

Article

Improving Road and Vehicle Safety Through Administrative Register Data: Sustainable Road Safety Analytics for Romania (2023–2025) via Dual Severity and Context Clustering

Dorin Tataru ¹, Artur Budzyński ^{2,*} and Andreea Cristina Tataru ¹

¹ Department of Mechanical, Industrial and Transport Engineering (IMIT), University of Petroșani, 20 Universității Street, Hunedoara County, 332006 Petroșani, Romania; dorintataru@upet.ro (D.T.); andreeatataru@upet.ro (A.C.T.)

² Department of Transport Systems, Traffic Engineering and Logistics, Faculty of Transport and Aviation Engineering, Silesian University of Technology, 8 Krasińskiego Street, 40-019 Katowice, Poland

* Correspondence: artur.budzynski@polsl.pl

Abstract

Road traffic injuries remain a central challenge for sustainable transport, public health, and mobility governance. The task of monitoring these injuries requires indicators that jointly capture harm severity, road and environmental context, and patterns of vehicle involvement at scale. Using harmonised English-language Romanian police crash exports (2023–2025), we build 92,790 records with 36 variables and estimate two complementary *k*-means typologies: a severity partition based on the fatality, injury, and vehicle-count fields (a register proxy for involvement, not vehicle-type attributes) and a context partition based on the road, environment, mechanism, and cause fields with one-hot encoding and TruncatedSVD. Reported tables and figures reproduce the archived MiniBatch pipeline for replication; for context, full-batch *k*-means clustering on the same embedding is the recommended default when cross-year prevalence stability is required (train–test TVD 0.039 versus 0.569 under MiniBatch). We report silhouette-guided choices ($k = 6$ severity, $k = 4$ context), cross-seed stability, feature ablations, and a 2023–2024 versus 2025 prevalence comparison. A Pearson χ^2 test on severity $\times$ context labels reveals strong statistical significance, yet Cramér’s *V* remains small—statistical association with limited practical coupling, consistent with complementary rather than redundant partitions. Limitations include police-reported injury counts; a coarse vehicle proxy; weak context geometry; and large MiniBatch context drift, which binds inference to within-year descriptive profiling unless analysts refit the model, add version labels, or adopt full-batch context clustering. The contribution is an integrated, reproducible profiling and governance workflow for dashboards and follow-on modelling—not a fixed multi-year cluster taxonomy.



Academic Editors: Marco Guerrieri and Sanjay Nimbalkar

Received: 23 April 2026

Revised: 29 June 2026

Accepted: 1 July 2026

Published: 6 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: safer roads; safer vehicles; sustainable road safety; sustainability of transport systems; road safety; crash data; unsupervised learning; clustering; Romania; machine learning; temporal drift diagnostics

1. Introduction

Road traffic injury remains a measurable burden for sustainable transport systems and population health, and national crash registers are often the only practical source for monitoring exposure and harm at scale [1]. In the global policy framework, transport safety is explicitly embedded in the Sustainable Development Goals through targets on

road-injury reduction and safe, inclusive mobility access [2,3]. Administrative registers also record infrastructure context and vehicle involvement at scale [1,4,5]. EU road-safety policy continues Vision Zero logic and encourages indicators that link infrastructure, behaviour, and vehicle interfaces rather than single-factor blame narratives [1,4–6]. Headline aggregates can nevertheless conceal heterogeneous risk profiles: similar yearly counts may combine very different mixes of severity, road context, and contributing circumstances [7,8]. Methodologically, roadway safety research has deep roots in count and discrete-outcome models for crash frequency and injury severity, with continuing discussion of unobserved heterogeneity, model comparison, and interpretation under administrative reporting constraints [7–9]. Recent intersection-focused severity studies extend this line by linking sight distance, road surface state, and weather to injury outcomes in random-parameters logit specifications with heterogeneity in means and variances [10].

Complementary to model-based inference, exploratory multivariate summaries—including clustering—are widely used to compress high-dimensional survey and register data into interpretable profiles [11,12]. Open software stacks have lowered the cost of reproducible pipelines for encoding, scaling, and partitioning large tables [13,14]. In Central and Eastern Europe, police-maintained incident registers (including SEWIK-style schemas in Poland) have supported factor analyses, severity modelling, and, increasingly, machine-learning studies, though cross-country comparability depends on coding practices and export conventions.

Related work links built-environment heterogeneity and hotspots to severity modelling [15], compares statistical and machine-learning severity pipelines [16], and synthesises data-driven safety modelling [17–19].

We treat the analysis as descriptive profiling, not causal inference: we harmonise Romanian Police releases for 2023–2025 [20]; build severity and context k -means partitions (reported tables follow the archived MiniBatch pipeline; full-batch k -means is recommended for context when cross-year stability is required—see Comparator Design); and report stability metrics, ablations, and a 2023–2024 versus 2025 split. Severity answers the question of “how harmful”; context answers the question of “under what circumstances” [21].

We ask whether the dual partition is interpretable for monitoring (RQ1), stable under restarts and ablations (RQ2), and consistent under the temporal split (RQ3). This study’s contribution to the field is a reproducible pipeline with transparent diagnostics and explicit interpretation limits—not a new clustering algorithm.

2. Materials and Methods

2.1. Software and Implementation Environment

Analyses were implemented in Python 3.11 (Python Software Foundation, Wilmington, DE, USA) using scikit-learn 1.4.2 (INRIA, France), pandas 2.2.2 (NumFOCUS, Austin, TX, USA), and NumPy 1.26.4 (NumFOCUS, Austin, TX, USA) on harmonised Romanian police exports.

2.2. Data Source and Preprocessing

The dataset was assembled from yearly English-language administrative exports covering Romanian road traffic crashes (2023–2025) [20].

After harmonisation, the merged dataset has 92,790 rows and 36 columns (1 January 2023 to 31 December 2025). Yearly exports are harmonised into one analytical table. The key steps were as follows:

- Unify the yearly schema by column position and canonical column names;
- Normalise column names to a lowercase style with words separated by underscores;
- Export a cleaned analytical table suitable for downstream modelling.

Missing Data (Beyond Severity Counts)

Categorical fields used in context clustering are coerced to strings, and missing levels are mapped to a literal `MISSING` token before one-hot encoding (`handle_unknown = "ignore"` in the encoder). Calendar fields (month, weekday, hour) are derived from date; invalid dates are coerced to missing time parts and imputed as zero in the numeric block together with year and number_of_lanes. This mirrors the harmonised analytical table used in the released scripts so that all downstream tables and figures as well as the temporal split use the same rules.

2.3. Severity Clustering

The feature set for severity clustering was as follows:

- Number of fatalities (people who died);
- Number of seriously injured participants;
- Number of slightly injured participants;
- Number of vehicles involved.

As a domain rule, missing values in the above count fields are interpreted as zero. Features are standardised and clustered with MiniBatch k-means [13,14]. The number of clusters k is selected in the range [2, 6] using the silhouette score [22].

2.4. Context Clustering

Inputs include selected road, weather, signalling, mechanism, and cause variables, along with temporal derivatives of the date (month, weekday, hour).

