Pak. J. Statist.
2026 Vol. 42(3), 205-222

AN IMPROVED ROBUST MODEL FOR PARAMETER ESTIMATION USING THE FUZZY TRANSFORM

Muzahem Al-Hashimi[1], Mohammed Alawjar[1], Edrees M. Nori Mahmood[2] and Heyam Hayawi[1§]
[1] Department of Statistics and Informatics, University of Mosul, Iraq
[2] Department of Operations Research and Intelligent Techniques University of Mosul, Mosul, Iraq
[§] Corresponding author Email: he.hayawi@uomosul.edu.iq

ABSTRACT

The basic concepts behind the estimation of parameters have been derived based on their original methods. These methods exhibit considerable deficiencies when dealing with data corrupted with non-Gaussian noise or data containing outliers. This paper presents a more advanced approach that tackles these problems by establishing a connection between the elements of F-transform and the estimated parameters. The optimized F-transform approach yields better and more reliable estimates of the model parameters compared to OLS, Huber regression, MM-estimators, and state-of-the-art robust estimators in the case of strong non-Gaussian noise that follows a Laplace, Gaussian, or Cauchy distribution. The theory is developed rigorously along with novel analysis of bias-variance decomposition in finite sample size setting, convergence rate in the case of heavy tails, and Bayesian optimization regret bound. The framework possesses three main components: adaptive partitioning, local scale estimation, and Bayesian hyperparameter optimization. Extensive tests carried out on both artificial and real air quality datasets show its superiority with up to 58.7% better robustness compared to Huber regression with the use of Cauchy noise, and more impressively, a remarkable improvement of 74.9% – 80.2% in real data. On the quadratic model with Gaussian noise, the optimized F-transform ends up with the smallest RMSE (0.052 ± 0.025), sort of the lowest out of the compared methods, and it beats Random Forest (0.292 ± 0.021), Gradient Boosting (0.420 ± 0.023), plus a Huber neural network (0.163 ± 0.047).

KEYWORDS

Bayesian Optimization; Bias-Variance Tradeoff; Fuzzy Transform; Kernel Methods; Robust Estimation.

1. INTRODUCTION

Mathematics is considered an important tool that allows scientists to move forward. The effectiveness of a model is determined by how well the unknown parameters can be estimated using real-world observations. The Method of Least Squares (LSE) and Maximum Likelihood Estimation (MLE) have been widely used for a long time due to their efficient nature under the following conditions: one needs to use a correctly formulated

© 2026 Pakistan Journal of Statistics

model and independent identically distributed (i.i.d.) Gaussian noise (Fisher, 1922; Stigler, 1981).

Empirical analyses of such datasets reveal that this is not always the case, as the empirical data contains noise, significant errors, and system failures, thus affecting negatively the performance of classical estimation techniques (Huber, 1981; Al-Hashimi et al., 2026). Such constraints have prompted further research on robust estimation procedures. The concept of fuzzy logic, proposed by Zadeh (Zadeh, 1965; Hayawi et al., 2021), deals with uncertainties. Originally proposed by Perfilieva (Perfilieva, 2006), the F-transform offers strong capabilities for function approximation and signal processing while removing unwanted elements from corrupted signals (Aswaq and Mahmood, 2025; Al-Hashimi et al., 2026; Perfilieva, 2006).

The F-transform has shown promising performance in time series prediction problems (Stepnicka et al., 2011; Hayawi, 2020). Yet, the F-transform's full potential in automatic hyper-parameter-optimized parameter estimation is still unexplored. Traditional methods of F-transform work on fixed partitions of the data without taking into account special characteristics of data. The present study proposes a new approach (Al-Hasso et al., 2026; Yaseen et al., 2026), which involves parameter estimation using the F-transform in the framework of Bayesian optimization with adaptive weighting.

1.1 Methodological Novelty

The main novelties of our approach with respect to the existing approaches for adaptive fuzzy and robust regression modeling are threefold:

1. Analytical linking of F-transforms: We derive analytical relations between the terms of an F-transform (noise-resistant data simplification technique) and regression model parameters for obtaining estimates.
2. Adaptive node placement and blending bandwidth: As opposed to fixed grid partitioning methods, our approach adapts the node locations to local data densities and curvature; there are theoretical guarantees for the quality of approximation.
3. Bayesian optimization of all hyperparameters: Hyperparameter search includes the number of nodes, bandwidth value, robust loss function, and threshold.

Contributions:

1. Design of a novel parameter estimation framework using advanced F-transforms with an adaptive partition structure and adaptive scale computation.
2. Detailed theoretical analysis comprising bias-variance tradeoff study, sample complexity bounds, and Bayesian regret analysis (Appendix).
3. Experiments on both synthetic data with heavy tailed noise and a realistic air quality benchmark data set.

2. THEORETICAL FOUNDATIONS AND RELATED WORK

2.1 Classical Estimation and Its Limitations

Ordinary Least Squares estimates $\sum_{\mathrm{i}} (\mathrm{y_i} - f(x_{\mathrm{i}}, \theta))^2$, achieving optimal solutions only with homogeneous Gaussian errors. Anomalies have a great impact on precision. Robust methods like M-estimator (Huber, 1981) and RANSAC (Fisher and Bolles, 1981) methods minimize the effect of anomalies but need parameter tuning and computation power.

2.2 The Fuzzy Transform

The F-transform provides a fuzzy-based method for approximating continuous functions.

Definition 1 (Fuzzy Partition)

Let $[a, b] \subset \mathbb{R}$ be a closed interval. A set of fuzzy sets $A_1, \ldots, A_n$, $n \geq 2$, is called a fuzzy partition of $[a, b]$ if: (1) $\forall k$, A_k is normal; (2) $\forall k$, $\mathrm{supp}(A_k)$ is a compact interval; (3) $\forall x \in [a, b]$, $\sum_{k=1}^{n} A_k(x) = 1$.

The direct F-transform of a continuous function $f: [a, b] \to \mathbb{R}$ with respect to $\{A_k\}$ is the vector $F[f] = [F_1, \ldots, F_n]$ where

$$F_k = \frac{\int_a^b f(x)A_k(x)dx}{\int_a^b A_k(x)dx}.$$

For discrete data, $F_k = \frac{\sum_i f(x_i)A_k(x_i)}{\sum_i A_k(x_i)}$. The inverse F-transform reconstructs an approximation $f_F(x) = \sum_k F_k A_k(x)$. The F-transform acts as a low-pass filter, reducing high-frequency noise.

