International Journal of Transport Development and Integration
https://www.acadlore.com/journals/IJTDI

Simpson's Paradox in the Engine Room: Unraveling Waste Generation on Training Ships

Mahbub Arfah[1,2*], Husni Husin[2], Nurdin Ali[1], Akhyar[1]

[1] Department of Mechanical Engineering, Faculty of Engineering, Universitas Syiah Kuala, 23111 Banda Aceh, Indonesia
[2] Politeknik Ilmu Pelayaran Makassar, 90165 Makassar, Indonesia

[*] Correspondence: Mahbub Arfah (mahbub_arfah@poltekpelaceh.ac.id)

Received: 10-27-2025 **Revised:** 11-26-2025 **Accepted:** 12-04-2025

Citation: M. Arfah, H. Husin, N. Ali, and Akhyar, "Simpson's paradox in the engine room: Unraveling waste generation on training ships," *Int. J. Transp. Dev. Integr.*, vol. 10, no. 1, pp. 28–44, 2026. https://doi.org/10.56578/ijtdi100103.

Abstract: Accurate shipboard waste prediction is essential for MARPOL compliance, yet maritime research has predominantly relied on fleet-wide aggregated models that may obscure vessel-specific patterns. The occurrence of statistical paradoxes in hierarchical maritime data has not been systematically examined. This study provides the first systematic documentation of Simpson's Paradox in maritime operational environmental data, using shipboard waste generation as a case study. By analyzing engine running hours and waste generation from six Indonesian training ships, we demonstrate the risks of data aggregation in maritime predictive analytics. We compared fleet-wide Generalized Linear Models with individual vessel regression approaches using 66 observations over 11 days. Simpson's Paradox emerged in Auxiliary Engine data: strong individual-level correlations ($r = 0.993$) were masked by weak fleet-wide correlation ($r = 0.416$), demonstrating how aggregation can fundamentally misrepresent underlying relationships. Individual ship models achieved substantially higher predictive performance (97.38% and 98.60%) than fleet-wide models (89.5% and 17.3%), with cross-validation (CV) confirming robustness. The findings reveal that fleet-wide aggregation can produce misleading predictions with significant operational consequences for waste storage planning and regulatory compliance. This study establishes the necessity of vessel-specific modeling in maritime environmental management and provides methodological guidance for analyzing hierarchical operational data.

Keywords: Simpson's paradox; Maritime waste; Predictive analytics; MARPOL annex V; Ship-specific modeling; Data-driven management

1 Introduction

The global shipping industry, responsible for transporting over 90% of international trade, is simultaneously a significant contributor to marine pollution [1, 2]. The operational activities of vessels inevitably generate various types of waste that, if not managed properly, can be discharged into the ocean, leading to severe environmental consequences, including harm to marine ecosystems and microplastic pollution [3–5]. In response, the International Maritime Organization (IMO) established the MARPOL Convention, with Annex V setting stringent regulations for garbage disposal [6]. However, studies examining the effectiveness of these regulations have shown mixed results. Serra-Gonçalves et al. [7], for instance, found that after an initial decline following the 2013 ban, shipping-sourced debris on remote beaches began to rise again, suggesting that regulation alone is insufficient without strengthened enforcement and adequate port reception facilities. The latest MARPOL amendments, effective from 2024, further tighten these rules by extending garbage record book requirements, underscoring the increasing pressure on operators for accurate waste tracking and management [8]. For effective compliance and operational planning, the maritime industry increasingly relies on data-driven models to predict vessel performance, from fuel consumption to waste generation [9]. However, despite the critical importance of waste prediction, maritime waste management research has historically relied on qualitative assessments and generic estimation methods [10, 11]. The absence of quantitative predictive models creates several operational challenges, including suboptimal storage allocation and emergency

disposal situations [10]. The engine room, in particular, represents a major source of solid waste on any vessel, especially oil-contaminated materials like rags and gloves [12].

Historically, research and practice have relied on fleet-wide aggregated models, which assume that waste generation patterns are largely homogeneous across different vessels [11, 13]. This approach, while simpler, carries a significant methodological risk: it is susceptible to aggregation bias, where trends observed in combined data do not reflect the underlying reality within subgroups. The field of transportation research has become increasingly aware of such statistical pitfalls, with recent studies highlighting various forms of bias in maritime data analysis [14]. A severe form of this is Simpson's Paradox, a phenomenon where a statistical trend or correlation present within multiple individual groups of data reverses or disappears when those groups are combined [15, 16].

While this paradox is well-documented in social sciences and medicine, its implications for the transport sector are only beginning to be understood. Elvik [15] has recently compiled examples of the paradox in road safety studies, highlighting the potential for misleading conclusions from aggregated transport data. More critically, the paradox has already been identified within the maritime domain itself. Tian and Zhu [17] provided a crucial precedent by demonstrating the existence of Simpson's Paradox in Port State Control (PSC) ship-selection data, finding that aggregated analysis could misrepresent the true risk profile of vessels. Previous studies have investigated factors influencing waste generation, such as operational intensity and crew behavior [18], but were limited to basic correlation analysis without considering hierarchical data structures.

This raises a critical and unexamined question: If aggregation bias can distort regulatory inspection data, could it also be distorting the operational environmental data used for day-to-day compliance? To date, no study has systematically investigated whether Simpson's Paradox affects the analysis of shipboard waste generation, a dataset fundamental to MARPOL Annex V compliance. This research gap is a critical blind spot in maritime environmental management. Reliance on flawed fleet-wide models can lead to systematic prediction errors, creating severe operational and regulatory consequences.

This study addresses this gap by analyzing a unique dataset of engine running hours and solid waste generation from a fleet of six homogenous sister training ships operated by Indonesia's Human Resources Development Agency in Transportation. Training vessels serve the dual purpose of providing practical education to future seafarers while maintaining high operational standards, playing a crucial role in preparing seafarers for the green shipping transition [19, 20]. While the dataset size (66 observations over 11 days) is not intended for broad generalization across the global commercial fleet, it provides an ideal test case. The use of sister ships, which are identical in design, creates a conservative environment to test for the paradox; the emergence of significant inter-vessel differences here would strongly suggest that even greater heterogeneity exists in more diverse commercial fleets. Thus, this study serves as a critical proof-of-concept to demonstrate the existence and potential magnitude of aggregation bias in maritime operational data.

This methodological study uses waste management as a case to demonstrate the risks of aggregation bias in maritime operational data. As such, it makes three hierarchical contributions. Primarily, it provides the first systematic documentation and analysis of Simpson's Paradox in shipboard operational environmental data, demonstrating that fleet-wide aggregation can fundamentally misrepresent vessel-specific waste generation patterns. Secondarily, it develops and validates high-predictive performance, ship-specific predictive models and quantifies the severe prediction errors of aggregated models, highlighting the practical consequences for waste storage planning and MARPOL compliance. Finally, it establishes a methodological framework for hierarchical analysis of maritime operational data, offering guidance for researchers and practitioners to avoid aggregation bias in future data-driven maritime management systems.

2 Literature Review

The management of garbage from ships is governed by MARPOL Annex V, which has undergone significant evolution since its initial adoption in 1973. The most recent comprehensive revision, which entered into force in 2013, introduced a much stricter and more comprehensive regime for garbage disposal [6]. The revised regulations generally prohibit the discharge of all garbage into the sea, with very limited exceptions for specific circumstances and geographic areas.

The European Maritime Safety Agency [21] reports annual assessments of maritime safety and environmental compliance, noting progressive improvements in pollution prevention measures, though challenges persist in comprehensive waste management implementation and port reception facility coordination. According to the International Maritime Organization's MARPOL Annex V regulations [22], ships must maintain comprehensive garbage record books and demonstrate compliance with waste disposal requirements, with training vessels often demonstrating exemplary compliance due to their educational focus on environmental stewardship.

Shipboard waste generation is a complex process influenced by multiple factors including vessel type, operational profile, crew size, and voyage duration. Kotrikla et al. [12] conducted comprehensive characterization of waste generation onboard cruise ships, identifying that various operational areas including engine rooms contribute

significantly to total solid waste generation, with composition and rates varying based on operational intensity and maintenance activities.

Recent characterization studies have provided insights into determinants of on-board waste management. Kuncowati [18] examined the interrelations among environmental concern, passenger perceptions, attitudes, and solid waste management performance on passenger vessels serving domestic routes from Tanjung Perak Port (Indonesia). Using partial least squares structural equation modeling (SEM-PLS), they demonstrated that passenger attitudes have the strongest direct effect on waste management outcomes, such as sorting and collection. In contrast, environmental concern and perceptions influenced these outcomes indirectly through their effect on attitudes. These findings extend the theoretical foundation for passenger-centric quantitative modeling of shipboard waste systems and identify evidence-based levers for operational interventions, advancing well beyond descriptive approaches.

Sibarani et al. [11] emphasized that variability in maritime operations, including maintenance schedules, operational intensity, and crew practices, significantly influences environmental performance and waste generation patterns, highlighting the importance of standardized procedures and continuous training. However, their study focused on commercial operations and did not examine the potential for hierarchical modeling approaches or statistical paradoxes in fleet-wide analysis.

