Journal of the Society of Automotive Engineers Malaysia
Volume 10, Issue 1, pp 97-103, January 2026
e-ISSN 2550-2239

Unveiling Risk Factors in Motorcycle Safety: A Data-Driven Analysis in Spain

G. Mensa*, D. Martín and P. Garriga

Applus+ IDIADA, P.I Albornar, S/N, 43710 Santa Oliva, Tarragona, Spain

*Corresponding author: genis.mensa@idiada.com

ORIGINAL ARTICLE ***Open Access***

Article History:

Received
28 Oct 2025

Accepted
15 Dec 2025

Available online
1 Jan 2026

ABSTRACT – *This study analyses motorcycle safety in Spain using an Exposure- Based Risk Index (IREx) methodology applied to 4,006,804 registered motorcycles and 30,875 road accidents from 2022. The research integrates regionally specific Vehicle Kilometers Travelled (VKT) estimation combining traffic census data, technical inspection information, and statistical modelling. Key findings reveal rider age as the predominant risk factor, with riders ≤24 years showing higher accident probability. Vehicles with high-displacement engines (>500cc) and Sport touring motorcycles present elevated mortality risks compared to other motorcycle categories. The methodology addresses critical gaps in exposure-based assessment through comprehensive Spanish traffic pattern analysis and refined risk stratification with transparent uncertainty quantification. Findings support ASEAN NCAP's motorcycle safety protocol development by establishing evidence- based risk profiles and providing actionable insights for targeted interventions, particularly for high-risk populations. This scalable framework offers an adaptable motorcycle risk assessment methodology for diverse regional contexts, supporting evidence-based road safety policies.*

KEYWORDS: Malaysia motorcycle safety, exposure-based risk index (IREx), vehicle kilometers travelled (VKT), road traffic accidents, risk assessment methodology, rider demographics

1. INTRODUCTION

Road traffic accidents represent one of the leading causes of mortality worldwide, with the World Health Organization (2022) identifying them as the eighth leading cause of death globally. Motorcycle incidents represent a disproportionately high percentage of fatalities despite motorcycles being a small fraction of the vehicle fleet. This vulnerability stems from rider exposure, limited protection, and complex interactions between human factors, vehicle characteristics, and traffic environments.

Current motorcycle safety research has critical knowledge gaps limiting intervention effectiveness. Traditional accident analysis fails to account for exposure variations across rider demographics and vehicle categories, leading to misleading risk assessments (European Commission, 2023). While foundational methodologies exist, significant limitations remain in regional applicability and data granularity.

The European context, particularly Spain, presents unique opportunities to address these methodological challenges. Spain's comprehensive vehicle registration systems, detailed accident reporting mechanisms, and diverse motorcycle usage patterns provide an ideal testbed for developing refined risk assessment methodologies. However, current Spanish motorcycle safety research lacks systematic exposure-based analysis that accounts for the complex interplay between rider age, vehicle characteristics, and regional traffic patterns. Current frameworks lack the risk stratification necessary for targeted safety interventions. This research aims to develop and validate a comprehensive exposure-based motorcycle risk assessment methodology tailored to regional traffic patterns and demographic characteristics.

2. METHODOLOGY

This study employed a comprehensive four-phase methodology to analyze the relationship between motorcycle characteristics and road accident risk in Spain. The approach was designed to address significant data quality challenges inherent in the national motorcycle fleet database while establishing a robust analytical framework. The methodology progressed from initial data quality assessment through advanced artificial intelligence applications, culminating in a standardized categorization system that enables reliable statistical analysis of accident patterns and risk factors. The methodology emphasizes data integrity, systematic validation, and the integration of multiple external sources to ensure the reliability of findings. Each phase built upon the previous one, creating an increasingly refined dataset suitable for drawing meaningful conclusions about motorcycle safety in Spain (Directorate-General for Traffic, 2022).

2.1 Data Quality Assessment and Initial Evaluation

The first phase involved comprehensive evaluation of the motorcycle fleet database quality, revealing significant challenges requiring systematic attention. The database contained exclusively static vehicle information, lacking dynamic data such as annual mileage, which limited analysis aspects.