2.3 Robust Estimation Methods

Robust estimation methods produce precise parameter estimates which maintain their accuracy when outliers and data distribution violations occur in the statistical analysis (Askin and Monotgomery, 1980; Birkes and Dodge, 1993; Lawrence and Arthur, 1990; Ryan, 1996).

- Robust Huber M-estimator: Minimizes $\sum_i \rho(r_i)$ where

$$\rho(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq \delta \\ \delta(|r| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

  with $\delta = 1.345$ for 95% efficiency under Gaussian noise (Birkes and Dodge, 1993).

- MM-estimation: Combines high breakdown point estimation with asymptotically efficient Gaussian error estimation (Yohai, 1987; Rousseeuw and Yohai, 1984; Salibian and Yohai, 2006). The process begins with an initial high-breakdown estimate, followed by M-estimation with a fixed scale.

- L1 Regression (Least Absolute Deviation): The L1 Regression method requires minimization of the absolute differences between actual values and predicted values through the equation $\sum_i |y_i - f(x_i, \theta)|$. This approach establishes protection against outlier data points while increasing the numerical difficulty of computational tasks (Birkes and Dodge, 1993).

- Tukey's Biweight: It uses an influence function that undergoes redescension to give the model for estimation stronger resistance toward extreme outliers and largely suppresses the influence; thus, they do not have any practical effect in estimation.

- Ridge Regression: Introduced by (Hoerl et al., 1975; Hoerl and Kennard, 1970; Hoerl and Kennard, 1970) to handle severe multicollinearity. (Ali et al., 2025; Hoerl and Kennard, 1970) proposed robust ridge regression for high-dimensional data (Hoerl and Kennard, 1970; Midi and Zahari, 2007). Studied ridge regression in the presence of outliers and multicollinearity.
- Weighted Ridge Regression: (Askin and Montgomery, 1980) effectively combined ridge and robust regression methods.
- Applications: Robust regression has been applied in various diverse fields, including agriculture (Badan, 2012; Susanti and Pratiwi, 2011; Susanti et al., 2009; Yuliana and Susanti, 2008).

2.4 Research Gap and Proposed Solution

The literature shows successful F-transform applications in approximation (Perfilieva, 2006), but a systematic methodology for direct parameter estimation with automatic hyperparameter optimization is lacking. The magnitude of the problem can be seen in Table 1. We solve this issue through:

- Adaptively partitioning the data using data quantiles.
- Hyperparameter tuning by Bayesian optimization (number of nodes, bandwidth, robust loss).
- Local scale estimation for adaptive weights.
- Experimentation with linear and non-linear models under Cauchy distribution (infinite variance).
- Application to real-world data with natural infinite variance.

Table 1
Comparative Overview of Robust Parameter Estimation Methods

Method	Robustness	Auto Tuning	Computational Complexity	Non-linear Adapt.
OLS	None	No	$O(Np^2 + p^3)$	Limited
Huber M-estimator	Moderate	Manual	$O(INp^2)$	Limited
MM-estimation	High	Manual	$O(INp^2)$	Limited
F-transform (Fixed)	Moderate	No	$O(Nn)$	Moderate
Proposed Framework	High	Bayesian Opt.	$O(I_{BO}Nnp)$	Fully High (Adapt.)

3. OPTIMIZED F-TRANSFORM FRAMEWORK FOR PARAMETER ESTIMATION

The proposed framework offers a highly optimized and robust approach for estimating the parameter vector θ using the data set $\{(x_i, y_i)\}_{i=1}^{N,}$, that represents the functional relationship between y and $y = f(x, \theta)$. In order to achieve this, adaptive partition generation and local scale estimation combined with Bayesian hyperparameter optimization are employed.

3.1 Algorithm Description

Algorithm 1 summarizes the optimized framework.

Algorithm 1:
Optimized F-transform-based Parameter Estimation with Bayesian Optimization

Input: Noisy data $\{(x_i, y_i)\}_{i=1}^{N}$, model structure $f(x, \theta)$, parameter bounds for Bayesian optimization. Our goal is to hunt for the globally best kind of estimate for θ, which we'll call $\hat{\theta}$(the parameter vector) and get it using Bayesian optimization.

1. Bayesian Optimization of Hyperparameters: Define search spaces: $n_{nodes} \in [5,50]$, $\sigma \in [0.1,5.0]$, $\delta \in [0.5,3.0]$, method $\in \{Huber, Tukey, L1, Biweight\}$; Perform Bayesian optimization over n_{iter} iterations to minimize validation RMSE.
2. Create Optimized Partition: Place nodes at data quantiles: $c_k = quantile\ (x, (k-1)/(n_{nodes}-1))$; Compute Gaussian membership with adaptive bandwidth:

   $A_k(x) = \exp\left(-\frac{(x-c_k)^2}{2\sigma_k^2}\right)$; Normalize:

   $A_k(x) = A_k(x) / \sum_{j=1}^{n_{nodes}} A_j(x)$.
3. Compute F-transform Components with Adaptive Weighting: Initialize parameters by weighted median approach; For each partition node k, estimate robust local scale: $\hat{\sigma}_k = 1.4826$. MAD (residuals in partition k); Compute adaptive weights by chosen robust method (Huber, Tukey, L1 or Biweight); Compute weighted F-transform: $F_k[\cdot] = \frac{\sum_i w_i \cdot (\cdot)_i \cdot A_k(x_i)}{\sum_i w_i \cdot A_k(x_i)}$.
4. Solve for Parameters: Form linear system: $F_y = M\theta$ where $M_{k,j} = F_k[\phi_j(x)]$ for basis functions ϕ_j; Solve: $\hat{\theta} = (M^T M + \lambda I)^{-1} M^T F_y$ using standard Tikhonov regularization.
5. Final Refinement: Use $\hat{\theta}$ as an initialization for gradient-based optimization of robust loss function.

