

MACHINE LEARNING

Beyond Incident Counts: A Multi-Dimensional Statistical Analysis of Consequence Intensity and Recovery Dynamics in Oil and Gas Pipelines

Ojeyemi Festus Oluwaseun^{1*}, Amusa Kamoli Akinwale¹, Adejumobi Isaiah Adediji¹, Owoeye Samuel Oluyemi²

¹Department of Electrical and Electronic Engineering, College of Engineering, Federal University of Agriculture, Abeokuta, Nigeria

²Department of Mechatronics Engineering, College of Engineering, Federal University of Agriculture, Abeokuta, Nigeria

Published October 2026

Abstract

Pipeline incident surveillance often uses event counts or one consequence measure to set priorities, obscuring differences between total burden, consequence concentration, and service disruption. This study analyses 1,681 audited incidents across five categories: Bunkering (BK), Hot Tapping (HT), Ruptured Line Leak (RLL), Sabotage Explosive (SE), and Valve Insertion (VI), and a separate series of 120 monthly aggregates from January 2015 to December 2024. The weighted Composite Risk Score (CRS) combined category mean recorded volume, mean downtime, mean repair cost, and the 95th percentile of recorded volume. Consequence intensity was calculated as the mean incident-level outcome per reported affected kilometre. The SE ranked highest on the CRS (100.00), while VI had the highest recorded-volume and repair-cost intensity and RLL had the highest downtime intensity. Welch ANOVA and Kruskal–Wallis tests did not detect category differences for recorded volume, downtime, repair cost, or affected length (all $p > 0.05$); this is not evidence of equivalence. Recorded volume was not significantly associated with downtime (Pearson $r = 0.0229$, $p = 0.348$; Spearman $\rho = 0.0255$, $p = 0.296$), and simple regression explained 0.052% of downtime variation. In the separate monthly series, all four aggregate measures peaked in 2019 and were lower in 2024; additive decomposition showed May elevations in volume, downtime, and affected length, while repair cost peaked seasonally in October. HAC-adjusted linear trends were not significant. The findings support outcome-specific reporting rather than treating counts or a composite score as a complete risk measure.

Keywords: affected-length intensity; composite risk score; consequence analysis; downtime; pipeline incidents

1 Introduction

Oil and gas pipelines operate across long corridors exposed to varied technical, environmental, and human threats. Their incidents include intentional interference, equipment or material failure, accidental release, and other causes. These mechanisms can lead to product loss, interruption of service, repair expenditure, safety consequences, and environmental damage. A count of reported incidents captures only the number of records; it does not describe the magnitude, concentration, or duration of their consequences.

Pipeline incident research increasingly treats causes and consequences as heterogeneous. Systems-based work separates failure frequency from social, environmental, and economic consequences, while resilience reviews emphasize that failure mechanisms vary in predictability and impact (Silva

et al., 2024; Yodo & Aragon, 2026). These distinctions matter for surveillance planning: a category with a high recorded total may not also have the highest consequence per affected length or the longest recovery-related downtime. Leak-detection reviews likewise distinguish incident history from measured surveillance performance; historical records alone cannot establish a sensor's probability of detection or false-alarm rate (Baroudi *et al.*, 2019).

Several analytical distinctions follow. First, *record frequency is not exposure-adjusted risk*. Counts or shares of records cannot estimate a failure rate without a denominator such as pipeline length, operating time, or another measure of exposure. Second, *absolute consequence and consequence intensity are different estimands*. Absolute volume or cost describes total burden per incident; dividing an incident outcome by its reported affected length describes concentration relative to that length. Third, *a composite score is a decision aid, not automatically a probability or inferential result*. Its ranking depends on the selected variables, normalisation, and weights. Finally, a non-significant test does not prove that incident categories are equivalent.

This study uses 1,681 audited incident records in five categories and evaluates recorded volume, downtime, repair cost, and affected length. It develops a weighted CRS and an affected-length intensity measure, tests for category differences with Welch ANOVA and Kruskal–Wallis procedures, and examines the unadjusted association between recorded volume and downtime. The work is intended to distinguish descriptive ranking from inferential evidence and to assess whether one consequence measure can serve as a proxy for another.