Categorical features are encoded with one-hot encoding (`handle_unknown = "ignore"`), numeric features are standardised, and dimensionality reduction is performed with TruncatedSVD (up to 40 components). MiniBatch k-means is then applied on the SVD embedding. We evaluate $k \in [3, 6]$ with the silhouette score [14,22] as a diagnostic guide. On the full harmonised register, `scripts/diagnose_context_k34.py` exports geometry and overlap diagnostics for $k = 3$ versus $k = 4$ (Section 2.7). The wider seed-stability sweeps in Figures 1–3 use a different aggregation protocol and can highlight other k values in their curves; they are reported as a measure of sensitivity, not as a mandate to change the reported label set within the archived pipeline.

On the archived MiniBatch $k = 4$ partition, the Calinski–Harabasz index is about 7003 and the Davies–Bouldin index about 3.08 on the production embedding. Label–field overlap remains weak for weather (NMI about 0.001) but non-trivial for primary cause (NMI about 0.16) and lighting (NMI about 0.10); the NMI for accident mechanism is about 0.06. These diagnostics support reading clusters as soft co-occurrence profiles, not crisp behavioural types (Tables 1 and 2).

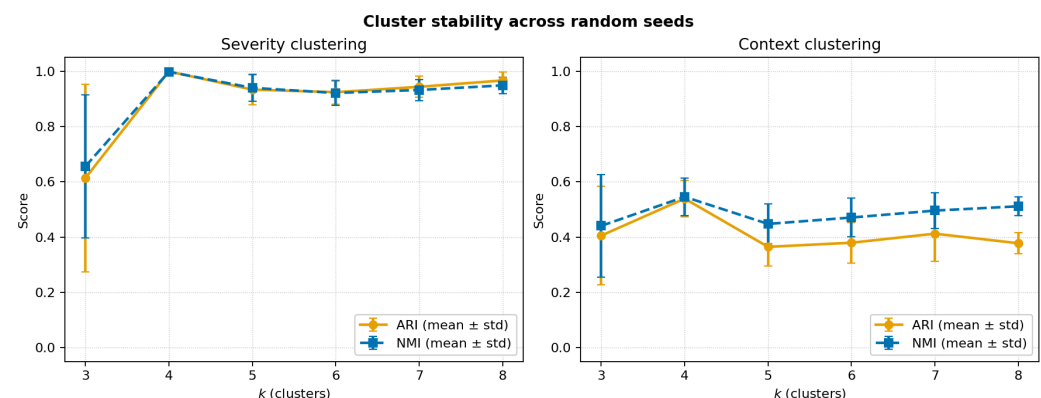


Figure 1. Cluster stability across k (mean and SD across random seeds).

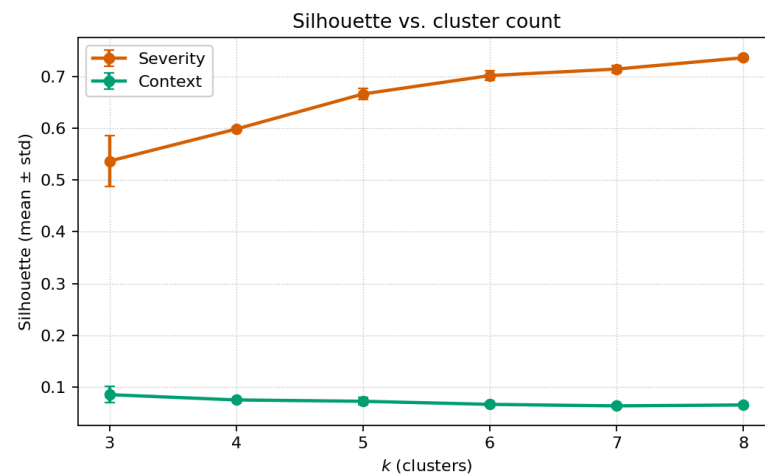


Figure 2. Silhouette sensitivity across k .

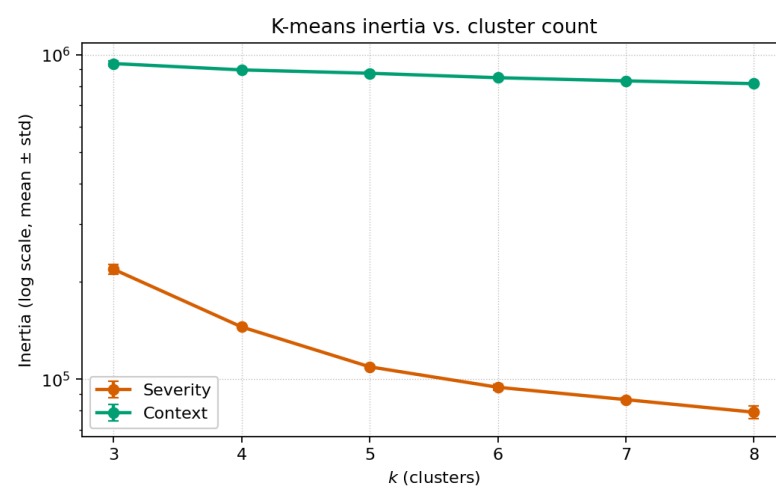


Figure 3. Inertia sensitivity across k .

Table 1. Auxiliary separation scores for the archived MiniBatch context partition (fixed $k = 4$) on the same SVD embedding used in production.

Metric	Value
Silhouette (SVD embedding; sample up to 15,000)	0.0744
Calinski–Harabasz	7003.3
Davies–Bouldin	3.0806
Largest cluster prevalence share	0.2963
Entropy of cluster sizes (nats)	1.3577

Table 2. Entropy overlap between context cluster labels and selected categorical register fields (natural units; NMI in the unit interval).

Register Field	H(Y) Nats	H(Y Cluster) Nats	NMI
primary_cause	3.411	3.033	0.159
accident_mechanism	2.586	2.471	0.058
weather_conditions	0.320	0.319	0.001
lighting_conditions	1.240	1.105	0.104

2.5. Context Feature Ablation Variants (Register Columns)

The ablation workflow instantiates three feature bundles on the harmonised extract before one-hot encoding, SVD compression (up to 40 components, `random_state = 42`), and MiniBatch k -means sweeps (Table 3).