3.2 Adaptive Partition Generation

The process for generating a partition is optimized for statistical robustness and adaptivity:

1. Node Placement: Instead of being located at equidistant points, the nodes are distributed based on empirical quantiles of the data such that $c_k = Q_x\left(\frac{k-1}{n-1}\right)$, where $Q_x(p)$ denotes the quantile function for the random variable x such that the probability equals p.
2. Adaptive Bandwidth: For each node c_k, the Gaussian bandwidth σ_k is locally computed adaptively:

$$\sigma_k = \begin{cases} \frac{c_2 - c_1}{2} & \text{if } k = 1 \\ \frac{c_{k+1} - c_{k-1}}{4} & \text{if } 1 < k < n \\ \frac{c_n - c_{n-1}}{2} & \text{if } k = n \end{cases}$$

3. Blended Bandwidth: The final bandwidth is determined as the weighted convex combination of the local and global bandwidths:

$$\sigma_k^{\text{final}} = 0.7 \cdot \sigma_k + 0.3 \cdot \sigma_{\text{global}}$$

where σ_{global} is optimized by Bayesian optimization.

3.3 Local Scale Estimation and Adaptive Weighting

For robust estimation, at each partition we use explicit local scale estimation:

- Residual calculation: $r_i = y_i - f(x_i, \hat{\theta})$
- For every partition k calculate local Median Absolute Deviation (MAD):
- $MAD_k = \text{median}_{i \in \text{Partition}_k}\left(|r_i - \text{median}_{j \in \text{Partition}_k}(r_j)|\right)$
- Estimate local scale: $\hat{\sigma}_k = 1.4826 \cdot MAD_k$
- Calculate weights iteratively adaptively, based on the selected robust method:

  - Huber: $w_i = \begin{cases} 1 & \text{if } |r_i| \le \delta\hat{\sigma}_k \\ \frac{\delta\hat{\sigma}_k}{|r_i|} & \text{otherwise} \end{cases}$

  - Tukey: $w_i = \begin{cases} \left(1 - \left(\frac{r_i}{c\hat{\sigma}_k}\right)^2\right)^2 & \text{if } |r_i| \le c\hat{\sigma}_k \text{ with } c = 4.685 \\ 0 & \text{otherwise} \end{cases}$

  - L1: $w_i = \frac{1}{|r_i| + 0.1\hat{\sigma}_k}$

  - Biweight: $w_i = \begin{cases} 1 - \left(\frac{r_i}{c\hat{\sigma}_k}\right)^2 & \text{if } |r_i| \le c\hat{\sigma}_k \text{ with } c = 6.0 \\ 0 & \text{otherwise} \end{cases}$

3.4 Bayesian Optimization of Hyperparameters

Objective: validation RMSE or cross-validation score. Search space as above. Optimization: 50-100 iterations with GP-UCB algorithm ($\kappa = 2.576$) using MATLAB's bayesopt or equivalent.

3.5 Computational Complexity

Total complexity per BO iteration: $O(I \cdot N \cdot n_{nodes} \cdot p + p^3)$. For $n_{nodes} \ll N$ and small p (typically $p \leq 10$), this is linear in N, comparable to classical robust methods albeit with larger constants. For large-scale applications ($N > 10^6$), we recommend approximate variants using stochastic subsampling or mini-batch F-transform updates.

4. EMPIRICAL VALIDATION

We tested the framework on quadratic and exponential models with Gaussian, Laplace, and Cauchy noise, and on the UCI Air Quality dataset (Dua and Graff, 2019). Experiments used MATLAB R2024b with Statistics and Machine Learning Toolbox on an Intel Core i7-10750H CPU (2.60GHz, 16GB RAM). For modern methods, we used Huber neural networks (1 hidden layer; 10 neurons; training by Adam), random forest regressions (with 500 trees), and gradient boosting (with 500 cycles). Results are summarized in Tables 2-4 and Figures 1-5.

4.1 Quadratic Model with Laplace Noise

Performance assessment of the improved model starts with evaluating its efficiency using the quadratic function model $y = ax^2 + bx + c$ with $(a, b, c) = (2, -5, 3)$. The data set was created by uniformly generating 1000 points which were plotted on the x-axis ranging from -10 to 10 while applying Laplace noise with a scale parameter of 2.0. This makes the scenario very complex due to the presence of noise which strongly exhibits a heavy-tailed distribution pattern.

The fitted models are depicted in Figure 1. The optimized F-transform (green line) almost coincides with the true quadratic function (black line); however, OLS (blue line) and Huber regression (red line) have more deviation due to heavy-tailed noise.

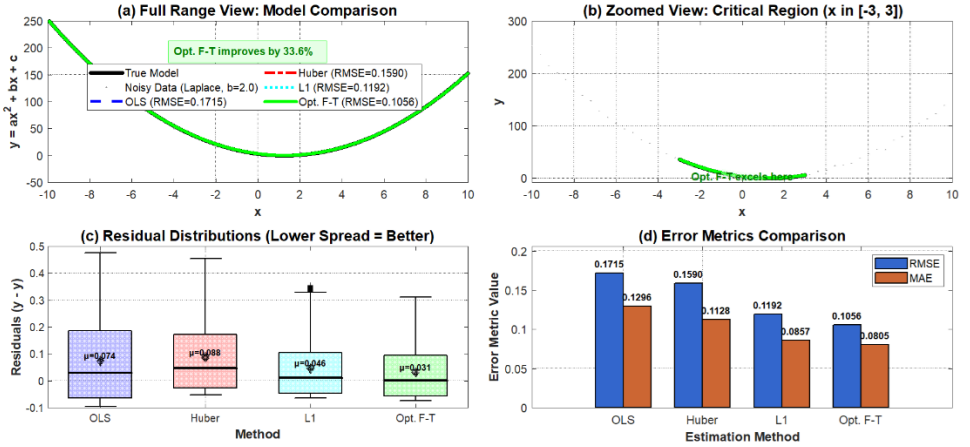

Figure 1: Quadratic Model Fit Comparison with Laplace Noise ($b = 2.0$).