The contribution is fourfold:

1. it reports an auditable incident-level denominator after excluding unresolved duplicate or ambiguous records;
2. it compares a relative composite ranking with outcome-specific intensity measures;
3. it tests whether incident category differentiates the measured outcomes and whether recorded volume is associated with downtime; and
4. it states the boundary between observed incident consequences and claims about occurrence risk, recovery processes, spatial vulnerability, or surveillance performance.

1.1 Research questions

1. How do the five incident categories rank under a weighted composite of selected consequence measures?
2. Do category rankings change when outcomes are expressed per reported affected kilometre?
3. Do the available data provide evidence of differences among categories in recorded volume, downtime, repair cost, or affected length?
4. Is recorded volume associated with downtime in the pooled incident records?

1.2 Hypotheses

For each measured outcome, the omnibus mean-comparison null hypothesis was that the category means were equal. The corresponding alternative was that at least one category mean differed. For the nonparametric comparison, the null was that the category outcome distributions did not differ in rank location under the Kruskal–Wallis test. For association, the null hypotheses were zero Pearson linear correlation and zero Spearman rank correlation between recorded volume and downtime.

These are distinct hypotheses. In particular, a non-significant omnibus test is not an equivalence test and does not establish identical category distributions.

1.3 Recent literature and study gap

Recent pipeline studies reinforce the need to keep incident occurrence, consequence, and recovery measures distinct. A systems-based framework separates social, environmental, and economic consequences across pipeline segments, while emphasizing that a composite risk score depends on how incident frequencies and consequence indicators are defined (Yodo & Aragon, 2026). A resilience framework for oil-pipeline incidents likewise uses event trees and multi-event indicators to examine disruption and recovery, noting the need for time-dependent system-performance data (Asha *et al.*, 2024). These studies support multidimensional analysis, but their scores or scenario outputs should not be confused with exposure-adjusted probabilities unless the underlying exposure and probability models support that interpretation.

Recent empirical evidence also shows that category and consequence associations can vary across datasets. A 2019–2024 crude-oil incident study in southeastern Türkiye reports cause-related differences in incident cost and identifies pipeline material as an important predictor of high-cost cases; however, it does not report downtime or restoration duration (Koçak *et al.*, 2026). By contrast, a U.S. gas-pipeline failure analysis covers a longer historical record and focuses on incident causes and failure-year prediction rather than post-incident recovery time (Silva *et al.*, 2024). These outcomes and populations differ from the present study, so their findings provide context rather than direct replications.

Methods for studying severity increasingly account for spatial and temporal heterogeneity. A study of U.S. natural-gas distribution incidents found that several correlates of severity, including incident response time, varied across space and time (Li *et al.*, 2021). A separate urban gas risk study compares day- and night-time population exposure in spatial risk maps; that temporal contrast concerns changing vulnerability, not a monthly incident-count trend (Rahimi *et al.*, 2024). For catastrophic liquid-pipeline events, extreme-value methods have been used because sparse upper-tail observations make catastrophic severity difficult to estimate from ordinary empirical distributions (Liu & Liang, 2021). Together, this work supports reporting upper-tail downtime and treating percentile estimates cautiously.

Most directly relevant to recovery, a 2026 U.S. hazardous-liquid study models shutdown occurrence and conditional shutdown duration as separate stages, reporting different associated factors for each (Ramezanpour & Hainen, 2026). This two-stage framing is consistent with the present finding that recorded product volume alone is not a useful pooled proxy for downtime. The current records, however, do not include the operational covariates or timestamped response stages needed to fit such a model. The contribution here is therefore narrower: an audited comparison of five incident categories (BK, HT, RLL, SE, and VI) across absolute outcomes, affected-length intensity, and downtime distributions, supplemented by a separately analysed monthly aggregate series.

2 Materials and methods

2.1 Study data and analysis population

The incident-level analysis covered five categories: BK, HT, RLL, SE, and VI. After the record audit, 1,681 records were retained for the primary analysis. The category counts were 517 BK, 642 HT, 71

RLL, 268 SE, and 183 VI. Another 437 ambiguous or duplicate rows were excluded from the primary statistical denominator but retained in the audit trail.

The available fields included a source field labelled stolen volume in barrels, downtime in hours, repair cost in US dollars, and affected segment length in kilometres. Because the incident categories include RLL, the term recorded volume is used in this paper to represent lost product for every category.

