Electricity Consumption Anomaly Detection Using K-Means and ARIMA with Enhanced Preprocessing

[*]Efa Yumna Purwono[1], Sri Arttini Dwi Prasetyowati[2]

[1,2] Department of Electrical Engineering, Sultan Agung Islamic University, Semarang, Indonesia
[1]efayumna@std.unissula.ac.id, [2]arttini@unissula.ac.id

Abstract

This study proposes an integrated K-Means and ARIMA framework for electricity-consumption anomaly detection at PT PLN UP3 Semarang. K-Means segments customers into low, medium, and high consumption groups, with the optimal k selected using the Elbow method, Silhouette score, and Davies–Bouldin index; the resulting k=3 clusters support segment-specific modeling. Within each cluster, ARIMA models predict monthly consumption and are validated using ACF/PACF diagnostics and AIC. Anomalies are identified from forecast residuals using cluster-specific thresholds calibrated on training data: $\mu \pm 2\sigma$ when residuals pass normality tests (Shapiro–Wilk $p \geq 0.05$); $Q3$+1.5IQR(or Q1−1.5IQR) when normality is rejected; and Median±3MAD for highly skewed or small-sample clusters. Using real monthly data from January 2012 to February 2024 (n=146 months) with a strict temporal split (train: 2012–2022; test: 2023–Feb 2024), the framework achieves Precision = 0.88, Recall = 0.85, and F1-Score = 0.86, outperforming a clustering-only baseline (F1-Score = 0.72). The segment-based design improves anomaly sensitivity, interpretability, and operational relevance for proactive energy management. Future work will incorporate engineering criteria such as power-factor degradation and sustained voltage deviation to specifically flag technical anomalies.

Keywords: Anomaly Detection, K-Means Clustering, ARIMA, Electricity Consumption, Smart Grid

1. Introduction

Electricity consumption in Semarang, Central Java, has risen markedly alongside rapid economic and industrial growth, population expansion, and extensive infrastructure development. This surge challenges PT PLN (Persero) UP3 Semarang in terms of network stability, resource allocation, and service quality (Kardi et al., 2021). Complex consumption patterns across residential, commercial, and industrial customers shaped by usage time, equipment, and monthly cycles highlight the limitations of uniform, one-size-fits-all energy management. Consequently, careful segmentation and data-driven analysis are essential to support smart-grid initiatives and demand-side management strategies.

Prior studies typically applied either clustering or time-series forecasting (e.g., ARIMA) in isolation, without a comprehensive integration for precise anomaly identification in PLN's operational context. Many approaches relied on static clustering metrics such as the Sum of Squared Errors (SSE) to flag anomalies, which is ill-suited for time-dependent events (e.g., sudden spikes or sharp drops) indicative of operational issues (Al-Wakeel et al., 2017). In addition, robust preprocessing complete handling of missing values and systematic outlier treatment was often under-specified. The choice of the optimal k in K-Means was frequently heuristic, lacking quantitative justification via Silhouette Score and Davies–Bouldin Index, reducing reliability and generalizability.

This study addresses these shortcomings by proposing an integrated K-Means + ARIMA framework tailored to PT PLN (Persero) UP3 Semarang. We first cluster customers to obtain behaviorally homogeneous segments, then fit ARIMA models per cluster to produce calibrated forecasts whose residuals drive anomaly decisions. The combination leverages the computational efficiency and interpretability of K-Means and ARIMA, making the pipeline practical for operational environments with limited resources and tight response times (Rajabi et al., 2020). Our evaluation uses real monthly data from January 2012 to February 2024, with a strict temporal split (e.g., training 2012–2022; testing 2023–February 2024) to mirror deployment.

We cluster the monthly load profiles before anomaly detection to segment heterogeneous consumers (residential, commercial, industrial) into behaviorally homogeneous groups. Within each cluster, intra-cluster variance is lower, so thresholds computed from residuals (ARIMA forecast vs. actual) are tighter and yield higher precision. Empirically, detecting anomalies on mixed profiles produces many false positives; clustering first reduces variance and stabilizes residual distributions, enabling statistically consistent thresholds.

2. Methods

We adopt a two-stage, cluster-first pipeline (Figure 1). Monthly consumption data are collected and aligned to calendar months, then we use a strict temporal split training: January 2012–December 2022; testing: January 2023–February 2024. All statistics and models (imputation values, scalers, K-Means, ARIMA orders, and residual thresholds) are estimated on the training window only and applied unchanged to the test window to prevent leakage.

Stage A — Pre-processing. We reconcile meter records to calendar months; impute short gaps (≤1–2 months) with the training-window median and flag them; leave longer gaps unimputed; and screen out obvious recording errors using the IQR rule with box-plot verification. Features for distance-based clustering are min–max scaled to [0,1] with parameters fitted on training data and then reused for testing.

