the identification of countries with extreme under 5 mortality rate. In addition, we evaluate the performance of the several regression estimators to diagnose the effect of outlying observations. We conduct a simulation study to compare the performance of the classical and robust regression estimators. We analysis under 5 mortality data to demonstrate the benefit of outlier diagnostics for the robust regression estimators in real life application.

MATERIALS AND METHODS

Data Description

The dataset used in this study was obtained from the United Nations Inter-agency Group for Child Mortality Estimation, a collaborative initiative of UNICEF, the World Health Organization, the United Nations Population Division, and the World Bank Group. The version analyzed corresponds to the 2025 update. These data are internationally recognized for standardized definitions, harmonized measurement procedures, and cross-country comparability.

The dataset consists of country-level observations and six variables: country identifier, under-five mortality rate (u5mr2023), infant mortality rate (imr2023), neonatal mortality rate (nmr2023), mortality rate for children aged 5–14 (m5_14_2023), and stillbirth rate (sbr2023). The under-five mortality rate serves as the response variable, while the remaining four mortality indicators are used as explanatory variables.

This study adopts a quantitative approach to identify countries with extreme under-five mortality rates and to evaluate the performance of classical and robust regression estimators. The robustness and efficiency of the competing methods are assessed using both real data and simulation experiments under the presence of outliers. Similar comparative methodological frameworks have been applied in recent robust regression studies to evaluate estimators resistance to contamination (Ismail and Rasheed, 2021; Gharehgozli, 2021).

The use of internationally harmonized mortality data improves the validity of cross-national comparisons and reduces measurement bias. This strengthens the reliability of statistical inference and enhances the reproducibility of the empirical results.

Descriptive Statistics and Data Preprocessing

Summary statistics were computed to describe the distributional characteristics of all variables. Measures reported include the mean, median, standard deviation, minimum, and maximum values. These statistics provide an overview of central tendency, variability, and dispersion across countries. Table 2 shows the descriptive statistics of mortality indicators

Prior to analysis, the dataset was screened for missing values, inconsistencies, and coding errors. Observations with incomplete mortality indicators were excluded to ensure comparability across variables. All numerical variables were verified for plausible ranges based on internationally accepted mortality thresholds.

Special attention was given to detecting extreme observations, as the presence of outliers may influence regression estimates. Therefore, preliminary inspection of cell wise and row wise outliers' diagnostics was conducted prior to model estimation.

Description of the Models Used

The relationship between child mortality and its determinants was modelled using a multiple linear regression framework. Let y_i denote the response variable representing child mortality for observation i. The model is specified as

$$y_i = \beta_o + \sum_{j=1}^{p} \beta_j x_{ij} + e_i, \quad i = 1,2,\ldots,n \tag{1}$$

where β_0 is the intercept term, β_j denotes the regression coefficient associated with the jth explanatory variable, e_i is a random error term, n represents the sample size, and p is the number of explanatory variables. The classical linear regression model assumes that the errors e_i are independently and identically distributed.

For this study, we specify a regression model to investigate the relationship between the under-five mortality rate and four mortality predictors measured at the country level in the year 2023. The study assumes a linear regression framework of the form:

$$U5MR_i = \beta_0 + \beta_1 IMR_i + \beta_2 NMR_i + \beta_3 M5{-}14_i + \beta_4 SBR_i + e_i \tag{2}$$

Where i is a country index, β_0 is the intercept term, $\beta_1, \ldots \beta_4$ are the coefficients of regression of each explanatory variables with under five mortality rate, e_i is the error term capturing unobserved influences on the dependent variable. Our choice of this model is guided by several demographic works that treats age specific mortality as component of overall child mortality (Ezeh et al., 2021; Mwanga et al., 2025).