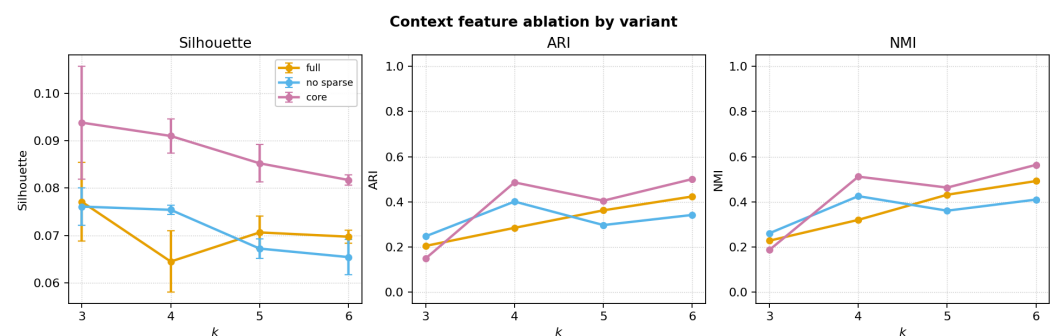
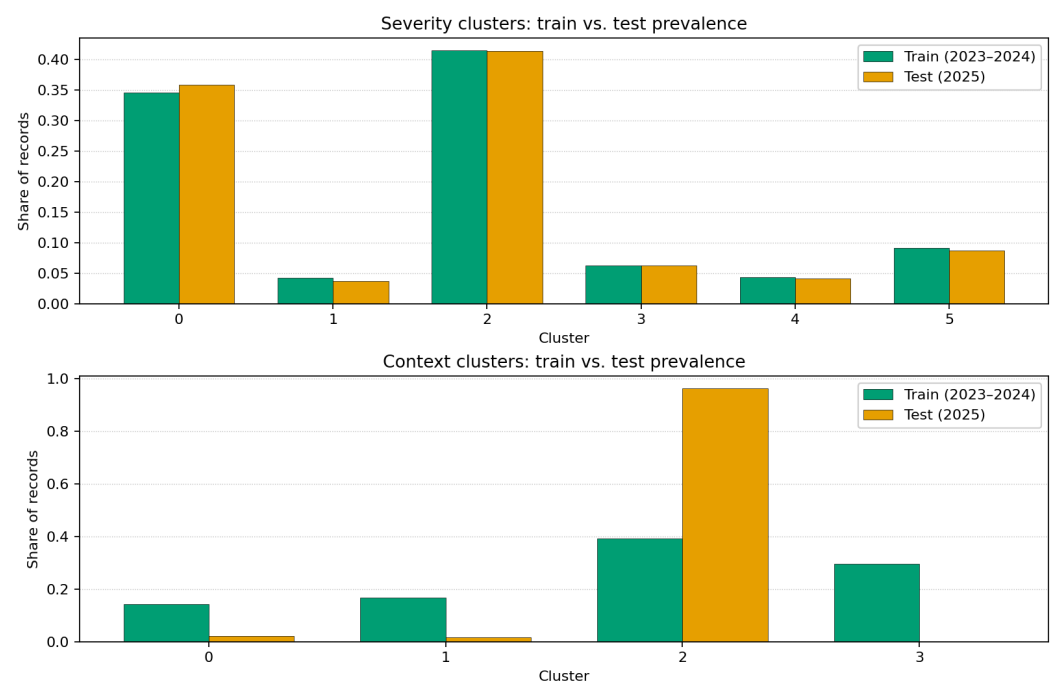
Table 3. Context feature bundles used in ablation diagnostics.

Variant	What Enters the Encoder
full	All 21 categorical context fields used in production clustering (from county through accident_mechanism as enumerated in the repository) plus numeric year; number_of_lanes; and calendar fields month, weekday, hour derived from date.
no_sparse	Same as full except that the sparsest categorical columns are removed: contributing_cause, feature_configuration, and lane_category.
core	Thirteen categorical focus fields (county, within_locality, road_category, road_feature, road_gradient, one_way, pavement_composition, surface_condition, pavement_grip, lighting_conditions, weather_conditions, primary_cause, and accident_mechanism) plus the same numeric/temporal block.

(The upstream register still labels the mechanism column accident_mechanism; elsewhere, we prefer “crash” terminology.)

2.6. Evaluation Metrics (Interpretation)

We use standard scikit-learn diagnostics [13,22–24]: silhouette and inertia on the embedded feature space; ARI/NMI between partitions from different random seeds at fixed k ; and total variation distance (TVD) between train- and test-year label histograms [25] (a larger TVD indicates a stronger prevalence shift). Figures 1–5 report stability sweeps, sensitivity curves, ablations, and the 2023–2024 versus 2025 split.

**Figure 4.** Context ablation comparison across k for silhouette, ARI, and NMI.**Figure 5.** Temporal generalization: train–test cluster-share comparison.

2.7. Choosing the Final Cluster Counts (k)

Throughout the manuscript, we distinguish (i) the *operational* cluster count used for reporting typologies and (ii) *diagnostic* sweeps over alternative k for stability and sensitivity. For severity, the operational search uses $k \in [2, 6]$ with silhouette as the primary score. For context, we use silhouette on the SVD-embedded representation over $k \in [3, 6]$ as a diagnostic, but we retain a fixed operational $k = 4$ for reporting so that the manuscript, tables, and temporal comparisons stay aligned. Separately, we evaluate seed stability for $k \in [3, 8]$ (Figures 1–3) to quantify how partitions change under stochastic k-means restarts and to compare severity and context behaviour. This diagnostic grid is intentionally wider than the operational silhouette search for severity ($k \in [2, 6]$) and is wider than the single reported context cardinality: it stress-tests sensitivity to k and to stochasticity without implying that every k in $[3, 8]$ is a candidate reporting choice.

We retain $k = 6$ (severity) and $k = 4$ (context) under an explicit hierarchy for context: (1) non-degenerate prevalence (no single cluster above about 30% national share on the full register); (2) agreement of geometry scores on the operational SVD embedding (Table 4); (3) interpretable modal contrasts and stronger label–field overlap at $k = 4$ (Tables 5 and 6); (4) cross-seed stability and ablation behaviour in the wider sweeps.

Reporting consistency across tables and temporal panels is a transparency constraint for replication, not a primary scientific reason to prefer $k = 4$.

For severity, when silhouette and restart-stability signals conflict within $k \in [2, 6]$, we prioritise interpretable harm strata and non-degenerate prevalence while reporting the stability figures without claiming a unique true k .

Table 4. Diagnostic context clustering at $k = 3$ vs. $k = 4$ (same preprocessing as production; geometry scores on the SVD embedding; silhouette on a random sample of up to 15,000 rows).

k	Rows	SVD Comp.	Silhouette	Calinski–Harabasz	Davies–Bouldin	Inertia	Max Share	Min Share	Largest ID	Smallest ID
3	92,790	40	0.0589	5727.0	3.4144	974,067.9925	0.3819	0.2832	2	0
4	92,790	40	0.0744	7003.3	3.0806	892,021.4127	0.2963	0.1557	0	2

Table 5. Modal primary-cause and accident-mechanism labels at $k = 3$ versus $k = 4$ (archived MiniBatch replication fit on the same SVD embedding).

k	Cluster	Modal Primary Cause (Share)	Modal Mechanism (Share)
3	0	failure to yield to vehicles (0.14)	side collision (0.27)
3	1	speed not adapted to road conditions (0.18)	side collision (0.16)
3	2	cyclist violations (0.16)	side collision (0.18)
4	0	failure to yield to vehicles (0.13)	side collision (0.19)
4	1	cyclist violations (0.15)	side collision (0.18)
4	2	failure to give priority to vehicles (0.15)	side collision (0.27)
4	3	speed not adapted to road conditions (0.14)	side collision (0.18)

Table 6. Normalised mutual information between context cluster labels and selected register fields at $k = 3$ versus $k = 4$ (same SVD embedding and archived MiniBatch fit; see also Table 4).