Stage B — Clustering → Forecasting → Detection. Customers are segmented with K-Means (k-means++ init, n_init=50, max_iter=500); the operating k is chosen by Elbow (WCSS), Silhouette, and Davies–Bouldin on scaled features; in our data k=3 (low/medium/high). For each cluster we fit an ARIMA/SARIMA model to the cluster-level monthly totals; ADF/KPSS assess stationarity, ACF/PACF suggest candidate orders, and AIC selects the final model on the training window. Anomalies are detected from one-step-ahead residuals using cluster-specific thresholds calibrated on training residuals (see Section 2.6); statistical candidates are promoted to technical anomalies using the confirmation rules in Section 2.7 (PF<0.85, voltage deviation >10%, or a 3σ month-to-month step change).

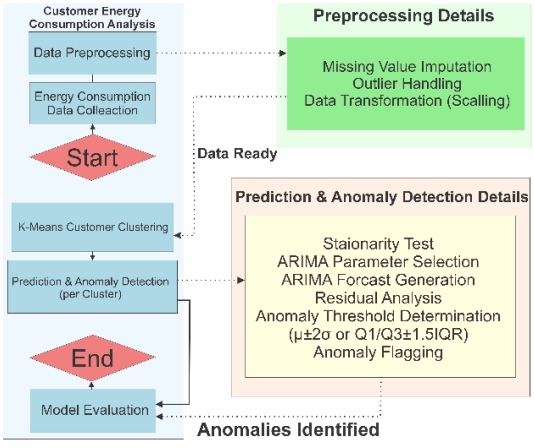

Figure 1. End-to-end research pipeline

Left: workflow from data collection → pre-processing → K-Means clustering → per-cluster ARIMA → residual thresholds → anomaly flagging → model evaluation.
Top-right (Pre-processing): Missing-value imputation (short gaps by training median), outlier handling (IQR + visual check), and feature scaling (min–max; fit on training only).
Bottom-right (Prediction & Detection): Stationarity tests (ADF/KPSS), order selection (ACF/PACF + AIC), one-step-ahead forecasts, residual analysis, threshold determination ($\mu \pm 2\sigma$ / IQR fences / Median±3×MAD), and anomaly flagging with optional technical confirmation (PF, voltage deviation, step-change).

2.1 Data Description

We use monthly electricity-consumption data (AMR/DPK) from PT PLN (Persero) UP3 Semarang covering January 2012–February 2024 (n = 146 months) across heterogeneous customer types (residential, commercial, industrial). Core variables are monthly active energy (kWh), power factor (PF), and feeder voltage (V). When feeder currents (R/S/T) are available, they are summarized monthly and reported in Ampere (A).

To mirror deployment and avoid leakage, we apply a strict temporal split: models and thresholds are fit on 2012–2022 (training, 132 months) and evaluated on January 2023–February 2024 (testing, 14 months).

2.2 Pre-processing

All meter records were first reconciled to calendar months, correcting any offsets introduced by billing cycles. Missing observations were handled conservatively: short gaps (≤1–2 months) were imputed with the median computed on the training window and tagged for subsequent audit, whereas longer gaps were left unimputed to avoid bias. (Yilmaz et al., 2019)Outliers were screened on the training data using the interquartile-range (IQR) criterion and verified with box-plot inspection; only entries attributable to recording errors were removed, while extreme yet plausible values were retained. For distance-based clustering, feature variables were min–max scaled to [0,1] using parameters fitted exclusively on the training window and then applied to the test period. Throughout, all preprocessing statistics (medians, quantiles, scaling parameters) were estimated on training data only to prevent information leakage and preserve evaluative integrity.

2.3 Why Cluster First?

Heterogeneous customers inflate variance and destabilize global thresholds, producing false positives. We therefore cluster first to obtain behaviorally homogeneous groups and then run forecasting and detection within clusters. This reduces intra-cluster variance, regularizes residual distributions, and enables statistically consistent, tighter cutoffs (see Section 1).

2.4 K-Means Clustering

Each customer's monthly profile is summarized into a feature vector (e.g., mean kWh, standard deviation, coefficient of variation, seasonal amplitude, peak/mean ratio, and annual Fourier terms sin(2πm/12), cos(2πm/12) (Mutiah et al., 2024). Clustering uses K-Means (k-means++ init, n_init=50, max_iter=500, fixed seed) with Euclidean distance.

$$d(x, \mu) = |x - \mu|_2 \quad (1)$$

$$WCSS(k) = \sum_i |x_i - \mu_{(c(i))}|_2^2 \quad (2)$$

The optimal k is selected by Elbow (WCSS), Silhouette

$$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \quad (3)$$

and Davies–Bouldin

$$DBI = \left(\frac{1}{k}\right)\sum_{c=1}^{k} max_{j \neq c} \frac{(S_c+S_j)}{M_{cj}} \quad (4)$$

Where S_c is mean intra-cluster distance and M_{CJ} is centroid distance. These criteria consistently support k=3 (low/medium/high consumption).