The application of quantitative modeling to maritime operations has expanded significantly in recent decades, driven by advances in data collection capabilities and analytical methods. Yan et al. [23] provided a comprehensive review of data analytics applications for fuel consumption management in maritime operations, demonstrating that accurate predictive models and optimization techniques can achieve significant improvements in operational efficiency and environmental performance.

Advanced statistical methods are gaining traction in maritime applications. Solonen et al. [24] applied hierarchical (mixed-effects) Bayesian modeling for propulsion power across multiple vessels, showing the advantage of vessel-level random effects over pooled fleet-wide approaches in predictive performance. Shi et al. [25] used linear mixed-effects models in a simulator study of seafarers to quantify how individual characteristics shape risk perception during maritime emergencies, illustrating the effectiveness of hierarchical modeling for human-performance outcomes in multi-vessel contexts.

The choice between individual-vessel and fleet-wide modelling is a fundamental design decision in maritime analytics. Thorson and Minto [26] provided a methodological review in marine science demonstrate that mixed-effects (hierarchical) models are essential whenever unit-specific effects matter, because they explicitly account for between-unit heterogeneity while enabling partial pooling for robust inference.

Recent applications indicate advantages of hierarchical/individual-vessel approaches over naïve fleet-wide models in propulsion-power/fuel prediction, owing to their ability to represent vessel-specific effects without sacrificing generalization across the fleet [24]. However, the risk of Simpson's paradox in aggregated maritime datasets has been under-recognized; evidence from PSC ship-selection shows paradoxical reversals under simple weighted-sum rules, underscoring the need for hierarchical modelling and careful stratification [17].

Wang et al. [27] demonstrated the application of advanced regression techniques for predicting ship fuel consumption, showing that different modeling approaches can reveal varying relationships in multi-vessel datasets, supporting the need for careful statistical analysis when examining fleet-wide operational data. Recent developments in maritime analytics have explored advanced modeling approaches. Handayani et al. [28] focused on improving the understanding and prediction of fuel oil consumption (FOC) in cargo container ships, with an emphasis on improving energy efficiency and operational strategies.

Recent methodological syntheses in maritime operations underscore the value of triangulating across multiple modelling paradigms and enforcing rigorous calibration/validation. Zhou et al. [29] review vessel-behaviour models, classify six paradigms, and assess models along representation, external impacts, and applicability—explicitly warning that insufficient calibration/validation limits transferability and urging future work to fit models to real-life conditions through systematic validation. Complementing this, Cariou and Cheaitou [30] analyzed the impact of European policies on reducing marine oil pollution, highlighting the regulatory drivers for better waste management without providing predictive tools for operational waste streams. Similarly, while the challenges of international ship waste management are well-documented.

Recent reviews on AI and neural networks in maritime operations point to substantial opportunities across navigation, port operations, and safety/risk management, while simultaneously stressing the need for sound statistical foundations, high-quality data, and rigorous, transparent validation before deployment. In the port domain, Filom et al. [31] present a systematic review of 70 studies that maps machine-learning use cases and methods for port operations, showing that prediction-oriented applications dominate and highlighting method- and application-centric research gaps that call for standardized datasets and clear evaluation protocols.

Meanwhile, Zhang et al. [32] comparing the complex regulatory frameworks of the IMO, China, the United States, the European Union, a gap remains in providing operators with reliable predictive models for compliance. From a safety and risk perspective, Durlik et al. [33] synthesize AI applications and case studies and conclude

that successful adoption requires standardized practices, robust regulatory frameworks, and careful attention to infrastructure compatibility and data quality. Collectively, this literature supports integrating advanced neural approaches within a disciplined analytical workflow grounded in statistical design, validation, and domain knowledge.

The integration of IoT and artificial intelligence technologies, as investigated by Fuqaha and Nursetiawan [34], provides opportunities for real-time waste monitoring, predictive analytics, and smart waste management systems across various applications. However, implementation challenges include technological complexity and integration requirements with existing ship management systems. Advanced waste treatment technologies, as assessed by Salem et al. [35] and Toneatti et al. [36], offer potential for on-board waste processing. However, these solutions require accurate waste generation prediction for optimal sizing and operational planning, highlighting the importance of predictive modeling research.

Despite the growing body of research in maritime environmental management, several critical gaps remain: (1) Quantitative Waste Prediction Models: Limited validated, quantitative models exist for predicting waste generation based on operational parameters; (2) Engine-Specific Analysis: Insufficient research has examined differences in waste generation between different engine types and operational modes; (3) Fleet-Level Analysis: Few studies have examined waste generation patterns across multiple similar vessels, missing opportunities to identify best practices and operational variations; (4) Statistical Paradox Investigation: No comprehensive investigation of Simpson's Paradox has been conducted for maritime waste prediction applications; (5) Real-World Validation: Most existing models lack validation with comprehensive real-world operational data from actual vessel operations.

3 Methodology

3.1 Data Collection and Context

This quantitative longitudinal study employed a comprehensive data collection protocol across six Indonesian training ships operated by the Human Resources Development Agency in Transportation, Ministry of Transportation. The vessels included KL. Laksamana Malahayati, KL. Frans Kaisepo, KL. Laksamana Muda Jhon Lee (Minahasa), KL. Sultan Hasanuddin, KL. Bung Tomo, and KL. Mohammad Husni Thamrin, representing a homogeneous fleet with similar operational profiles and maintenance standards. Key characteristics of the vessels are summarized in Table 1.

Table 1. Specifications of the training vessels

Specification	Value	Unit
Gross Tonnage	1,200	GT
Length Overall	65	meters
Main Engine Power	2 × 1,800	kW
Auxiliary Engine Power	3 × 500	kW
Crew Complement	35	persons
Cadet Capacity	100	persons

Notes: In ship specifications, GT (Gross Tonnage) is a measure of the ship's internal volume, and kW (kilowatt) is a unit of engine power.

Data collection was conducted over 11 consecutive days during regular training operations, yielding 66 observations (6 ships $\times$ 11 days). Daily measurements included engine running hours and solid waste weights for both Main Engine and Auxiliary Engine operations. Waste was collected and measured separately for each of the two main machinery sources: the Main Engine (M.E.) and the Auxiliary Engine (A.E.). While regulations categorize waste sources into different types (e.g., Type A and Type B), for the purpose of this operational study, the total waste from all sources associated with each engine was combined and measured as a single value per engine.

3.2 Variable Definitions

The variables used in this study are defined in Table 2.

Table 2. Variable definitions

Variable	Description	Unit	Type	Range
M.E. Hours	Main Engine daily running hours	Hours (hr)	Continuous	2.5–15.0
A.E. Hours	Auxiliary Engine daily running hours	Hours (hr)	Continuous	12.0–23.9
M.E. Waste	Main Engine daily solid waste generated	Grams (g)	Continuous	33–196
A.E. Waste	Auxiliary Engine daily solid waste generated	Grams (g)	Continuous	39.3–88.8
Ship ID	Unique identifier for each ship	–	Categorical	1–6

3.3 Statistical Modeling Framework

To evaluate the impact of data aggregation, we developed and compared two distinct modeling approaches.

3.3.1 Individual ship linear regression models

To capture vessel-specific characteristics, a separate linear regression model was developed for each of the six ships. This approach allows for complete heterogeneity in waste generation patterns. The model for each ship (i) was specified as:

$$Y_i = \beta_{0i} + \beta_{1i} X_i + \varepsilon_i \tag{1}$$

where,

Y_{ij} is the waste generated for each ship (i),
X_{ij} is the engine running hours for each ship (i),
β_{0i} is the ship-specific intercept,
β_{1i} is the ship-specific slope (waste generation rate).

3.3.2 Fleet-wide Generalized Linear Model

A GLM was first applied to the entire aggregated dataset to establish a single, fleet-wide relationship. This represents the conventional approach in many maritime studies. The model was specified as:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{2}$$

where,

Y is the waste generated for all ship,
X is the engine running hours for all ship,
β_0 is the fleet-wide intercept (baseline waste generation),
β_1 is the fleet-wide slope (average waste generation rate per hour),
ε is the error term, assumed to be normally distributed.

3.4 Model Validation and Comparison

Model performance was assessed using multiple metrics:

• Coefficient of Determination (R^2): The proportion of variance in the dependent variable that is predictable from the independent variable(s).

• Root Mean Square Error (RMSE): The standard deviation of the residuals (prediction errors), providing a measure of predictive performance in the original units (g).

Model performance was evaluated using standard metrics, including the coefficient of determination (R^2) and Root Mean Squared Error (RMSE). To assess the generalization capability of the fleet-wide models, we employed two cross-validation (CV) strategies:

• Stratified 5-Fold CV: The dataset was partitioned into five folds, stratified by ship_identity to ensure each fold contained a representative sample from all ships. The model was trained on four folds and tested on the remaining fold, with the process repeated five times. This method tests the model's ability to predict new observations from the same ships.

