

A socio-technical systems (STS) analysis of road fatalities in Gauteng province (2015–2019)

F N Masuku, M Sinclair

Road safety in Gauteng represents a persistent and complex socio-technical challenge, shaped by dynamic interactions between human factors, engineering design, institutional governance, and socio-economic contexts. This study employs a socio-technical systems (STS) framework to analyse the multifaceted contributors to road traffic fatalities in the province for 2015 to 2019; data from 2020 was excluded due to covid-19 mobility distortions that would have confounded analysis of underlying systemic risk factors. Utilising a mixed methods, convergent design, the research integrates qualitative institutional data with quantitative fatality and exposure metrics to model the interplay of social and technical subsystems. Key findings reveal that temporal variables (time of day, seasonal weather patterns), infrastructural and operational factors (road class, posted speed limits), and socio-economic pressures (unemployment rates, population density, GDP per capita) are interconnected within the STS and collectively exacerbate fatality risk. The Spatially Aware Poisson Random Forest (SAPRF) model achieved MAE ≈ 0.031 , RMSE ≈ 0.109 , and $R^2 \approx 0.958$, substantially outperforming standard random forest ($R^2 \approx 0.872$) and Poisson regression ($R^2 \approx 0.06$). The novelty lies in the first application of SAPRF in a South African road safety context, integrating STS theory with spatially aware machine learning. The study concludes that prevailing behaviour-centric interventions are insufficient and argues for a paradigm shift toward integrated, multi-level policy and design interventions targeting the root systemic couplings and feedback mechanisms inherent in the road transport system.

Keywords: socio-technical systems, road safety, Gauteng, infrastructure, modelling, fatality risk

INTRODUCTION

Road fatality rates in South Africa, and within the province of Gauteng specifically, remain at a critically elevated and unacceptable level. This constitutes a significant public health crisis and imposes severe economic burdens on communities, a challenge emblematic of many low- and middle-income countries (LMICs). Prevailing road safety interventions frequently adopt a reductionist focus, targeting individual road user behaviour in isolation. However, a more comprehensive understanding is demanded by the rapidly evolving and interconnected landscape of contemporary risks. This landscape is characterised by dynamic pressures such as climate variability, shifting demographics, evolving mobility patterns, and the disruptive integration of new technologies. Within this complex environment, fatal crash outcomes are demonstrably shaped

by the intricate interplay of these social and technical factors. Socio-technical systems (STS) theory offers a robust, holistic framework for analysing such complexity, as it explicitly examines the interdependent relationships between human actors, physical infrastructure, governing institutions, technological artefacts, and broader socio-economic contexts.

This study employs an STS lens to diagnose the systemic weaknesses contributing to persistent roadway risk in Gauteng between 2015 and 2019. Methodologically, the research advances a novel synthesis by integrating foundational STS principles with quantitative spatial and statistical modelling techniques. This integrated approach enables the systematic analysis and fusion of multidimensional data streams including social variables (e.g. population density, gender), economic indicators (e.g. GDP per capita, unemployment,

TECHNICAL PAPER

JOURNAL OF THE SOUTH AFRICAN INSTITUTE OF CIVIL ENGINEERING

ISSN 1021-2019 (print) | ISSN 2309-8775 (online)

Vol 68 No 3, September 2026, Pages 37–47, Paper 1969



DR FREEMAN NKULULEKO MASUKU holds a PhD and a Master of Engineering in Civil Engineering from Stellenbosch University, as well as an MBA from Wits Business School. He is Impact Area Manager: Transport Infrastructure Engineering at the CSIR Smart Mobility unit in Pretoria. With more than 20 years of experience in

transport engineering, infrastructure planning, and road safety, he previously served as Chief Director of Transport Policy and Planning at the Gauteng Department of Roads and Transport. His research focuses on applying machine learning to road safety analysis, particularly the influence of social, technical, environmental, and meteorological factors on fatal crash risk.

Contact details:

Department of Civil Engineering
Stellenbosch University
Private Bag X1
Matieland, 7600
South Africa
freemanmasuku@gmail.com



PROF MARION SINCLAIR is an Associate Professor in Transportation in the Department of Civil Engineering at Stellenbosch University. Her specialist field is road safety, focusing on human factors contributing to road risk, as well as on societal problems emerging from traffic crashes and injuries. She has spent the last

20+ years working in this field, shared between international road safety planning and implementation, and research and teaching of road safety problems and solutions in South Africa. In her current post she is responsible for teaching road safety to undergraduate and postgraduate civil engineering students, and for carrying out research and commercial projects.

ORCID ID: 0000-0001-9029-5911

Contact details:

Department of Civil Engineering
Stellenbosch University
Private Bag X1
Matieland, 7600
South Africa
msinclair@sun.ac.za

gross fixed capital formation), and geographical factors (e.g. road hierarchy, land use, influence of townships) into a unified predictive-analytical model.

The primary objective is to identify critical factors influencing road safety within an integrated socio-technical system, moving beyond isolated variables to model interactions and emergent effects. Three gaps motivate this study: no published work has applied a spatially aware Poisson ensemble to road fatalities in Gauteng; existing South African safety research relies on parametric methods unable to capture nonlinear STS interactions (Washington *et al* 2020); and the integration of STS theory with advanced machine learning for road safety remains absent from the South African literature. This study addresses all three by combining STS theory, spatial statistics, and ensemble machine learning in a unified framework.

LITERATURE REVIEW

According to the World Health Organization, road traffic crashes (RTCs) claim approximately 1.19 million lives globally each year (WHO 2024). This burden is starkly unequal: while high-income countries possess extensive road networks, they account for only 8% of these fatalities. In contrast, over 92% of all road deaths occur in LMICs (WHO 2024). South Africa exemplifies this disparity, reporting a disproportionately high rate of RTCs and fatalities, a finding consistent across global and national data (RTMC 2023). According to the South African Road Traffic Management Corporation’s annual publication, a total of 12 545 fatalities were recorded nationally between January and December 2021 data (RTMC 2023). Provincial data for this period reveals that Gauteng (2 293 fatalities), KwaZulu-Natal (2 143 fatalities), and the Western Cape (1 088 fatalities) were among the most severely affected regions, with Gauteng recording the highest number of deaths (see Table 1).

The evolution of technology and automation has fundamentally reconfigured the nature of human work, shifting the

primary demands from physical labour to tasks requiring sophisticated cognitive engagement. This shift necessitates an understanding of work not in isolation, but embedded within a complex matrix of technological tools, physical environments, and socio-cultural dimensions, including psychosocial climate, established work practices, procedural norms, and their evolution across both space and time (Carayon *et al* 2015). Transport, as a critical domain of human activity, has been theorised as a typical STS – a complex collection of interconnected elements such as physical artefacts, engineered infrastructure, formal regulations, specialised knowledge, supply networks, and embedded cultural meanings. Consequently, modern transport systems are intricate structures wherein technological, social, economic, political, legal, and environmental components continuously interact to produce mobility services.

The traditional conceptual model for deconstructing these systems has centred on a foundational triad: infrastructure, vehicles, and users. However, scholars like Auvinen and Tuominen (2014) convincingly argue for the inclusion of a fourth, critical element: management. This addition explicitly accounts for the essential governance, regulatory, and oversight functions that orchestrate the system. It is the dynamic, reciprocal influence among these core elements – infrastructure, vehicles, users, and management – that causes the transport system to operate as a cohesive, adaptive, and interconnected whole. A socio-technical systems approach, which meticulously examines the interdependencies among technology, individuals, organisations, and broader societal factors, therefore provides a powerful and appropriate analytical framework. It is particularly suited to interrogating persistent disparities in road safety performance across different regions, even those with ostensibly comparable technological capabilities but markedly divergent safety outcomes.

Kling (1991) provides a foundational definition of a socio-technical system as an integrated configuration of people, equipment, data, and varied resources encompassing financial, skill-based, and

status-related assets alongside documents, messages, legal arrangements, and political forces, all of which dynamically shape collective system behaviour. This comprehensive lens is indispensable for explaining the stark variations in transport safety outcomes between high-income countries and LMICs, where profound differences in institutional governance, infrastructure quality, enforcement capacity, and prevailing cultural norms create fundamentally different risk landscapes (Horton *et al* 2005).