Register Field	NMI ($k = 3$)	NMI ($k = 4$)
accident_mechanism	0.044	0.058
lighting_conditions	0.057	0.104
primary_cause	0.087	0.159
weather_conditions	0.003	0.001

2.7.1. Context Cluster Validity at $k = 3$ Versus $k = 4$ (Full Register)

On this embedding, all three geometry scores in Table 4 favour $k = 4$ (higher silhouette and Calinski–Harabasz index; lower Davies–Bouldin index). The trade-off is balance: $k = 3$ prevents the smallest stratum from falling to about 16% of records but concentrates about 38% of mass in one cluster. We therefore do not dismiss silhouette in favour of $k = 4$; we report that the full-table geometry diagnostics align with $k = 4$, while the five-seed stability sweep can still highlight $k = 3$ under a different subsampling protocol (Results).

At $k = 4$, priority/yield, cyclist-violation, and speed-not-adapted themes occupy distinct modal cause labels across four strata; at $k = 3$, the largest cluster merges cyclist violations with other mass. Modal purities remain modest (about 0.13–0.27); thus, neither partition defines crisp behavioural types.

At $k = 4$, overlap with `primary_cause` and `lighting_conditions` is higher than at $k = 3$; weather remains negligible under both cardinalities. This supports retaining $k = 4$ when the goal is to stratify the register by circumstance fields that policy readers inspect, while still treating NMI values as modest (not majority-exclusive clusters).

2.7.2. Relating Severity $k = 6$ to the Stability and Sensitivity Sweeps

Readers may juxtapose three facts from Figures 1 and 2 and the generated stability summary: for severity, the mean cross-seed ARI is highest at $k = 4$ (about 1.00 on five seeds), and the mean silhouette increases with k in the diagnostic grid and peaks at $k = 8$ (about 0.74), while the operational severity search is intentionally capped at $k \in [2, 6]$ with silhouette as the primary score among reporting candidates. On the same five-seed protocol underlying the figures, the mean silhouette for severity rises from about 0.60 at $k = 4$ to about 0.67 at $k = 5$ and about 0.70 at $k = 6$, whereas the mean cross-seed ARI remains high but below the $k = 4$ plateau (about 0.93 at $k = 6$). We therefore retain $k = 6$ as a deliberate trade-off: it improves geometric separation in the severity feature space relative to $k = 4$ while still exhibiting strong restart agreement, and it avoids the finer $k = 8$ split that lies outside the operational reporting cap and can fragment prevalence for dashboard use. Finally, the ARI value 0.780 for severity in Table 7 uses the baseline script’s two-seed aggregation at the operational k ; it need not coincide with the five-seed sweep mean at the same k (Section 2.9, “How to read” paragraph).

Table 7. Methodological comparator summary (identical inputs; diagnostics reported as severity/context).

Method	k	Silhouette	ARI/NMI Stability	Temporal TVD	Interpretation Note
MiniBatch k-means (archived replication)	6/4	0.535/0.035	ARI 0.780/0.249; NMI 0.819/0.294	0.012/0.569	Reproduces published tables; not recommended default for cross-year context
Full-batch k-means (recommended for context)	6/4	0.557/0.041	ARI 0.906/0.985; NMI 0.909/0.975	0.154/0.039	Preferred when context prevalence must transfer across years
Agglomerative (SVD space)	6/4	0.541/0.017	ARI 1.000/1.000; NMI 1.000/1.000	0.067/0.097	Deterministic labels; lower context separation
Gaussian mixture (feasible setting)	6/4	0.460/0.026	ARI 0.615/0.235; NMI 0.722/0.269	0.366/0.071	Highest severity drift in this comparison

2.8. Comparator Design for Methodological Baselines

We compare MiniBatch k -means with full-batch k -means, agglomerative clustering on the SVD embedding, and a Gaussian-mixture baseline on identical inputs [13,14]. *Reported typology tables* (severity and context cluster summaries, cross-tabulation, dashboard illustration) reproduce the *archived MiniBatch pipeline* so that readers can match the repository

artefacts exactly. That choice is for *replication transparency*, not because MiniBatch is the scientifically preferred estimator for every deployment goal.

Table 7 summarises comparator diagnostics on the dense TruncatedSVD embedding (at most 40 components), where full-batch k -means is computationally feasible at $n \approx 9.3 \times 10^4$. For the *context* partition, full-batch k -means attains markedly higher cross-seed ARI/NMI and markedly lower train–test TVD than MiniBatch under the same preprocessing (context TVD **0.039 vs. 0.569**). We therefore **recommend full-batch k -means as the default for context** when year-to-year stability of label prevalence is the dominant monitoring concern. MiniBatch context labels in the main tables should be read as a frozen reproducibility run and sensitivity anchor; severity under MiniBatch shows small temporal TVD (0.012) and remains the archived severity line in this manuscript. Jurisdictions can adopt full-batch k -means as a drop-in replacement on the same SVD features without changing upstream encoding.

Feasibility Note (Dense Embedding Stage)

One-hot expansion of the full categorical design matrix is memory-intensive, but after TruncatedSVD compression, the partition step is a dense $n \times d$ problem with modest d ; both MiniBatch and full-batch k -means are therefore feasible for that step at $n \approx 9.3 \times 10^4$. The gap in Table 7 is therefore interpretable as an estimator choice on the *same* geometry, not as a difference in feature construction.

2.9. Baseline Comparison Template

Table 7 reports direct method comparisons on identical preprocessed inputs. In each metric cell, the first value corresponds to the severity partition and the second to the context partition.

How to Read the ARI/NMI and TVD Cells

Temporal TVD uses the same definition as the temporal-generalisation summary below: models are fitted on 2023–2024, and cluster prevalences are compared to assignments on 2025 (total variation distance between normalised label histograms). The ARI/NMI column reports *cross-seed agreement at the operational k* using the diagnostic protocol bundled with the baseline script (two high-variance seeds, averaged) and therefore *does not* need to match the peak values quoted in the stability sweep, where five seeds are averaged separately for each $k \in [3, 8]$.

3. Results

Cluster typologies below use **archived MiniBatch labels** ($k = 6$ severity, $k = 4$ context; SVD components: 40) so that the tables match the repository export. Comparator diagnostics above define the recommended **full-batch** default for context when cross-year stability is required.

Table 8 maps each experiment to its metrics, the reporting setting, and where the results appear in the text.

Table 8. Core quantitative outcomes.

Experiment	Main Metrics	Setting	Where Reported
Severity clustering	Silhouette + stability	$k = 6$	Table 9; Section 3.3
Context clustering	Silhouette + stability	$k = 4$	Tables 10–12; Section 3.4
Severity $\times$ context linkage Stability analysis	Pearson χ^2 , Cramér's V ARI/NMI	Archived MiniBatch labels $k \in [3, 8]$	Table 13 Figures 1–3
Context ablation	Silhouette/ARI/NMI	full/no_sparse/core	Figure 4
Temporal generalization	Train–test TVD	Train 2023–2024; test 2025	Figure 5
Illustrative monitoring tiles	Prevalence shares; row-normalised severity–context mix; yearly context mix	Archived MiniBatch labels	Figure 6

Table 9. Severity clusters: summary statistics.