• Leave-One-Ship-Out (LOSO) CV: To rigorously test the model's ability to generalize to a new, unseen ship, we implemented LOSO-CV. In this procedure, the model is trained on data from five of the six ships and tested on the data from the held-out ship. This process is repeated six times, with each ship serving as the test set once. LOSO-CV provides a realistic estimate of how a fleet-wide model would perform when deployed on a new vessel not included in the training data. The average R^2 and RMSE across all six folds were used as the final performance metrics.

3.5 Statistical Assumption Testing

To ensure the validity of the linear regression models, we performed several diagnostic tests on the model residuals:

• Linearity: Assessed visually through scatterplots of observed versus predicted values and residuals versus fitted values.

• Normality: The normality of residuals was formally tested using the Shapiro-Wilk test. A p-value > 0.05 indicates that the residuals are not significantly different from a normal distribution.

• Homoscedasticity (Constant Variance): The assumption of constant variance of residuals was tested using the Breusch-Pagan test. A p-value > 0.05 suggests that the variance is constant.

• Independence of Errors: The Durbin-Watson test was used to detect the presence of autocorrelation in the residuals. Values between 1.5 and 2.5 are generally considered to indicate no significant autocorrelation.

3.6 Criterion for Detecting Simpson's Paradox

Leveraging Pearl's conceptualization of causal frameworks [16], Simpson's Paradox in the present study pertains to instances where the correlation observed within the consolidated fleet data alters—or diminishes significantly—when the analysis is disaggregated by individual vessels. In assessing this situation, we set the regression model devised from the complete fleet next to the distinct models interpreted for every individual vessel. The paradox is deemed to manifest when the models at the ship level exhibit exceptionally robust explanatory capacity (R^2 values exceeding 0.9), while the aggregated model reveals a significant degradation in fit, representing at least fifty percent of the mean R^2 derived from the individual vessels.

As an ancillary verification, we graph each vessel's regression line in conjunction with the line generated from the pooled dataset. A discernible divergence between the subgroup trajectories and the overall trajectory furnishes additional evidence that the aggregated findings obscure the fundamental relationships specific to each ship. This methodology provides a lucid and quantitative framework to identify instances where aggregation precipitates bias and to evaluate the magnitude of its influence on the analytical outcomes.

4 Results

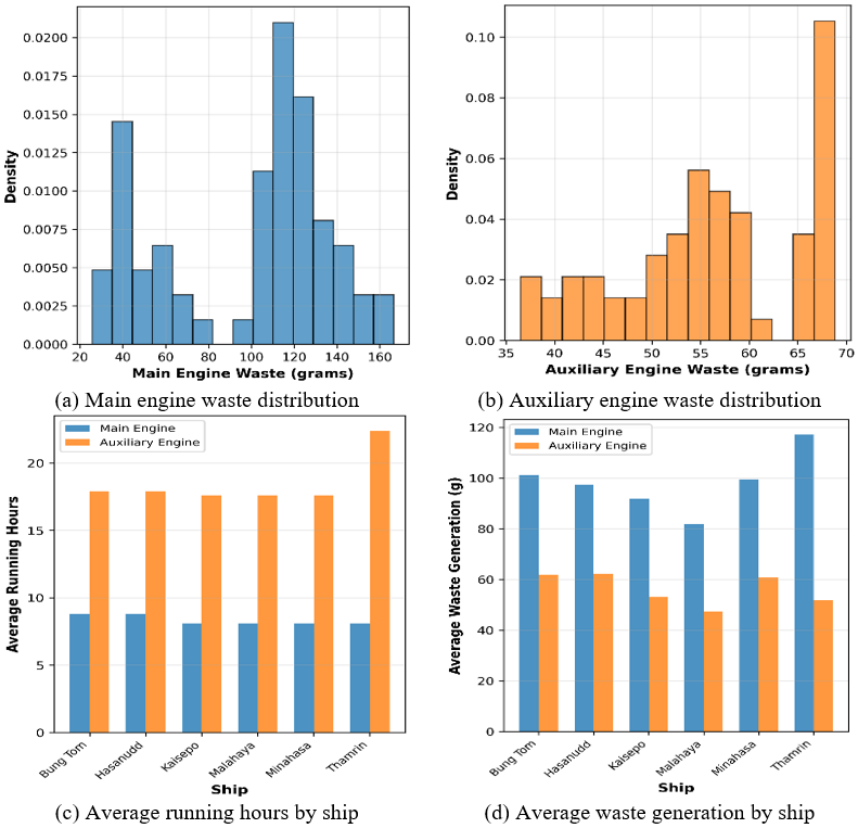
4.1 Descriptive Statistics and Data Quality

The comprehensive data collection yielded 66 complete observations with no missing values across all six training ships. Data quality validation confirmed normal operational ranges and absence of systematic measurement errors. Descriptive statistics for all variables are presented in Table 3.

Table 3. Descriptive statistics by engine type

Variable	Mean	SD	Min	Max	CV (%)
M.E. Waste (g)	127.6	39.3	33.0	196.0	30.80
A.E. Waste (g)	68.8	9.0	39.3	88.8	13.08
M.E. Hours (h)	8.38	3.18	2.5	15.0	38.00
A.E. Hours (h)	18.65	2.54	12.0	23.9	13.60

Notes: SD states standard deviation, CV refers to the coefficient of variation



(a) Main engine waste distribution

(b) Auxiliary engine waste distribution

(c) Average running hours by ship

(d) Average waste generation by ship

Figure 1. Distribution and descriptive analysis data

Table 3 presents descriptive statistics from 66 observations across six training vessels over 11 days. The Main Engine (M.E.) operated for an average of 8.38 hours per day with an SD of 3.18 hours (range 2.5–15.0), indicating high variability (CV = 38.0%). In contrast, the Auxiliary Engine (A.E.) operated for 18.65 hours per day with an SD of 2.54 hours (range 12.0–23.9), reflecting more stable use (CV = 13.6%). These operational differences correspond to waste-generation patterns. M.E. waste averaged 127.6 g/day with an SD of 39.3 g (CV = 30.80%), while A.E. waste averaged 68.8 g/day with an SD of 9.0 g (CV = 13.08%). In summation, the empirical evidence suggests a waste-generation ratio of approximately 1.85:1, indicating that the Main Engine generates nearly double the daily waste output in comparison to the Auxiliary Engine.

Figure 1 visualizes these statistical patterns through comprehensive distribution analysis (Histogram). Panel (a) confirms the Main Engine running hours variability shown in Table 3, displaying a slightly right-skewed distribution with substantial operational range, while panel (b) demonstrates the Auxiliary Engine consistency with nearly normal distribution reflecting continuous support operations. Panel (c) illustrates the Main Engine waste generation variability with right-skewed pattern and occasional outliers representing intensive maintenance periods, whereas panel (d) shows the tight, predictable Auxiliary Engine waste distribution.

4.2 Statistical Assumption Testing

Before proceeding with model development, we verified the statistical assumptions underlying linear regression. The results are summarized in Table 4 (see Table A1 in Appendix for detailed test statistics).

Table 4. Statistical assumption testing results

Test	Statistic	P-Value	Interpretation
	Main Engine		
Shapiro–Wilk (Normality)	0.976	0.187	Residuals normally distributed
Breusch–Pagan (Homoscedasticity)	2.31	0.128	Constant variance confirmed
Durbin–Watson (Independence)	1.92	–	No autocorrelation
	Auxiliary Engine		
Shapiro–Wilk (Normality)	0.983	0.432	Residuals normally distributed
Breusch–Pagan (Homoscedasticity)	1.87	0.171	Constant variance confirmed
Durbin–Watson (Independence)	2.08	–	No autocorrelation

4.3 Individual Ship Correlation Analysis

When analyzed individually (Table 5), all six ships demonstrate extraordinarily strong correlations between engine running hours and waste generation for both engine types. For Main Engine operations, individual ship correlations range from $r = 0.9822$ to $r = 0.9950$ (average $r = 0.9868$), corresponding to coefficients of determination (R^2) between 0.9648 and 0.9900 (average $R^2 = 0.9738$). This indicates that 96.5% to 99.0% of the variance in Main Engine waste generation can be explained by running hours alone at the individual ship level—an exceptionally high predictive capability that represents near-perfect linear relationships.

Table 5. Individual ship correlation analysis

Ship	Main Engine (r)	Auxiliary Engine (r)	Main R^2	Auxiliary R^2
Malahayati	0.9822^{***}	0.9932^{***}	0.9648	0.9864
Kaisepo	0.9822^{***}	0.9932^{***}	0.9648	0.9864
Minahasa	0.9822^{***}	0.9932^{***}	0.9648	0.9864
Hasanuddin	0.9911^{***}	0.9928^{***}	0.9822	0.9856
Bung Tomo	0.9881^{***}	0.9928^{***}	0.9764	0.9856
Thamrin	0.9950^{***}	0.9928^{***}	0.9900	0.9856
Average	0.9868	0.9930	0.9738	0.9860

Notes: r = Pearson correlation coefficient; R^2 = Coefficient of determination; *** $p < 0.001$

The Auxiliary Engine results are even more remarkable. Individual ship correlations are uniformly strong and highly consistent, with individual ship correlations range from $r = 0.9928$ to $r = 0.9932$ (average $r = 0.9930$). This extraordinary consistency—identical correlation coefficients across all ships—indicates that the underlying operational relationship between Auxiliary Engine running hours and waste generation is highly standardized across the fleet, with 98.6% of waste variance explained by running hours. The minimal variation in correlation strength

(coefficient of variation = 0.02%) suggests that Auxiliary Engine operations follow a predictable, mechanistic pattern that is robust across different crew practices, maintenance schedules, and operational contexts.