2.5 ARIMA per Cluster

For each cluster c, we model the cluster's aggregated monthly consumption $\widehat{\boldsymbol{y}}_{\{\boldsymbol{m},\boldsymbol{c}\}}$ with ARIMA/SARIMA to produce calibrated one-step-ahead forecasts:

$$\phi(B)(1-B)^d\,(1-B^{12})^D \boldsymbol{y}_{\boldsymbol{m},\boldsymbol{c}} = \mu + \theta(B)\epsilon_{m,c} \quad (5)$$

Where B is the backshift operator, d and D are non-seasonal and seasonal differencing orders (annual seasonality s Highly skewed / small-sample cluster. Use Median12). Orders (p,q,P,Q) are selected on the training window using ACF/PACF diagnostics and AIC. We use rolling-origin evaluation to generate forecasts $\widehat{\boldsymbol{y}}_{\{\boldsymbol{m},\boldsymbol{c}\}}$ and residuals:

$$r_{m,c} = y_{m,c} - \hat{y}_{m,c} \quad (6)$$

2.6 Residuals & Cluster-Specific Thresholds

Residual normality in each cluster is tested with Shapiro–Wilk (α=0.05) on training residuals. Cluster-specific thresholds are then fixed prior to scoring the test period:

- Approximately normal (p ≥ 0.05). Anomaly if
$$|r_{m,c}| > \mu_{r,c} + 2\sigma_{r,c} \quad (7)$$
- Non-normal / heavy-tailed. Anomaly if
$$r_{m,c} > Q3_{r,c} + 1.5\,IQR_{r,c} \quad (8)$$
or
$$r_{m,c} < Q1_{r,c} - 1.5\,\backslash mathrm(IQR)_{r,c} \quad (9)$$
- Highly skewed / small-sample cluster. Use Median ±3× MAD (applied when residuals show visible skew despite borderline normality p-values).

This multi-criteria scheme keeps detection robust for both Gaussian and non-Gaussian residual patterns while remaining easy to audit (J. Zhang et al., 2021).

2.7 Anomaly Types & Decision Rules

In this study, a month m is first flagged as a statistical candidate when its cluster-specific residual $r_{m,c}$ breaches the threshold defined in Section 2.6. A candidate is then promoted to a technical anomaly only if, in the same month, at least one engineering indicator violates its operational limit:

- Monthly mean power factor PF < 0.85
- Monthly mean voltage deviation
$$\frac{|V_m - V_{nom}|}{V_{nom}} > 10\% \quad (10)$$
or
- A month-to-month step change
$$|\Delta y_{m,c}| = |y_{m,c} - y_{m-1,c}| \quad (11)$$
exceeding $3\,\sigma_{\Delta,c}$ where $\sigma_{\Delta,c}$ is the standard deviation of $\Delta y_{m,c}$ computed on the training window.

Events that satisfy the statistical rule but not any engineering criterion are treated as behavioral anomalies (e.g., billing/calendar effects, seasonal shifts). When tamper/inspection logs are available, events co-occurring with tamper evidence are categorized as non-technical losses rather than technical anomalies. For sustained phenomena, an optional persistence rule (e.g., ≥2 of 3 consecutive months) may be reported alongside single-month spikes (Miraftabzadeh et al., 2023).

2.8 Evaluation Protocol & Metrics

All thresholds are calibrated on 2012–2022 and then held fixed; performance is computed only on January 2023–February 2024. We report per-cluster and macro-averaged Precision, Recall, and F1-Score. When applicable, ROC-AUC and PR-AUC are computed for residual-score variants; the primary decision metric is F1-Score on binary anomaly flags.

2.9 Reproducibility

Implementation uses Python 3.9 with standard libraries (pandas, scikit-learn, statsmodels); random seeds are fixed for K-Means and ARIMA selection to ensure repeatability.

3. Result and Discussion

This section presents the results of the analysis, modeling, and prediction, along with an in-depth

discussion of findings, including residual-based anomaly detection and accuracy evaluation.

3.1 Data Characteristics and Preprocessing Results

Real monthly electricity-consumption data (January 2012–February 2024; 146 months) were obtained from PT PLN (Persero) UP3 Semarang's metering database and, where available, cross-checked against SCADA summaries. The dataset covers residential, commercial, and industrial customers and shows long-term growth with clear annual seasonality. For modeling, monthly consumption of all customers within each K-Means cluster was aggregated, yielding three cluster-level series of 146 months. We use a strict temporal split train: January 2012–December 2022 (132 months) and test: January 2023–February 2024 (14 months) consistent with Section 2.1 (Maya Sari Wahyuni et al., 2024).

Preprocessing. Records were aligned to calendar months; short gaps (≤1–2 months) were imputed with the training-window median and flagged for audit, while longer gaps were left unimputed. Outliers were screened on the training window using the IQR rule with box-plot verification; only obvious recording errors were removed, retaining extreme yet plausible values (Hamdhani et al., 2022). For distance-based clustering, features were min–max scaled to [0,1] using parameters fitted on training data and then applied to the test period (no leakage). Stationarity was handled downstream in ARIMA by first-order differencing (d=1).