Three primary categories of data inconsistencies were identified: typographical errors and nomenclature inconsistencies in brand and model denominations, including spelling variations, inconsistent capitalization, and simple mistakes hindering proper categorization; anomalous power-to-weight ratio values with technically impossible records such as zero ratios and infinite values indicating calculation errors; and engine displacement data with implausible values exceeding 3000cc, suggesting systematic data entry errors or unit conversion problems.

This phase established the foundation for subsequent data cleaning efforts by documenting the scope and nature of quality issues. The researchers developed recommendations for future database maintenance, including implementation of data validation systems at entry points, establishment of acceptable ranges for numerical values, creation of standardized catalogues for brands and models, and development of periodic data quality verification protocols.

2.2 External Database Integration and Validation

The second phase focused on leveraging external databases to validate and correct identified inconsistencies through systematic cross-referencing. Three primary sources were selected: the BIKEZ_KAGGLE database (38,684 motorcycle models), NHTSA database (29,164 models), and EURONCAP database (791 models). The integration process centered on cross-referencing Vehicle Identification Number (VIN) codes between the national motorcycle fleet and NHTSA data, with additional validation through brand and manufacturer matching. This approach aimed to provide authoritative corrections for nomenclature inconsistencies while establishing verification protocols.

However, VIN codes proved unreliable in the original dataset, leading to much lower matching success rates than anticipated. While some brand corrections were achieved, the inability to verify model information highlighted the need for more sophisticated methodologies. The challenges provided valuable insights into cross-database integration complexity and limitations of traditional matching approaches with imperfect data. These findings informed the decision to pursue advanced computational approaches, recognizing that simple cross-referencing was insufficient. Partial success nonetheless established important validation protocols and demonstrated the value of external source verification.

2.3 Advanced Computational Processing and Artificial Intelligence Applications

The third phase implemented sophisticated computational techniques combining natural language processing, machine learning, and artificial intelligence to address data quality challenges that traditional methods couldn't resolve. This represented a significant methodological advancement employing cutting-edge techniques for systematic data cleaning at scale.

The process began with comprehensive data preprocessing, including selection of relevant columns, systematic transformations (conversion to uppercase, power-to-weight ratio calculations, duplicate removal), and column renaming to establish homogeneous schemas.

VIN-based imputation utilized the NHTSA_VIN_EXTRACTOR library to extract brand information directly from vehicle identification numbers, providing high-confidence corrections for matched records while identifying those requiring more sophisticated treatment. Semantic imputation employed multiple master databases with hierarchical prioritization, utilizing the Jaro-Winkler algorithm for string similarity calculations against validated records from EURONCAP, BIKEZ, NHTSA, and manually curated databases. Two similarity thresholds were established: 0.90 for direct matches and 0.70 for cases requiring AI support.

Large Language Model integration provided intelligent correction capabilities for ambiguous cases, constructing detailed context prompts and suggesting probable corrections with appropriate origin labelling. Pseudonym substitution addressed common model variations through predefined dictionaries, while hierarchical model imputation operated on filtered datasets grouped by validated brands, reducing computational complexity while maintaining accuracy

2.4 Comprehensive Categorization and Final Data Structuring

The final phase consolidated all previous improvements into a comprehensive categorization system that enabled meaningful statistical analysis of motorcycle risk factors. This phase involved systematic validation labelling, external categorical integration, machine learning-based classification, and systematic discretization of continuous variables into analytically useful categories.

Data consolidation involved merging vehicle fleet and accident datasets with systematic validation labelling. Each record received a validation status indicating whether it had no validation, partial validation covering only brands, or complete validation covering both brands and models. Date formatting was standardized across datasets, and accident status indicators were added to enable comparative analysis between general fleet and accident-involved vehicles.

External categorical integration mapped validated brand and model combinations to motorcycle categories using the comprehensive BIKEZ database. This process involved direct matching where possible, supplemented by manual exploration using updated web sources, distributor information, and motorcycle association data to capture categories do not present in the primary external databases.