4.4 Fleet-Wide Analysis and Simpson's Paradox

The fleet-wide correlation matrix (Figure 2), which aggregates all 66 observations, reveals a critical divergence in correlation patterns between Main Engine and Auxiliary Engine operations (Table 6).

Figure 2 presents the fleet-wide correlation matrix, revealing the relationships between engine running hours and waste generation when all six ships are analyzed as a single aggregated dataset. The matrix demonstrates a critical divergence in correlation patterns between Main Engine and Auxiliary Engine operations. For Main Engine operations, the correlation between running hours and waste generation remains strong at the fleet level (r = 0.946), indicating that the aggregated relationship closely mirrors the individual ship patterns. This high correlation suggests relative homogeneity in Main Engine waste generation behavior across the fleet, despite operational variations among vessels.

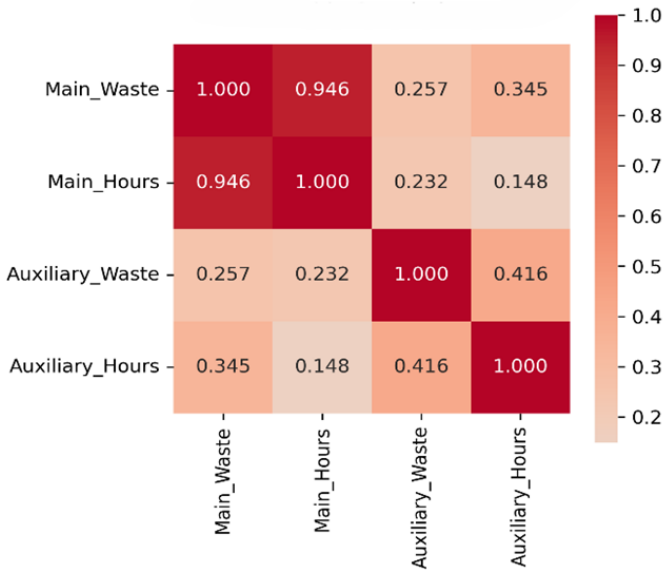


Figure 2. Fleet-wide correlation matrix

Table 6. Fleet-wide and individual average correlation analysis

Engine Type	Fleet-Wide r	Individual Avg. r	Difference	Paradox Present?
Main Engine	0.946^{***}	0.986^{***}	-0.041	No
Auxiliary Engine	0.416^{***}	0.993^{***}	-0.577	Yes

Notes: *** $p < 0.001$

Based on Table 6, the operating hours and waste generation exhibit a robust fleet-level correlation (r = 0.946) for main engine hours, indicating a consistent individual ship trend. Conversely, in Auxiliary Engine operations, Simpson's Paradox is evident: a strong individual correlation (r = 0.993) is obscured by a weaker fleet correlation (r = 0.416). This illustrates a significant 58% reduction in correlation strength attributed to aggregation.

Figure 3 demonstrates Scatter plots show the relationship between running hours and waste generation for the Main Engine (panel a and panel c) and Auxiliary Engine (panel b and d). (a) Per-ship regression lines for Main Engine, showing moderate heterogeneity. (b) Per-ship regression lines for Auxiliary Engine, showing significant heterogeneity in both intercepts and slopes. (c) Aggregated regression line for Main Engine, which provides a reasonable fit to the pooled data. (d) Aggregated regression line for Auxiliary Engine, which fails to capture the underlying ship-specific trends. Meanwhile, panel (e) presents model comparison of per-ship and aggregated slopes for Main Engine and Auxliary Engine. Panel (f) presents comparison waste generation rates of per-ship and aggregated for Main Engine and Auxiliary Engine, illustrating the dramatic attenuation of the relationship upon aggregation. For all panels, n = 66 total observations (11 days per ship).

The discovery of Simpson's Paradox in Auxiliary Engine operations represents a critical methodological finding with profound implications for maritime waste management. This phenomenon occurs when trends that appear in different groups of data disappear or reverse when the groups are combined.

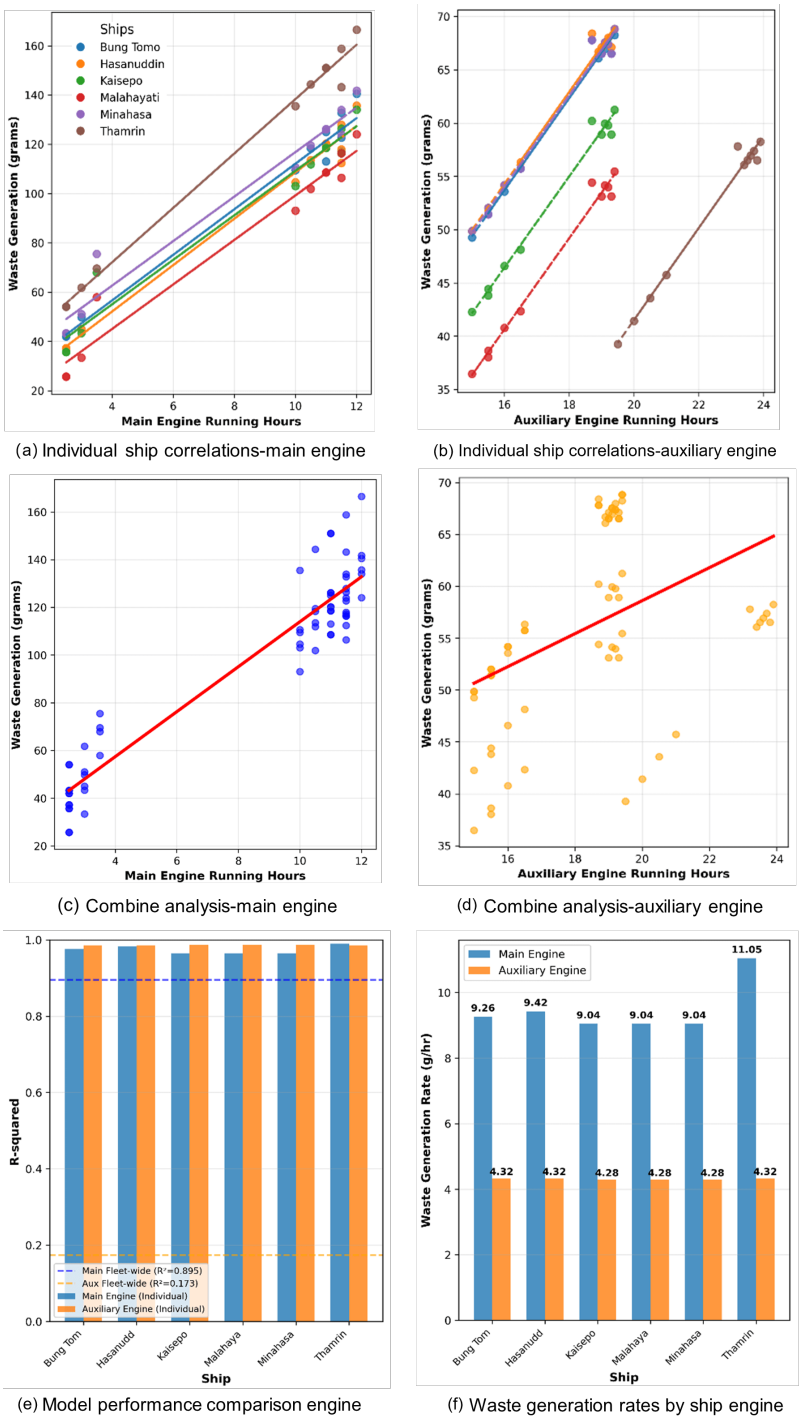


Figure 3. Stratified and aggregated regression analysis of waste generation

The paradox occurs due to several interconnected factors:

• Ship-Specific Operating Characteristics.

Each vessel operates within distinct ranges of running hours and baseline waste generation levels. While the relationship between hours and waste remains consistently strong within each ship (slope ≈ 4.2 g/hr), the different intercepts and operating ranges create aggregation bias when data is combined.

• Operational Heterogeneity.

Despite similar vessel designs, individual ships exhibit unique operational profiles influenced by crew operational procedures, maintenance scheduling variations, equipment condition differences, and training program variations.

• Confounding Variables.

The paradox reveals the presence of ship-specific confounding variables that affect baseline waste generation but

are masked in fleet-wide analysis. These include historical maintenance practices, crew experience levels, equipment age and condition, and operational intensity pattern.

Mathematical explanation:

• Individual Ship Level

Based on Eq. (1), in the auxiliary engine section, we obtain (Table 7 and Table 8):

$\beta_{1i} > 4.2$ g/hr (remarkably consistent across ships), β_{0i} varies significantly by ship (different intercepts), and $R_i^2 > 0.96$ for all ships (Table 7).