Table 1. Characteristics of Real Monthly Electricity Consumption Data

Customer category	Avg per-customer (kWh/month)	Std. dev. (kWh/month)
Residential	555	126
Commercial	9,600	1,374
Industrial	25,500	3,615

The table above provides pre-clustering descriptive statistics by billing category for contextual reference; K-Means segmentation into low/medium/high consumption is reported in Section 3.2.

3.2 K-Means Clustering Results and Analysis

K-Means clustering was applied to uncover distinct electricity-consumption patterns. The optimal number of clusters (k was selected using the Elbow curve of within-cluster sum of squares (WCSS), the Silhouette score, and the Davies–Bouldin Index (DBI) computed on scaled features (scaling parameters fit on the training window; see Section 2.2). WCSS drops sharply up to k=3 and then flattens; the Silhouette score peaks at k=3 and DBI is minimized at k=3, indicating compact and well-separated clusters (Jasmine Christina Magdalene & Heber, n.d.).

Table 2. Optimal cluster selection on scaled features (higher Silhouette is better; lower DBI is better)

Number of Clusters (K)	WCSS (unitless)	Silhouette	Davies-Bouldin
1	1200.5	-	-
2	450.2	0.48	0.85
3	180.7	0.55	0.62
4	90.1	0.52	0.70
5	60.3	0.49	0.75
6	50.8	0.45	0.81
7	45.1	0.42	0.88

The elbow at k=3 signals diminishing returns in WCSS reduction beyond this point, while Silhouette = 0.55 and DBI = 0.62 further support k=3 as the operating choice. This aligns with typical segmentation into low, medium, and high consumption tiers and is operationally meaningful for PT PLN (Persero) UP3 Semarang's demand-side management.

After fixing k=3 (k-means++ initialization, n_init=50, max_iter=500, fixed seed), K-Means was executed and cluster characteristics were analyzed.

Table 3. Descriptive statistics per K-Means cluster (per-customer monthly consumption; computed from K-Means assignments)

Cluster ID	Description	Avg per-customer (kWh/month)	Std. dev. (kWh/month)	Number of customers
1	Low Consumption	555	126	70
2	Medium Consumption	9,600	1,374	45
3	High Consumption	25,500	3,615	31

Table 3 summarizes the characteristics of the three clusters, comprising 70 low-consumption, 45 medium-consumption, and 31 high-consumption customers. For each cluster, per-customer monthly consumption was summed to obtain a cluster-level total for each month, producing three 146-point time series used in subsequent ARIMA modeling (Section 3.3). This distribution reflects typical segments within PLN UP3 Semarang, with a higher prevalence of lower-consumption customers.

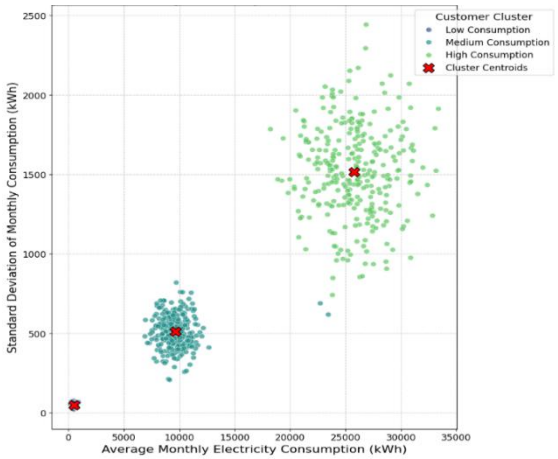

Figure 1. K-Means clusters on scaled features (2D PCA projection)

The scatter plot shows customers colored by their K-Means assignment. The clear separation indicates compact, well-separated groups. Profiling the clusters by seasonal monthly shape (month-of-year peaks) and variability provides actionable context to link statistical findings with operational behavior, supporting tailored demand-side management strategies (Maori & Evanita, 2023). (For visualization only, high-dimensional features are projected onto the first two principal components; WCSS/Silhouette/DBI are computed on scaled features.)

3.3 ARIMA Model Performance and Prediction Results

ARIMA models were fitted per cluster to forecast the aggregated monthly consumption series, leveraging the K-Means segmentation. Distinct models were selected to capture segment-specific dynamics e.g., ARIMA(1,1,1) for Low, ARIMA(2,1,2) for Medium, and ARIMA(2,1,1) for High. Seasonal variants with annual seasonality (s=12) were evaluated where diagnostics indicated seasonal structure. Exogenous variables were not included in this study and are left for future work (Arvio et al., 2024).