Cluster	Avg. Fatalities	Avg. Serious Injuries	Avg. Slight Injuries	Avg. Vehicles	Count
0	0	0	1.024	1	38,433
1	0	0	1.004	2	32,484
2	0.010	1.075	0.294	1.424	8353
3	1.099	0.097	0.361	1.550	3768
4	0.021	0.115	3.712	1.952	3976
5	0.002	0.021	1.158	3.309	5776

Table 10. Context clusters: casualty averages (post hoc profiling).

Cluster	Count	Avg. Fatalities	Avg. Serious Injuries	Avg. Slight Injuries
0	27,498	0.045	0.105	1.047
1	23,366	0.049	0.106	1.025
2	14,448	0.022	0.090	1.004
3	27,478	0.059	0.118	1.091

Table 11. Context clusters: modal categorical descriptors.

Cluster	Primary Cause (Mode)	Weather (Mode)	Lighting (Mode)
0	failure to yield to vehicles	normal	daylight
1	cyclist violations	normal	daylight
2	failure to give priority to vehicles	normal	daylight
3	speed not adapted to road conditions	normal	daylight

Table 12. Context clusters: modal purity (share of records matching the cluster modal label).

Cluster	Primary Cause (Mode Share)	Mechanism (Mode Share)
0	0.132	0.193
1	0.147	0.180
2	0.154	0.273
3	0.142	0.183

Table 13. Cross-classification of archived MiniBatch severity and context cluster labels (counts per record). Pearson $\chi^2 = 1155.7$, $df = 15$, $p < 10^{-16}$; Cramér's $V = 0.064$ (descriptive association on categorical labels).

Severity	Ctx. 0	Ctx. 1	Ctx. 2	Ctx. 3
Sev. 0	11,640	10,412	5019	11,362
Sev. 1	9484	7575	6400	9025
Sev. 2	2423	2144	1203	2583
Sev. 3	1091	1005	290	1382
Sev. 4	1173	920	387	1496
Sev. 5	1687	1310	1149	1630

3.1. Key Quantitative Results

3.2. Cluster Typologies

Table 9 summarises severity clusters by average casualty counts and vehicle involvement (means are taken on the original count scale after clustering on standardised inputs). Tables 10 and 11 summarise context clusters, first with post hoc casualty averages and then with modal categorical descriptors for readability (Table 12 reports modal purity).

Modal labels summarise the most frequent categorical level inside each cluster. Because the English export predominantly records daytime crashes in benign weather, the modes of daylight/normal largely reflect marginal distributions of the register, not an algorithmic artefact; we therefore interpret mechanism/cause contrasts as the substantive signal. Mode shares (purity of the modal label within each cluster) are moderate—typically 0.13–0.27 for the modal primary cause or mechanism—and so clusters are not single-level duplicates of the margin. Full lift tables remain in `tables/context_cluster_lift.xlsx` for auditability.

3.2.1. Linkage Between Severity and Context Partitions

The severity and context partitions are estimated independently on the same records, and so a natural question is whether their labels co-vary beyond chance. Table 13 reports the cross-tabulation of archived severity and context cluster labels together with a Pearson chi-square test and Cramér’s V as a descriptive effect size for categorical co-membership.

When Cramér’s V is small, the Pearson test can still reject independence because sample size is large: the combination signals *statistically detectable but practically weak* association between severity and context labels. That pattern matches the design intent: the partitions emphasise different register dimensions (count-based harm proxies versus road–environment–mechanism configuration). A small Cramér’s V therefore indicates that the label systems are not redundant encodings of the same latent axis, while still allowing *local* concentration in individual contingency cells for targeted review.

Local Cells (Pearson Residuals)

Under independence, expected counts follow the usual row–column product rule. The largest positive Pearson residuals appear at the severity–context pair (Sev. 1, Ctx. 2) with a residual of about +19 (observed 6400 versus expected about 5058), indicating more co-occurrence than independence would predict for low-severity mass inside context cluster 2. Large negative residuals include those of (Sev. 3, Ctx. 2) and (Sev. 0, Ctx. 2), where fatality-heavy or bulk low-harm strata are less represented in context cluster 2 than the baseline condition of independence. The full cell-level table of observed counts, expected counts, and residuals is exported alongside the repository (`tables/cross_tab_severity_context_residuals_full.csv`; top absolute residuals in `tables/cross_tab_severity_context_residuals_top12.csv`) (Table 14).

Table 14. Largest Pearson residuals for the severity–context contingency table (local departures from independence).

Severity	Context	Observed	Expected	Pearson Resid.
Sev. 1	Ctx. 2	6400	5058	18.870
Sev. 0	Ctx. 2	5019	5984.300	−12.480
Sev. 3	Ctx. 2	290	586.700	−12.250
Sev. 4	Ctx. 2	387	619.100	−9.330
Sev. 4	Ctx. 3	1496	1177.400	9.280
Sev. 5	Ctx. 2	1149	899.400	8.320
Sev. 3	Ctx. 3	1382	1115.800	7.970
Sev. 0	Ctx. 1	10,412	9678	7.460
Sev. 1	Ctx. 1	7575	8180	−6.690
Sev. 1	Ctx. 3	9025	9619.500	−6.060
Sev. 5	Ctx. 1	1310	1454.500	−3.790
Sev. 2	Ctx. 2	1203	1300.600	−2.710

Table 13 remains a transparency check rather than a claim of strong coupling.

3.2.2. Numerical Summaries from the Analysis Pipeline

Stability checks. Cluster robustness was evaluated by repeating clustering over multiple random seeds ($n = 5$) for each k in $[3, 8]$. Stability was quantified using the pairwise Adjusted Rand Index (ARI) and Normalised Mutual Information (NMI).

For severity clustering, the highest mean ARI was observed at $k = 4$ (ARI = 1.000, NMI = 0.998). The best silhouette average was obtained at $k = 8$ (silhouette = 0.736).

For context clustering, the highest mean ARI was observed at $k = 4$ (ARI = 0.539, NMI = 0.546). The best silhouette average was obtained at $k = 3$ (silhouette = 0.085).

Figures 1–3 visualise the full stability and sensitivity sweeps.

Context ablation. Ablation experiments compared three context feature variants: full, no-sparse, and core. The variant core achieved the best silhouette at $k = 3$ (silhouette = 0.094, ARI = 0.149, NMI = 0.187).

The variant full achieved the best silhouette at $k = 3$ (silhouette = 0.077, ARI = 0.205, NMI = 0.228).

The variant no_sparse achieved the best silhouette at $k = 3$ (silhouette = 0.076, ARI = 0.247, NMI = 0.261).

Overall, the strongest trade-off between separation and stability was observed for core ($k = 3$).

Temporal generalization. Temporal transfer was evaluated by fitting models on 2023–2024 and assigning clusters to 2025 records.

For severity clustering ($k = 6$), the silhouette changed from 0.696 (train) to 0.705 (test), with distribution shift TVD = 0.012.