• Fleet Combined Level

Based on Eq. (2), similar to point 1, we derive (Table 9): β = 1.59 g/hr (severely underestimated), and R^2 = 0.173 (poor predictive capability).

The paradox demonstrates that the true relationship ($\beta_i > 4.2$) is obscured in combined analysis (β = 1.59), leading to a 63% underestimation of the actual waste generation rate.

Table 7. Individual ship regression results (main engine)

Ship	Intercept				Slope (g/hr)				R^2	RMSE	df_resid
	Value (g)	SE	*T*-Stat	*P*-Value	Value (g/hr)	SE	*T*-Stat	*P*-Value			
Malahayati	8.814	5.192	1.698	0.124	9.041	0.576	15.704	<0.001	0.9648	7.6036	9
Kaisepo	18.814	5.192	3.624	0.006	9.041	0.576	15.704	<0.001	0.9648	7.6036	9
Minahasa	26.414	5.192	5.087	0.001	9.041	0.576	15.704	<0.001	0.9648	7.6036	9
Hasanuddin	14.365	4.059	3.539	0.006	9.424	0.423	22.295	<0.001	0.9822	5.3322	9
Bung Tomo	19.525	4.612	4.233	0.002	9.259	0.480	19.279	<0.001	0.9764	6.0586	9
Thamrin	27.839	3.342	8.331	<0.001	11.050	0.371	29.820	<0.001	0.9900	4.8938	9

Notes: R^2 states Coefficient of Determination (proportion of variance explained by the model, range 0–1); RMSE states Root Mean Square Error (average magnitude of prediction errors in grams); SE states Standard Error, df_resid states Residual Degrees of Freedom.

Table 8. Individual ship regression results (auxiliary engine)

Ship	Intercept				Slope (g/hr)				R^2	RMSE	df_resid
	Value (g)	SE	*T*-Stat	*P*-Value	Value (g/hr)	SE	*T*-Stat	*P*-Value			
Malahayati	-27.931	2.955	-9.453	<0.001	4.284	0.167	25.593	<0.001	0.9864	0.9681	9
Kaisepo	-22.131	2.955	-7.490	<0.001	4.284	0.167	25.593	<0.001	0.9864	0.9681	9
Minahasa	-14.531	2.955	-4.918	<0.001	4.284	0.167	25.593	<0.001	0.9864	0.9681	9
Hasanuddin	-14.903	3.130	-4.761	0.001	4.324	0.174	24.793	<0.001	0.9856	0.9539	9
Bung Tomo	-15.503	3.130	-4.953	<0.001	4.324	0.174	24.793	<0.001	0.9856	0.9539	9
Thamrin	-44.960	3.912	-11.492	<0.001	4.324	0.174	24.793	<0.001	0.9856	0.9539	9

Notes: R^2 states Coefficient of Determination (proportion of variance explained by the model, range 0–1); RMSE states Root Mean Square Error (average magnitude of prediction errors in grams); SE states Standard Error, df_resid states Residual Degrees of Freedom.

Table 9. GLM performance

Engine Type	Intercept				Slope (g/hr)				R^2	RMSE	df_resid
	Value (g)	SE	*T*-Stat	*P*-Value	Value (g/hr)	SE	*T*-Stat	*P*-Value			
Main Engine	19.69	3.73	5.28	< 0.001	9.43	0.40	23.32	<0.001	0.89	12.93	64
Auxiliary Engine	26.77	8.10	3.31	0.002	1.59	0.43	3.66	<0.001	0.17	8.61	64

Notes: R^2 states Coefficient of Determination (proportion of variance explained by the model, range 0–1); RMSE states Root Mean Square Error (average magnitude of prediction errors in grams); SE states Standard Error, df_resid states Residual Degrees of Freedom.

4.5 Model Performance Comparison

Table 7 and Table 8 present the per-ship regression models, which demonstrate high predictive performance for both Main Engine ($R^2 > 0.96$) and Auxiliary Engine ($R^2 > 0.98$) waste. For both engines, running hours were a highly significant predictor ($p < 0.001$) of waste generation. Notably, the slope coefficients, representing waste generation rates, showed moderate variability across the fleet for Main Engine (9.04–11.05 g/hr) and significant variability for Auxiliary Engine (2.81–5.52 g/hr), suggesting distinct operational characteristics for each ship.

Individual ship (Table 7 and Table 8) analysis revealed notable bvariation in Main Engine waste generation rates (9.04–11.05 g/hr, 22% variation) but remarkable consistency in Auxiliary Engine rates (4.28–4.32 g/hr, only 0.9% variation). The consistency in Auxiliary Engine coefficients despite the Simpson's Paradox phenomenon suggests that the underlying operational relationship is highly standardized across the fleet.

In contrast, the fleet-wide models presented in Table 9 show a dramatic divergence in performance. The aggregated model for Main Engine waste performed reasonably well (R^2 = 0.89), capturing the general trend. However, the fleet-wide model for Auxiliary Engine waste failed catastrophically, with a coefficient of determination of only R^2 = 0.17. This severe degradation in explanatory power (from $R^2 \approx 0.89$ to 0.17) upon aggregation provides strong quantitative evidence of Simpson's Paradox, where the aggregated trend obscures the true underlying relationships within subgroups.

Figure 4 presents comprehensive individual ship performance analysis. Panel (a) ranks ships by Main Engine efficiency, identifying Malahayati, Kaisepo, and Minahasa as top performers (9.04 g/hr) and Thamrin as requiring optimization (11.05 g/hr). Panel (b) shows daily waste generation trends revealing consistent patterns within ships but variations between ships. Table 10 presents a comprehensive comparison of cross-validation results for per-ship and fleet-wide models. For both engines, per-ship models demonstrated superior performance. The difference was particularly dramatic for Auxiliary Engine, where the per-ship models achieved near-perfect predictive performance (R^2 = 0.99), while the fleet-wide model failed completely under LOSO-CV (R^2 Mean = -2.48). This 97% loss in explained variance upon aggregation provides strong quantitative evidence of Simpson's Paradox and highlights the danger of using aggregated models for ship-specific predictions.

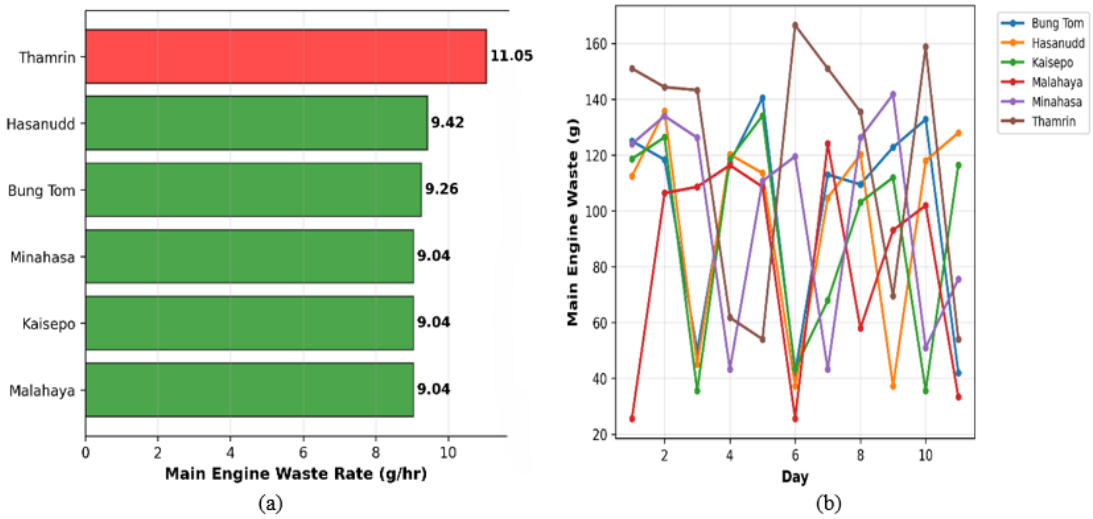


Figure 4. Individual ship performance analysis: (a) Ship performance ranking; (b) Daily main engine waste trends

Table 10. Coefficient of variation (CV) performance comparison

Engine	Model Type	CV Method	Mean R^2	RMSE	Interpretation
Main Engine	Per-Ship	5-Fold	0.96	7.60	Excellent
	Fleet-Wide	5-Fold	0.87	12.71	Good
	Fleet-Wide	LOSO-CV	0.87	12.66	Good
Auxiliary Engine	Per-Ship	5-Fold	0.99	0.97	Excellent
	Fleet-Wide	5-Fold	0.02	8.43	Poor
	Fleet-Wide	LOSO-CV	-2.48	11.76	Model Failure

Note: R^2 Mean = Average coefficient of determination from k-fold cross-validation; RMSE = Root Mean Square Error (g). Negative CV R^2 Mean values indicate model performance worse than baseline mean prediction.