Ordinary Least Squares Estimator

The Ordinary Least Squares (OLS) estimator is obtained by minimizing the sum of squared residuals:

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{3}$$

Where $\widehat{\beta}_{OLS}$ represent the estimates of regression coefficients. The OLS is unbiased, consistent, and efficient within the class of linear unbiased estimators (Portnoy, 2022). However, because squared residuals are minimized, the estimator is highly sensitive to extreme observations and influential points. The OLS is implemented in R using lm function.

Least Absolute Deviations Estimator

The Least Absolute Deviations (LAD) estimator minimizes the sum of absolute residuals:

$$\hat{\beta}_{LAD} = \arg\min_{\beta} \sum_{i=1}^{n} \left| y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right| \quad (4)$$

The LAD estimator targets the conditional median of the response variable and is more robust to outliers than OLS. It performs well when the error distribution is heavy-tailed or asymmetric (Giloni et al., 2015; Genc and Lukman, 2025). This estimator is implemented using the rq function in R.

Least Trimmed Squares Estimator

The Least Trimmed Squares (LTS) estimator minimizes the sum of the smallest h squared residuals:

$$\hat{\beta}_{LTS} = \arg\min_{\beta} \sum_{i=1}^{h} r_{(i)}^2(\beta) \quad ()$$

Where

$$r_i(\beta) = y_i - \beta_o - \sum_{j=1}^{p} \beta_j x_{ij}, \quad ()$$

And

$$r_{(1)}^2 \le r_{(2)}^2 \le \cdots \le r_{(n)}^2, \quad ()$$

Denote the ordered squared residuals. The trimming constant is defined as

$$h = \left\lfloor \frac{n + p + 1}{2} \right\rfloor \quad (5)$$

The LTS estimator achieves a breakdown point of up to 50%, meaning that nearly half the observations can be contaminated without invalidating the estimator. This makes it highly robust to outliers and leverage points, though its statistical efficiency is reduced when the data are uncontaminated (Huang et al., 2015). This is implemented using ltsReg under the robustbase package.

MM Estimator

The MM estimator combines high breakdown robustness with high asymptotic efficiency. It is defined as the solution to

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}}{\hat{\sigma}} \right) x_{ij} = 0 \quad (6)$$

Where $\psi(\cdot)$ a bounded influence function such as Tukey's bisquare is, $\hat{\sigma}$ is a robust estimate of scale, and the initial parameter estimates are obtained from a high-breakdown estimator such as LTS. The estimator is implemented using rlm function under the MASS package in R. MM estimators retain resistance to outliers while achieving high statistical efficiency under normally distributed errors. They provide stable coefficient estimates across varying levels of contamination and represent a widely recommended approach for robust regression analysis (Pitselis, 2002; Rusgiyono, 2021).

RESULTS AND DISCUSSION

Correlation, multicollinearity, and spatial outlier diagnostics

The correlation, variance inflation factor, and outlier diagnostics based on covMcd and cellMCD approaches are applied to check the nature of the association between the mortality variables, confirm the severity of multicollinearity, and check the presence of rowwise and cellwise outliers. The correlation matrix reveals strong positive associations among the mortality indicators, indicating that countries with high mortality in one age category tend to exhibit high mortality across other age groups. Infant and neonatal mortality show extremely high correlation, suggesting substantial redundancy in information. Variance Inflation Factor results confirm severe multicollinearity. Infant and neonatal mortality exhibit VIF values above 70, far exceeding the conventional threshold of 10, indicating serious collinearity problems. The stillbirth rate also shows high multicollinearity, while mortality among children aged 5–14 demonstrates moderate collinearity. We used a spatial visualization technique to construct the world map using the ggplot2 package in R. Country classification as risk zone or not was done using rowwise and cellwise outlier detection techniques (Filzmoser and Gregorich, 2020; Raymaekers and Rousseeuw, 2024; Mwanga et al., 2025). For these tasks, the covMcd and cellMcd functions in R are used. Both rowwise MCD and cellwise CellMCD detected 83 countries as risk zones. MCD detected 9 countries and CellMCD flagged 12 as risk zones globally; see Figure 1. These findings imply that classical regression estimators may produce biased or unreliable estimates due to the effect of outliers (Jiao et al., 2024; Raza et al., 2024), which in turn affects the inferences and interpretations of results. Consequently, robust regression methods are justified, as they provide more reliable estimates under contamination. Figure 1 show a clear distribution of multivariate mortality profile worldwide.