Factors in road crash fatalities

Developing effective countermeasures for RTCs necessitates a systemic understanding of their complex, multifactorial causes. Traditional approaches often isolate individual risk elements, yet a socio-technical perspective reveals that causality emerges from the dynamic interplay between diverse social, technical, economic, and environmental subsystems. Pertinent socio-technical features influencing RTC outcomes encompass demographic variables (e.g. gender distribution, population density), economic indicators (e.g. gross domestic product (GDP), unemployment rates), vehicle fleet characteristics, roadway infrastructure (e.g. classification, condition), and contextual factors such as weather, temporal patterns (e.g. time of day, season), and enduring spatial legacies (a critical consideration in South Africa given the persistent distortions from its apartheid-era urban planning).

The multidimensional nature of crash causation is well documented in the literature. Qin *et al* (2019) demonstrated that crash frequency is shaped by interacting demographic, environmental, and institutional factors, including total land area, population density, vehicle ownership, gender and racial composition, and the density of alcohol outlets. Building directly on the analytical framework and variables identified by Qin *et al* (2019), systemic crash risk factors can be organised into the following interdependent domains:

- **Infrastructure and engineering:** Road classification, geometric design, pavement condition, and intersection density.
- **Road environment and exposure:** Weather conditions, traffic volume and

Table 1 Number of road fatalities per province 2021

Year	Eastern Cape	Free State	Gauteng	KwaZulu-Natal	Limpopo	Mpumalanga	Northern Cape	North West	Western Cape
2021	1 533	799	2 561	2 409	1 400	1 261	403	908	1271

composition, vehicle population, and fleet safety standards.

- **Socio-demographic and economic context:** Population age and gender structure, educational attainment, Human Development Index (HDI), unemployment levels, and income inequality.
- **Behavioural and institutional factors:** Alcohol availability and consumption patterns, enforcement effectiveness, and post-crash response capacity.
- **Spatio-temporal context:** Geographic location, land-use patterns, and temporal variation such as peak travel periods and weekends.

Available methods of analysis

The integration and analytical synthesis of these heterogeneous data streams present a significant opportunity. As Mehdizadeh *et al* (2020) argue, advanced data analytics applied to this socio-technical pattern hold considerable potential to decode complex crash etiologies, moving beyond description towards predictive modelling and the development of targeted, systemic safety interventions.

This predictive modelling paradigm, however, is demarcated by a fundamental methodological divergence. Classical statistical approaches, including negative binomial, Poisson, and multiple linear regression provide interpretable, parametric frameworks but are inherently limited in their ability to capture the non-linearities, high-dimensional interactions, and spatio-temporal dependencies that define contemporary transport systems as complex socio-technical entities (Washington *et al* 2020). In response, advanced computational models have emerged as robust alternatives. By leveraging machine learning (ML) and artificial intelligence (AI), these data-adaptive techniques autonomously learn intricate, non-linear relationships from multifaceted datasets without imposing restrictive statistical assumptions, yielding superior predictive performance (Yan & Shen 2022). Their capacity to process large-scale, complex data solidifies their role as indispensable tools in modern traffic safety analysis.

Within this advanced analytical repertoire, tree-based ML models, such as decision trees and their ensemble derivatives, have proven particularly effective for tasks like severity classification and strategic traffic management. These models excel at revealing latent patterns

and high-order interactions within data, facilitating the identification of previously obscured risk relationships (Vincent *et al* 2024). Consequently, the transition toward ML-based predictive analytics encompassing random forests, gradient boosting, and support vector machines has gained substantial momentum. Empirical validation of this shift is evident; for instance, integrated applications of ML models with high-resolution geospatial data have demonstrated predictive accuracy exceeding 85% in crash severity forecasting (Tejashwini *et al* 2024). This progression underscores a broader evolution toward predictive analytics capable of grappling with systemic complexity. Random forest (RF) ensembles discover non-linear interactions and rank predictors by variance-reduction importance without distributional assumptions (Yan & Shen 2022). A limitation of standard RF is that it optimises squared error, mismatched to over-dispersed count data. Replacing this with Poisson deviance aligns the learning objective with crash count distributions and improves calibration (Lord & Mannering 2010, Mathlouthi 2015). Adding spatial covariates further accounts for geographic autocorrelation without imposing a parametric spatial structure (Aguero-Valverde & Jovanis 2006). The resulting SAPRF synthesises these strengths and forms the primary analytical tool of this study.

METHODOLOGY

This study operatively implemented a mixed-method STS analytical framework to examine road safety dynamics. This approach facilitated the concurrent analysis of quantitative, objective metrics and qualitative, contextual factors, recognising their interdependence within the system. Quantitative data comprised fatality records from the Road Traffic Management Corporation for the period 2015 to 2020 and the Gauteng Department of Community Safety, integrated with socio-economic indicators, meteorological data, detailed road infrastructure inventories, and demographic statistics.

Central to the analysis was the explicit classification of all data into technical and social subsystems, a categorisation fundamental to the STS lens. This structured classification enabled a systematic investigation of the interactions between engineered artifacts and human/social elements.

Technical subsystem factors pertained to the physical and operational components of the transport system:

- **Vehicle fleet dynamics:** Population growth, spatial distribution, and composition by type (e.g. light/heavy passenger vehicles, light/heavy freight vehicles, motorcycles), including the safety implications of vehicle age distributions.
- **Road infrastructure and design:** Network hierarchy (Class 1–6), structural characteristics of intersections, and design standards influencing safety performance.
- **Traffic exposure and infrastructure condition:** Annual average daily traffic volumes as a core exposure metric and the visual condition index as an indicator of pavement quality and maintenance.
- **Spatial-physical configuration:** The geographic layout of the road network and the location of key land uses, such as townships and formal traffic analysis zones (TAZs).

Social subsystem factors encompassed human, institutional, economic, and environmental influences:

- **Demographic and behavioural vulnerabilities:** Population characteristics, including the age and gender distribution of fatalities, highlighting groups at disproportionate risk.
- **Economic context:** Indicators such as GDP per capita, unemployment rates, the Gini coefficient (inequality), and gross fixed capital formation, which shape both individual travel behaviours as well as public infrastructure investment capacity.
- **Land use and activity generators:** The spatial distribution of schools and liquor outlets, which critically influence localised traffic patterns, pedestrian activity, and alcohol-related risk.
- **Environmental conditions:** Meteorological factors (e.g. rainfall, temperature) that interact with the technical subsystem to modify road surface friction and visibility.
- **Historical-institutional legacy:** The enduring impact of past social policies, particularly apartheid-era spatial planning that created peripheral townships disconnected from economic centres, fundamentally structuring mobility patterns and exposure risks in Gauteng.