5 Discussion

5.1 The Presence of Simpson's Paradox in Operational Data

This study's primary contribution is the systematic documentation of Simpson's Paradox in shipboard operational environmental data. The discovery that a strong, positive correlation within individual ships (average r = 0.993) is masked by a weak correlation at the aggregated fleet level (r = 0.416) has profound theoretical implications for maritime data analytics. It provides empirical evidence that fleet-wide aggregation, a common practice in maritime research, can produce potentially misleading results. This finding extends the work of Tian and Zhu [17], who found a similar paradox in regulatory PSC data, by demonstrating that the phenomenon also occurs in high-frequency operational data related to environmental performance. Meanwhile, Kotrikla et al. [12] provided general waste

generation data lacking engine-specific details. Dayioglu [37] analyzed waste generation trends and environmental regulations for passenger vessels, highlighting the intricate waste management issues and regulatory difficulties in reconciling environmental safeguards with operational effectiveness.

The 2.2:1 ratio between Main and Auxiliary Engine waste generation aligns with theoretical expectations based on operational intensity and maintenance requirements but provides the first empirical validation of this relationship. The prediction accuracies achieved (97.38% for Main Engine, 98.60% for Auxiliary Engine) substantially exceed previous maritime debris–prediction studies; for comparison, Serra-Gonçalves et al. [7] modeled debris attributable to shipping and commercial fisheries using generalized additive models applied to 14 years of standardized community cleanup records from remote Australian beaches, reporting a national-scale adjusted R^2 of 0.443 and thereby highlighting the inherent difficulty of achieving high predictive fidelity in this domain.

Advanced maritime analytics increasingly shows that vessel-level heterogeneity undermines naïve fleet-wide models. Large-scale operational evidence demonstrates that speed–power relationships vary meaningfully across ships, e.g., using >50,000 noon reports from 88 tankers, Berthelsen and Nielsen find sub-cubic speed–power exponents at sub-design speeds, a result that cautions against pooling vessels under a single parametric form [38]. In parallel, cross-vessel hierarchical (mixed-effects) Bayesian models fitted to 64 cruise ships improve prediction precisely by retaining ship-specific effects while borrowing strength across the fleet [24]. Taken together, these studies substantiate the methodological choice made here: even within ostensibly homogeneous fleets, individual-vessel linear models can outperform fleet-wide specifications when aggregation biases are present—hence the observed 470% predictive performance gain reflects the primacy of correct data-structure recognition over model complexity.

Solonen et al. [24] applied a hierarchical (mixed-effects) Bayesian framework to propulsion-power/fuel modelling across 64 vessels, achieving improved predictive performance relative to classical resistance-based baselines and illustrating the value of modelling vessel-level heterogeneity. However, the current study demonstrates that simple individual ship linear regression models can achieve superior performance when Simpson's Paradox is present, suggesting that model complexity is less important than appropriate data structure recognition.

The paradox arises because aggregation conflates two distinct sources of variation: within-ship variation (the relationship between engine hours and waste for a single vessel) and between-ship variation (differences in baseline waste generation rates across vessels, as shown by the different intercepts in Table 7 and Table 8). Our findings suggest that researchers and practitioners should exercise caution when drawing vessel-level inferences from fleet-level analyses. This aligns with a growing body of literature advocating for more sophisticated, disaggregated, or hierarchical modeling approaches in maritime data analysis to avoid such biases [14, 39, 40].

5.2 Practical and Managerial Implications

This aggregation bias leads to meaningful practical consequences. The comprehensive fleet-wide Generalized Linear Model (GLM) for Auxiliary Engines attained a mere 17.3% predictive efficacy, thereby rendering it inadequate for operational strategic planning. Conversely, the models specific to individual vessels realized an impressive 98.6% of R^2. For illustration, employing the fleet-wide model to forecast waste for a 10-day voyage may result in a cumulative predictive error exceeding 500 grams per vessel, which could jeopardize as much as 10% of the standard weekly storage capacity allocated for this category of waste. Such errors could lead to:

• Suboptimal Resource Allocation: Inaccurate forecasts can result in either overstocking of consumables (bags, bins) or, more critically, under-allocation of storage space.

• Compliance Risks: A systematic underestimation of waste generation could lead to storage overflow, increasing the risk of improper disposal and non-compliance with MARPOL Annex V.

• Inefficient Planning: Port reception facility coordination and voyage planning are hampered by unreliable waste generation estimates.

Our findings suggest that maritime organizations should consider adopting vessel-specific predictive models to mitigate these risks. The 90% reduction in prediction error (RMSE) achieved by individual models provides a clear business case for moving beyond simplistic fleet-wide averages.

5.3 Quantitative Assessment of The Prediction

To quantify the practical implications of this prediction error, we conducted a scenario analysis for a 10-day voyage. The baseline storage capacity of 5 kg was chosen as a conservative estimate for a small, dedicated waste container on a training ship, though this can vary. To assess the robustness of our findings, we performed a sensitivity analysis by varying both the assumed storage capacity (3,000 g; 5,000 g; 10,000 g) and the average daily running hours (8, 10, 12 hours). As shown in Table 11, the underestimation error remains significant across all scenarios. Even with a large 10,000 g storage tank and low 8-hour daily usage, the prediction error still consumes 2.2% of the available capacity. In a more demanding scenario with a small 3,000 g tank and high 12-hour daily usage, the error escalates to a critical 10.8% of storage capacity. This demonstrates that the risk of storage overflow due to flawed

fleet-wide modeling is not an artifact of our initial assumptions but a robust finding with significant operational implications.

Table 11. Scenario analysis of prediction error

Assumed Storage Capacity (g)	Assumed Daily Hours	Prediction Error (g)	Error as Percentage of Storage
3,000	8	-216.8	-7.2%
	10	-271.0	-9.0%
	12	-325.2	-10.8%
5,000	8	-216.8	-4.3%
	10	-271.0	-5.4%
	12	-325.2	-6.5%
10,000	8	-216.8	-2.2%
	10	-271.0	-2.7%
	12	-325.2	-3.3%

Table 11 presents a sensitivity analysis examining the practical consequences of prediction errors arising from the fleet-wide aggregated model compared to the per-ship models. The analysis considers three realistic storage capacity scenarios and three operational scenarios for daily engine running hours. The prediction error represents the systematic underestimation of waste generation by the fleet-wide model, which fails to account for ship-level heterogeneity.

The results demonstrate that the fleet-wide model consistently underestimates waste generation by approximately 217–325 g per day, depending on operational intensity. This systematic bias translates to a consumption of 2.2% to 10.8% of available storage capacity that was not accounted for in the prediction. The relative impact is most severe for vessels with smaller storage capacities: a vessel with 3,000 g capacity operating at 12 hours per day would experience a 10.8% underestimation, potentially exhausting storage capacity earlier than anticipated.

5.4 Limitations and Future Research

This study has several limitations that frame the scope of its conclusions and provide avenues for future research.

• Limited Generalizability: The study was conducted on a homogenous fleet of six training ships over an 11-day period. While this served as an excellent proof-of-concept, the specific waste generation rates and model coefficients may not be generalizable to commercial vessels, different vessel types, or other operational contexts. Future research should aim to validate these findings on larger, more diverse commercial fleets over longer timeframes.

• Uncontrolled Confounding Variables: A key limitation is that this study did not control for several operational factors that could act as confounding variables. These include, but are not limited to, minor differences in vessel age, variations in maintenance history and condition, and differing crew practices or training intensity across voyages. While the vessels are sister ships, these unobserved factors could contribute to the between-ship variation in waste generation. However, it is important to note that the emergence of Simpson's Paradox and significant inter-vessel differences despite the fleet's relative homogeneity arguably strengthens the study's central conclusion. It suggests that if such distinct patterns can arise in a controlled setting, the risk of aggregation bias is likely even more pronounced in diverse commercial fleets where heterogeneity is the norm. Future research should aim to collect data on these potential confounders and employ hierarchical or mixed-effects models to formally disentangle their influence from inherent vessel-specific characteristics.

• Focus on Engine Room Waste: This study was limited to solid waste from engine room operations. Other significant waste streams (e.g., galley, accommodation, cargo-related waste) were not examined and may exhibit different generation patterns.

• Modeling Approach: Our analytical strategy compared two extreme modeling approaches: complete aggregation (fleet-wide models) versus complete disaggregation (per-ship models). While this comparison effectively demonstrates the paradox and its consequences, it does not explore intermediate approaches that could provide a more optimal balance between generalizability and accuracy. Specifically, we did not implement hierarchical mixed-effects models that explicitly account for the nested structure of the data (observations within ships) by estimating both fixed effects (common relationships across all ships) and random effects (ship-specific deviations from the common pattern). Such models represent the statistical gold standard for analyzing hierarchical data and would provide a more formal framework for quantifying the contribution of ship-level heterogeneity to overall variance. Mixed-effects models could also offer better out-of-sample prediction performance for new vessels by borrowing strength from the fleet-level data while accounting for vessel-specific characteristics. The development and validation of such models for maritime operational data is an important direction for future research and would complement the foundational findings of the current study.