The analysis is record-based. The available evidence does not establish a pipeline-length or operating-time exposure denominator; a verified event position, or a timestamped sequence of detection, dispatch, repair, and return-to-service milestones. Consequently, the study does not estimate exposure-adjusted occurrence rates, spatial hotspot risk, sensor performance, or recovery trajectories.

2.2 Record audit and inclusion

Records with sufficiently clear event identity and valid outcome fields were retained for the primary analysis. Rows identified as ambiguous duplicates or cross-file duplicates were excluded because the available identifiers did not establish whether each row represented an independent physical event. Exclusion therefore denotes unresolved event identity, not proof that a record is false. The excluded rows remain necessary for a transparent audit and future source-level adjudication.

A total of 437 rows were excluded. The 437 excluded rows are a material feature of the analysis. If ambiguity or duplicate reporting is associated with category or outcome magnitude, the retained records may differ systematically from the full record set. The results should therefore be interpreted as applying to the 1,681 retained records, not automatically to all source rows or to all pipeline incidents.

2.3 Descriptive analysis and composite risk score

The CRS combined four category-level quantities: mean recorded volume, mean downtime, mean repair cost, and the 95th percentile of recorded volume. Each component was normalised to the maximum value among the five categories, then assigned a weight of 0.40, 0.30, 0.20 and 0.10, respectively. For category k , the raw score was:

$$CRS_k^{\text{raw}} = 0.40N_{V,k} + 0.30N_{D,k} + 0.20N_{C,k} + 0.10N_{V95,k} \quad (1)$$

where $N_{V,k}$, $N_{D,k}$, $N_{C,k}$, and $N_{V95,k}$ are the normalised mean recorded volume, mean downtime, mean repair cost, and 95th-percentile recorded volume. The reported score was rescaled so the highest category scored 100:

$$CRS_k = 100 \frac{CRS_k^{\text{raw}}}{\max_j (CRS_j^{\text{raw}})} \quad (2)$$

The weights express a chosen screening preference; they were not estimated from outcome data or validated against operational decisions. The CRS is therefore a relative ranking among these five categories under this particular specification. It is not an incident probability, exposure-adjusted risk, causal effect, or statistically established severity hierarchy. Because the score is normalised across only five groups, changing the category set or weights can change the result. Weight and specification sensitivity should be reported before operational use.

2.4 Affected-length intensity

For outcome X , category k , and affected length L_i , intensity was calculated as the arithmetic mean of the incident-level ratios:

$$I_{X,k} = \frac{1}{n_k} \sum_{i \in k} \frac{X_i}{L_i} \quad (3)$$

Here, n_k is the number of retained incidents in category k , and L_i is the reported affected length for incident i . The reported units were barrels per kilometre, hours per kilometre, and US dollars per kilometre. The measure is the mean of incident-level ratios; it is not the ratio of category totals or the ratio of category means. It describes consequence relative to reported affected length and is not an incident rate per kilometre or a spatial vulnerability estimate. The measure also depends on the accuracy and comparability of affected-length reporting.

2.5 Statistical tests

Welch ANOVA was used to compare category means while allowing unequal group sizes and variances. For each outcome, the test evaluated the null that all five category means were equal. Kruskal–Wallis testing was used as a rank-based omnibus comparison. Its epsilon-squared effect estimate was calculated as:

$$\epsilon^2 = \max \left(0, \frac{H - K + 1}{N - K} \right) \quad (4)$$

where H is the Kruskal–Wallis statistic, $K = 5$ is the number of categories, and $N = 1,681$ is the total retained sample size. Values displayed as 0.000 are rounded and, under this implementation, estimates below zero are truncated to zero. They should not be interpreted as exact proof that category explains no variation.

The significance threshold was 0.05. Because none of the reported omnibus tests was significant, no category-specific post-hoc comparisons were interpreted. The supplied results do not include confidence intervals for category contrasts, full Welch denominator degrees of freedom, or a prospective power/equivalence analysis. Thus, the strength and precision of possible differences cannot be fully assessed from the summary statistics reported here.