Global classification of countries based on covMcd and cellMCD reveal strong spatial clustering of high risk countries (outliers) in part of Latin America, Asia, and Africa. Similar results has been confirmed by several studies for example empirical reports from UN IGME (2024) and the World Health Organization (2026) demonstrate that the Global South bears the vast majority of the global under-five mortality burden. Specifically, sub-Saharan Africa and Southern/Central Asia remain highly vulnerable, combined representing over 80% of global child fatalities. Furthermore, systematic spatial and economic assessments by Chao et al. (2018) and contemporary multi-country analyses (Shitu et al., 2025) prove that deep structural inequalities cause high-risk U5MR pockets to form strong spatial clusters across Latin America, Asia, and Africa. The high risk countries are classified into four outliers groups: regular, both, cellMCD only, and covMCD only. Countries identified by both methods defined extreme mortality profile are influence by rowwise and cellwise multivariate outliers. Only covMCD are influence by rowwise outliers, only cellMCD are influence by cellwise outliers, and regular defined countries with control mortality profile. The simultaneous coexistence of the covMCD only and cellMCD only outliers indicate that both rowwise and cellwise mechanism are present. This suggest that purely rowise outlier approaches are limited in this context. Tables 1 and 2 present the description of variables and descriptive statistics of 2023 child mortality dataset. Table 1 show a clear heterogeneity and strong skewness across mortality variables with under five mortality ranging from 1-115 deaths per 1000 per live births. This also confirm the rationale for applying robust regression estimators rather than the classical OLS.

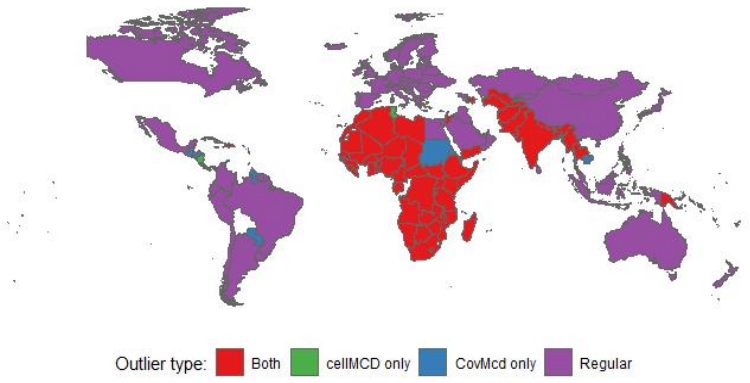

Figure 1: Global Classification of Countries by Multivariate Outlier Status Using Cellmcd and Covmcd. Countries Are Classified As Regular (Purple), Outliers Detected By Cellmcd Only (Green), Outliers Detected By Covmcd Only (Blue), or Outliers Detected By Both Methods (Red)

Table 1: Description of Variables Used in the Analysis

Variable	Type	Description
Country	Categorical	Country name or identifier
u5mr2023	Numeric	Under-five mortality rate (response)
imr2023	Numeric	Infant mortality rate
nmr2023	Numeric	Neonatal mortality rate
m5_14_2023	Numeric	Mortality rate ages 5–14
sbr2023	Numeric	Stillbirth rate

Table 2: Descriptive Statistics of Mortality Indicators (2023)

Variable	Mean	Median	Std. Dev.	Min	Max
Under-five mortality	24.491	15.173	24.847	1.429	114.789
Infant mortality	18.670	13.281	16.625	1.367	72.616
Neonatal mortality	11.611	8.510	9.748	0.618	40.241
Mortality ages 5–14	5.147	2.994	5.829	0.436	30.100
Stillbirth rate	9.974	8.200	7.437	1.180	34.899