Machine learning classification addressed missing categories through development of a Random Forest classifier using technical variables including engine displacement, vehicle weight, and power output. The classifier was trained on a 67% subset of records with known categories and validated on the remaining 33%, representing over 1.3 million motorcycles. Variable standardization prevented magnitude bias, and the resulting model achieved high overall accuracy with confidence metrics indicating that 75% of predictions exceeded 92% certainty and all remaining predictions exceeded 80% certainty.

Systematic variable discretization transformed continuous variables into meaningful categorical ranges suitable for risk analysis. Engine displacement was categorized into small (0-125cc), medium-low (125-500cc), medium-high (500-750cc), and very high (750cc+) displacement ranges. Rider age was grouped into categories of 18-24 years, 25-34 years, 35-44 years, and 44+ years to capture developmental and experiential differences. Vehicle age categories ranged from very recent (0-2 years) through recent (3-5 years), medium age (6-10 years), old (11-15 years), very old (16-20 years), to classic or prolonged use (20+ years).

The final categorization achieved impressive coverage rates, with 80% of the vehicle fleet and 86% of accident-involved vehicles successfully categorized. This comprehensive categorization system provided the foundation for all subsequent statistical analyses, ensuring that risk assessments could be conducted across meaningful vehicle and rider categories while maintaining statistical validity through proper treatment of data quality issues identified and addressed throughout the four-phase methodology.

2.5 Estimation of Kilometers Travelled

To estimate the total kilometers travelled for each category, data from mandatory ITV (Inspección Técnica de Vehículos) inspections have been used. During these inspections, the yearly travelled distance is read from the odometer.

In order to impute the missing values, averages were computed for each motorcycle type (sport, naked bike, scooter, etc.), for motorcycles of four years, to control for age. This resulted in 4,818,979 km read from 826 different vehicles. Each motorcycle was then given its corresponding average. A potential limitation of this imputation is failing to take into account other crucial data, such as driver age or motorcycle age.

3. RESULTS AND DISCUSSION

3.1 Overall Risk Index Performance and Population Characteristics

Analysis of Spain's motorcycle fleet showed significant safety variations across rider and vehicle categories. While low engine displacement scooters dominate the fleet (68% of Spanish motorcycle registrations), high-displacement sport and touring motorcycles (12% of the fleet) are disproportionately represented in severe accidents. This mismatch between fleet composition and accident severity highlights the importance of exposure-based risk assessment methods (Nombela & Infantes, 2011).

3.2 Age Impact on Motorcycle Accident Risk

The comprehensive examination of IREx values across different ages demonstrates consistent and statistically significant risk patterns that transcend vehicle categories and usage characteristics. The mortality rate analysis based on kilometers travelled clearly identifies that riders aged ≤24 years exhibit the highest exposure to fatal motorcycle accidents. This demographic demonstrates fatality IREx values consistently exceeding 30 per 100 million kilometers across multiple motorcycle categories, placing young riders firmly within the high-risk classification threshold. In Figure 1, the fatal injury rate per average kilometer travelled is shown.