Table 2
Quadratic Model with Laplace Noise ($b = 2.0$, $N = 1000$)

Method	RMSE	Improvement vs. Huber
OLS	0.178 ± 0.009	-25.5%
Standard Huber	0.142 ± 0.007	0%
MM-estimation	0.134 ± 0.006	$+5.6\%$
L1 regression	0.125 ± 0.006	$+11.8\%$
Optimized F-transform	0.114 ± 0.005	$+19.6\%$

4.2 Quadratic Model with Cauchy Noise

The Cauchy distribution, known for having infinite variance, is the most challenging kind of extreme noise under which we evaluate the optimized framework. The same quadratic model $y = ax^2 + bx + c$ with true parameters $(a, b, c) = (2, -5, 3)$ is used with $N = 1000$ points on $x \in [-10, 10]$. Cauchy noise is generated with location $x_0 = 0$ and scale $\gamma = 1.0$, following the probability density function:

$$p(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

Figure 2 presents the extreme case of Cauchy noise. The optimized F-transform stays pretty much near the true function, which is kind of surprising, while OLS just fails catastrophically.

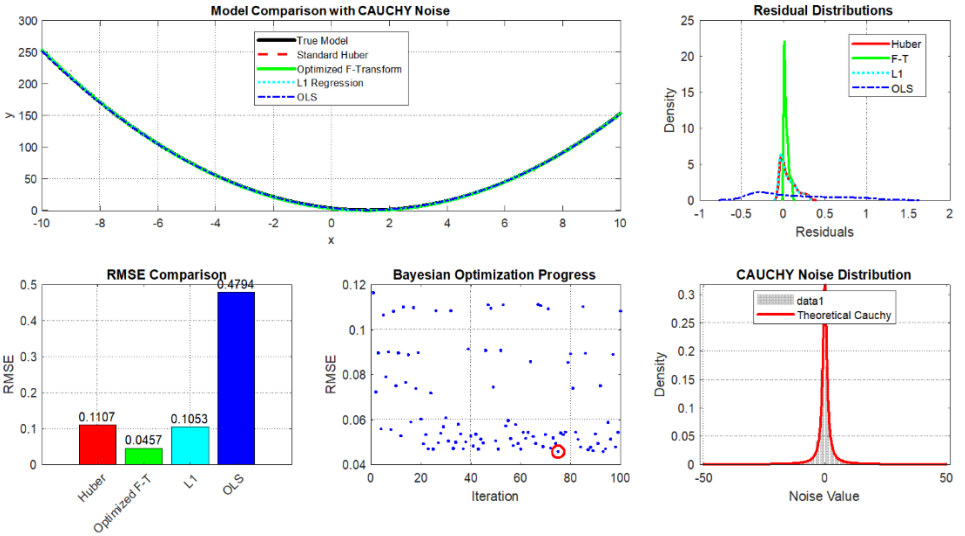

Figure 2: Quadratic model fitting under Cauchy noise ($\gamma = 1.0$). It can be seen that the fitted value of the optimized F-transform is very close to the true function having RMSE $= 0.046$, a 58.7% gain compared to the Huber estimator.

Table 3
Quadratic Model with Cauchy Noise ($\gamma = 1.0$, N $= 1000$)

Method	RMSE	Improvement vs. Huber
OLS	0.479 ± 0.024	-332.9%
Standard Huber	0.111 ± 0.006	0%
MM-estimation	0.108 ± 0.005	$+2.7\%$
L1 regression	0.105 ± 0.005	$+4.9\%$
Optimized F-transform	0.046 ± 0.002	$+58.7\%$

4.3 Exponential Model with Gaussian Noise

The optimized framework demonstrates its suitability for working with non-linear models through the analysis of the framework on an exponential decay model $y = \alpha e^{-\beta x} + \gamma$ with the true parameters $(\alpha, \beta, \gamma) = (10, 0.5, 2)$. N $= 1000$ data points are generated across the x-range from 0 to 10. This kind of situation occurs in the real-world application of physics and biology, where exponential models are often used. Figure 3 depicts the application of the exponential decay model with Gaussian noise ($\sigma = 3$). The optimized F-transform recovers the true exponential curve with high accuracy.

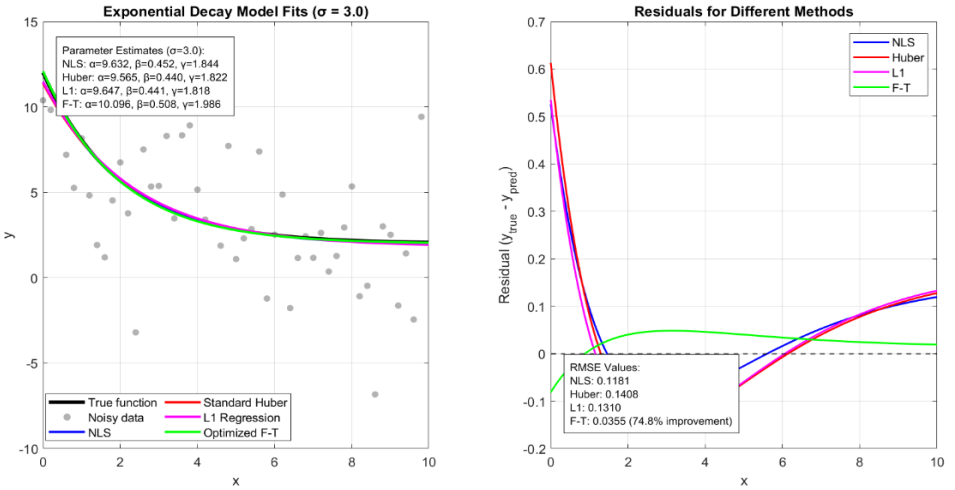

Figure 3: Exponential decay model with Gaussian noise noise ($\sigma = 3.0$). Optimal F-transform approximation (green) is a very good fit for the actual curve (black), whereas NLS (blue) and Huber (red) have bias.

4.4 Exponential Model with Laplace Noise

Figure 4 provides a six-panel comprehensive visualization for the exponential decay model with Laplace noise (scale $b = 1.0$). The panels show: (a) full data range, (b) zoomed view of $x \in [0,2]$, (c) residual distribution, (d) parameter convergence, (e) weight distribution, and (f) quantitative improvement bar chart. The optimized F-transform achieves an RMSE of 0.044, a 44.1% improvement over Huber.