Coefficient Estimates

The OLS, LAD, LTS, and MM was estimated for the five variables to obtain the best estimates. Table 3 displays the results for all the studied estimators in this paper. It can be observed that the OLS produced the largest coefficients values whereas the robust methods shrinks and stabilizes them. All estimators show that infant mortality has positive relationship with U5MR but OLS and LAD produced the highest slope value, indicating strong sensitivity to extreme infant mortality pairs. The MM and LTS produced a slightly smaller coefficients, suggesting a more conservative average marginal effect provided that outliers are downweighted or truncated. For NMR variables all the methods produced negative slopes but MM estimators push this effect close to zero, showing that robust estimator reduces the spurious negative relationship provided contaminated points are controlled. Similar negative slope values was realised for SBR variable and positive for M5-14 variable. Overall, these results show that the robust MM and LTS estimators yield more promising and stable effect for IMR, NMR, M5-14, and SBR compared to the OLS and LAD estimators in present of cellwise and rowwise contamination. These results agree with finding of (Filzmoser et al. 2020, Su et al. 2024).

Table 3: Regression Coefficient Estimates under Different Estimation Methods

Variable	OLS	MM	LTS	LAD
Intercept	−2.092	0.087	0.140	−0.507
IMR (2023)	1.428	1.132	1.199	1.341
NMR (2023)	−0.374	−0.055	−0.144	−0.337
M5–14 (2023)	0.925	0.399	0.414	1.140
SBR (2023)	−0.051	−0.067	−0.091	−0.210

OLS and LTS significance levels are based on t-tests. Significance for LAD estimates is inferred from confidence intervals. The MM estimator prioritizes robustness and efficiency rather than classical hypothesis testing.

Model Fit and Robustness Diagnostics

Table 4 summarizes goodness of fit measures and residual dispersion across the different estimation methods. The OLS model exhibits a very high coefficient of determination (R^2) but also a large residual spread, indicating sensitivity to outlying observations. The MM estimator substantially reduces the residual scale while maintaining stable coefficient estimates, confirming strong robustness to contamination. The LTS estimator achieves the smallest residual dispersion and the highest explained variance after trimming extreme observations, albeit at the cost of a reduced effective sample size. The LAD estimator yields comparatively larger coefficient magnitudes for key predictors, reflecting its focus on conditional medians rather than conditional means. Across the robust estimators (MM, LTS, and LAD), IMR and M5–14 remain consistently positive and influential, while NMR and SBR exhibit stable negative effects. These findings suggest that robust estimators provide more reliable inference than OLS in the presence of contaminated mortality data. The superior performance of the MM estimator further supports its suitability for public health datasets characterized by outliers and leverage points.

Table 4: Model Fit Statistics and Residual Dispersion

Metric	OLS	MM	LTS	LAD
Residual SE Scale	4.169	0.804	0.575	1.111
Degrees of freedom	194	194	139	194
R^2	0.972	–	0.997	–
Adjusted R^2	0.972	–	0.997	–
Max |Residual|	18.28	32.93	1.53	21.03

Simulation Study

An extensive Monte Carlo simulation study was conducted to evaluate the finite-sample performance, robustness properties, and predictive accuracy of the proposed regression framework under multiple contamination mechanisms. The estimators considered were Ordinary Least Squares (OLS), MM-estimation, Least Trimmed Squares (LTS), Least Absolute Deviations (LAD), and Weighted LAD (WLAD).