By disaggregating factors into these two interdependent subsystems, the analysis

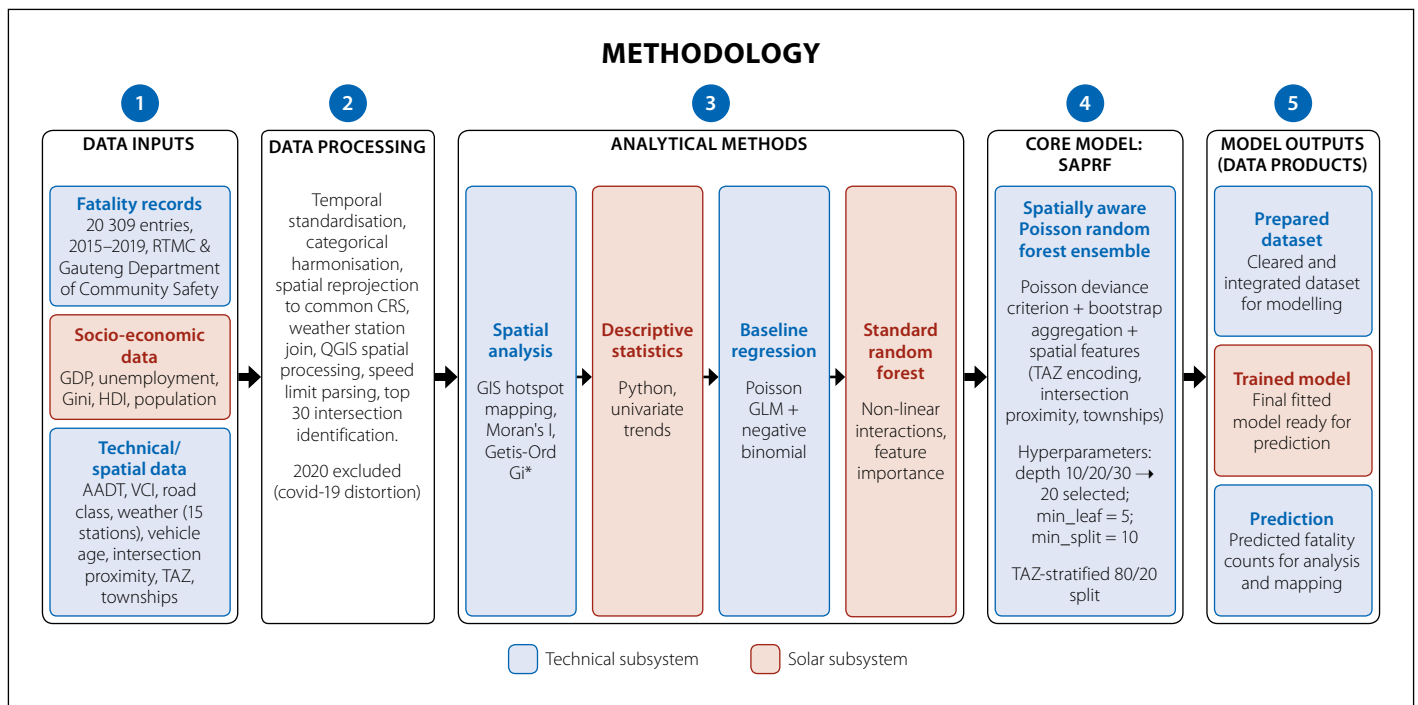


Figure 1 Integrated socio-technical methodological framework

moves beyond siloed examination, enabling a holistic investigation of the socio-technical interactions that collectively produce systemic risk.

Data collection and processing

A comprehensive data preprocessing pipeline was implemented to ensure consistency and analytical integrity across all datasets prior to modelling. The primary fatality dataset, comprising 20 309 entries, underwent temporal standardisation, with disparate date and time formats converted into a single, uniform schema. Categorical variables, including demographic descriptors, road user types, and vehicle classifications, were harmonised from varied source terminologies into common, consistent categories. To align with the 2015 to 2020 study period, historical economic data series (2010 to 2020) were extracted and formatted to cover the relevant years. Meteorological data from 15 weather stations, capturing temperature and rainfall, were standardised into decimal formats and spatially joined to the analysis units based on proximity to the nearest station. All geospatial data was reprojected to a common coordinate reference system to enable accurate spatial integration and analysis.

The pre-processed data was subsequently analysed using a sequential, multi-method analytical strategy designed to leverage both exploratory and inferential techniques:

- **Spatial analysis and hotspot identification:** Geographic information

systems (GIS) were employed to visualise the spatial distribution of fatalities and to statistically identify high-risk clusters or hotspots using spatial auto-correlation measures.

- **Descriptive statistical analysis:** Python was used to summarise key trends, distributions, and univariate relationships within the data, providing a foundational understanding of the dataset.
 - **Traditional statistical modelling:** To establish baseline relationships and account for the count nature of the outcome variable, generalised linear models, specifically Poisson and negative binomial regressions, were applied. These models helped quantify the association between predictor variables and fatality counts while controlling overdispersion.
 - **Advanced machine learning modelling:** A standard RF model was implemented to capture complex, non-linear interactions among variables and to provide an initial ranking of predictive feature importance.
 - **Integrated socio-technical modelling:** The core analysis employed the SAPRF, which synthesises the strengths of the above methods. This model integrates Poisson-based loss for count data, ensemble learning for robust prediction, and embedded spatial features to explicitly account for geographic dependencies and exposure.
- This layered methodological approach was explicitly designed to address the socio-technical interplay of risk factors, from

identifying high-risk areas and temporal trends to modelling the complex, systemic relationships that underpin fatality outcomes in Gauteng (see Figure 1).

The study focused on the province of Gauteng (see Figure 2), leveraging its diverse urban and rural road networks to provide insights applicable to similar contexts.

Spatially Aware Poisson Random Forest Ensemble (SAPRF)

Data analysis was conducted using an SAPRF. This advanced model extends the standard RF framework by implementing a Poisson deviance criterion for node splitting, making it explicitly suitable for modelling over-dispersed and zero-inflated count data such as traffic fatalities (Lord & Mannering 2010). Its spatial awareness is engineered through the direct incorporation of geospatial covariates, including proximity to intersections, TAZ target-encoded indicators, and binary township location variables to explicitly capture local heterogeneity in infrastructure and risk exposure.

This integrated methodology overcomes key limitations of conventional machine learning and regression techniques, which frequently fail to account for spatial auto-correlation and nuanced exposure effects, thereby aligning with foundational principles of spatial econometrics and safety modelling (Aguero-Valverde & Jovanis 2006). The model leverages an ensemble of decision trees trained on a comprehensive feature set encompassing traffic exposure metrics (annual average daily traffic,

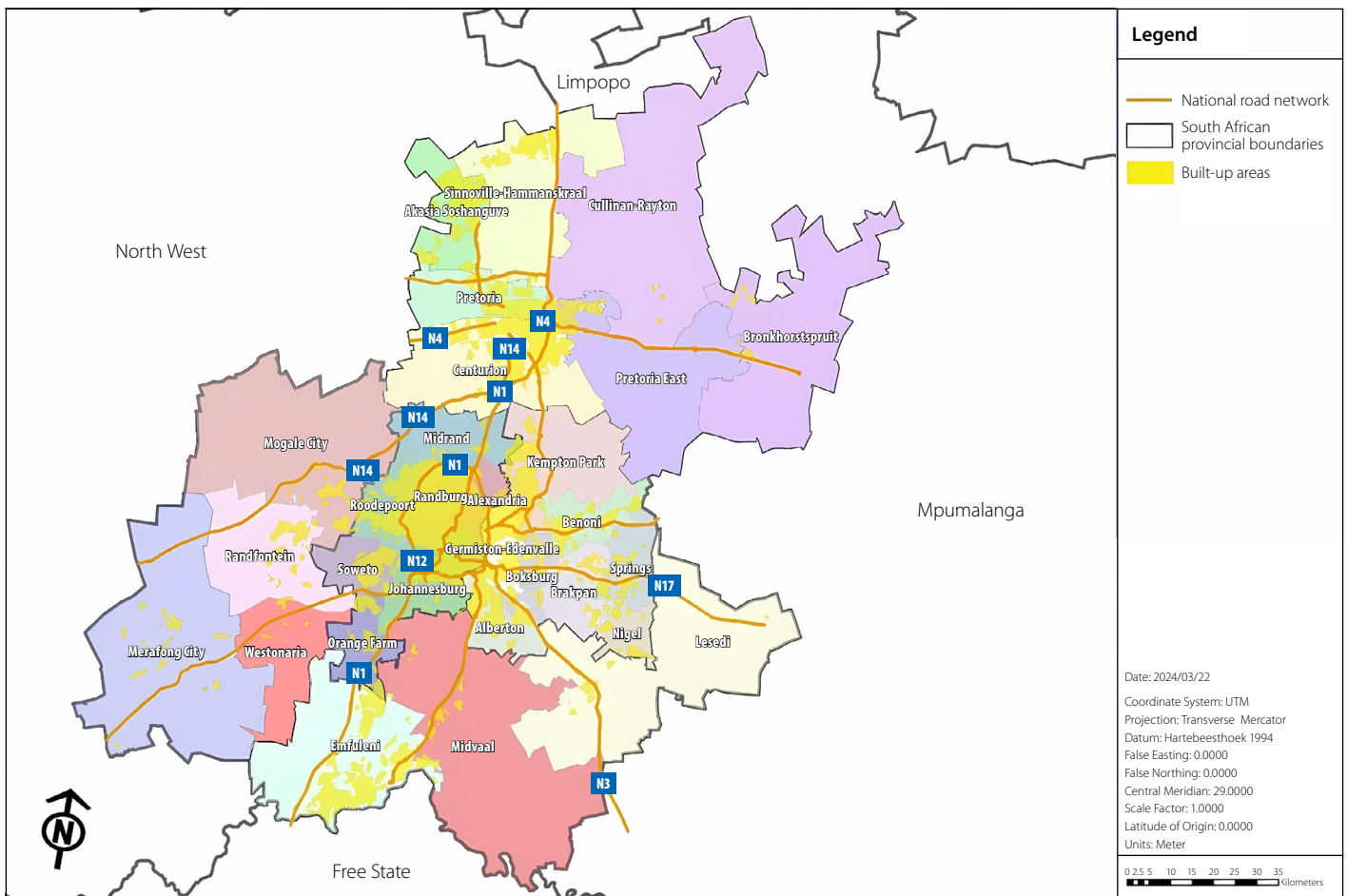


Figure 2 Map of Gauteng province

visual condition index), meteorological data, socio-economic and demographic indicators, and vehicle fleet characteristics. Crucially, exposure is explicitly parameterised to normalise predictions, countering bias from variations in traffic volume or population density (Shauna *et al* 2015).