For context clustering ($k = 4$), the silhouette changed from 0.086 (train) to 0.134 (test), with distribution shift TVD = 0.569.

3.3. Narrative Summary of Severity Clusters

Table 9 reveals a highly skewed severity structure on the original count scale. Clusters 0 and 1 concentrate the majority of records and combine negligible fatality and serious-injury averages with roughly one slight injury per event and one or two involved vehicles. Cluster 3 carries the highest mean number of fatalities (1.099) and therefore represents the dominant tail-risk stratum in this partition. Cluster 2 isolates serious-injury-dominated outcomes (mean serious injuries 1.075 with comparatively few slight injuries). Clusters 4 and 5 elevate slight-injury and vehicle-involvement averages, respectively, providing interpretable proxies for high-contact crash contexts even though the register does not encode vehicle technologies.

The severity partition separates bulk low-harm events from rarer high-harm profiles for dashboards and labelled strata in supervised pipelines, as descriptive segments rather than mechanistic crash types.

3.4. Narrative Summary of Context Clusters

Tables 10–12 show weaker separation in casualty averages than in the severity partition. Both fatality and injury means are broadly similar across context clusters, indicating that this partition captures co-occurrence patterns in categorical descriptors more than severity extremes.

Instead, we summarise contrasts using *modal* contributing circumstances (Table 11, read together with Table 12). Because modal purities for primary cause or mechanism typically fall between about 0.13 and 0.27, the exemplar labels identify the most frequent levels inside each cluster rather than majority-exclusive identities. Representative contrasts include strata where failures to yield or give priority are modal versus strata where speed not adapted to road conditions is modal as well as a stratum where cyclist-related violations

are modal. Where daylight or benign weather dominates in the export, interpretation relies on mechanism and cause contrasts rather than weather alone. Context clusters summarise co-occurrence in the register; they do not support causal claims about modal labels.

3.5. Illustrative Monitoring Layout

Figure 6 assembles four tiles that could sit on a national road-safety monitoring page next to drill-down tables: prevalence of severity strata, prevalence of context strata, context mix within each severity stratum (row percentages), and coarse calendar-year shifts in context prevalence. The layout is *illustrative* rather than a deployed product interface; it shows how the typology exports translate into compact views for decision support.

3.6. Robustness and Temporal Shift

Stability curves, sensitivity plots, ablations, and the train–test split give a conservative view of behaviour. High ARI/NMI at some k describes restart sensitivity, not a mandate to change the operational k (Section 2.7).

For the **archived** MiniBatch context partition at $k = 4$, the train–test TVD is **0.569** (Table 7; Figure 5): most national probability mass on context IDs shifts between the 2023–2024 fit and 2025 assignments, and so a frozen label map is a weak multi-year taxonomy. Year-to-year comparability of context prevalence at national scale is **not** supported by these MiniBatch labels unless analysts refit, relabel, or switch to full-batch context clustering. Full-batch k -means on the same embedding yields context TVD **0.039**; instability is therefore partly estimator-specific. We recommend that deployments prioritising cross-year stability of context strata adopt full-batch k -means on the same SVD features and treat the MiniBatch context typology in the main tables as a **reproducibility snapshot**, not a stable monitoring taxonomy. The TVD of severity under MiniBatch is 0.012; thus, the archived severity partition does not show the same transport issue. Under the frozen context model, 2025 assignments concentrate in one stratum in the temporal workbook (ctx_test_profile), and so the test-year profile is dominated by a single label. TVD may reflect a genuine change in the mix, reporting drift, or a change in coding; separating these drivers will require data beyond this export.

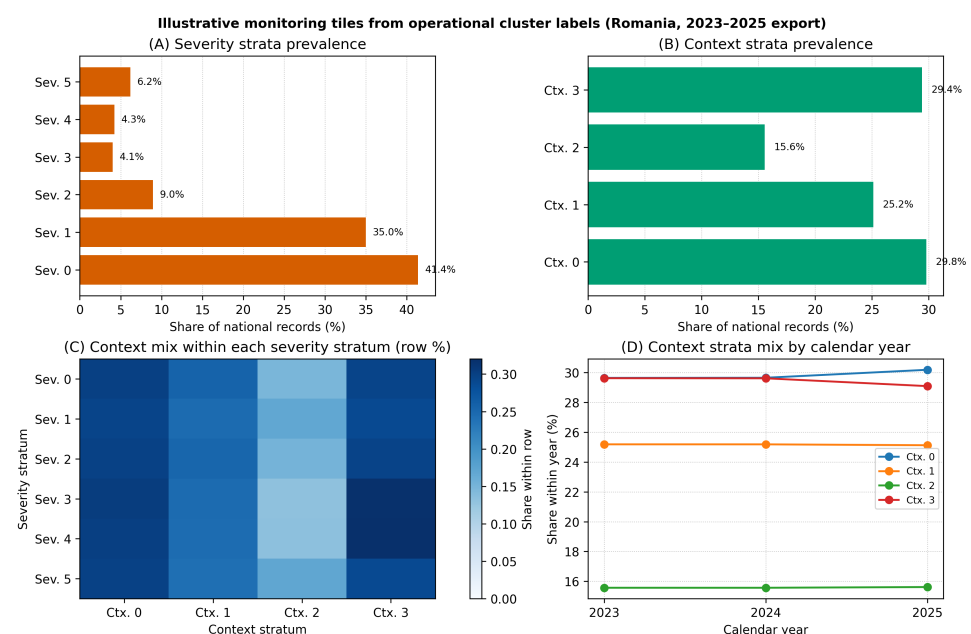


Figure 6. Illustrative monitoring tiles built from archived MiniBatch severity and context labels (Romania, harmonised 2023–2025 export). Panels (A,B) show national prevalence shares; panel (C) row-normalises the severity–context contingency for within-severity context mix; panel (D) tracks context prevalence by calendar year for high-level temporal monitoring.

4. Discussion

The workflow offers a transparent way to profile harmonised Romanian police exports with two complementary partitions: severity (harm proxies) and context (infrastructure, environment, and mechanism fields). The outputs are exploratory typologies for monitoring, not validated risk predictors [11,21].

4.1. Sustainable Transport Monitoring

Under SDG-oriented road-safety policy, sustainable use of administrative registers requires reproducible analytics *and* honest temporal governance: report drift, version labels, and the selection of estimators matched to the monitoring goal [1–4]. This manuscript delivers that combination for Romania (2023–2025)—harmonisation, dual typologies, stability/ablation/temporal diagnostics, and explicit limits—instead of promising immutable cluster IDs across years.

Partitioning inherits the usual fragilities of k -means on heterogeneous administrative data: Euclidean geometry after mixed-type encoding, sensitivity to scaling and rare categories, and partial identifiability across restarts [14,24]. Cross-seed agreement (ARI/NMI) is a restart diagnostic at fixed k , not external validation [23]. Register inference also inherits administrative limits: reporting drift, evolving coding, and English exports that may smooth categories [7,8]. Modal descriptors hide intra-cluster heterogeneity; police records under-represent near misses and encode heterogeneous local practice.