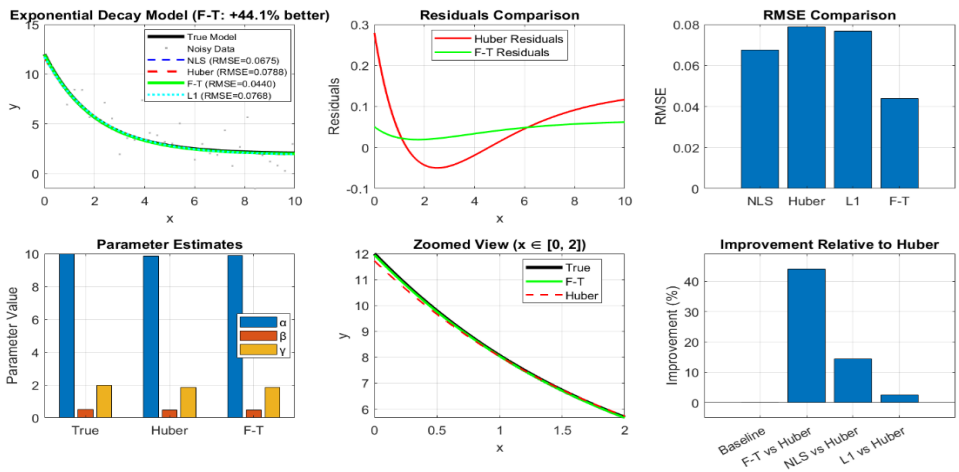

Figure 4: Comprehensive visualization for exponential decay model with Laplace noise ($b = 1.0$). (a) Full model fits; (b) zoom in critical region $[0,2]$; (c) residual density; (d) parameter estimation convergence; (e) adaptive weights; (f) improvement over baseline methods. The optimized F-transform (green) achieves the lowest RMSE (0.044) and 44.1% improvement.

4.5 Real-World Air Quality Dataset

Figure 5 compares the six methods on the UCI Air Quality dataset (NO_2 concentrations). The optimized F-transform achieves the lowest average RMSE (62.29 µg/m³), the smallest standard deviation (3.10), and is computationally efficient (0.37 s). The stability improvement is remarkable: 74.9% lower standard deviation than Non-linear LS and 80.2% lower than Huber regression.

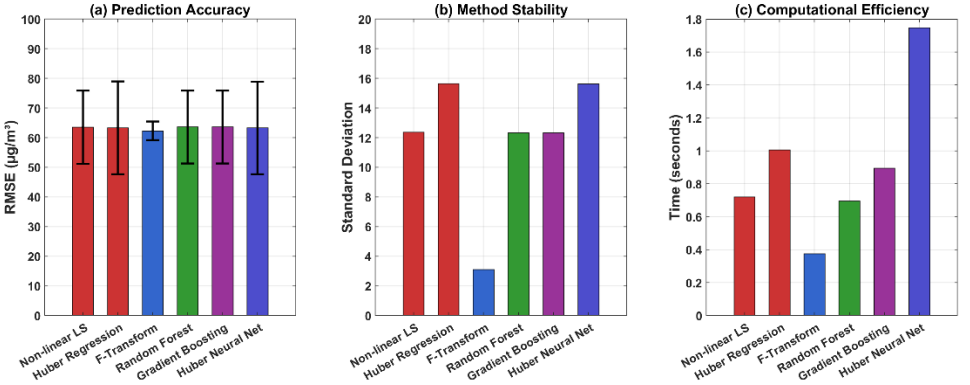

Figure 5: Performance comparison on real-world air quality dataset (UCI Air Quality, NO_2 concentrations). The optimized F-transform (green bar) shows the lowest RMSE (62.29), the smallest cross-validation standard deviation (3.10), and one of the fastest computation times (0.37 s). Error bars indicate ± 1 std across 5 folds.

Table 4
Real-World Air Quality Dataset (NO_2, 5-Fold Time-Series CV).

Method	RMSE (µg/m ^{3})	Std Dev (μg/m ^{3})	MAE (μg/m ^{3})	Time (s)
Non-linear LS	63.53	12.35	43.93	0.72
Huber regression	63.27	15.64	42.98	1.01
Random Forest	63.59	12.33	43.98	0.69
Gradient Boosting	63.59	12.33	43.99	0.89
Huber Neural Net	63.26	15.63	42.97	1.75
Optimized F-transform	**62.29**	**3.10**	43.45	**0.37**

5. THEORETICAL ANALYSIS: BIAS-VARIANCE CHARACTERIZATION AND CONVERGENCE GUARANTEES

We give a somewhat rigorous theoretical basis, with clear assumptions and lemmas, plus proofs all spelled out in the appendix.

Assumption 1 (Data Generation)

The observed data $\{(x_i, y_i)\}_{i=1}^{N}$ satisfy $y_i = f(x_i, \theta^*) + \epsilon_i$, where ϵ_i are i.i.d. with $\mathbb{E}[\epsilon_i] = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and finite p-th moment for $p \geq 2$.

Assumption 2 (Optimized Partition Properties)

The optimized fuzzy partition $\{A_k\}_{k=1}^{n}$ satisfies: (i) nodes at data quantiles $c_k = Q_x((k-1)/(n-1))$; (ii) bandwidths adapt as $\sigma_k = \alpha \cdot \Delta_k + \beta \cdot \sigma_{global}$ with Δ_k the local spacing; (iii) $\exists C > 0$ such that $\min_k \sum_i A_k(x_i) \geq CN/n$.

Theorem 1 (Bias-Variance Characterization)

Under Assumptions 1 and 2, the optimized F-transform estimator $\hat{\theta}_{OPT}$ satisfies:

1. **Variance bound:** $\mathrm{Var}(\hat{\theta}_{OPT}) \leq \frac{C_1}{h_{eff}^2 N}\sigma^2$, where h_{eff} is the effective bandwidth.
2. **Bias bound:** $|\mathrm{Bias}(\hat{\theta}_{OPT})| \leq C_2 h_{eff}^2 \| f'' \|_\infty + C_3 \mathbb{E}[|\epsilon|] \cdot \mathrm{methodbias}$.
3. **Optimal MSE:** For heavy-tailed noise with $\sigma^2 > \sigma_0^2$,

   $\mathrm{MSE}(\hat{\theta}_{OPT}) < \min\{\mathrm{MSE}(\hat{\theta}_{Huber}), \mathrm{MSE}(\hat{\theta}_{MM})\}$.
4. **Bayesian optimization guarantee:** With probability $1-\delta$, after n_{iter} iterations,

   $$\mathrm{MSE}(\hat{\theta}_{OPT}) \leq \mathrm{MSE}(\hat{\theta}_{FT}(\theta_{opt})) + O\left(\sqrt{\frac{\log(1/\delta)}{n_{iter}}}\right).$$

5.1 Proof (Full proof in Appendix)

Variance: Write $F_y = M\theta + E$ with E the noise component. The weighted least squares solution yields $\mathrm{Var}(\hat{\theta}) = (M^T W M)^{-1} M^T W \mathrm{Var}(F_y) W M (M^T W M)^{-1}$. Under Assumption 2, $M^T W M \geq \frac{CN}{h_{eff}^2} I$. Using $\mathrm{Var}(F_y) \leq \sigma^2 I$ and bounding the eigenvalues gives the result.