Data Generating Model

Data were generated from the linear regression model

$$y_i = X_i^T \beta + e_i, \; i = 1, 2, \ldots, n \quad (7)$$

Where x_i denotes the covariate vector (including intercept), β is the true parameter vector, and e_i is the random error term. The design matrix X was fixed and obtained from the empirical dataset according to the specified regression formula. This preserves the realistic correlation structure among predictors and avoids artificial covariate generation. Two true parameter configurations were considered: β_{MM} and β_{LTS} obtained from the MM and LTS fits. The errors are generated from $N(0,1)$ and from a Student-t distribution with 5 degrees of freedom. The uncontaminated response is generated using the linear model $y_i^0 = x_i'\beta + e_i$. For the contamination set, let $\in$ denote the contamination rate. For a given contamination proportion $\in$, let be n_c the number of contaminated observations, where $n_c = \in n$ with n denoting the sample size. To mimic realistic data corruption patterns, cellwise perturbation was introduced. For each contaminated row, a fraction γ of predictor cells (excluding the intercept) was randomly selected. For those cells,

$$x_{ij}^{con} = x_{ij} + \delta z_{ij} \quad (8)$$

Where δ>0controls the magnitude of contamination and z_{ij} are random shocks. We consider three different outlier scenarios: response outliers (Y-outliers), leverage outliers (X-outliers), and combined contamination (Y+X-outliers). The performance of each method is evaluated over 500 replications using the average coefficient root mean squared error (RMSE) defined by:

$$RMSE_j = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\beta}_j^{(r)} - \beta_j\right)^2} \quad (9)$$

And prediction RMSE on a fixed test set as:

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \quad (10)$$

Results are reported under two data-generating mechanisms: true regression coefficients obtained from the MM estimator and from the LTS estimator, see Tables 5 and 6. Lower values indicate better predictive performance. Both response and predictor perturbations were applied simultaneously to the same index set in the Y+X-outlier scenario, representing the most adversarial contamination case. This contamination differs from classical rowwise contamination and reflects modern epidemiology and public health data corruption structures. Across all contamination scenarios and both true-parameter specifications, the MM estimator consistently achieved the lowest coefficient RMSE and prediction RMSE. The LTS and LAD estimators provided partial robustness but were uniformly outperformed by MM, particularly under combined response and leverage contamination. In contrast, OLS exhibited substantial degradation in both estimation and prediction accuracy in the presence of outliers. All through the simulation we set the contamination rate to $\in = 0.10$, so that 10% of observations are contaminated in each replication; for each contaminated row, we perturbed $\gamma = 0.5$ of the predictor cells (excluding the intercept), with contamination magnitude $\delta = 3$ standard deviation. However, the real-data analysis and the simulation results indicate that traditional OLS regression is severely affected by multivariate outliers and multicollinearity. In contrast, the robust MM estimator yields stable coefficient estimates with small prediction error in both the global under-five mortality dataset and the simulation study, suggesting that it is well suited for policy-relevant child mortality modeling and analysis.

Table 5: Average Coefficient RMSE across Outlier Scenarios under Different True-Parameter Settings (Lower Is Better)

True β	Scenario	OLS	MM	LTS	LAD
MM-based	Y-outliers	0.089	0.055	0.064	0.066
MM-based	X-outliers	0.129	0.053	0.064	0.065
MM-based	Y+X-outliers	0.162	0.056	0.066	0.070
LTS-based	Y-outliers	0.088	0.058	0.061	0.064
LTS-based	X-outliers	0.151	0.055	0.069	0.074
LTS-based	Y+X-outliers	0.170	0.051	0.061	0.063

Table 6: Prediction RMSE across Outlier Scenarios under Different True-Parameter Settings (Lower Is Better)

True β	Scenario	OLS	MM	LTS	LAD
MM-based	Y-outliers	0.278	0.174	0.206	0.209
MM-based	X-outliers	0.420	0.177	0.212	0.216
MM-based	Y+X-outliers	0.510	0.177	0.215	0.221
LTS-based	Y-outliers	0.288	0.174	0.198	0.198
LTS-based	X-outliers	0.473	0.167	0.218	0.225
LTS-based	Y+X-outliers	0.559	0.164	0.193	0.207