Relative to mainstream roadway analytics, severity modelling and discrete-outcome frameworks remain the reference for principled inference on injury outcomes [7–9]; our profiles are complementary inputs to monitoring before confirmatory modelling [1,3].

For Romanian road authorities, the typologies are best used as *monitoring layers* rather than automated enforcement triggers: severity clusters can prioritise counties or periods for deeper forensic review of high-fatality strata, while context clusters help structure communication around recurring mechanism/cause motifs (priority violations, speed–environment mismatches, and cyclist-related codes). Because the English export lacks geocoded coordinates, *spatial* prioritisation at link or corridor scale is not supported by this workflow; operational triage remains administrative or thematic until coordinates or network referencing can be joined under separate data agreements. National programmes such as infrastructure audits, seasonal enforcement waves, or targeted education can align dashboards with the observed modal mixes, always subject to local validation because the register does not capture interventions or exposure denominators.

Operational Value When Context Transport Is Weak

(1) Within-year descriptive profiling (thematic campaigns, drill-down QA, and strata for follow-on models) remains defensible when modal contrasts and prevalence are read cautiously each release. (2) Fixed national KPI systems on frozen MiniBatch context IDs are not supported at TVD 0.569. (3) Deployments requiring stable cross-year context prevalence should use full-batch k -means on the same SVD features (TVD 0.039), periodic refits, or explicit label versioning (Table 7; Figure 5). Clusters are monitoring strata: users should track prevalence, review modes, and validate them before interventions. Table 15 lists uses and overinterpretation risks.

Table 15. Operational interpretation matrix for cluster use.

Cluster Family	Potential Operational Use	Main Risk of Overinterpretation
Severity (low-harm dominant groups)	Monitoring baseline prevalence, allocating routine prevention communication, building reference strata for annual dashboards	Treating descriptive prevalence as evidence that no intervention is needed in these strata
Severity (high-harm tail groups)	Early warning for rare but severe profiles, prioritizing deeper case review and targeted enforcement planning	Interpreting cluster membership as causal mechanism or individual-level risk certainty
Context (modal priority/yield vs. speed themes)	Structuring thematic safety campaigns and infrastructure audits according to recurring circumstance patterns	Inferring that modal cause labels fully explain observed injuries
Context (modal cyclist-violation theme)	Supporting monitoring of vulnerable road users and cross-checking with local cycling network diagnostics	Translating administrative labels directly into blame assignment without contextual validation

4.2. Limitations and Boundary Conditions

Several boundary conditions follow directly from the register and the modelling choices. First, clusters summarise multivariate co-occurrence in the register; we do not estimate causal links from cluster labels to injury outcomes in this workflow. Second, the context-cluster silhouette on the archived embedding is about 0.074 (Table 4), indicating weak spherical separation; clusters are soft co-occurrence profiles useful for within-year thematic stratification when modal contrasts are read cautiously, not crisp behavioural types. The Calinski–Harabasz and Davies–Bouldin indices on the same embedding favour $k = 4$ over $k = 3$, but the NMI between labels and register fields remains modest (primary cause about 0.16; mechanism about 0.06). Mixed-data methods (k -prototypes, Gower-distance medoids, FAMD-class reductions, latent-class models) are named as priority extensions requiring aligned k , temporal, and stability reporting; they are not benchmarked here to keep the replication stack scikit-learn-centric. Third, temporal drift for **archived MiniBatch context** shows train–test TVD 0.569 (2023–2024 vs. 2025), which rules out silent reuse of those IDs as a multi-year taxonomy; full-batch k -means on the same features yields context TVD 0.039 and is the recommended cross-year default. Fourth, the harmonised export used here does not include geocoded coordinates or linkable linear referencing for network segmentation. Without such fields, *spatial* clustering and standard hotspot or network-risk workflows at link–node or corridor-buffer scales are effectively unavailable: practitioners cannot merge clusters with road-graph geometry, conflict points, or continuous exposure surfaces inside this pipeline. Administrative fields such as county support coarse tabulation and jurisdictional triage but do *not* substitute for crash locations on the network; thus, many map-centred road-manager use cases remain out of scope despite the typologies being useful for national thematic monitoring and list-based drill-down. Fifth, vehicle safety variables such as vehicles’ age, type, mass, advanced driver assistance systems, or consumer test ratings are not present in the harmonised tables; the vehicle-involvement signal is therefore a count-based proxy rather than a direct vehicle-safety measurement programme. Finally, police-reported injury counts inherit under-reporting, especially for slight injuries, classification heterogeneity across counties, and local coding drift—issues that can jointly bias modal descriptors and prevalence trends [7,8]. We do not treat cluster labels as substitutes for field audits or hospital-linked validation.

4.3. Public Value and Implementation Path

The clustering outputs can be translated into citizen-facing communication and budget-facing management tools. Context clusters support local risk messaging (for example, recurring priority or speed patterns), while severity clusters identify where social harm is concentrated.

For road authorities, this creates a direct bridge from diagnosis to intervention and monitoring under constrained budgets. Table 16 operationalises this bridge by linking cluster patterns to intervention focus, public-facing outputs, and measurable KPIs.

Table 16. Implementation bridge from cluster profiles to public value.

Cluster Pattern	Candidate Intervention Focus	Public-Facing Output	KPI for Follow-Up
High-harm severity tail	Targeted treatment of high-severity corridors and junctions (speed management, conflict-point redesign, visibility upgrades)	Local “high-severity risk” bulletin with location and period focus	Change in severe/fatal crash share at treated sites
Low-harm but high-volume severity groups	Low-cost, wide-coverage prevention package (signage consistency, lane guidance, communication campaigns)	Seasonal preventive guidance for common crash scenarios	Change in total incident volume in matched areas
Priority/yield context pattern	Junction operations and right-of-way compliance package (markings, control refresh, enforcement coordination)	“Priority-risk” map and plain-language driver guidance	Change in priority-related cluster prevalence
Speed-related context pattern	Corridor speed-calming and compliance strategy (engineering + enforcement scheduling)	Area-level speed-risk alerts for residents and commuters	Change in speed-related cluster prevalence and temporal drift

Natural extensions include semi-supervised hybrids, explicit spatial and temporal hierarchies, geocoded linkage or small-area dashboards once coordinates or network referencing become consistently available alongside the register, model-based clustering for categorical data (including latent-class and finite-mixture treatments of multivariate indicators), mixed-data reductions and clustering (k -prototypes, Gower-distance medoids, and FAMD-class factor scores) benchmarked with the same temporal and stability reporting used here, and tighter linkage to countermeasure evaluation when intervention records can be joined safely to crash microdata.

5. Conclusions

This project provides a reproducible unsupervised pipeline for profiling road traffic crashes in Romania. The dual view (severity and context) separates bulk low-harm structure from contextual co-occurrence patterns; k is chosen with explicit silhouette, stability, ablation, and temporal diagnostics. Typologies support within-year monitoring and further modelling as descriptive strata, not mechanistic crash taxonomies or fixed multi-year label systems. Herein, “sustainable road safety analytics” means transparent register workflow plus governance (drift reporting, refit cadence, label versioning, and full-batch context when cross-year stability is required) in line with SDG-oriented transport policy [1–3]. The present study yields three takeaways: (1) Dual typologies aid within-year dashboards. (2) Stability and TVD should be reported with any partition. (3) Archived MiniBatch context labels (TVD 0.569) must not be reused as a cross-year taxonomy; full-batch context (TVD 0.039), refits, or versioned labels should be used instead.