2.6 Association and simple regression

Pearson and Spearman correlations were used to examine, respectively, linear and monotonic association between recorded volume V_i and downtime D_i . A simple ordinary least-squares model was also fitted:

$$D_i = \beta_0 + \beta_1 V_i + e_i \quad (5)$$

where β_0 is the intercept, β_1 is the unadjusted slope, and e_i is the residual. This model does not adjust for category, repair complexity, event location, access, response delay, damage severity, or other potential determinants of downtime. Its R^2 is a measure of in-sample linear fit, not out-of-sample predictive accuracy.

2.7 Separate monthly aggregate analysis

A second dataset contained 120 valid monthly observations from January 2015 to December 2024 for recorded volume, downtime (DT), repair cost, and affected length. This series was analysed

separately from the 1,681 incident records; its observations are monthly aggregates, not additional incident-level cases. Annual totals were calculated from the monthly values. Additive decomposition with a 12-month seasonal period was used to summarize trend, seasonal component, and residual. Seasonally adjusted linear trends were tested with Newey–West heteroskedasticity-and-autocorrelation-consistent (HAC) standard errors; Kendall trend tests were also applied, with Benjamini–Hochberg adjustment across the five metrics. Residual anomaly screening used a classical threshold of 2.0 and a robust median-absolute-deviation threshold of 3.5. ADF tests were applied to decomposition residuals only. These analyses are descriptive and model-based; they do not establish causes for monthly changes.

3 Results

3.1 Audit and category composition

The primary analysis included 1,681 records. The HT contributed 642 records (38.2% of the retained sample), BK had 517 (30.8%), SE accounted for 268 (15.9%), VI had 183 (10.9%), and RLL had 71 (4.2%). These proportions describe the composition of the retained records. Without exposure denominators and a reporting model, they should not be interpreted as comparative incident rates or probabilities.

3.2 Composite ranking and affected-length intensity

The SE ranked first on the CRS at 100.00, followed by HT (98.43), BK (96.82), VI (95.67), and RLL (95.42). The score used Mean Recorded Volume (MRV) and an upper-tail volume measure, while the category-level CRS values were relative to the maximum component values observed among the five categories. Table 1 presents details of the CRS and ranking of incident categories with respect to mean recorded volume, and mean downtime.

Table 1. Composite ranking of the incident records.

Incident	Records	MRV (bbl)	Mean DT (h)	CRS (%)	Rank
SE	268	161.95	18.09	100.00	1
HT	642	157.82	17.80	98.43	2
BK	517	154.15	18.66	96.82	3
VI	183	149.26	18.37	95.67	4
RLL	71	148.88	18.32	95.42	5

MRV, mean recorded volume; DT, downtime; CRS, composite risk score.

The affected-length results produced a different ordering. The VI had the highest recorded-volume intensity (88.66 bbl/km) and repair-cost intensity (USD 5,861.02/km). The RLL recorded the highest downtime intensity (11.01 h/km) while the SE had the lowest reported intensity for all three measures: recorded volume intensity (RVI), downtime intensity (DTI), and repair-cost intensity (RCI), despite ranking first on the CRS. Table 2 depicts obtained results of ranking with respect to affected-length of the pipeline.

Table 2. Ranking of incident categories with respect to RVI, DTI and RCI.

Incident category	Records	RVI (bbl/km)	DTI (h/km)	RCI (USD/km)
BK	517	80.70	8.43	5,068.83
HT	642	70.94	8.23	4,853.99
RLL	71	79.85	11.01	4,688.78
SE	268	62.69	7.47	4,019.42
VI	183	88.66	9.18	5,861.02

RVI, recorded-volume intensity; DTI, downtime intensity; RCI, repair-cost intensity.

The difference between the two rankings is expected because the CRS summarises category-level absolute outcomes, whereas the intensity measures divide each incident outcome by its reported affected length before averaging. However, neither of the two measures subsumes the other.

3.3 Downtime distribution

Downtime summaries show differences between the central and upper portions of the recorded distributions. The BK had the longest median downtime (16.50 h) while the RLL had the largest 95th-percentile downtime (45.65 h), despite its smaller sample size. The VI also recorded a high 95th percentile (44.86 h). Table 3 presents details of the results with respect to downtime distribution.

Table 3. Variation of downtime with incident category.