The SAPRF was specified as an RF with Poisson deviance for node splitting, optimising for count data likelihood:

$$D(y, \hat{y}) = 2(y \log \frac{y}{\hat{y}} - (y - \hat{y})) \quad (1)$$

Where:

y is the observed fatalities

$\hat{y}$ is the predicted value, a positive real number representing the expected count from the model

$D(y, \hat{y})$ is the deviance, a measure of how well the model fits data.

The SAPRF was selected over standalone Poisson regression or standard RF models because it uniquely synthesises the Poisson framework's appropriateness for count data, the ensemble's capacity to model complex, non-linear interactions in high-dimensional data, and spatial feature integration to address location-specific risk. This makes

it an ideal analytical tool for the Gauteng context, where complex interactions between spatial, temporal, and socio-economic factors necessitate a model with both high predictive accuracy and interpretable, actionable outputs for safety planning. The training dataset used for the SAPRF was identical to that prepared for the baseline RF model.

To validate the model's predictive accuracy and generalisability beyond the data on which it was trained, a holdout validation strategy was employed. The complete dataset was randomly split into a training subset (80%) used to build the SAPRF model and a separate, unseen test subset (20%) reserved for final, unbiased evaluation. This procedure ensures that reported performance metrics reflect the model's capability to generalise to new data, a critical step for generating reliable insights for policy application.

Modelling process

The modelling process was structured into three sequential phases to ensure methodological rigor, from data preparation to model deployment and interpretation.

Phase 1: Data processing

All source datasets were cleaned, integrated, and standardised for the 2015 to

2019 study period, guided by the socio-technical framework. A key pre-processing decision was the exclusion of data from the year 2020 from the final analysis. This was necessary to avoid the severe distortions in traffic patterns, exposure, and fatality rates caused by the stringent covid-19 lockdown measures implemented in South Africa, which would have confounded the analysis of underlying systemic risk factors. Variables were explicitly selected to represent both social and technical subsystems:

- **Social variables:** Population, unemployment rate, GDP per capita, HDI, a binary township indicator, and vehicle age distribution shares.
- **Technical variables:** Traffic exposure (total vehicles), road hierarchy class, distance to the nearest intersection, meteorological conditions (temperature, rainfall), temporal effects (time of day, day of week), and TAZ identifiers for spatial encoding.

Specific transformations ensured consistency: the Gini coefficient was harmonised from a master file; annual population totals were imported from yearly Excel files; unemployment data was reshaped into a time-series format; vehicle age was converted to proportional shares (0–5, 5–10, 10–15, >20 years); weather data

was aggregated to daily means; and HDI values were manually compiled. Geospatial processing in QGIS derived key features, including intersection proximity and binary township flags. Speed limits were extracted from descriptive strings using a custom parsing function. Fatality locations were spatially joined to the road network, and the top 30 high-fatality intersections were identified for subsequent analysis.

Phase 2: Model training and validation

The pre-processed dataset was partitioned into an 80% training set (see Figure 2) and a 20% validation set, stratified by a TAZ identifier to prevent spatial leakage. A sensitivity analysis confirmed less than 3% RMSE difference versus a random split, validating this approach. Hyperparameters were tuned via grid search cross-validation: maximum tree depth evaluated at 10, 20, and 30 (depth 20 selected); minimum samples per leaf = 5; minimum samples per split = 10. The SAPRF was implemented using Breiman's ensemble framework employing bootstrap aggregation and random feature selection. Feature scaling was applied to continuous variables. A log (+1) transformation was applied to the target variable for exploratory visualisation only; the SAPRF model operates on raw count data, with the Poisson deviance criterion handling the skewed, zero-inflated distribution. There is therefore no conceptual inconsistency between this preprocessing step and the Poisson learning objective.

RESULTS

Annual fatality trends were analysed by aggregating individual data points by year. Figure 3 illustrates the resulting annual fatality counts for the study period. The year 2019 recorded the highest number of fatalities, followed sequentially by 2016, 2017, 2018, 2015, and finally 2020. The trend shows an increase from 2015 to 2016, a slight decrease in 2017, relative stability in 2018, a sharp peak in 2019, and a significant decline in 2020.

This pronounced drop in 2020 is directly attributable to the onset of the covid-19 pandemic. The first case was confirmed in South Africa in March 2020, after which varying levels of government-mandated travel restrictions were implemented. These restrictions drastically reduced traffic exposure, leading to the observed decrease in fatalities. Consequently, the core socio-technical systems analysis for this study focuses on the pre-pandemic period (2015

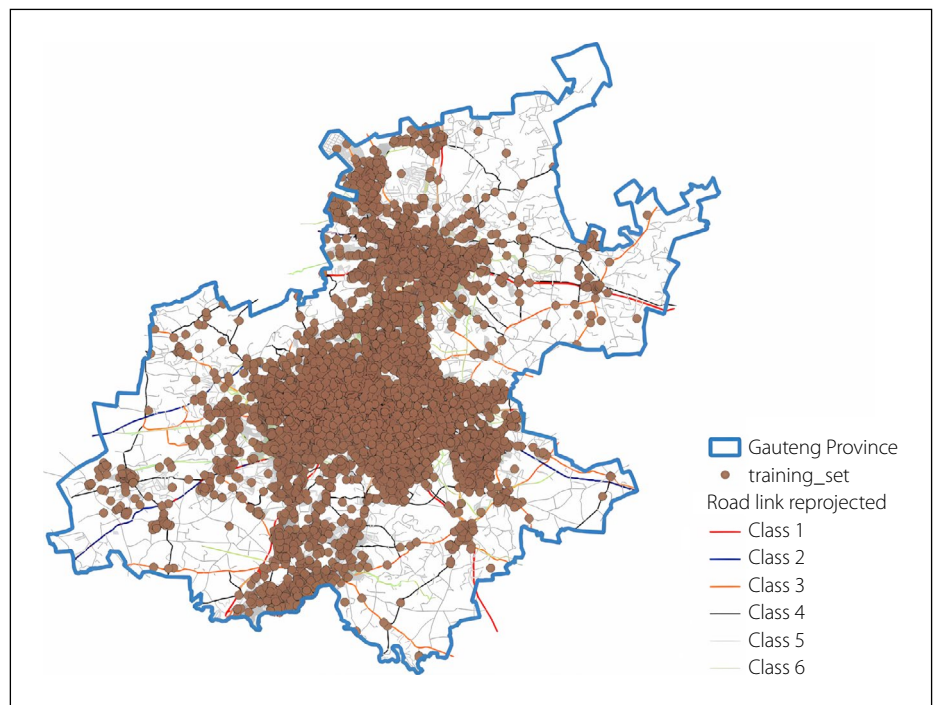


Figure 3 Map of SAPRF fatalities training set

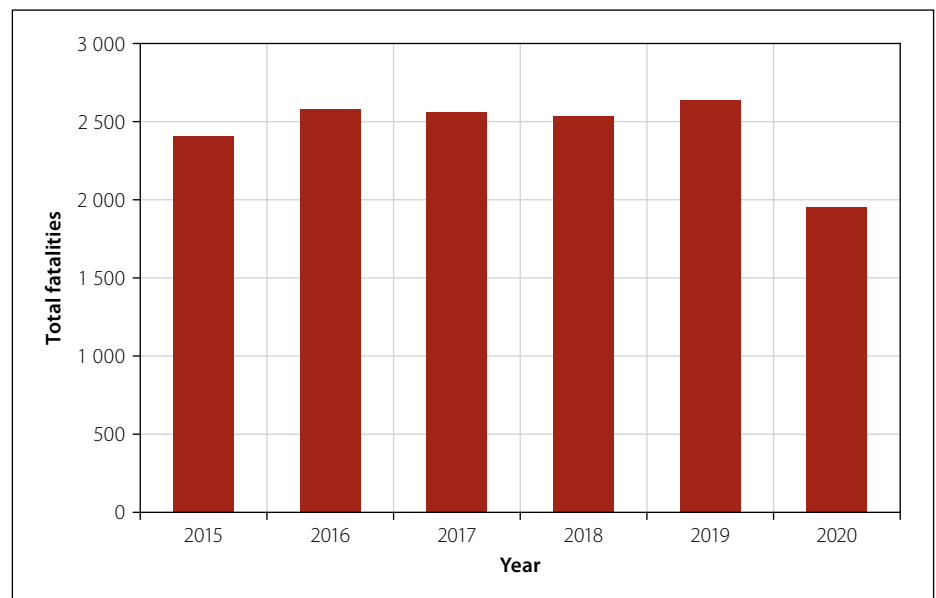


Figure 4 Total fatalities per month 2015 to 2020

to 2019) to examine typical systemic risk factors. The anomalous patterns of 2020 are treated separately in a subsequent temporal analysis of monthly and weekly fatality data for that year, which contextualises the impact of the lockdown phases.