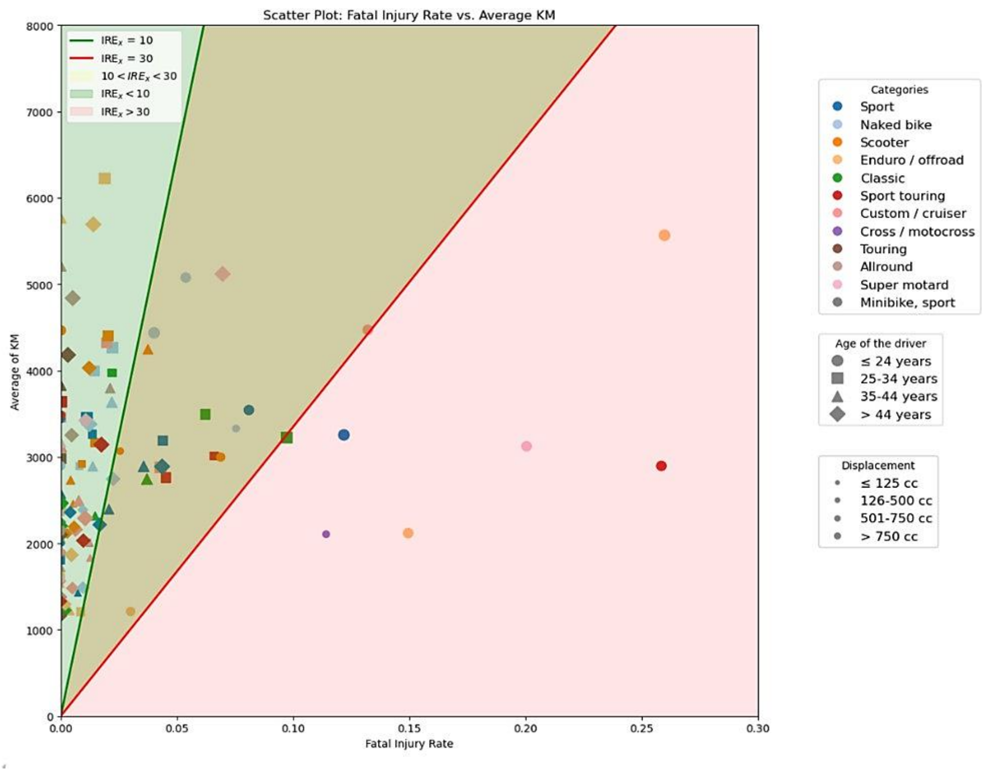

FIGURE 1: Scatter plot showing fatal injury rate per average kilometer travelled

Similarly, the serious injury rate analysis based on kilometers travelled confirms that riders aged ≤24 years demonstrate the highest exposure to severe non-fatal motorcycle accidents. Young riders exhibit serious injury IREx values exceeding 300 per 100 million kilometers, with some motorcycle categories for supermotard vehicles operated by this age group. The consistency of age-related risk elevation across both fatal and serious injury outcomes indicates systematic factors affecting young rider safety rather than random variation or measurement artifacts. The progressive risk reduction observed in successive age decades supports developmental theories linking cognitive maturation, risk perception capabilities, and experiential learning to improved safety outcomes.

3.3 Motorcycle Type and Risk Analysis

Vehicle-specific risk assessment reveals significant safety disparities across motorcycle categories, challenging conventional safety assumptions. Sport touring motorcycles emerge as the highest mortality risk category in Spain. This elevated risk appears linked to their dual-purpose nature, combining high-performance capabilities with comfort features that encourage extended high-speed riding and long-distance touring. Supermotard motorcycles demonstrate the highest serious injury IREx values. Their unique risk profile reflects design characteristics combining off-road capabilities with on-road operation, potentially encouraging aggressive riding behaviors. While not leading fatality rates, supermotard motorcycles dominate serious injury statistics, with accident patterns characterized by high-frequency, moderate-severity incidents.

3.4 Engine Displacement Analysis

Engine displacement analysis revealed clear risk escalation patterns transcending traditional boundaries. Six of the seven highest mortality risk profiles, as in Figure 2(a), corresponded to medium-high displacement motorcycles (500cc or more). For sport motorcycles, see Figure 2(b), engine displacement emerged as a determining risk factor with threshold effects at approximately 500cc, above which fatality risk increases exponentially. Sport motorcycles below this threshold demonstrated fatality IREx values of 20-30, while those exceeding 500cc achieved values of 45-75 depending on rider age. High-displacement motorcycles showed consistent overrepresentation in fatal accidents, while representing only 12% of the registered fleet. This disproportionate representation persisted after exposure adjustment, indicating genuine risk elevation rather than usage pattern artifacts.