Incident category	Records	Median downtime (h)	90th percentile (h)	95th percentile (h)
BK	517	16.50	36.94	44.72
HT	642	14.30	35.73	41.17
RLL	71	13.00	41.50	45.65
SE	268	15.15	36.09	43.50
VI	183	14.90	37.72	44.86

It ought to be noted that these summaries describe recorded downtime, not the sequence of restoration activities. The records available for this analysis do not distinguish detection, verification, dispatch, isolation, repair, and return-to-service times.

3.4 Category comparisons

Welch ANOVA detected no statistically significant differences among incident categories for any of the four tested outcomes. Similarly, Kruskal–Wallis tests gave the same qualitative conclusion. Table 4 presents results obtained from the comparison of incident categories.

Table 4. Comparison of incident categories.

Outcome	Welch ANOVA		Kruskal–Wallis test		
	<i>F</i> statistic	<i>p</i> -value	<i>H</i> statistic	<i>p</i> -value	Epsilon-squared
Recorded volume	0.559	0.692	0.962	0.915	0.000
Downtime	0.301	0.877	2.017	0.733	0.000

Outcome	Welch ANOVA		Kruskal–Wallis test		
	F statistic	p-value	H statistic	p-value	Epsilon-squared
Repair cost	0.558	0.693	2.586	0.629	0.000
Affected length	0.484	0.747	2.714	0.607	0.000

All p-values exceeded 0.05. The appropriate interpretation is that these analyses did not detect category differences in the retained sample. They do not demonstrate that the categories have identical distributions, that category has no practical relevance, or that the outcomes are equivalent. The displayed epsilon-squared estimates are rounded and truncated to 3 decimal places.

3.5 Recorded volume and downtime

Next the relationship between the recorded volume and downtime was investigated via the use of correlation coefficients. It was found that the bivariate association estimates were close to zero and not statistically significant. Table 5 presents the results for both Pearson (P) and Spearman (S) correlations. In Table 5, I and sl denote the intercept and slope, respectively.

Table 5. Correlation and regression between recorded volume and downtime.

Measure	Correlation		Regression		
	P	S	I (h)	sl (h/bbl)	R ²
Estimate	0.023	0.026	17.746	0.00288	0.00052
p-value	0.348	0.296	-	0.348	-

P, Pearson correlation; S, Spearman correlation; I, intercept; sl, slope.

The simple model explained 0.0524% of the observed downtime variation. Within this pooled, unadjusted analysis, recorded volume was not a useful linear proxy for downtime. This finding does not show that volume is unrelated to downtime under every operating condition, nor does it rule out category-specific, nonlinear, or confounded associations.

3.6 Volume-to-downtime ratios

Two descriptive ratios were available by category: the ratio of category mean recorded-volume to category mean downtime (MRV/MDT), and the mean of incident-level volume-to-downtime (MILV/DT) ratios. These ratios answer different arithmetic questions and are not recovery rates or estimates of throughput during restoration. Obtained results are shown in Table 6.

Table 6. Volume-to-downtime ratios.

Incident category	MRV/MDT (bbl/h)	MILV/DT (bbl/h)
BK	8.26	17.01
HT	8.86	18.03
RLL	8.13	14.02
SE	8.95	17.16
VI	8.12	16.57

MRV/MDT, ratio of category mean recorded volume to category mean downtime; MILV/DT, mean of incident-level volume-to-downtime ratios.

The HT had the highest mean of event-level ratios, while the SE had the highest ratio of group means. This ordering difference reflects the effect of averaging ratios versus taking a ratio of averages; neither quantity measures repair efficiency or causal recovery speed.

3.7 Temporal results from the monthly aggregate series

The independent monthly series contained 120 observations from January 2015 to December 2024. Annual totals for all four metrics were highest in 2019. Comparing 2019 with 2024, recorded volume fell by 53.4%, downtime by 43.4%, repair cost by 54.6%, and affected length by 53.4%. These are descriptive differences between aggregate totals, not causal effects or exposure-adjusted trends. Table 7 presents the analysis results.

Table 7. Results of analysis of monthly aggregates.