5.1. Reproducibility

The analysis is implemented as a reproducible computational workflow (data harmonization, clustering, and manuscript compilation). Analysis scripts used for figure generation and table export are available from the corresponding author upon reasonable request, subject to provider terms and institutional policies. Raw police exports are not redistributed; the reported derived tables and figures can be reproduced from the documented workflow once equivalent source exports are obtained from the official releases.

Key Modelling Hyperparameters

The key modelling hyperparameters were as follows (MiniBatch *k*-means unless noted). Severity clustering uses `n_clusters = 6`, `batch_size = 2048`, `n_init = 20` in the operational clustering script and `n_init = 30` in the temporal generalisation script, `random_state = 42` for the primary labels reported in tables. Context clustering uses `n_clusters = 4`, `batch_size = 2048`, `n_init = 20`, `TruncatedSVD(n_components = min(40, p-1), random_state = 42)` after one-hot encoding with `handle_unknown="ignore"`. *Stability diagnostics* repeat clustering with seeds `{7,21,42,84,168}` across $k \in [3,8]$ (`cluster_stability.py`). *Ablation diagnostics* use seeds `{7,42,168}` (`context_ablation.py`). These values are fixed in the repository so that reruns reproduce the exported `.tex` summaries.

Author Contributions: Conceptualization, D.T. and A.C.T.; methodology, software, formal analysis, investigation, data curation, visualization, writing—original draft preparation, and writing—review and editing, A.B.; validation, resources, supervision, project administration, D.T. and A.C.T.; writing—review and editing, D.T. and A.C.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Underlying micro-level crash exports were obtained from the Romanian Police English releases; redistribution of raw records may be restricted by the data provider's terms, and so we report derived aggregates, tables, and figures here. The analysis scripts and harmonisation workflow used to build these outputs are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ARI	adjusted Rand index
NMI	normalised mutual information
TVD	total variation distance
SVD	singular value decomposition
<i>k</i> -means	<i>k</i> -means clustering (MiniBatch variant used here)
APC	article processing charge

References

1. World Health Organization. Global Status Report on Road Safety 2018. 2018. Available online: <https://www.who.int/publications/i/item/9789241565684> (accessed on 17 April 2026).
2. United Nations. Sustainable Development Goal 3: Ensure Healthy Lives and Promote Well-Being for All at All Ages. 2015. Available online: <https://sdgs.un.org/goals/goal3> (accessed on 17 April 2026).
3. United Nations. Sustainable Development Goal 11: Make Cities and Human Settlements Inclusive, Safe, Resilient and Sustainable. 2015. Available online: <https://sdgs.un.org/goals/goal11> (accessed on 17 April 2026).
4. European Parliament. Resolution on the EU Road Safety Policy Framework 2021–2030: Recommendations on Next Steps Towards Vision Zero. 2021. Available online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021IP0407> (accessed on 17 April 2026).
5. European Commission. Commission Staff Working Document: EU Road Safety Policy Framework 2021–2030—Next Steps Towards “Vision Zero”. 2020. Available online: <https://op.europa.eu/en/publication-detail/-/publication/d7ee4b58-4bc5-11ea-8aa5-01aa75ed71a1> (accessed on 17 April 2026).
6. Pomoni, M. Smart Crosswalks for Advancing Road Safety in Urban Roads: Conceptualization and Evidence-Based Insights from Greek Incident Records. *Future Transp.* **2025**, *5*, 180. [CrossRef]
7. Mannering, F.L.; Bhat, C.R. Analytic methods in accident research: Methodological frontier and future directions. *Anal. Methods Accid. Res.* **2014**, *1*, 1–22. [CrossRef]
8. Savolainen, P.T.; Mannering, F.L.; Lord, D.; Quddus, M.A. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accid. Anal. Prev.* **2011**, *43*, 1666–1676. [CrossRef] [PubMed]
9. Lord, D.; Mannering, F.L. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transp. Res. Part A Policy Pract.* **2010**, *44*, 291–305. [CrossRef]
10. Ullah, I.; Farid, A.; Ksaibati, K. Examining the Effects of Sight Distance, Road Conditions, and Weather on Intersection Crash Severity: A Random Parameters Logit Approach with Heterogeneity in Means and Variances. *Safety* **2025**, *11*, 117. [CrossRef]
11. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv.* **1999**, *31*, 264–323. [CrossRef]
12. Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data Sci.* **2015**, *2*, 165–193. [CrossRef]
13. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
14. Sculley, D. Web-scale k-means clustering. In Proceedings of the 19th International Conference on World Wide Web, New York, NY, USA, 26–30 April 2010; pp. 1177–1178. [CrossRef]
15. Islam, M.K.; Reza, I.; Gazder, U.; Akter, R.; Arifuzzaman, M.; Rahman, M.M. Predicting Road Crash Severity Using Classifier Models and Crash Hotspots. *Appl. Sci.* **2022**, *12*, 11354. [CrossRef]
16. Infante, P.; Jacinto, G.; Afonso, A.; Rego, L.; Nogueira, V.; Quaresma, P.; Sáias, J.; Santos, D.; Nogueira, P.; Silva, M.; et al. Comparison of Statistical and Machine-Learning Models on Road Traffic Accident Severity Classification. *Computers* **2022**, *11*, 80. [CrossRef]
17. Skaug, L.; Nojournian, M.; Dang, N.; Yap, A. Road Crash Analysis and Modeling: A Systematic Review of Methods, Data, and Emerging Technologies. *Appl. Sci.* **2025**, *15*, 7115. [CrossRef]
18. Wang, C.; Fiorentini, N.; Riccardi, C.; Losa, M. Data-Driven Road Traffic Safety Modeling: A Comprehensive Literature Review. *Appl. Sci.* **2026**, *16*, 149. [CrossRef]
19. Bokaba, T.; Doorsamy, W.; Paul, B.S. Comparative study of machine learning classifiers for modelling road traffic accidents. *Appl. Sci.* **2022**, *12*, 828. [CrossRef]
20. General Inspectorate of the Romanian Police. Road Traffic Accidents: Public Statistical Releases (English Exports). Export Years 2023–2025. 2026. Available online: <https://www.politiaromana.ro/en> (accessed on 17 April 2026).
21. Hennig, C.; Meila, M.; Murtagh, F.; Rocci, R. (Eds.) *Handbook of Cluster Analysis*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2015. [CrossRef]
22. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
23. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
24. Vinh, N.X.; Epps, J.; Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. Mach. Learn. Res.* **2010**, *11*, 2837–2854.
25. Gibbs, A.L.; Su, F.E. On choosing and bounding probability metrics. *Int. Stat. Rev.* **2002**, *70*, 419–435. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.