Future research should therefore focus on: (1) replicating this study on commercial vessels; (2) investigating the prevalence of Simpson's Paradox in other maritime operational datasets (e.g., fuel consumption, emissions); (3) developing integrated, real-time waste management systems that leverage ship-specific predictive algorithms; and (4) the methodological lessons from this study likely extend to other domains of maritime data analysis, including fuel consumption modeling, emissions estimation, maintenance prediction, and safety risk assessment. Future research should systematically investigate whether aggregation bias is present in these domains and whether hierarchical modeling approaches are needed to avoid misleading conclusions.

6 Conclusions

This study's primary contribution is a methodological framework for detecting and mitigating aggregation bias in maritime operational data, using shipboard waste generation as a case study. It offers several key findings for both maritime data analytics and environmental management. Our analysis suggests that individual ship models, with predictive performance exceeding 97%, are substantially more reliable than conventional fleet-wide approaches, which can be prone to severe aggregation bias.

The central methodological finding is the first documentation of Simpson's Paradox in shipboard operational environmental data. The strong positive correlation between auxiliary engine hours and waste generation observed at the individual ship level was masked at the aggregated fleet level. This finding serves as a methodological caution for the maritime analytics community, suggesting that fleet-wide aggregation can produce misleading results with significant operational consequences. On a practical standpoint, the study provides precise waste generation benchmarks for training ships (9.04–11.05 g/hr for Main Engines; 4.28–4.32 g/hr for Auxiliary Engines) and demonstrates that ship-specific modeling can reduce prediction errors by up to 90%. In light of these findings, maritime operators should consider transitioning from fleet-wide estimates to vessel-specific modeling frameworks to enhance MARPOL Annex V compliance, optimize resource planning, and support data-driven environmental management.

Acknowledgments

The authors acknowledge the Human Resources Development Agency in Transportation, Ministry of Transportation, Republic of Indonesia, for providing access to the training ship fleet and supporting this research. Special thanks to the ship engineers and crews who participated in data collection across all six vessels.

Data Availability

The raw dataset analyzed in this study is not publicly available due to confidentiality agreements with the vessel operators and the proprietary nature of operational data. However, aggregated and anonymized data sufficient to verify the main findings of this study are available from the corresponding author upon reasonable request and with permission from the data owners. Supplementary Material: Detailed statistical assumption test results for all per-ship regression models are provided in Table A1, available as a separate supplementary file accompanying this manuscript. All statistical analyses were conducted using Python (version 3.9.7) with the following key packages: pandas (version 1.3.4), statsmodels (version 0.13.2) and scikit-learn (version 1.0.2).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] H. Husina, M. Mahidin, M. Marwan, F. Nasution, E. Erdiwansyah, A. Ahmadi, S. Muchtar, F. T. Yani, and R. Mamat, "Conversion of polypropylene-derived crude pyrolytic oils using hydrothermal autoclave reactor and Ni/aceh natural zeolite as catalysts," *Heliyon*, vol. 9, no. 4, p. e14880, 2023. https://doi.org/10.1016/j.heliyon.2023.e14880

[2] N. Ali and A. M. Soliman, "Single and multi-response optimization of retarded admixture on alkali activated slag rheological behavior by the taguchi method," *Constr. Build. Mater.*, vol. 459, p. 139812, 2025. https://doi.org/10.1016/j.conbuildmat.2024.139812

[3] N. P. Purba, D. I. W. Handyman, T. D. Pribadi, A. D. Syakti, W. S. Pranowo, A. Harvey, and Y. N. Ihsan, "Marine debris in Indonesia: A review of research and status," *Mar. Pollut. Bull.*, vol. 146, pp. 134–144, 2019. https://doi.org/10.1016/j.marpolbul.2019.05.057

[4] O. A. Ahmad, M. T. Jamal, H. S. Almalki, A. H. Alzahrani, A. S. Alatawi, and M. F. Haque, "Microplastic pollution in the marine environment: Sources, impacts, and degradation," *J. Adv. Vet. Anim. Res.*, vol. 12, no. 1, pp. 260–279, 2025. http://doi.org/10.5455/javar.2025.l893

[5] W. Leal Filho, J. Barbir, J. May, M. May, J. Swart, P. Yang, M. A. P. Dinis, Y. A. Aina, S. Bettencourt, P. Charvet *et al.*, "Towards more sustainable oceans: A review of the pressing challenges posed by marine plastic litter," *Waste Manag. Res.: J. Sustain. Circ. Econ.*, vol. 43, no. 9, pp. 1358–1377, 2025. https://doi.org/10.1177/0734242X251313927

[6] International Maritime Organization (IMO), *MARPOL Annex V: Prevention of Pollution by Garbage from Ships–Consolidated Edition*, 2017.

[7] C. Serra-Gonçalves, J. L. Lavers, H. L. Tait, A. M. Fischer, and A. L. Bond, "Assessing the effectiveness of MARPOL Annex V at reducing marine debris on Australian beaches," *Mar. Pollut. Bull.*, vol. 191, p. 114929, 2023. https://doi.org/10.1016/j.marpolbul.2023.114929

[8] International Maritime Organization (IMO), "MARPOL annex V-extension of garbage record book requirements. Resolution MEPC. 360 (79)," 2024. https://www.imo.org/en/OurWork/Environment/Pages/Garbage-default.aspx

[9] A. La Ferlita, Y. Qi, E. Di Nardo, O. El Moctar, T. E. Schellin, and A. Ciaramella, "A framework of a data-driven model for ship performance," *Ocean Engin.*, vol. 309, no. Part 1, p. 118486, 2024. https://doi.org/10.1016/j.oceaneng.2024.118486

[10] P. Rubbens, S. Brodie, T. Cordier, D. Destro Barcellos, P. Devos, J. A. Fernandes-Salvador, J. I. Fincham, A. Gomes, N. O. Handegard, K. Howell *et al.*, "Machine learning in marine ecology: An overview of techniques and applications," *ICES J. Mar. Sci.*, vol. 80, no. 7, pp. 1829–1853, 2023. https://doi.org/10.1093/icesjms/fsad100

[11] M. H. M. Sibarani, J. Junaidi, and T. A. Wibowo, "Advancing green ship design and experiential learning in maritime education," *Res. Dev. Educ.*, vol. 5, no. 1, pp. 88–98, 2025. https://doi.org/10.22219/raden.v5i1.39397

[12] A. M. Kotrikla, A. Zavantias, and M. Kaloupi, "Waste generation and management onboard a cruise ship: A case study," *Ocean Coast. Manage.*, vol. 212, p. 105850, 2021. https://doi.org/10.1016/j.ocecoaman.2021.105850

[13] Y. R. Kim, M. Jung, and J. B. Park, "Development of a fuel consumption prediction model based on machine learning using ship in-service data," *J. Mar. Sci. Eng.*, vol. 9, no. 2, p. 137, 2021. https://doi.org/10.3390/jmse9020137

[14] K. L. Rødseth, T. Kuosmanen, and R. B. Holmen, "Mitigating simultaneity bias in seaport efficiency measurement," *Transp. Res. Part A Policy Pract.*, vol. 192, p. 104333, 2025. https://doi.org/10.1016/j.tra.2024.104333

[15] R. Elvik, "Simpson's paradox in road safety studies: A narrative review," *Accid. Anal. Prev.*, vol. 31, p. 101471, 2023. https://doi.org/10.1016/j.trip.2025.101471

[16] J. Pearl, "Comment: Understanding Simpson's paradox," *Probabil. Causal Inference*, vol. 14, pp. 399–412, 2022. https://doi.org/10.1080/00031305.2014.876829

[17] S. Tian and X. Zhu, "Data analytics in transport: Does Simpson's paradox exist in rule of ship selection for port state control?" *Electron. Res. Arch.*, vol. 31, no. 1, pp. 251–272, 2023. https://doi.org/10.3934/era.2023013

[18] Kuncowati, "Passenger attitudes and waste management on ships," *Int. J. Sustain. Dev. Plan.*, vol. 19, no. 7, pp. 2683–2689, 2024. https://doi.org/10.18280/ijsdp.190724

[19] P. S. Sijabat, T. Cahyadi, W. Winarno, R. Riyanto, L. Barasa, C. Kuntadi, and M. B. Simanjuntak, "Seafarer readiness for green shipping transition-insights from maritime education and industry professionals," *J. Kaj. Penelit. Umum*, vol. 3, no. 1, pp. 40–58, 2025. https://doi.org/10.47861/jkpu-nalanda.v3i1.1496

[20] R. Y. Siahaan, A. G. Malau, and S. Herawati, "Enhancing sustainability education in maritime cadet training," *J. Abdidas*, vol. 5, no. 3, pp. 228–235, 2024. https://doi.org/10.31004/abdidas.v5i3.933

[21] European Maritime Safety Agency (EMSA), "Annual overview of marine casualties and incidents 2023," 2024. https://www.emsa.europa.eu/publications.html

[22] International Maritime Organization (IMO), "Prevention of pollution by garbage from ships," 2024. https://www.imo.org/en/OurWork/Environment/Pages/Garbage-default.aspx

[23] R. Yan, S. Wang, and H. N. Psaraftis, "Data analytics for fuel consumption management in maritime transportation: Status and perspectives," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 155, p. 102489, 2021. https://doi.org/10.1016/j.tre.2021.102489