Metric	Total			Change between 2019 and 2024 (%)
	2015	2019	2024	
Recorded volume (bbl)	30,277.4	48,070.5	22,407.0	-53.4
Downtime (h)	3,348.2	5,371.8	3,038.5	-43.4
Repair cost (USD)	1,903,276	3,339,636	1,514,980	-54.6
Affected length (km)	564.15	942.10	438.82	-53.4

The series was non-monotonic: totals rose to the peak in 2019, declined sharply in 2020, recovered partly in 2021–2022, and declined again in 2023–2024. Additive decomposition showed the largest positive seasonal component occurred in May for recorded volume (+15.88%), downtime (+19.24%), and affected length (+16.89%). Repair cost had its largest positive component in October (+16.26%). The most negative components were February for volume (−12.68%) and affected length (−15.54%), and April for downtime (−14.65%) and repair cost (−22.44%). These are seasonal components relative to the overall mean, not forecasts or evidence that calendar month causes incidents.

Seasonally adjusted HAC linear trends were not significant for any metric after adjustment. The seasonally adjusted Kendall tests showed downward rank ordering for recorded volume and affected length, but not for downtime or repair cost. Detail of the results is presented in Table 8.

Table 8. Results of HAC and Kendall tests on seasonally adjusted metric data.

Metric	HAC slope per year	HAC adjusted p	Kendall adjusted p	Summary
Recorded volume	−80.37 bbl	0.333	0.027	Kendall decline; no supported linear trend
Downtime	−3.72 h	0.434	0.291	No supported trend
Repair cost	−USD 3,379	0.434	0.201	No supported trend
Affected length	−1.31 km	0.354	0.029	Kendall decline; no supported linear trend

HAC, heteroskedasticity-and-autocorrelation-consistent (Newey–West).

Classical residual screening flagged 21 metric-month observations in total: four for volume, six for downtime, six for repair cost, and five for affected length. None exceeded the robust MAD threshold of 3.5. May 2022 was the only month flagged across all four metrics by the classical rule. These are review candidates, not confirmed anomalous incidents. The ADF tests indicated stationarity in decomposition residuals ($p < 2 \times 10^{-8}$); this result applies to residuals, not to the original monthly series.

4 Discussion

4.1 One ranking does not represent every consequence dimension

The first main finding is that the category with the highest composite score was not the category with the greatest consequence per affected kilometre. The SE ranked first on the CRS, while VI led in recorded-volume and repair-cost intensity and RLL led in downtime intensity. This is not a contradiction: the metrics answer different questions. Absolute measures describe the recorded burden associated with incidents, whereas per-length ratios describe concentration relative to the affected length reported for each event.

The distinction is important for surveillance design. A relative composite may help organize a management review queue, but decisions based on it should retain the underlying dimensions and explain the chosen weights. A score based on five category averages can conceal within-category variation, and a small difference between two normalized values does not itself establish a meaningful operational separation. The score also does not include exposure, event probability, asset criticality, environmental consequence, or uncertainty. It should therefore be presented as a screening index, not as a complete risk estimate.

This interpretation is consistent with systems-based incident analyses that keep occurrence and consequence dimensions analytically distinct as reported by [Yodo and Aragon \(2026\)](#). It also accords with the broader resilience literature's treatment of pipeline failure mechanisms as heterogeneous, while recognizing that other incident studies may measure different outcomes from those analysed here ([Silva et al., 2024](#)).

4.2 What the non-significant tests mean

Neither Welch ANOVA nor Kruskal–Wallis testing detected category differences for recorded volume, downtime, repair cost, or affected length. The agreement between the two procedures supports a cautious conclusion: the available sample did not provide evidence of category separation on these measured outcomes under these tests. It does not prove that the categories are equivalent. Establishing equivalence would require a pre-specified smallest effect of practical interest, suitable confidence intervals, and an equivalence-testing design.

The sample is also unbalanced: the RLL group contains 71 incidents, compared with 642 HT incidents. Unequal sample sizes do not invalidate Welch ANOVA, but they affect precision, especially for distribution tails. The largest reported 95th-percentile downtime belongs to the smaller RLL group and should be interpreted with appropriate uncertainty. Confidence intervals or bootstrap intervals would improve the reporting of group means, percentile estimates, intensity values, and CRS components.

4.3 Downtime is not a surrogate for product loss

The correlation and regression results show no meaningful pooled bivariate association between recorded volume and downtime in this dataset. The slope is small and the model explains almost none of the in-sample linear variation. Therefore, a surveillance or planning model should not use recorded volume alone as a stand-in for downtime.