Analysis of the pre-pandemic period (2015 to 2019) reveals a distinct seasonal trend in road fatalities. The data indicates that mid-year and late-year months were associated with elevated risk, while the early months consistently recorded lower proportions of annual fatalities. Specifically, July accounted for the highest proportion of fatalities at 10.1%, followed by August (9.6%), October (9.3%), and December (9%). Conversely, January recorded the lowest

proportion (5.5%), with February (7%), March (7.8%), and April (7.8%) also representing periods of relatively lower incidence.

This seasonal pattern suggests that road safety risk is influenced by temporal factors beyond random distribution. The peaks in July, August, and December may correlate with periods of increased long-distance travel, such as holiday migrations and festive season journeys, combined with potentially shorter daylight during winter weather conditions in the mid-year months. The consistent trough in early-year fatalities may reflect reduced travel volumes after post-holiday periods. The annual breakdown of monthly fatalities is presented in Figure 5.

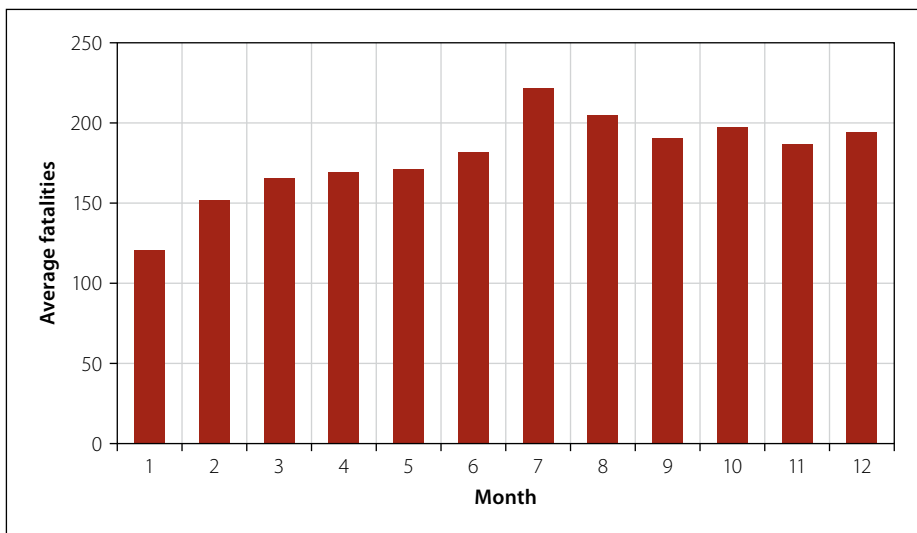


Figure 5 Average monthly fatalities (2015 to 2019)

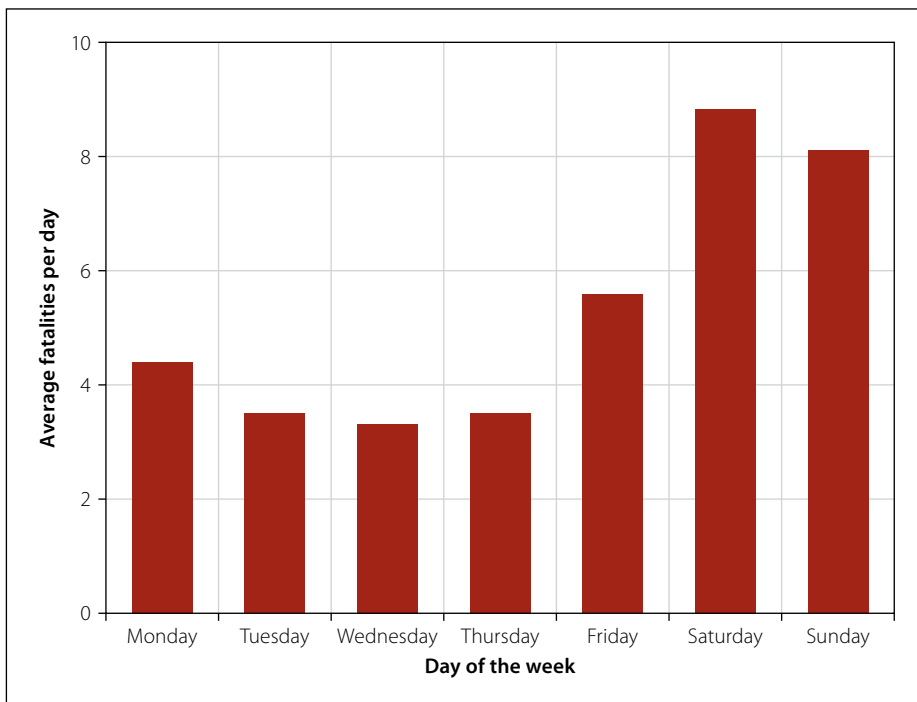


Figure 6 Average daily fatalities by day of week (2015 to 2019)

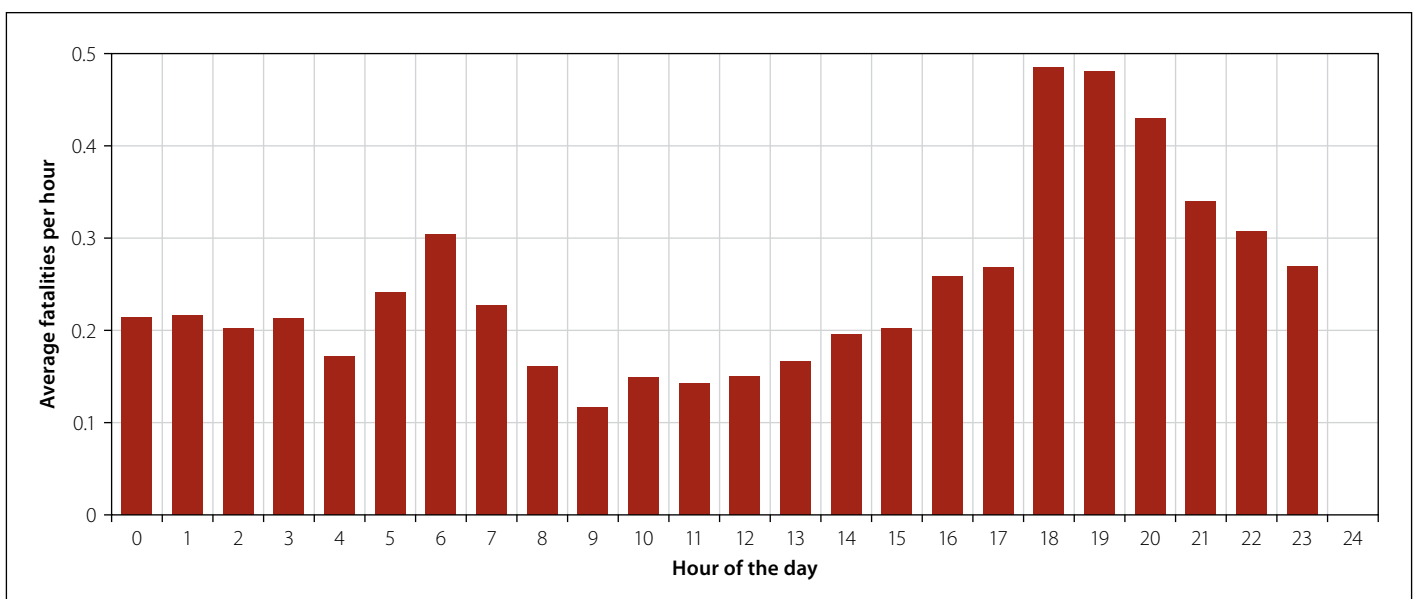


Figure 7 Fatalities per hour of the day (2015 to 2019)

An analysis of daily fatality distribution for the 2015 to 2019 period reveals a pronounced concentration of fatalities on weekends. This is a common phenomenon worldwide, so it is not unexpected, however the scale of concentration is unusually high. Saturdays and Sundays collectively accounted for 45.5% of all fatalities, with Saturday representing the peak at 23.7%, closely followed by Sunday at 21.8%. Weekdays demonstrated significantly lower proportions, with Thursday (15%) being the highest among them. The lowest fatality rates were recorded on Tuesday (8.9%), while Monday and Wednesday shared identical proportions (9.4% each). Friday occupied an intermediate position at 11.8%. This pattern underscores the heightened risk associated with recreational travel, differing traffic compositions during weekend periods, and the potential impact of increased alcohol consumption on weekends. The annual breakdown of fatalities by day of the week is detailed in Figure 6.