CONCLUSION

In public health practice, identifying the target population remains a major challenge for donors, policymakers, and governments at various levels. At the same time, identifying the key factors associated with a high risk of a particular epidemiological disease is another major challenge. Therefore, appropriate statistical models are essential for sound decision-making, the identification of vulnerable areas, and the allocation and distribution of interventions. To address these problems, the global under-five mortality dataset was analysed to identify high-risk zones of under-five mortality worldwide. The concept of multivariate outliers, operationalised via the CovMCD and cellMCD methods, was adapted and applied. Regression methods were also applied to determine the influence of selected factors on the under-five mortality rate. The OLS, LAD, LTS, and MM estimators were used. The analysis confirms that robust estimators, particularly MM, substantially outperform traditional OLS in modelling under-five mortality data contaminated by outliers, skewness, and heteroscedasticity in global health datasets. Spatial categorization via CovMCD and cellMCD effectively flagged the majority of African countries and parts of Asia and Latin America as the highest-risk zones on the world map highlighting urgent intervention priorities. The OLS produces biased coefficients and invalid inference in such data, MM maintains high efficiency (85%+) and a 50% breakdown point, making it ideal for public health applications where policy decisions depend on accurate risk factor identification. The simulation results further demonstrate the superior robustness and predictive performance of the MM estimator under multiple contamination scenarios. This study fills a critical gap, as the existing literature rarely applies multivariate outlier detection and classification techniques to U5MR spatial risk mapping. This paper focused on rowwise regression estimators and demonstrated their performance. Future work could explore methods that solve the combined problem of cellwise and rowwise outliers; for example, developing a cellwise-weighted LAD estimator to isolate cell contamination effects in epidemiological modelling would be an excellent contribution to public health and robust statistics.

REFERENCES

Cai, J., Hu, W., Yang, Y., Yan, H., & Chen, F. (2024). Outlier detection in spatial error models using modified thresholding-based iterative procedure for outlier detection approach. *BMC Medical Research Methodology*, *24*(1), Article 89. doi.org

Chao, F., You, D., Pedersen, J., Hug, L., & Alkema, L. (2018). National and regional under-5 mortality rate by economic status for low-income and middle-income countries: A systematic assessment. *The Lancet Global Health*, *6*(5), e535–e547. doi.org

Currit, N. (2002). Inductive regression: Overcoming OLS limitations with the general regression neural network. *Computers, Environment and Urban Systems*, *26*(4), 335–353. doi.org

Dodge, Y. (1997). LAD regression for detecting outliers in response and explanatory variables. *Journal of Multivariate Analysis*, *61*(1), 144–158. doi.org

Ezeh, O. K., Ogbo, F. A., Odumegwu, A. O., Oforkansi, G. H., Abada, U. D., Goson, P. C., Ishaya, T., & Agho, K. E. (2021). Under-5 mortality and its associated factors in northern Nigeria: Evidence from 22,455 singleton live births (2013–2018). *International Journal of Environmental Research and Public Health*, *18*(18), Article 9899. doi.org

Filzmoser, P., & Gregorich, M. (2020). Multivariate outlier detection in applied data analysis: Global, local, compositional and cellwise outliers. *Mathematical Geosciences*, *52*(8), 1049–1066. doi.org

Filzmoser, P., Höppner, S., Ortner, I., Serneels, S., & Verdonck, T. (2020). Cellwise robust M regression. *Computational Statistics & Data Analysis*, *147*, Article 106944. doi.org

Fitrianto, A., & Xin, S. H. (2022). Comparisons between robust regression approaches in the presence of outliers and high leverage points. *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, *16*(1), 243–252. doi.org

Genç, M., & Lukman, A. (2025). Weighted LAD-Liu-LASSO for robust estimation and sparsity. *Computational Statistics*, 1–30. Advance online publication. doi.org