The result does not identify what drives downtime. Repair complexity, access, damage extent, availability of parts and crews, isolation requirements, verification delay, and reporting definitions are plausible operational factors, but they were not measured in the analysis described here. A stronger recovery model would require timestamped milestones and covariates, and might treat downtime as a duration outcome with right-skew, censoring, or competing restoration states where appropriate. Until those data are available, separate outcome models are a design recommendation rather than a demonstrated predictive improvement.

4.4 The meaning and limits of affected-length intensity

The intensity metric brings the affected-length field into the consequence analysis without pretending it supplies event location. It identifies categories where the average incident-level outcome is large relative to the reported length affected. It does not indicate where incidents occur, how likely they are per kilometre, or where sensors should be installed. Spatial risk and sensor placement require verified positions, pipeline geometry, exposure, and a detection model. Pipeline risk studies that use spatial statistics rely on such locational information ([Hou et al., 2020](#)).

The mean of incident-level ratios is also sensitive to small denominators and to measurement error in affected length. Future reporting should include the distribution of event-level ratios and compare the mean of ratios with the ratio of aggregate totals, since these are different summaries. Any records with missing, zero, or non-positive affected length should be explicitly enumerated for the intensity analysis.

4.5 Data audit and measurement definitions

Excluding 437 ambiguous or duplicate rows reduces a clear risk of double counting, but exclusion may also create selection bias if records become ambiguous for reasons related to event type, severity, or reporting source. A one-per-cluster sensitivity analysis could test the dependence of results on those exclusions, but its numerical results were not available for this manuscript. It should be run and reported before claiming that the primary conclusions are insensitive to duplicate handling.

The source field labelled “stolen volume” also requires a category-aware definition. A ruptured line leak may involve released product rather than theft. The most defensible cross-category variable name is “recorded product volume,” with the source definition and classification described. Similarly, downtime needs a consistent operational definition, including its start and end points. Repair cost should be defined as nominal or inflation-adjusted, and its currency year and cost boundary should be reported.

4.6 Implications for surveillance and research

The findings support an evidence-aware, multidimensional surveillance workflow:

- use the CRS only as a transparent relative screening index, with component values and weight sensitivity shown;
- report absolute outcomes alongside affected-length intensity and downtime quantiles;
- avoid treating record counts as exposure-adjusted risk;
- do not infer spatial concentration from affected length;
- collect detection, verification, dispatch, isolation, repair, and return-to-service timestamps; and
- retain an audit trail for duplicate, ambiguous, and excluded records.

These recommendations concern analytical design. The present data do not establish surveillance-system probability of detection, false-alarm rate, or field response benefit. Those require labelled sensor streams and operational validation; incident records alone cannot provide those performance estimates (Baroudi *et al.*, 2019).

4.7 Comparison with recent incident and recovery studies

The literature shows why the non-significant category tests in this dataset should not be generalized into a claim that cause never matters. In a 192-event crude-oil series from southeastern Türkiye, incident cost differed by cause, and multivariable analysis associated pipeline material with high-cost outcomes (Koçak *et al.*, 2026). That finding differs from the present non-significant repair-cost tests, but the datasets, categories, cost definitions, and covariate information differ. It is a reason to test contextual predictors in the present data if available, not evidence that one analysis invalidates the other.

Recent recovery research also supports separating disruption occurrence from duration. The 2026 hazardous-liquid study models the probability of shutdown and duration conditional on shutdown as two stages, with different reported correlates for each (Ramezanpour & Hainen, 2026). A multi-event resilience framework similarly relies on system-performance and time-dependent information to represent recovery (Asha *et al.*, 2024). The present volume–downtime regression cannot substitute for these models: it uses one unadjusted predictor and has no timestamped response covariates. Thus, the finding is not that recovery is unpredictable, but that recorded volume alone did not explain downtime in this sample.

The additional monthly results also need careful comparison with prior temporal-risk work. Spatial risk can vary with changing population exposure across day and night (Rahimi *et al.*, 2024), and incident-severity correlates may vary over space and time (Li *et al.*, 2021). Those studies address changing vulnerability or local covariate relationships, whereas the current secondary analysis examines four monthly aggregate outcome series. Its May and October seasonal components should therefore be treated as data-specific monitoring signals requiring validation, not as universal seasonal pipeline-risk patterns.