The temporal analysis reveals a robust diurnal pattern in fatality occurrence, characterised by two pronounced peak periods. The morning peak, defined as 05:00 to 08:00, accounted for 15.8% of total fatalities from 2015 to 2019. A significantly longer afternoon peak, spanning 17:00 to 21:00, was responsible for 35.1% of fatalities during the same period. This pattern demonstrated notable consistency across individual years, with the morning peak proportion ranging from 13.4% to 16.4% and the afternoon peak varying between 31.4% and 35.7%.

These distributions, visualised in Figure 7, point to a systemic relationship between commuting traffic flows, visibility conditions,

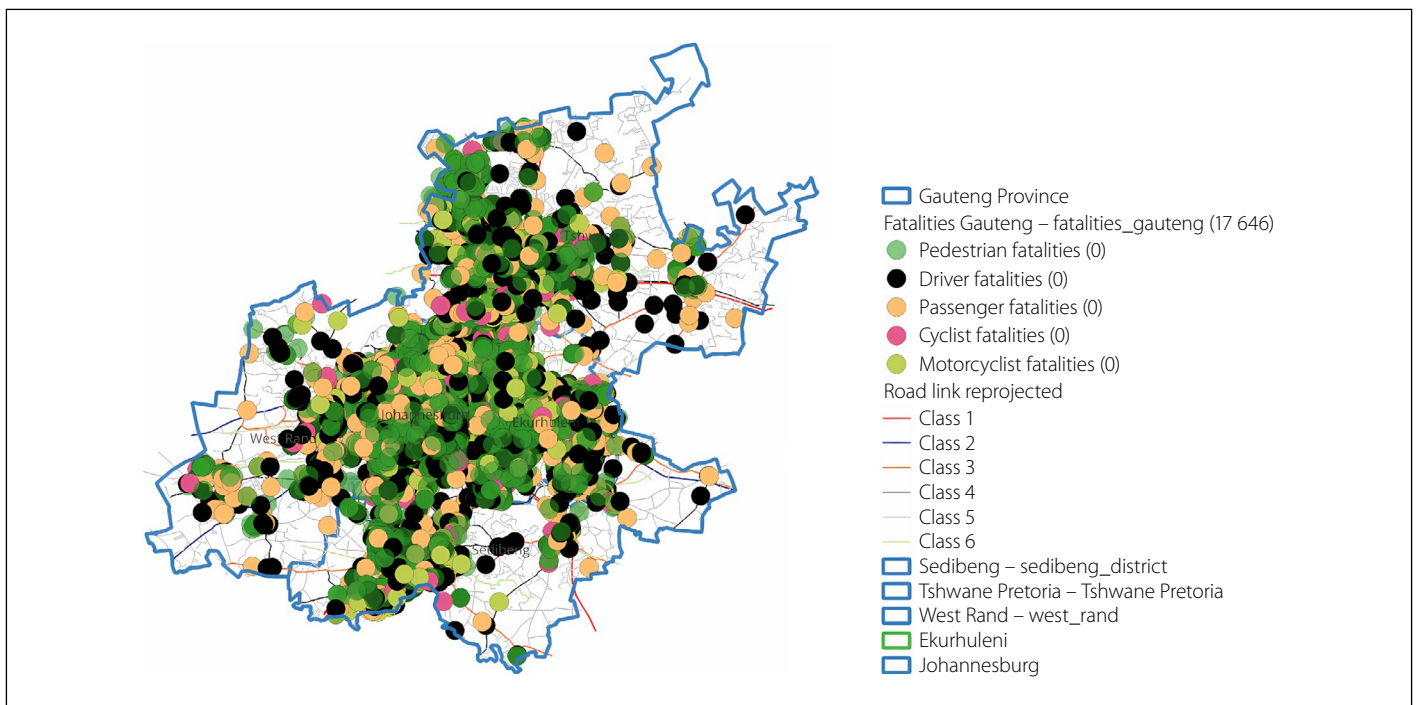


Figure 8 Map of road fatalities in Gauteng (2015 to 2020)

and driver fatigue with elevated fatality risk. The predominance of the extended afternoon peak, encompassing the evening commute and early nighttime, suggests a compounding of risk factors including higher traffic volumes, potential driver impairment, and reduced visibility. The consistency of this pattern underscores the necessity of temporally targeted interventions, such as dynamic speed enforcement, enhanced road lighting, and public awareness campaigns during these high-risk windows, to disrupt the systemic conditions that propagate fatalities.

Spatially aggregating to the district municipality scale and using the conventional per 100 000 metric reveals a distinct geographic pattern. The West Rand District (29.39) and Sedibeng District (27.88) exhibited the highest regional fatality rates, markedly exceeding the rates for Tshwane (18), Johannesburg (15.53), and Ekurhuleni (15.11). This spatial disparity highlights pronounced inequities in road safety outcomes across Gauteng, suggesting that risk is not uniformly distributed but concentrated in specific regional systems, potentially reflecting underlying variations in infrastructure quality, enforcement regimes, or socio-economic vulnerabilities.

Analysis of Gauteng road fatalities (2015 to 2019) reveals a persistent and inequitable crisis defined by specific, high-risk demographics and temporal patterns. The burden is distributed unevenly across road user groups:

- Pedestrians bear the greatest toll, constituting 46% to 49% of all fatalities.

This crisis is most acute for children (aged 0 to 14), among whom 86.3% of fatalities are pedestrians.

- Drivers account for the second-largest share (23% to 25%), with risk heavily concentrated during late weekend nights.
- Passengers represent 20% to 24% of fatalities, disproportionately affecting young adults (aged 15 to 19).
- An analysis of modal share against fatalities data reveals a critical overrepresentation of vulnerable road users in fatal collisions. Despite minimal mode shares (cyclists 0.2% to 0.3%; motorcyclists 0.1%), these groups constitute 1.9% and 4.3% of fatalities respectively.

Fatality risk peaks predictably during weekend evenings and nights across all major user groups. This evidence confirms that Gauteng's road safety challenge is not uniform but a series of interconnected, population-specific epidemics. Achieving meaningful reductions requires moving beyond generic measures to precision interventions that directly target the most vulnerable users, highest-risk times, and most dangerous locations.

The descriptive spatio-temporal and socio-technical patterns identified above – including pronounced peaks during evenings and weekends, pedestrian dominance, intersection and arterial hotspots, and municipal disparities – provide a foundational diagnosis of systemic risk factors in Gauteng. To quantify these interactions, capture non-linear relationships, and generate interpretable predictors for engineering and policy

applications, a predictive modelling phase was undertaken. This employed the SAPRF model, which integrates Poisson deviance-based learning with ensemble methods and explicit spatial encoding. The following sections present the model's performance evaluation against benchmarks (classical Poisson/negative binomial regression and standard RF) and detailed feature importance analysis derived exclusively from the trained SAPRF. These results directly build on the descriptive findings by revealing the relative influence and interactions of infrastructure, exposure, temporal, and contextual variables in driving fatality outcomes.

Performance of the SAPRF model

The SAPRF model demonstrated superior overall performance and offered the most theoretically coherent treatment of the data. While classical Poisson regression provided interpretable incidence rate ratios, its restrictive log-linear assumption failed to capture the complex non-linearities and interactions inherent in the crash system, resulting in a poor fit ($MAE \approx 0.21$, $RMSE \approx 0.51$, $R^2 \approx 0.06$). A standard RF model substantially improved predictive accuracy by automatically learning complex patterns ($MAE \approx 0.043$, $RMSE \approx 0.189$, $R^2 \approx 0.872$), as it is not constrained by parametric assumptions.