Gharehgozli, O. (2021). An empirical comparison between a regression framework and the synthetic control method. *The Quarterly Review of Economics and Finance*, *81*, 70–81. doi.org

Giloni, A., Simonoff, J. S., & Sengupta, B. (2006). Robust weighted LAD regression. *Computational Statistics & Data Analysis*, *50*(11), 3124–3140. doi.org

Horrace, W. C., & Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, *90*(3), 321–327. doi.org

Huang, D., Cabral, R., & De la Torre, F. (2015). Robust regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(2), 363–375. doi.org

Hu, Y., Fang, X., Zeng, W., Kutterer, H., & Li, D. (2024). Statistical robust estimation of spatial symmetric transformations based on Mahalanobis distance. *IEEE Transactions on Geoscience and Remote Sensing*, *62*, 1–10. doi.org

Isiko, I., Jacob, E. C., Mwesigwa, A., Olot, H., Okoro, L. N., Asingwire, J. M., Izunwanne, M. J. P., Ikwara, E. A., Ajudua, S. C., Gmanyami, J. M., & Alonge, O. (2025). Socioeconomic, demographic and environmental factors associated with under-five mortality among children in Kenya: Analysis of the 2022 Kenya Demographic and Health Survey. *BMC Pediatrics*, *25*(1), Article 509. doi.org

Ismail, M. I., & Rasheed, H. A. (2021). Robust regression methods/A comparison study. *Turkish Journal of Computer and Mathematics Education*, *12*(14), 2939–2949. doi.org

Jiao, X., Pretis, F., & Schwarz, M. (2024). Testing for coefficient distortion due to outliers with an application to the economic impacts of climate change. *Journal of Econometrics*, *239*(1), Article 105547. doi.org

Kaombe, T. M., & Manda, S. O. (2023). A novel outlier statistic in multivariate survival models and its application to identify unusual under-five mortality sub-districts in Malawi. *Journal of Applied Statistics*, *50*(8), 1836–1852. doi.org

Kefale, B., Jancey, J., Gebremedhin, A. T., Belay, D. G., Pereira, G., & Tessema, G. A. (2026). Under-five mortality estimation methods: A methodological systematic review. *Annals of Epidemiology*, *113*, 71–77. doi.org

Khan, D. M., Ali, M., Ahmad, Z., Manzoor, S., & Hussain, S. (2021). A new efficient redescending M-estimator for robust fitting of linear regression models in the presence of outliers. *Mathematical Problems in Engineering*, *2021*, Article 3090537. doi.org

Khotimah, K., Sadik, K., & Rizki, A. (2020). Study of robust regression modeling using MM-estimator and least median squares. *Journal of Physics: Conference Series*, *1497*(1), Article 012014. doi.org

Kiviet, J. F., & Phillips, G. D. (1996). The bias of the ordinary least squares estimator in simultaneous equation models. *Economics Letters*, *53*(2), 161–167. doi.org

Lakshmi, R., & Sajesh, T. (2023). Empirical study on robust regression estimators and their performance. *Reliability: Theory & Applications*, *18*(2), 466–478. doi.org

Lei, L., & Wooldridge, J. (2022). *What estimators are unbiased for linear models?* (ArXiv: 2212.14185). arXiv Preprint. arxiv.org

Musa, M., Asiribo, O., Dikko, H., Usman, M., & Sani, S. (2020). Modelling the determinants of under-five child mortality rates using Cox proportional hazards regression model. *FUDMA Journal of Sciences*, *4*(4), 401–408. doi.org

Mwanga, M. K., Mirau, S. S., Tchuenche, J. M., & Mbalawata, I. S. (2025). Bayesian prediction of under-five mortality rates for Tanzania. *Franklin Open*, *10*, Article 100221. doi.org

Pitselis, G. (2002). Application of GM and MM estimators to regression credibility. In *Proceedings of the 2nd Conference in Actuarial Science and Finance* (pp. 85–104). University of the Aegean. aegean.gr