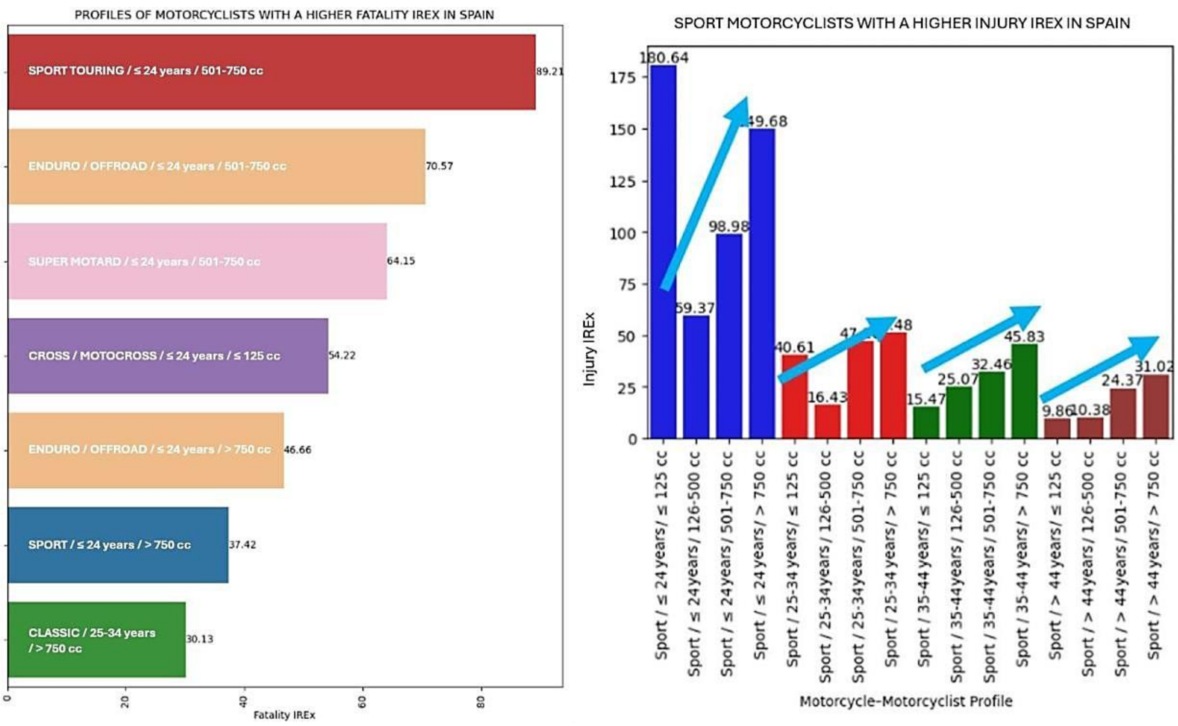

FIGURE 2: (a) Motorcyclists profiles with higher fatality IREx; (b) Sport motorcyclists' injury IREx by age and engine

4. CONCLUSION

This comprehensive exposure-based analysis of Spain's motorcycle safety provides critical evidence for targeted intervention strategies and international safety protocol development. The study's analysis of 4,006,804 motorcycles and 30,875 accidents demonstrates that rider age represents the dominant risk factor, with riders ≤24 years exhibiting higher accident probability across all vehicle categories, transcending motorcycle type and engine displacement boundaries.

Complex motorcycle-specific risk patterns challenge conventional safety assumptions, with sport touring motorcycles unexpectedly demonstrating the highest mortality rates and engine displacement threshold effects at 500cc, indicating non-linear risk escalation. High engine displacement motorcycles represent only 12% of the fleet but account for 34% of fatalities, emphasizing the critical importance of exposure-based assessment for accurate policy targeting.

Synergistic risk interactions between young riders and high-performance motorcycles produce risk amplification effects exceeding expectations, providing clear intervention priorities combining licensing restrictions, training requirements, and technology-based solutions.

The study's findings directly support ASEAN NCAP's motorcycle safety objectives by establishing quantitative vulnerability assessment frameworks that can inform regional protocol development. The observed demographic-mechanical risk relationships provide evidence-based foundations for standardized rider vulnerability classifications across diverse ASEAN contexts. The methodology's scalable framework enables adaptation to varying regional fleet compositions, licensing systems, and infrastructure conditions while maintaining consistent risk assessment standards.

The IREx methodology offers ASEAN NCAP a robust tool for developing region-specific safety protocols that account for local riding patterns while supporting harmonized safety evaluation procedures. The identification of high-risk rider-vehicle combinations provides actionable targets for coordinated regional safety interventions, directly advancing ASEAN NCAP's mission to reduce motorcycle fatalities through systematic, evidence-based safety assessment and protocol implementation.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the Dirección General de Tráfico (DGT) of Spain for providing comprehensive access to the national motorcycle registration database, accident records, and vehicle inspection data that enabled this research. Special recognition is extended to the DGT's National Road Safety Observatory and Motorcycle Working Group for their methodological guidance and validation support. This research was conducted under DGT Project 3DGT6A000172.