Finally, the 95th-percentile volume and downtime summaries are useful complements to means, but they are not stable estimates of catastrophic risk by themselves. Extreme-value studies emphasize that sparse upper-tail pipeline events can produce substantial uncertainty in severity estimates (Liu & Liang, 2021). The present analysis does not fit an extreme-value model, so its 95th percentiles remain descriptive sample quantiles.

5 Conclusion

This study shows why pipeline incident prioritisation should extend beyond counts and single outcomes. In the audited sample of 1,681 incidents, Sabotage Explosive ranked highest under the specified weighted CRS, but Valve Insertion had the highest recorded-volume and repair-cost intensity per affected kilometre, and Ruptured Line Leak had the highest downtime intensity. Welch ANOVA and Kruskal–Wallis tests did not detect statistically significant category differences; these results do not establish equivalence. Recorded volume also showed no significant pooled association with downtime and explained almost none of its linear variation.

The separate 120-month series peaked in 2019, showed metric-specific seasonal components, and had no significant seasonally adjusted HAC linear trends. The observed downward Kendall ordering for volume and affected length is not evidence of a steady causal decline. Incident-level and monthly aggregate findings are complementary but have different denominators and should remain separate.

The practical conclusion is methodological: retain outcome-specific reporting, label composite scores as relative screening tools, distinguish affected-length intensity from spatial risk, and improve event identity and response-time data. A stronger recovery analysis requires timestamped response and restoration stages and multivariable duration modelling. Until those fields, category definitions, and duplicate-sensitivity results are verified, findings should be treated as evidence about the retained historical records, not as a universal severity hierarchy or predictive surveillance model.

References

- [1] Asha LN, Yodo N, Huang Y (2024). A quantitative approach to evaluating multi-event resilience in oil pipeline incidents. *CivilEng*, 6(1), 1. <https://doi.org/10.3390/civileng6010001>
- [2] Baroudi U, Al-Roubaiey AA, Devendiran A (2019). Pipeline leak detection systems and data fusion: A survey. *IEEE Access*, 7, 97426–97439. <https://doi.org/10.1109/ACCESS.2019.2928487>
- [3] Hou P, Yi X, Dong HP (2020). A spatial statistic based risk assessment approach to prioritize the pipeline inspection of the pipeline network. *Energies*, 13(3), 685. <https://doi.org/10.3390/en13030685>
- [4] Koçak TK, Karakoç Ş, Aygöl H (2026). Investigating crude oil pipeline incidents in southeastern Türkiye: Causes, consequences, and the environmental cost analysis. *Engineering Perspective*, 6(2), 122–132. <https://doi.org/10.64808/engineeringperspective.1799368>
- [5] Li X, Penmetsa P, Liu J, Hainen A, Nambisan S (2021). Severity of emergency natural gas distribution pipeline incidents: Application of an integrated spatio-temporal approach fused with text mining. *Journal of Loss Prevention in the Process Industries*, 69, 104383. <https://doi.org/10.1016/j.jlp.2020.104383>
- [6] Liu SL, Liang Y (2021). Statistics of catastrophic hazardous liquid pipeline accidents. *Reliability Engineering & System Safety*, 208, 107389. <https://doi.org/10.1016/j.res.2020.107389>
- [7] Rahimi F, Sadeghi-Niaraki A, Ghodousi M, Abuhmed T, Choi SM (2024). Temporal dynamics of urban gas pipeline risks. *Scientific Reports*, 14(1), 5509. <https://doi.org/10.1038/s41598-024-56136-9>
- [8] Ramezanpour E, Hainen A (2026). Statistical modeling of the probability and duration of hazardous liquid pipeline shutdowns: A hurdle regression approach. *Infrastructures*, 11(5), 177. <https://doi.org/10.3390/infrastructures11050177>
- [9] Silva A, Evangelista L, Ferreira C, Valença J, Mendes MP (2024). Towards resilient pipeline infrastructure: Lessons learned from failure analysis. *Discover Applied Sciences*, 6(11), 585. <https://doi.org/10.1007/s42452-024-06273-7>
- [10] Yodo N, Aragon LG (2026). Assessing sustainability risks in natural gas pipeline infrastructure: A systems-based incident analysis framework for social, environmental, and economic impacts. *Environment Systems and Decisions*, 46(2), 17. <https://doi.org/10.1007/s10669-026-10081-2>