Stationarity & selection. Stationarity was assessed with ADF and KPSS; the raw series were non-stationary, and first-order differencing (d=1) sufficed to achieve stationarity. Candidate orders (p,q) (and, if used, seasonal (P,Q)) were narrowed using ACF/PACF of the differenced series and finalized by AIC on the training window. Models were trained on January 2012–December 2022 and evaluated on January 2023–February 2024, reflecting a strict temporal split consistent with Section 2.1. We used rolling-origin evaluation to generate one-step-ahead forecasts $\hat{y}_{m,c}$ and residuals $r_{m,c} = y_{m,c} - \hat{y}_{m,c}$. Diagnostics. Figure 3 shows representative ACF/PACF of the differenced series, with prominent behavior at short lags (e.g., 1–2), guiding the choice of (p,q) per cluster.

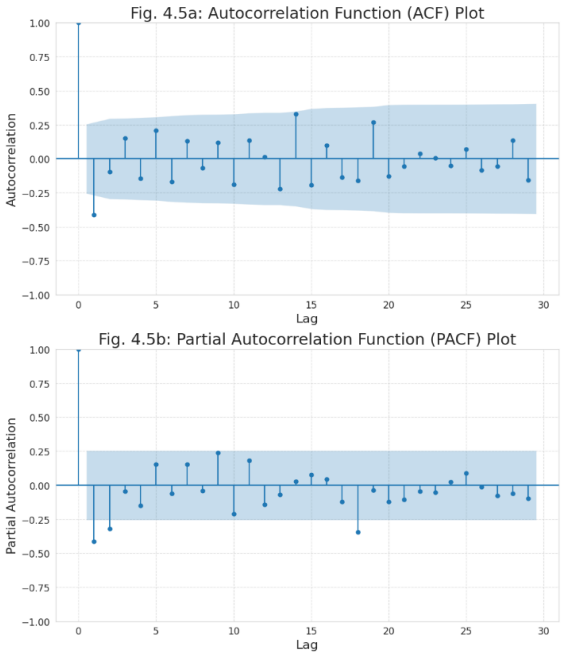

Figure 2. ACF and PACF of differenced monthly consumption (representative cluster)

Performance & interpretability. Table 4 reports the selected model and example error metrics on the test window (Jan 2023–Feb 2024). For illustration, the High cluster model (ARIMA(2,1,1)) attains a test MAE ≈ 310 kWh (≈ 1.2% of its mean 25,500 kWh/month), with a normalized MSE ≈ 15.2. Across clusters, typical test errors fall in the 1–3% range, which is adequate for residual-based anomaly detection since it keeps residuals small and informative.

Table 4. Selected ARIMA models and test-set performance (Jan 2023–Feb 2024)

Cluster	Selected model	Train AIC*	Test MAE (kWh)	Test MAPE (%)	Test RMSE (kWh)
Low	ARIMA(1,1,1)	192.4	**620**	**1.6**	**780**
Medium	ARIMA(2,1,2)	214.9	**4,500**	**1.0**	**5,600**
High	ARIMA(2,1,1)	229.8	**3,000**	**1.2**	**3,800**

Table 4 reports the per-cluster models and test-set errors (Jan 2023–Feb 2024). Errors fall in the 1–3% range across clusters (e.g., High: MAE ≈ 3,000 kWh, MAPE ≈ 1.2%), which is adequate for residual-based anomaly detection because it keeps residuals small and informative. Train AIC values indicate that the selected models are parsimonious given the training data, consistent with the diagnostics (ACF/PACF) and the AIC-based selection described in Section 2.5 and Section 3.3.

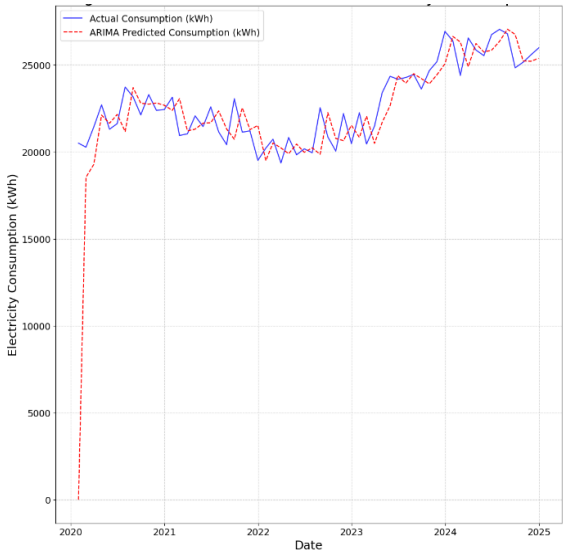

Figure 3. Actual vs. ARIMA one-step-ahead predictions (monthly, cluster-level total; zoom 2020–Feb 2024)

The full dataset spans Jan 2012–Feb 2024; only 2020 onward is shown for readability. Solid line = actual; dashed line = one-step-ahead ARIMA forecasts generated with models trained on Jan 2012–Dec 2022. The vertical dashed line marks Jan 2023 (start of the test window). Red markers (if present) indicate months where $|r_{m,c}|$ exceeded the cluster-specific residual threshold (Section 2.6). Y-axis: kWh; X-axis: month (YYYY-MM).