REFERENCES

Directorate-General for Traffic (2022). Main figures on road traffic accidents in Spain. National Road Safety Observatory, Spain.

European Commission (2023). Road safety thematic report – Motorcycles. European Road Safety Observatory, Directorate General for Transport, Brussels.

Nombela, M. & Infantes, E. (2011). Definition of a methodology to define a risk index for motorcyclists according to their exposure. Proceedings of the 22nd International Technical Conference on the Enhanced Safety of Vehicles (ESV) (Paper No. 11-0403).

World Health Organization (2022). Powered two-and three-wheeler safety: a road safety manual for decision-makers and practitioners (2nd ed.). World Health Organization, Geneva.

SUPPLEMENTARY

TABLE S1: Injury IREx by motorcycle type, age, and engine displacement

		Engine displacement (cc)			
Motorcycle type	**Age (years)**	**≤ 125**	**126-500**	**501-750**	**> 750**
Enduro / offroad	**≤ 24**	304.72	136.42	70.57	139.98
	25 - 34	19.04	31.25	85.21	12.03
	35 - 44	22.26	29.58	48.88	10.54
	> 44	12.65	7.36	28.22	18.45
Naked bike	**≤ 24**	128.33	102.09	67.08	36.21
	25 - 34	35.41	49.57	34.93	31.15
	35 - 44	11.13	14.22	29.50	34.24
	> 44	9.90	35.35	34.02	23.00
Scooter	**≤ 24**	136.36	114.75	225.45	-
	25 - 34	34.59	14.81	50.37	-
	35 - 44	23.61	23.37	13.25	-
	> 44	24.08	19.76	23.49	-
Sport	**≤ 24**	180.64	59.37	98.98	149.68
	25 - 34	40.61	16.43	47.26	51.48
	35 - 44	15.47	25.07	32.46	45.83
	> 44	9.86	10.38	24.37	31.02
Sport touring	**≤ 24**	-	108.93	89.21	-
	25 - 34	-	87.57	49.01	137.11
	35 - 44	-	153.30	29.84	39.69
	> 44	-	22.90	26.32	38.93
Super motard	**≤ 24**	600.27	-	192.46	-
	25 - 34	-	-	39.48	-
	35 - 44	-	-	43.74	-
	> 44	-	-	18.83	13.60

TABLE S2: Fatality IREx by motorcycle type, age, and engine displacement

		Engine displacement (cc)			
Motorcycle type	**Age (years)**	**≤ 125**	**126-500**	**501-750**	**> 750**
Enduro / offroad	**≤ 24**	0.00	24.80	70.57	46.66
	25 - 34	0.00	6.94	0.00	3.01
	35 - 44	0.00	3.29	0.00	0.00
	> 44	0.00	1.47	2.45	2.46
Naked bike	**≤ 24**	22.65	0.00	10.59	9.05
	25 - 34	2.72	0.00	3.58	5.19
	35 - 44	0.00	0.00	4.76	6.04
	> 44	3.96	6.43	8.19	3.63
Scooter	**≤ 24**	8.31	22.95	0.00	-
	25 - 34	3.07	4.56	4.58	-
	35 - 44	2.16	1.56	8.83	-
	> 44	1.32	2.62	3.03	-
Sport	**≤ 24**	0.00	0.00	22.84	37.42
	25 - 34	0.00	4.11	13.68	3.22
	35 - 44	5.16	0.00	8.66	12.34
	> 44	0.00	1.73	7.56	15.09
Sport touring	**≤ 24**	-	0.00	89.21	-
	25 - 34	-	21.89	16.34	0.00
	35 - 44	-	0.00	0.00	0.00
	> 44	-	0.00	4.78	5.56
Super motard	**≤ 24**	0.00	-	64.15	-
	25 - 34	-	-	0.00	-
	35 - 44	-	-	0.00	-
	> 44	-	-	3.14	13.60