The SAPRF synthesises the strengths of both approaches. It retains the ensemble's flexibility to model high-order interactions (e.g. between distance-to-intersection, time-of-day, and rainfall) but uses Poisson deviance as its splitting criterion. This

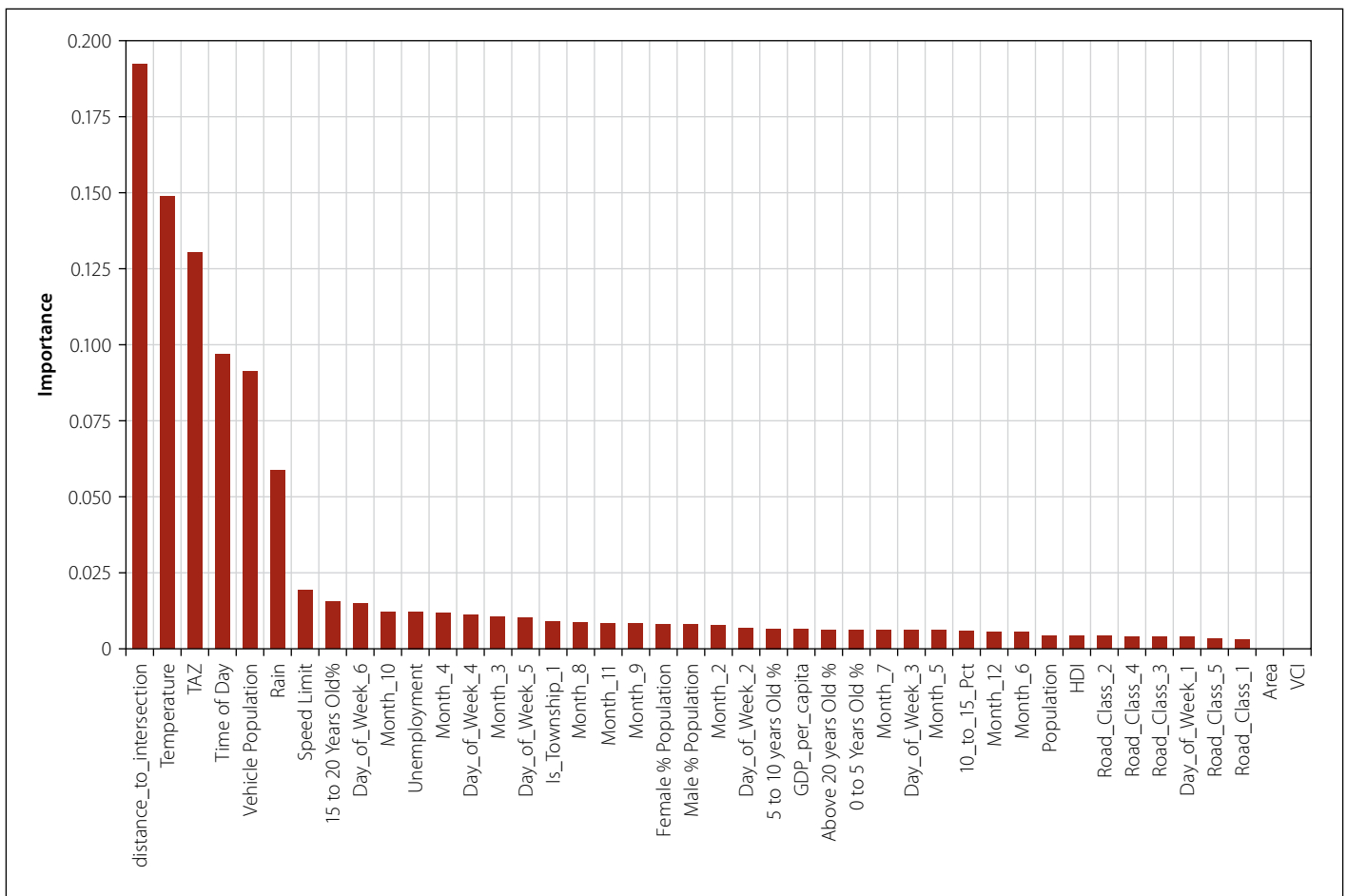


Figure 9 Feature importance

fundamental alignment of the learning objective with the underlying count data process ensures strictly positive predictions and superior statistical calibration. Consequently, the SAPRF achieved the optimal balance of flexibility and appropriateness, delivering the best model fit (MAE $\approx$ 0.031, RMSE $\approx$ 0.109, $R^2 \approx$ 0.958) for the task. Moran's I applied to the spatial residuals ($I = 0.2684$; $p = 0.001$) confirms that residual clustering is geographically structured, driven in part by apartheid-era spatial planning legacies. Larger residuals occur at isolated high-fatality TAZs, consistent with the difficulty of predicting rare extreme-count events.

The feature importance results from the SAPRF model articulate a coherent socio-technical narrative of risk. Proximity to an intersection emerged as the dominant predictor, followed by temperature which serves as a proxy for winter and shorter daylight hours and the spatial baseline encoded by TAZ. Substantial contributions were also made by time of day, vehicle volumes, rainfall, and posted speed limits. Secondary, yet meaningful, signals arose from vehicle fleet age composition, socio-economic indicators, and calendar effects (month and day of the week). Critically, the model captured the nonlinearities and

interactions among these features, revealing where and when fatalities concentrate. This provides a robust, evidence-based foundation for designing spatially and temporally targeted interventions.

A detailed breakdown of the key model features and their interpreted roles is as follows:

- **Distance to intersection (0.1923):** The strongest individual predictor. Risk increases sharply within a short distance of intersections, consistent with their role as complex conflict points for turning, crossing, and merging manoeuvres.
- **Temperature (0.1487):** A very strong predictor, likely acting as a proxy for light conditions as winters are associated with short daylight conditions.
- **TAZ encoding (0.1304):** Captures stable spatial baselines related to unobserved local factors such as land-use mix and enforcement culture.
- **Time of day (0.0971):** Confirms a clear diurnal pattern, with peak risk during night and late evening hours.
- **Vehicle population (0.0914):** Represents a localised exposure signal, indicating that higher vehicle densities increase the opportunity for severe conflicts.

- **Rainfall (0.0588):** Exhibits a hump-shaped relationship, where initial rain increases risk due to friction loss, while heavy rain suppresses activity and exposure.
- **Day of week:** Weekend days (Day 6 – Saturday: 0.0153; Day 7 – Sunday implied) and Friday (Day 5: 0.0104) show elevated importance, aligning with alcohol-related and leisure travel patterns.
- **Speed limit (VOPRT) (0.0195):** A secondary but notable factor, with its effect often interacting with road class and intersection proximity.
- **Socio-economic and demographic indicators:** Variables such as unemployment (0.0124), gender population shares ($\sim$ 0.008), and GDP per capita (0.0066) provided consistent but modest signals, suggesting they contextualise risk rather than drive it directly after controlling for core engineering and exposure factors.
- **Vehicle fleet age:** The share of vehicles aged 15 to 20 years (0.0157) was the strongest fleet-age indicator, likely proxying for a cohort with outdated crashworthiness. Newer vehicle shares (0 to 5 years: 0.0065) were associated with lower risk, reflecting modern safety standards.

■ **Spatial and institutional context:** 'Is_Township' (0.0093) captures place-based factors (pedestrian activity, lighting, informal crossings and network form) typical of township environments. After accounting for this spatial context, broader controls such as road class and total population contribute relatively little, indicating that risk is driven less by nominal hierarchy and more by the interaction of infrastructure, land use, and mobility patterns in specific places. This supports place-sensitive design standards, pedestrian-priority treatments, and differentiated investment for township and peri-urban contexts. This hierarchy confirms that, while fundamental engineering and exposure variables form the backbone of risk, their impact is powerfully activated and modulated by temporal, behavioural, and contextual social factors – a finding central to the socio-technical systems analysis.

DISCUSSION

The findings of this study demonstrate that road safety in Gauteng is not merely a problem of individual driver error but a symptom of systemic failure. Fatality risk emerges from the complex interaction of technical vulnerabilities, social-temporal patterns, and entrenched socio-spatial inequities. While conventional interventions often target road user behaviour in isolation, this analysis reveals that such campaigns are addressing symptoms within a flawed system. Sustainable reductions in fatalities require moving beyond isolated measures to implement holistic reforms that directly address the interconnected structural weaknesses identified.