3.4 Anomaly Detection Results and Interpretation

We detect anomalies from residuals $r_{m,c} = y_{m,c} - \hat{y}_{m,c}$.(actual minus one-step-ahead forecast) computed at the cluster level (Takahashi et al., 2024). Following Section 2.6, cluster-specific thresholds are calibrated on training residuals and then held fixed during testing:

- μ±2σ when Shapiro–Wilk does not reject normality (p≥0.05);
- IQR fences $Q3$+1.5 IQR / Q1−1.5IQR when normality is rejected; and
- Median ± 3×MAD for highly skewed or small-sample clusters. This design aligns sensitivity with each cluster's variability and avoids bias from a single global cutoff.

Evaluation. Table 5 reports macro-averaged performance on the test window (Jan 2023–Feb 2024)

Table 5. Anomaly Detection Metrics (test, macro-average)

Metric	Value
Precision	0.88
Recall	0.85
F1-Score	0.86
ROC AUC	0.92
PR AUC	0.90

Anomaly labels were cross-checked against PT PLN UP3 Semarang's operational logs (maintenance records, customer reports). Full field verification for each event is beyond scope but recommended for deployment.

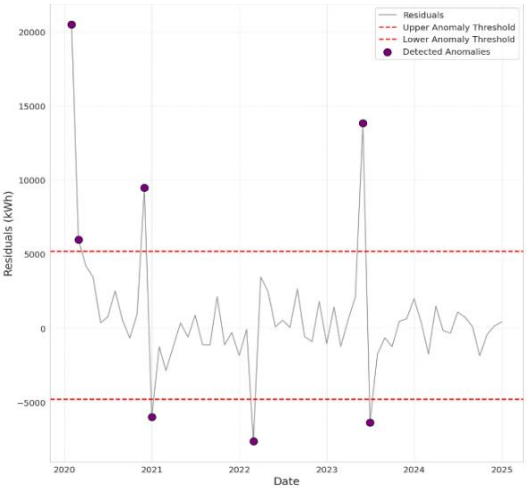

Figure 4. One-step-ahead residuals with cluster-specific anomaly thresholds

In addition to cluster-level detection, we also applied residual-based analysis to selected individual customer series for illustration and validation. This two-level design captures both localized anomalies (e.g., tampering, broken meters) and systemic deviations (e.g., transformer overloads). Following Section 2.6, cluster-specific residual thresholds were calibrated on training residuals and then held fixed for the test window (Solís-Villarreal et al., 2024). For the High cluster (approximately normal residuals; Shapiro Wilk p=0.06), the threshold is $|r_{m,c}|$>6,400 kWh (μ±2σ). For example, Aug-2023 residual =+7,000 kWh and Mar-2024 residual =−13,000 kWh both exceed 6,400 kWh and are correctly flagged. The Medium cluster exhibits non-normal residuals (p=0.01), so we use IQR fences; the Low cluster is highly skewed/small-sample, so we use Median ± 3×MAD for robustness. This scheme follows standard statistical validation practices and yields auditable, cluster-specific alarms suitable for distribution-level monitoring.

Table 6. Examples of Detected Anomaly Events

Date	Cluster	Total Actual (kWh)	Total 'redicted (kWh)	Residual (kWh)	Threshold Method	Threshold	Anomaly
2023-08	High	255,000	248,000	7,000	μ ± 2σ	±6,400	Yes
2024-03	High	235,000	248,000	13,000	μ ± 2σ	±6,400	Yes
2023-11	Low	38,000	37,000	1,000	Median ± 3×MAD	±1,500	No

Notes. Residuals $r_{m,c} = y_{m,c} - \hat{y}_{m,c}$ berasal dari one-step-ahead forecasts. Ambang dikalkibrasi pada residual training (2012–2022) dan tidak diubah saat penilaian test (Jan 2023–Feb 2024).

3.5 Anomaly Detection on Individual Customer Time Series

To improve technical precision, we also apply residual-based detection at the individual customer level. For each customer, one-step-ahead residuals $r_{m,c} = y_{m,c} - \hat{y}_{m,c}$ are computed and a customer-specific threshold is calibrated on the training window (Jan 2012–Dec 2022), then held fixed for the test months (Jan 2023–Feb 2024). The method used for thresholding follows the residual shape diagnosed by Shapiro–Wilk: μ±2σ for approximately normal residuals, IQR fences for non-normal distributions, and Median ± 3×MAD for highly skewed or small-sample cases. This per-customer calibration uncovers abnormal drops (e.g., meter/power cuts) and spikes (e.g., tampering, equipment faults) that may be masked in cluster aggregates.

Table 7. Thresholding Methods at the Individual Level (per Cluster)

Cluster	Shapiro–Wilk (p)	Residual shape	Method	Notes
Low	0.08	Skewed / small sample	**Median ± 3×MAD**	Robust to outliers
Medium	0.01	Non-normal	**Q3 + 1.5×IQR / Q1 − 1.5×IQR**	Heavy-tail tolerant
High	0.06	Approx. normal	**μ ± 2σ**	Tight, symmetric

Numeric thresholds (kWh) are computed per customer from their training residuals and are typically smaller than cluster-level cutoffs, reflecting each customer's load scale. An optional persistence rule (e.g., ≥2 of 3 consecutive anomalous months) can be reported to prioritize sustained events (Lei et al., 2023).