Portnoy, S. (2022). Linearity of unbiased linear model estimators. *The American Statistician*, *76*(4), 372–375. doi.org

Prahutama, A., & Rusgiyono, A. (2021). Robust regression with MM-estimator for modelling the number maternal mortality of pregnancy in Central Java, Indonesia. *Journal of Physics: Conference Series*, *1943*(1), Article 012148. IOP Publishing. doi.org

Raymaekers, J., & Rousseeuw, P. J. (2026). Challenges of cellwise outliers. *Econometrics and Statistics*, *38*, 6–25. https://doi.org/10.1016/j.ecosta.2024.02.002

Raza, A., Talib, M., Noor-ul Amin, M., Gunaime, N., Boukhris, I., & Nabi, M. (2024). Enhancing performance in the presence of outliers with redescending M-estimators. *Scientific Reports*, *14*(1), Article 13529. doi.org

Rousseeuw, P. J., & Bossche, W. V. D. (2018). Detecting deviating data cells. *Technometrics*, *60*(2), 135–145. doi.org

Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, *12*(1), 29–45. doi.org

Shatz, I. (2024). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods*, *56*(2), 826–845. doi.org

Shitu, S., Weldesamuel, B., Guesh, M., Abrha, S., Woldegerima, H., & Assefa, N. (2025). Spatial patterns and determinants of under-five mortality in Sub-Saharan Africa: A multi-country analysis. *BMC Public Health*, *25*(1), Article 341. doi.org

Silva, R. (2012). Child mortality estimation: Consistency of under-five mortality rate estimates using full birth histories and summary birth histories. *PLoS Medicine*, *9*(8), Article e1001296. doi.org

Stocker, T. C. (2008). *On the asymptotic properties of the OLS estimator in regression models with fractionally*

integrated regressors and errors [Doctoral dissertation, University of Konstanz]. KOPS Institutional Repository. uni-konstanz.de

Su, P., Tarr, G., Müller, S., & Wang, S. (2024). CR-lasso: Robust cellwise regularized sparse regression. *Computational Statistics & Data Analysis*, *197*, Article 107971. doi.org

Tagoe, E. T., Agbadi, P., Nakua, E. K., Duodu, P. A., Nutor, J. J., & Aheto, J. M. K. (2020). A predictive model and socioeconomic and demographic determinants of under-five mortality in Sierra Leone. *Heliyon*, *6*(3), Article e03508. https://doi.org/10.1016/j.heliyon.2020.e03508

Ugrinowitsch, C., Fellingham, G. W., & Ricard, M. D. (2004). Limitations of ordinary least squares models in analyzing repeated measures data. *Medicine & Science in Sports & Exercise*, *36*(12), 2144–2148. doi.org

United Nations Inter-agency Group for Child Mortality Estimation. (2024). *Levels & trends in child mortality: Report 2024*. United Nations Children's Fund. childmortality.org

Van Malderen, C., Amouzou, A., Barros, A. J., Masquelier, B., Van Oyen, H., & Speybroeck, N. (2019). Socioeconomic factors contributing to under-five mortality in Sub-Saharan Africa: A decomposition analysis. *BMC Public Health*, *19*(1), Article 760. doi.org

Wang, L. (2003). Determinants of child mortality in LDCs: Empirical findings from Demographic and Health Surveys. *Health Policy*, *65*(3), 277–299. doi.org

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Academic Press. doi.org

World Health Organization. (2026). *Child mortality (under 5 years)*. WHO Fact Sheets. who.int

Yildirim, V., & Mert Kantar, Y. (2020). Robust estimation approach for spatial error model. *Journal of Statistical Computation and Simulation*, *90*(9), 1618–1638. doi.org

Zdaniuk, B. (2024). Ordinary least-squares (OLS) model. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 4867–4869). Springer. doi.org