The STS framework, operationalised through spatial and statistical modelling, provides a diagnostic map of these systemic faults. The dominant influence of intersection proximity underscores a critical technical failure in conflict-point design, a flaw that is substantially amplified during evening and weekend periods of heightened social activity. This risk is not evenly distributed but is geographically concentrated, as shown by the strong predictive power of TAZ encoding and township indicators, revealing landscapes where historical planning legacies compound vulnerability. Furthermore, the significant signal from an aging vehicle fleet points to a parallel institutional-economic failure in safety regulations. These elements do not operate

in isolation; they interact dynamically. In township environments, fatality risk emerges from the convergence of multiple interacting factors. A substandard intersection, when combined with high pedestrian activity, weekend evening travel, limited lighting, an ageing vehicle fleet, and alcohol consumption, create conditions in which crash likelihood and severity are jointly elevated. These interactions produce predictable high-risk scenarios that cannot be explained by any single factor in isolation.

Consequently, the interaction between infrastructure degradation and weak enforcement capacity creates a permissive environment where risk is systematised, particularly for vulnerable road users. This analysis confirms that targeted behaviour-change campaigns, while potentially beneficial, are ultimately insufficient against such deeply embedded structural forces. The path forward requires a paradigm shift towards integrated, system-level interventions. This entails not only engineering safer intersections and roads but also synchronising enforcement with high-risk temporal windows, accelerating vehicle fleet renewal through economic and regulatory mechanisms, and prioritising infrastructure upgrades in high-vulnerability zones. Only by treating road safety as a complex socio-technical system can Gauteng disrupt the reinforcing patterns of risk and achieve a sustainable reduction in fatalities. These findings are consistent with the RTMC (2023), which similarly identifies infrastructure deficits and socio-spatial inequities as primary contributors to the national fatality burden. The SAPRF's superior performance over standard regression ($R^2 = 0.958$ versus 0.06) reinforces Washington *et al*'s (2020) assertion that conventional parametric models cannot capture spatio-temporal dependencies in complex transport systems – supporting precision, system-level interventions over generic behavioural campaigns.

CONCLUSIONS AND RECOMMENDATIONS

Building on the system-level interpretation developed in the preceding discussion, this study concludes that road traffic fatalities in Gauteng are emergent outcomes of interacting socio-technical factors rather than isolated behavioural failures. By integrating spatial statistics with RF and SAPRF models, the analysis demonstrates that fatality risk is produced through nonlinear

interactions between infrastructure design, temporal exposure, vehicle fleet characteristics, and socio-spatial context.

While descriptive spatial analysis identifies where and when fatalities cluster, the RF and SAPRF results explain how these clusters arise. In particular, SAPRF shows that intersection proximity, weekend evening travel, adverse conditions, and older vehicles combine to create predictable high-risk system states, especially within township and peri-urban environments. These findings indicate that nominal road class or traffic volume alone are insufficient predictors of risk once place-based context is accounted for, underscoring the need for system-level approaches to road safety.

Consistent with the STS framing adopted in this study, effective road safety interventions in Gauteng should be designed to recalibrate interactions between technical infrastructure, road users, institutional arrangements, and the broader socio-spatial environment, rather than addressing individual components in isolation. The following recommendations target the key system interactions identified through the spatial and modelling analyses:

1. Socio-technical, context-sensitive infrastructure design

Prioritise system redesign of high-risk intersections and arterial corridors, particularly within township and peri-urban interface zones. Engineering interventions should be co-designed with consideration of pedestrian behaviour, land-use context, lighting conditions, and typical vehicle characteristics, ensuring that infrastructure is forgiving of predictable human error.

2. Adaptive operational and enforcement strategies aligned with system states

Align traffic enforcement, operations, and public awareness initiatives with empirically identified high-risk temporal and spatial system states, such as weekend evenings and adverse weather conditions. This integrates social behaviour, enforcement capacity, and technical controls into a coordinated response.

3. Interventions at the human-technology interface

Address vehicle fleet safety as a socio-technical issue by promoting fleet modernisation, particularly within public and informal transport sectors. Regulatory and economic instruments should be used to

improve the safety performance of vehicles operating in high-exposure environments, reducing injury severity regardless of driver compliance.

4. Institutional integration and data-driven governance

Embed GIS-based hotspot analysis and predictive modelling tools, such as SAPRE, within routine provincial and municipal road safety governance. This supports coordinated decision making across transport engineering, law enforcement, urban planning, and public health institutions, strengthening the system's capacity to respond adaptively to emerging risk patterns.

LIMITATIONS

While this study provides a system-level diagnosis of road safety risk in Gauteng, several limitations should be acknowledged. The analysis relies on secondary institutional datasets, which may not fully capture informal operational practices, enforcement intensity, or inter-agency coordination dynamics. Although the modelling framework identifies where and when risk concentrates and which factors interact most strongly, it does not establish definitive causal mechanisms for individual crash events. In addition, the exclusion of 2020 data was necessary to avoid covid-19-related distortions, but limits inference to pre-pandemic system behaviour. Finally, while the spatial encodings used are context-specific to Gauteng, the socio-technical systems framework and analytical approach are transferable to other jurisdictions with comparable data availability.

ACKNOWLEDGEMENTS

The authors acknowledge the Road Traffic Management Corporation and the Gauteng Department of Community Safety for providing access to road traffic fatality and safety-related data used in this study. The authors further acknowledge the South African Weather Service for the provision of meteorological data, the Council for Scientific and Industrial Research for technical and analytical support, and the Gauteng Department of Economic Development for access to relevant socio-economic indicators. Any interpretations, analyses, or conclusions presented in this paper are those of the authors and do not necessarily reflect the views or positions of the data-providing institutions.

REFERENCES

- Aguero-Valverde, J & Jovanis, P P. 2006. Spatial analysis of fatal injury crashes in Pennsylvania. *Accident Analysis & Prevention*, 38(3): 618–625.
- Auvinen, H & Tuominen, A. 2014. Future transport systems: long-term visions and socio-technical transitions. *European Transport Research Review*, 6(3): 343–354.
- Carayon, P, Hancock, P, Leveson, N, Noy, I, Sznclwar, L & Van Hootegem, G, 2015. Advancing a sociotechnical systems approach to workplace safety—developing the conceptual framework. *Ergonomics*, 58(4): 548–564.
- Horton, K, Davenport, E & Wood-Harper, T. 2005. Exploring sociotechnical interaction with Rob Kling: five “big” ideas. *Information Technology & People*, 18(1): 50–67.
- Kling, R. 1991. Computerization and social transformations. *Science, Technology, & Human Values*, 16(3): 324–367.
- Lord, D & Mannering, F. 2010. The statistical analysis of crash-frequency data: A review and assessment

- of methodological alternatives. *Transportation research part A: policy and practice*, 44(5): 291–305.
- Mathlouthi, W. 2015. *Trees and random forests for nonhomogeneous count data processes*. HEC Montreal, Canada.
- Mehdizadeh, A, Cai, M, Hu, Q, Alamdar Yazdi, M A, Mohabbati-Kalejahi, N, Vinel, A, Rigdon S E, Davis, K C & Megahed, F M. 2020. A review of data analytic applications in road traffic safety. Part 1: Descriptive and predictive modeling. *Sensors*, 20(4): 1107.
- Qin, X, Shaon, M R R, Schneider, R J, Noyce, D A, Bill, A, Farid, Y Z & Lord, D. 2019. *Identifying highly correlated variables relating to the potential causes of reportable Wisconsin traffic crashes*. Madison: Wisconsin Department of Transportation.
- RTMC. 2023. *State of Road Safety in South Africa, January 2022 to December 2022*. Centurion: Road Traffic Management Corporation.
- Shauna, L, Hallmark, Y Q, Hawkins, N & Smadi, O. 2015. Crash modification factors for dynamic speed feedback signs on rural curves. *Journal of Transportation Technologies*, 5(1).
- Tejashwini, A G, Singh, C K, Sinha, C, Karnani, D & Pandey, A. 2024. Road accident severity prediction using random forest algorithm. *International Journal for Research in Applied Science and Engineering Technology*, 12(4): 1205–1210.
- Vicent, J F, Curado, M, Oliver, J L & Pérez-Sala, L. 2024. A novel approach to predict the traffic accident assistance based on deep learning. *Neural Computing and Applications*, 37(7): 5343–5368.
- Washington, S, Karlaftis, M G, Mannering, F & Anastasopoulos, P. 2020. *Statistical and econometric methods for transportation data analysis*. New York: CRC Press.
- WHO. 2024. Alcohol. World Health Organisation. Available at: <https://www.who.int/news-room/fact-sheets/detail/alcohol>.
- Yan, M & Shen, Y, 2022. Traffic accident severity prediction based on random forest. *Sustainability*, 14(3): 1729.