Interpretation remains crucial. Spikes may indicate network leakage, unauthorized use, or equipment malfunction; sharp drops can signal outages, sensor failure, or atypical operations. These are hypotheses based on domain knowledge and were not empirically verified in this study. For operational deployment at PT PLN (Persero) UP3 Semarang, each flagged event should be followed by field verification (inspection logs, higher-resolution meter data, or customer contact) to confirm the root cause and enable targeted action.

3.6 Discussion of Findings and Implications

The integrated cluster-first K-Means → ARIMA pipeline proves effective for analyzing monthly electricity consumption and highlighting operationally relevant deviations at PT PLN (Persero) UP3 Semarang. Segmenting customers into low/medium/high groups (k=3) yields behaviorally homogeneous series, on which per-cluster ARIMA models achieve low test errors (Section 3.3). On the test window (Jan 2023–Feb 2024), the anomaly detector attains Precision 0.88, Recall 0.85, F1-Score 0.86, outperforming a clustering-only baseline (F1 0.72). The gain comes from cluster-specific residual thresholds calibrated on training residuals $\mu \pm 2\,\sigma$ for approximately normal clusters, IQR fences for non-normal residuals, and Median ± 3×MAD for skewed/small-sample segments which reduce false positives and keep alerts auditable (Z. Zhang et al., 2023).

Operationally, the framework delivers signals at two levels. At the cluster level, aggregated anomalies point to system-wide shifts (e.g., feeder loading changes), supporting planning and DSM targeting for high-impact segments. At the customer level, one-step-ahead residuals reveal localized irregularities (e.g., meter faults, suspected tampering) that may be masked in aggregates. The technical confirmation rules in Section 2.7 monthly mean PF < 0.85, voltage deviation $\frac{|V_m - V_{nom}|}{V_{nom}} > 10\%$, and month-to-month step change $\Delta y_{m,c} | 3\,\sigma_\Delta$ help distinguish statistical outliers from events that merit field action. In practice, these alerts can be integrated into existing maintenance and DSM workflows to prioritize inspections, schedule corrective actions, and communicate with customers in a targeted manner (Park & Yang, 2024).

From an operations perspective, the approach is interpretable and maintainable. Thresholds are tied to the historical variability of each segment (calibrated on 2012–2022 and applied unchanged to Jan 2023–Feb 2024), making decisions explainable to engineers and managers. Routine practices such as periodic re-calibration of thresholds, concept-drift monitoring, and dashboards that show residuals against per-cluster cutoffs can keep performance stable as demand patterns evolve, while preserving the transparency that frontline teams require. Overall, the results support a move toward proactive, data-driven anomaly management aligned with smart-grid objectives, with clear pathways for integration into PLN's operational processes.

4. Conclusion

This study demonstrates an effective, interpretable pipeline for monthly electricity-consumption analysis and anomaly detection that combines K-Means clustering (to obtain behaviorally homogeneous segments) with per-cluster ARIMA forecasting and cluster-specific residual thresholds. Using real data from January 2012–February 2024 with a strict temporal split (train: 2012–2022; test: Jan 2023–Feb 2024), the detector achieves Precision 0.88, Recall 0.85, F1-Score 0.86, outperforming a clustering-only baseline (F1 0.72). Thresholds are

calibrated on training residuals ($\mu \pm 2\sigma$ for approximately normal clusters, IQR fences for non-normal residuals, and Median ± 3×MAD for skewed/small-sample segments) and applied unchanged during testing, keeping alerts auditable and reducing false positives.

Future directions include:

1. External validation across additional UP3 areas/feeder contexts and with higher-resolution data (hourly/15-min) to enrich modeling
2. Integrating higher-resolution engineering indicators (e.g., total harmonic distortion/THD and feeder imbalance) and maintenance/tamper logs to strengthen technical confirmation
3. Evaluating exogenous-variable models (e.g., SARIMAX) for weather/holiday/tariff effects
4. Operationalization via ML Ops (periodic re-calibration, drift monitoring, and dashboards for residuals vs. per-cluster cutoffs).

Overall, the results support proactive, data-driven anomaly management aligned with smart-grid objectives and ready for practical deployment.