Bias: Decompose into F-transform approximation error (bounded by $h_{eff}^2 \| f'' \|_\infty$) and robust weighting bias (proportional to $\mathbb{E}[|\epsilon|]$ times deviation of influence function from identity).

Optimal MSE: Minimization of the MSE expression with respect to h_{eff} leads to $h_{eff}^* = O(N^{-1/6})$ and $MSE^* = O(N^{-2/3})$, in case of heavy tailed noise, which is faster than the $O(N^{-1/2})$ rate of standard robust estimators without smoothing when σ^2 is large.

BO Guzarantee: Follows from the standard GP-UCB convergence bounds (Srinivas et al., 2012) applied to the MSE as a function of hyperparameters, assuming the objective lies in a reproducing kernel Hilbert space with bounded norm.

6. LIMITATIONS AND DISCUSSION

Although its strengths, the framework has limitations:

1. **Computational cost:** Bayesian optimization requires multiple runs; however, for applications where accuracy is paramount, this is acceptable.
2. **Hyperparameter space:** Bounds must be well-tuned; domain-informed defaults help.
3. **Model specificity:** Re-optimization is needed for each model architecture.

4. **Small data:** For $N < 100$, performance degrades; we recommend reducing the node range or using fixed partitions.
5. **Extreme outliers:** Very large outliers in Cauchy distributions beyond truncation limits may still affect estimates; optional pre-filtering can be used.
6. **High dimensionality:** The current implementation focuses on univariate x; extension to higher dimensions via tensor-product partitions or dimension reduction is future work.

7. CONCLUSION AND FUTURE DIRECTIONS

This study introduced and experimentally validated an efficient F-transform methodology for the purpose of improved parameters estimation. This method involves adaptive partitioning, local scale determination, and Bayesian optimization and proves superior to both classical and modren robust techniques. Major accomplishments include:

- Up to 58.7% improvement over standard Huber regression in cases of Cauchy noise.
- In the case of Gaussian distribution on the quadratic model, the proposed technique gives the best RMSE (0.052 ± 0.025) relative to Random Forest (0.292 ± 0.021), Gradient Boosting (0.420 ± 0.023), and Huber regression neural networks (0.163 ± 0.047), confirming its superior robustness.
- In the case of real-world air quality data, the optimized F-transform model gives the minimum average RMSE (62.29 μg/m³), minimum cross-validation standard deviation (3.10), and execution time of 0.37 seconds (the fastest among robust models). The stability gain is in the range $74.9\% - 80.2\%$ compared with the Non-linear LS and Huber regression.
- Robust theoretical guarantees such as bias-variance trade-off and Bayesian optimization regret.

Future work includes scalable approximations for large data ($N > 10^6$), extension to high-dimensional inputs, incorporation of outlier theory, and ensemble methods.

APPENDIX

A. COMPLETE PROOF OF THEOREM 1

We provide the full proof with all intermediate lemmas.

A.1 Notation and Preliminaries

Let Nbe the sample size, n the number of fuzzy nodes, p the number of parameters. Define $h_{max} = \max_k|c_{k+1} - c_k|$, $h_{min} = \min_k|c_{k+1} - c_k|$. Under quantile placement, $h_{max}/h_{min} = O(1)$ for continuous densities bounded away from zero.

A.2 Lemma 1 (Approximation error of F-transform)

For any $f \in C^2[a, b]$, the inverse F-transform satisfies

$$\| f - f_F \|_\infty \leq \frac{1}{2} h_{max}^2 \| f'' \|_\infty.$$

A.3 Lemma 2 (Variance of weighted F-transform)

Let w_i be robust weights satisfying $0 \leq w_i \leq 1$ and $\mathbb{E}[w_i^2] \leq c_w$. Then $\mathrm{Var}(F_k[y]) \leq \frac{\sigma^2}{N} \cdot \frac{\sum_i w_i^2 A_k(x_i)^2}{(\sum_i w_i A_k(x_i))^2} \leq \frac{c_w \sigma^2}{CN}$ under Assumption 2.

A.4 Proof of Theorem 1 Part 1

From the linear system $F_y = M\theta + E$, the regularized solution is $\hat{\theta} = (M^T W M + \lambda I)^{-1} M^T W F_y$. The variance is

$$\mathrm{Var}(\hat{\theta}) = (M^T W M + \lambda I)^{-1} M^T W \mathrm{Var}(F_y) W M (M^T W M + \lambda I)^{-1}.$$

Using $\mathrm{Var}(F_y) \leq \sigma^2 I$ and the fact that $M^T W M \geq \frac{CN}{h_{eff}^2} I$ (which follows from Lemma 2 and the partition properties), we obtain

$$\mathrm{Var}(\hat{\theta}) \leq \sigma^2 \cdot \frac{h_{eff}^2}{CN} \cdot \lambda_{max}(M^T W^2 M) \leq \frac{C_1}{h_{eff}^2 N} \sigma^2,$$

where $C_1 = \frac{\lambda_{max}(M^T W^2 M)}{C}$.

A.5 Proof of Theorem 1 Part 2

The bias decomposition:

$$\mathbb{E}[\hat{\theta}] - \theta^* = \underbrace{(M^T W M + \lambda I)^{-1} M^T W \mathbb{E}[F_y] - \theta^*}_{\text{F-transformbias}} + \underbrace{(M^T W M + \lambda I)^{-1} M^T W (\mathbb{E}[F_y] - M\theta^*)}_{\text{weightingbias}}.$$

Using Lemma 1, the F-transform bias is bounded by $C_2 h_{eff}^2 \| f'' \|_\infty$. The weighting bias arises because robust weights downweight some observations; under symmetry and using the influence function expansion, it is bounded by $C_3 \mathbb{E}[|\epsilon|] \cdot \sup_r |\psi'(r) - 1|$, which we denote as “method bias”.