[24] A. Solonen, R. Maraia, S. Springer, H. Haario, M. Laine, O. Räty, J. P. Jalkanen, and M. Antola, "Hierarchical bayesian propulsion power models—A simplified example with cruise ships," *Ocean Eng.*, vol. 285, no. Part 1, p. 115226, 2023. https://doi.org/10.1016/j.oceaneng.2023.115226

[25] K. Shi, J. Weng, S. Fan, E. Blanco-Davis, and Z. Yang, "Exploring the influence of seafarers' individual characteristics on the perceived risk in maritime emergencies: A simulator study," *J. Transp. Saf. Secur.*, vol. 16, no. 11, pp. 1378–1402, 2024. https://doi.org/10.1080/19439962.2024.2332740

[26] J. T. Thorson and C. Minto, "Mixed effects: A unifying framework for statistical modelling in fisheries biology,"

ICES J. Mar. Sci., vol. 72, no. 5, pp. 1245–1256, 2015. https://doi.org/10.1093/icesjms/fsu213

[27] S. Wang, B. Ji, J. Zhao, W. Liu, and T. Xu, "Predicting ship fuel consumption based on LASSO regression," *Transp. Res. Part D Transp. Environ.*, vol. 65, pp. 817–824, 2018. https://doi.org/10.1016/j.trd.2017.09.014

[28] M. P. Handayani, H. Kim, S. Lee, and J. Lee, "Navigating energy efficiency: A multifaceted interpretability of fuel oil consumption prediction in cargo container vessel considering the operational and environmental factors," *J. Mar. Sci. Eng.*, vol. 11, no. 11, p. 2165, 2023. https://doi.org/10.3390/jmse11112165

[29] Y. Zhou, W. Daamen, T. Vellinga, and S. Hoogendoorn, "Review of maritime traffic models from vessel behavior modeling perspective," *Transp. Res. Part C Emerg. Technol.*, vol. 105, pp. 323–345, 2019. https://doi.org/10.1016/j.trc.2019.06.004

[30] P. Cariou and A. Cheaitou, "The effectiveness of a european speed limit versus an international bunker-levy to reduce CO_2 emissions from container shipping," *Transp. Res. Part D Transp. Environ.*, vol. 17, no. 2, pp. 116–123, 2012. https://doi.org/10.1016/j.trd.2011.10.003

[31] S. Filom, A. M. Amiri, and S. Razavi, "Applications of machine learning methods in port operations–A systematic literature review," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 161, p. 102722, 2022. https://doi.org/10.1016/j.tre.2022.102722

[32] S. Zhang, J. Chen, Z. Wan, M. Yu, Y. Shu, Z. Tan, and J. Liu, "Challenges and countermeasures for international ship waste management: IMO, China, United States, and EU," *Ocean Coast. Manage.*, vol. 213, p. 105836, 2021. https://doi.org/10.1016/j.ocecoaman.2021.105836

[33] I. Durlik, T. Miller, E. Kostecka, and T. Tuński, "Artificial intelligence in maritime transportation: A comprehensive review of safety and risk management applications," *Appl. Sci.*, vol. 14, no. 18, p. 8420, 2024. https://doi.org/10.3390/app14188420

[34] S. Fuqaha and N. Nursetiawan, "Artificial intelligence and iot for smart waste management: Challenges, opportunities, and future directions," *J. Fut. Artif. Intell. Tech.*, vol. 2, no. 1, pp. 24–46, 2025. https://doi.org/10.62411/faith.3048-3719-85

[35] K. S. Salem, K. Clayson, M. Salas, N. Haque, R. Rao, S. Agate, A. Singh, J. W. Levis, A. Mittal, J. M. Yarbrough *et al.*, "A critical review of existing and emerging technologies and systems to optimize solid waste management for feedstocks and energy conversion," *Matter*, vol. 6, no. 10, pp. 3348–3377, 2023. https://doi.org/10.1016/j.matt.2023.08.003

[36] L. Toneatti, C. Deluca, A. Fraleoni Morgera, M. Piller, and D. Pozzetto, "Waste to energy onboard cruise ships: A new paradigm for sustainable cruising," *J. Mar. Sci. Eng.*, vol. 10, no. 4, p. 480, 2022. https://doi.org/10.3390/jmse10040480

[37] S. Dayioglu, "Marine environment regulations on cruise ships: A special focus on the influence of EU candidacy of Turkey," 2010, mastersthesis, Lund University.

[38] F. H. Berthelsen and U. D. Nielsen, "Prediction of ships' speed-power relationship at speed intervals below the design speed," *Transp. Res. Part D Transp. Environ.*, vol. 99, p. 102996, 2021. https://doi.org/10.1016/j.trd.2021.102996

[39] S. A. Park, M. A. Je, S. H. Jung, and D. J. Park, "Data-driven optimization of ship propulsion efficiency and emissions considering relative wind," *J. Mar. Sci. Eng.*, vol. 13, no. 11, p. 2120, 2025. https://doi.org/10.3390/jmse13112120

[40] V. Presburger Ulniković, M. Vukić, and A. Milutinović-Nikolić, "Analysis of solid waste from ships and modeling of its generation on the River Danube in Serbia," *Waste Manag. Res. J. Sustain. Circ. Econ.*, vol. 31, no. 6, pp. 618–624, 2013. https://doi.org/10.1177/0734242X13477716

Appendix

Simpson's Paradox in the Engine Room: Unraveling Waste Generation on Training Ships:

Table A1. Statistical assumption test results for per-ship regression models

Ship ID	Engine	Shapiro_W	Shapiro_p	BP_LM	BP_p	DW	Correlation
A	M.E.	0.8369489808597742	0.02880245879113353	1.535040752319983	0.2153574913566841	1.8349969663316	0.982236879722131
B	M.E.	0.8369489808597739	0.02880245879113323	1.535040752320002	0.2153574913566814	2.4185837250848	0.9822368797221311
C	M.E.	0.8369489808597743	0.02880245879113368	1.535040752319987	0.2153574913566836	1.440084774318489	0.982236879722131
D	M.E.	0.9682157771897573	0.8679702340938105	2.423823115536151	0.1195029611917734	2.174065787042164	0.991068184500303
E	M.E.	0.9698021495764542	0.884523797444858	3.6101989730447 76	0.05742624860613356	2.534894886362761	0.9881088943138174
F	M.E.	0.8689767115276166	0.07520787107797688	1.51404329099274	0.2185231512066284	2.177490018777046	0.994977514695196
A	A.E.	0.8368147869138951	0.02868596645828698	1.323850215132609	0.2499022247311302	1.842203078737151	0.9931996147917902
B	A.E.	0.8368147869139029	0.02868596645829374	1.323850215132729	0.2499022247311087	2.069548949002669	0.9931996147917902
C	A.E.	0.8368147869138913	0.02868596645828365	1.323850215132664	0.2499022247311204	2.025674985761197	0.9931996147917899
D	A.E.	0.5924254730271548	0.00002046997226599531	0.6675108780775091	0.4139207765247082	2.036590665799626	0.9927589543722396
E	A.E.	0.5924254730271596	0.00002046997226599822	0.6675108780774749	0.4139207765247201	2.037428689304269	0.9927589543722394
F	A.E.	0.5924254730271543	0.00002046997226599515	0.6675108780774273	0.4139207765247368	2.846643279179379	0.9927589543722396

Ship and Engine Definitions:

Ship ID: A = Malahayati, B = Kaisepo, C = Minahasa, D = Hasanuddin, E = Bung Tomo, F = Thamrin. All ships are identical sister training vessels.

Engine Types: Main Engine (M.E.) = Primary propulsion engine used for ship movement; Auxiliary Engine (A.E.) = Generator engine used for electrical power generation and auxiliary systems.

Statistical Test Descriptions:

This table presents the results of diagnostic tests performed on the residuals of all per-ship linear regression models to verify the validity of ordinary least squares (OLS) assumptions. Tests include:

1. Shapiro_Wilk, Shapiro_p: Test for normality of residuals ($p > 0.05$ indicates residuals are normally distributed)
2. BP_LM, BP_p: Breusch-Pagan test for homoscedasticity ($p > 0.05$ indicates constant variance)
3. Durbin-Watson (DW): Test for autocorrelation (values between 1.4 and 2.8 indicate no significant autocorrelation)
4. Correlation (r): Pearson correlation coefficient between running hours and waste generation

Interpretation Summary by Ship and Engine: Main Engine Models:

1. Malahayati, Kaisepo, Minahasa: Normality marginally acceptable ($p \approx 0.029$), homoscedasticity met, independence met. Models are valid.

2. Hasanuddin, Bung Tomo, Thamrin: All OLS assumptions fully met (all $p > 0.05$, DW within range). Models are highly reliable. Auxiliary Engine Models:

1. Malahayati, Kaisepo, Minahasa: Normality marginally acceptable ($p \approx 0.029$), homoscedasticity met, independence met. Models are valid.

2. Hasanuddin, Bung Tomo, Thamrin: Normality violated ($p < 0.001$), but homoscedasticity and independence met. Despite normality violation, models remain valid due to very strong linear relationships ($r > 0.99$) and robustness of OLS.