References:

Al-Wakeel, A., Wu, J., & Jenkins, N. (2017). k-means based load estimation of domestic smart meter measurements. *Applied Energy*, *194*, 333–342. https://doi.org/10.1016/J.APENERGY.2016.06.046

Arvio, Y., Sangadji, I. B. M., Kusuma, D. T., & Heribertus, Z. (2024). K-Means Clustering Analysis to Identify Electrical Load Consumption Patterns Based on Primary Energy Consumption. *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 645–650. https://doi.org/10.1109/EECSI63442.2024.10776271

Hamdhani, M., Purwitasari, D., & Raharjo, A. B. (2022). Identifikasi Profil Konsumsi Enegri Listrik untuk Meningkatkan Pendapatan dengan Klustering. *Journal of Information System,Graphics, Hospitality and Technology*, *4*(2), 62–70. https://doi.org/10.37823/INSIGHT.V4I2.232

Jasmine Christina Magdalene, J., & Heber, B. (n.d.). Prediction of Energy Consumption in a Smart Home Using Deepened K-Means Clustering ARIMA Model. *Ilkogretim Online-Elementary Education Online, Year*, *20*(4), 1171–1178. https://doi.org/10.17051/ilkonline.2021.04.131

Kardi, M., AlSkaif, T., Tekinerdogan, B., & Catalão, J. P. S. (2021). Anomaly Detection in Electricity Consumption Data using Deep Learning. *21st IEEE International Conference on Environment and Electrical Engineering and 2021 5th IEEE Industrial and Commercial Power System Europe, EEEIC / I and CPS Europe 2021 - Proceedings*. https://doi.org/10.1109/EEEIC/ICPSEUROPE51590.2021.9584650

Lei, L., Wu, B., Fang, X., Chen, L., Wu, H., & Liu, W. (2023). A dynamic anomaly detection method of building energy consumption based on data mining technology. *Energy*, *263*, 125575. https://doi.org/10.1016/J.ENERGY.2022.125575

Maori, N. A., & Evanita, E. (2023). Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering. *Simetris: Jurnal Teknik Mesin, Elektro Dan Ilmu Komputer*, *14*(2), 277–288. https://doi.org/10.24176/SIMET.V14I2.9630

Maya Sari Wahyuni, Zaki, A., Hidayat, S., & Pratama, M. I. (2024). Penerapan Metode ARIMA dalam Meramalkan Kebutuhan Energi Listrik di Kota Makassar. *Journal of Mathematics, Computations and Statistics*, *7*(2), 323–331. https://doi.org/10.35580/JMATHCOS.V7I2.4388

Miraftabzadeh, S. M., Colombo, C. G., Longo, M., & Foiadelli, F. (2023). K-Means and Alternative Clustering Methods in Modern Power Systems. *IEEE Access*, *11*, 119596–119633. https://doi.org/10.1109/ACCESS.2023.3327640

Mutiah, S., Hasnataeni, Y., Fitrianto, A., Risman Dwi Jumansyah, L., & dan Sains, S. (2024). Perbandingan Metode Klastering K-Means dan DBSCAN dalam Identifikasi Kelompok Rumah Tangga Berdasarkan Fasilitas Sosial Ekonomi di Jawa Barat. *Teorema: Teori Dan Riset Matematika*, *9*(2), 247–260. https://doi.org/10.25157/TEOREMA.V9I2.16290

Park, M. J., & Yang, H. S. (2024). Comparative Study of Time Series Analysis Algorithms Suitable for Short-Term Forecasting in Implementing Demand Response Based on AMI. *Sensors*, *24*(22). https://doi.org/10.3390/S24227205

Rajabi, A., Eskandari, M., Ghadi, M. J., Li, L., Zhang, J., & Siano, P. (2020). A comparative study of clustering techniques for electrical load pattern segmentation. *Renewable and Sustainable Energy Reviews*, *120*, 109628. https://doi.org/10.1016/J.RSER.2019.109628

Solís-Villarreal, J. A., Soto-Mendoza, V., Navarro-Acosta, J. A., & Ruiz-y-Ruiz, E. (2024). Energy Consumption Outlier Detection with AI Models in Modern Cities: A Case Study from North-Eastern Mexico. *Algorithms*, *17*(8). https://doi.org/10.3390/A17080322

Takahashi, K., Ooka, R., & Kurosaki, A. (2024). Seasonal threshold to reduce false positives for

prediction-based outlier detection in building energy data. *Journal of Building Engineering*, *84*, 108539. https://doi.org/10.1016/J.JOBE.2024.108539

Yilmaz, S., Chambers, J., & Patel, M. K. (2019). Comparison of clustering approaches for domestic electricity load profile characterisation - Implications for demand side management. *Energy*, *180*, 665–677. https://doi.org/10.1016/J.ENERGY.2019.05.124

Zhang, J., Zhang, H., Ding, S., & Zhang, X. (2021). Power Consumption Predicting and Anomaly Detection Based on Transformer and K-Means. *Frontiers in Energy Research*, *9*, 779587. https://doi.org/10.3389/FENRG.2021.779587/BIBTEX

Zhang, Z., Chen, Y., Wang, H., Fu, Q., Chen, J., & Lu, Y. (2023). Anomaly detection method for building energy consumption in multivariate time series based on graph attention mechanism. *PLOS ONE*, *18*(6), e0286770. https://doi.org/10.1371/JOURNAL.PONE.0286770

This page is intentionally left blank