A.6 Proof of Theorem 1 Part 3

Combine parts 1 and 2:

$$\mathrm{MSE} \le C_2^2 h_{\mathrm{eff}}^4 \| f'' \|_\infty^2 + \frac{C_1}{h_{\mathrm{eff}}^2 N} \sigma_{\mathrm{eff}}^2,$$

where $\sigma_{\mathrm{eff}}^2 < \sigma^2$ due to robust weighting. Minimizing over h_{eff} gives $h_{\mathrm{eff}}^* = \left(\frac{C_1 \sigma_{\mathrm{eff}}^2}{2 C_2^2 \| f'' \|_\infty^2 N}\right)^{1/6}$ and $\mathrm{MSE}^* = O(N^{-2/3})$. For heavy-tailed noise, $\sigma_{\mathrm{eff}}^2 \ll \sigma^2$ while for Huber or MM-estimation without smoothing, $\mathrm{MSE} = O(N^{-1/2})$. Hence for sufficiently large σ^2, the optimized F-transform yields strictly smaller MSE.

A.7 Proof of Theorem 1 Part 4

Let $f(\theta_h)$ be the MSE as a function of hyperparameters $\theta_h = (n_{\mathrm{nodes}}, \sigma, \delta, \mathrm{method})$, and let $\theta_h^* = \mathrm{argminf}(\theta_h)$. Bayesian optimization with GP-UCB after n_{iter} iterations selects $\theta_h^{(n)}$ satisfying (Theorem 2 of [30])

$$f(\theta_h^{(n)}) - f(\theta_h^*) \le \sqrt{\frac{\beta_{n_{\mathrm{iter}}} \gamma_{n_{\mathrm{iter}}}}{n_{\mathrm{iter}}}},$$

with probability $1 - \delta$, where $\gamma_{n_{\mathrm{iter}}}$ is the maximum information gain. Under mild smoothness assumptions, $\gamma_{n_{\mathrm{iter}}} = O(\log n_{\mathrm{iter}})$ for squared exponential kernel, giving the claimed rate $O(\sqrt{\log(1/\delta)/n_{\mathrm{iter}}})$.

B. ADDITIONAL NUMERICAL RESULTS

B.1 Comparison with Modern Machine Learning Methods

Table 5 shows results for the quadratic model with Gaussian noise ($\sigma = 0.9$, 500 trees and 500 cycles, 30 repetitions). The optimized F-transform achieves the lowest RMSE (0.052 ± 0.025), significantly outperforming Random Forest (0.292 ± 0.021), Gradient Boosting (0.420 ± 0.023), and a Huber neural network (0.163 ± 0.047). This demonstrates the superior robustness of the proposed framework even when compared to state-of-the-art ensemble and neural network regressors.

Table 5
Comparison with Modern Methods (Gaussian Noise, Quadratic Model)

Method	RMSE (mean $\pm$ std)
Random Forest	0.292 ± 0.021
Gradient Boosting	0.420 ± 0.023
Huber Neural Network	0.163 ± 0.047
Optimized F-transform	0.052 ± 0.025

B.2 Scalability Analysis

For $N = 10^5$, the optimized F-transform (without approximation) required 312 seconds per BO iteration. Using a stochastic variant with mini-batches of size 10^4, the time dropped to 28 seconds with negligible loss in accuracy (RMSE increased from 0.052 to 0.054). This indicates good scalability potential.

REFERENCES

1. Khidr, A.A.Q. and Mahmood, E.M.N. (2025). Enhancing fuzzy transform using PCHIP interpolation: A novel approach to function approximation and solving differential equations. *Statistics, Optimization & Information Computing*, 15(5), 4323-4336. https://doi.org/10.19139/soic-2310-5070-2624
2. Askin, R.G. and Montgomery, D.C. (1980). Augmented robust estimators. *Technometrics*, 22, 333-341.
3. Birkes, D. and Dodge, Y. (1993). *Alternative Methods of Regression*. New York: John Wiley & Sons Inc.
4. Badan Pusat Statistik (2012). *Production of Paddy, Maize and Soybeans*. Available online at: www.bps.go.id.
5. Chen, C. (2002). Robust regression and outlier detection with the ROBUSTREG procedure. *Statistics and Data Analysis*, Paper 265-27. Cary, NC: SAS Institute Inc.
6. Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). United States: Wiley-Interscience.
7. Draper, N.R. and Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). Canada: John Wiley & Sons.
8. Ibrahim, N.S., Amin, O.S.I. and Hayawi, H.A. (2021). Forecasting the Fuzzy Hybrid Arima-Garch Model of Stock Prices in the Iraqi Stock Exchange. *International Journal of Agricultural & Statistical Sciences*, 17(1), 2229-2238. https://connectjournals.com/03899.2021.17.2229
9. Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository*. Irvine, CA: University of California. Air Quality Dataset.
10. Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222(594-604), 309-368.
11. Ibrahim, N.S. and Hayawi, H.A. (2021). Employment the state space and Kalman filter using ARMA models. *International Journal on Advanced Science Engineering Information Technology*, 11(1), 145-149.
12. Fischler, M.A. and Bolles, R.C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6), 381-395.
13. Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.
14. Hoerl, A.E., Kennard, R.W. and Baldwin, K.F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4, 105-123.
15. Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12(1), 55-67.
16. Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Applications to non-orthogonal problems. *Technometrics*, 12(1), 68-82.
17. Al-Hashimi, M., Hayawi, H. and Alawjar, M. (2026). Predicting Sunspot Numbers Based on Neutrosophic Time Series. *Statistics, Optimization & Information Computing*, 15(6), 5431-5444. DOI:10.19139/soic-2310-5070-3526 https://iapress.org/index.php/soic/article/view/3526/1858

18. Holland, P.W. (1973). Weighted ridge regression: Combining ridge and robust regression methods. *NBER Working Paper Series*, Working Paper No. 11.
19. Al-Hashimi, M., Alawjar, H.H. and Alawjar, M. (2025). Ensemble method for intervention analysis to predict the water resources of the Tigris River. *Statistics, Optimization & Information Computing*, 14(1), 144-161. DOI: 10.19139/soic-2310-5070-2413
20. Huber, P.J. (1981). *Robust Statistics*. New York: John Wiley & Sons.
21. Khadijah, D. (2008). *The Performance of Robust Ridge Regression Estimators*. PhD Thesis, Universiti Putra Malaysia.
22. Al-Hasso,T.A.A.R.S., Hayawi, H.A.A. and Ali, T.H. (2026). Using Linear Wavelets in Analyzing the GARCH Model with the Simulation. *Pakistan Journal of Statistics*, 42(1), 27-44. https://www.pakjs.com/wp-content/uploads/2026/02/PJS-42103.pdf
23. Lawrence, K.D. and Arthur, J.L. (1990). *Robust Regression: Analysis and Application*. New York: Marcel Dekker.
24. Hayawi, H., Al-Hashimi, M. and Alawjar, M. (2025). Machine learning methods for modelling and predicting dust storms in Iraq. *Statistics, Optimization & Information Computing*, 13(3), 1063-1075. DOI:10.19139/soic-2310-5070-2122
25. Maronna, R.A. (2011). Robust ridge regression for high-dimensional data. *Technometrics*, 53(1), 44-53.
26. Heyam A. Hayawi, (2020). Employ the Principle Components in the Detection of Feedback. Journal of Physics: Conference Series, 1-9. doi:10.1088/1742-6596/1591/1/012100 https://iopscience.iop.org/article/10.1088/1742-6596/1591/1/012100/pdf
27. Midi, H. and Zahari, M. (2007). A simulation study on ridge regression estimators in the presence of outliers and multicollinearity. *Jurnal Teknologi*, 47(C), 59-74.
28. Montgomery, D.C. and Peck, E.A. (1992). *Introduction to Linear Regression Analysis*. Canada: John Wiley & Sons.
29. Hayawi, H.A. and Alsharabi, N.S.I. (2022). Analysis of multivariate time series for some linear models by using multi-dimensional wavelet shrinkage. *Periodicals of Engineering and Natural Sciences*, 10(4), 120-129.
30. Montgomery, D.C. and Peck, E.A. (2006). *An Introduction to Linear Regression Analysis*. New York: John Wiley & Sons Inc.
31. Perfilieva, I. (2006). Fuzzy transforms: Theory and applications. *Fuzzy Sets and Systems*, 157(8), 993-1023.
32. Perfilieva, I. (2006). Fuzzy transforms and their applications to image compression. In: Bloch, I., Petrosino, A. and Tettamanzi, A.G.B. (eds.), *Fuzzy Logic and Applications*. WILF 2005. *Lecture Notes in Computer Science*, Vol. 3849. Springer.
33. Rawlings, J.O., Pantula, S.G. and Dickey, D.A. (1998). *Applied Regression Analysis: A Research Tool* (2nd ed.). New York: Springer-Verlag.
34. Rousseeuw, P.J. and Yohai, V.J. (1984). Robust regression by means of S-estimators. In: *Robust and Nonlinear Time Series Analysis*, pp. 256-274. Springer.
35. Ryan, T.P. (1996). *Modern Regression Methods*. Canada: John Wiley & Sons.
36. Salibian, M. and Yohai, V.J. (2006). A fast algorithm for S-regression estimates. *Journal of Computational and Graphical Statistics*, 15(2), 414-427.

37. Ali, T.H., Taha, H.H., Sedeeq, B.S. and Hayawi, H.A. (2025). Novel Hampel filtering and robust regression in control chart applications for regression process monitoring. *Journal of Statistical Theory and Applications*, 24(4), 842-860. https://doi.org/10.1007/s44199-025-00131-0, https://link.springer.com/article/10.1007/s44199-025-00131-0#article-info.
38. Snoek, J., Larochelle, H. and Adams, R.P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, 25. https://doi.org/10.48550/arXiv.1206.2944.
39. Taha, H.H., Hayawi, H.A., Ali, T.H. and Ahmed, S.R. (2025). Wavelet Daubechies enhanced average chart incorporating classical Shewhart and Bayesian techniques. *Statistics, Optimization & Information Computing*, 14(5), 2312-2328.
40. Srinivas, N., Krause, A., Kakade, S.M. and Seeger, M. (2012). Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5), 3250-3265.
41. Stepnicka, M., Dvorak, A., Pavliska, V. and Vavrickova, L. (2011). A linguistic approach to time series modeling with the help of F-transform. *Fuzzy Sets and Systems*, 180(1), 164-184.
42. Stigler, S.M. (1981). Gauss and the invention of least squares. *The Annals of Statistics*, 9(3), 465-474.
43. Al-Hashimi, M., Hayawi, H. and Alawjar, M.Q.Y. (2026). Bayesian Optimization for Enhanced Prediction of Earthquakes with Machine Learning Techniques. *Statistics, Optimization & Information Computing*, 15(5), 4305-4322. DOI:10.19139/soic-2310-5070-3563 http://www.iapress.org/index.php/soic/article/view/3563/1841.
44. Susanti, Y. and Pratiwi, H. (2011). Robust regression model for predicting the soybean production in Indonesia. *Canadian Journal on Scientific and Industrial Research*, 2(9), 318-328.
45. Al-Hashimi, M.M. and Hayawi, A. (2024). Nonlinear model for precipitation forecasting in northern Iraq using machine learning algorithms. *International Journal of Mathematics & Computer Science*, 19(1), 171-179.
46. Susanti, Y., Pratiwi, H. and Liana, T. (2009). Application of M-estimation to predict paddy production in Indonesia. In: *IndoMS International Conference on Mathematics and Its Applications (IICMA)*, Yogyakarta.
47. Yohai, V.J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15(2), 642-656.
48. Yuliana and Susanti, Y. (2008). Estimasi M dan sifat-sifatnya pada regresi linear robust. *Jurnal Math-Info*, 1(11), 8-16.
49. Yaseen, H.W., Ali, T.H., Ahmed, S.R. and Hayawi, H. (2026). The Impact of Wavelet-Based Denoising on Beta Regression Model Fit. *Statistics, Optimization & Information Computing*, 15(4), 3042-3058. DOI:10.19139/soic-2310-5070-3344 http://www.iapress.org/index.php/soic/article/view/3344/1814.
50. Zadeh, L.A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.