Evaluating Segmentation and Temporal Resolution in Highway Crash Forecasting: The BR-040 Corridor

João P. Tozetto[1], Isadora Nunes[1], Bernardo Soares[1], Elisa T. Albergaria[1], Diego R. C. Dias[2(✉)], Vinícius F. S. Mota[2], and Leonardo C. D. Rocha[1]

[1] Universidade Federal de São João del Rei (UFSJ), São João del Rei, Brazil
{joaopedrocunhaolx,isamiranda222017,bernardosfferraz}@aluno.ufsj.edu.br, {etuler,lcrocha}@ufsj.edu.br

[2] Universidade Federal do Espírito Santo (UFES), Vitória, Brazil
{diego.dias,vinicius.f.mota}@ufes.br

Abstract. Spatiotemporal predictive models are key tools for anticipating highway crash volume and, consequently, for supporting road safety actions and operational planning, ranging from allocating enforcement resources to prioritizing infrastructure interventions. In this paper, we present a detailed evaluation of spatiotemporal models for forecasting crash counts along one of Brazil's most important highways, the BR-040. For this study, we used open data from the Brazilian Federal Highway Police for the period 2017 to 2024. An important contribution of the paper is that, instead of partitioning highways into fixed-length segments, we applied a spatial segmentation based on the socioeconomic characteristics of cities along the BR-040, which enabled analysis of meaningful territorial heterogeneity. We considered three models from different algorithmic families in our evaluation: SARIMAX, Random Forest, and Extra Trees Regression. We evaluated them in four experimental scenarios that combine two temporal granularities (monthly and weekly) with two attribute configurations (with and without holiday and non-working-day variables). The effectiveness of the models was evaluated using three accuracy metrics (R^2, MAE and RMSE), while computational efficiency was measured through total execution time. Our results show a robust trade-off between effectiveness and computational cost: while the SARIMAX model minimizes MAE/RMSE, it incurs the highest computational cost, especially at weekly granularity. On the other hand, tree-based ensemble methods incur the lowest runtimes but do not achieve the best results in all evaluated scenarios. We also evaluated the impact of segmentation strategies for specific road sections and different temporal granularities, and our results provide strong evidence for monitoring and prioritizing critical segments at specific times on Brazilian federal highways.

Keywords: Spatiotemporal prediction models · spatial segmentation by socioeconomic conditions · machine learning · road accident prediction

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 584–600, 2027.
https://doi.org/10.1007/978-3-032-30488-9_36

1 Introduction

In Brazil, the high number of road accidents has significantly impacted public health, overburdening emergency services and resulting in high economic costs, as well as many deaths and permanent injuries. Recent studies indicate that these road accidents represent approximately 3% of the Gross Domestic Product (GDP) in most countries. [17]. One strategy to improve this situation is to develop decision-support tools, especially spatiotemporal prediction models, leveraging the widespread availability of open data on highway incidents. In Brazil, the agency responsible for monitoring federal highways is the Federal Highway Police (PRF), which maintains large public databases with accident records that can be used for empirical studies. However, transforming raw data into useful information to support decision-making requires several methodological steps, such as defining spatial and temporal granularity, addressing biases, including variations in traffic exposure, and adopting validation protocols compatible with the operational context.

In this article, we present an evaluation of spatiotemporal prediction models for predicting accident occurrences on BR-040, one of the longest and most important highways in Brazil. Prior studies using PRF data and/or Brazilian roadway subsets demonstrate the potential of machine learning (ML) techniques for hotspot analysis and risk pattern discovery, while also highlighting sensitivity to choices in preprocessing, representation, and spatial partitioning [2]. In this context, rather than partitioning highways into fixed-length segments, we propose and apply a spatial segmentation of cities along the BR-040 based on their socioeconomic characteristics, thereby enabling analysis of meaningful territorial heterogeneity. Based on this partitioning, we organize segment-level time series and investigate how different temporal granularities and explanatory characteristics affect performance and computational cost, with an emphasis on configurations that are feasible for recurrent monitoring.

Recent advances in neural architectures can capture spatial and temporal dependencies, including recurrent networks with graph convolutions and spatiotemporal convolution variants [11,19]. In this context, we present a methodological experimental evaluation comparing three model families commonly used for applied forecasting and regression: a statistical time-series model with autoregressive and moving-average components and support for exogenous regressors (SARIMAX) and two tree-based ensemble methods (Random Forest and Extra Trees Regression). To evaluate the impact of temporal granularity and the presence of holidays on the predictive performance of these models, we consider two temporal granularities (monthly and weekly) and two attribute configurations (with and without holiday and non-working-day variables), yielding four scenarios. We consider complementary metrics of predictive effectiveness: the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). For efficiency, we measure the execution time. We used all these metrics to evaluate the trade-off between effectiveness and efficiency across different scenarios.

In our results, we observe that SARIMAX achieves the lowest MAE and RMSE across scenarios, but at a higher runtime, particularly under weekly granularity, where the cost grows to the order of hundreds of seconds as the number of segments increases. On the other hand, the ensemble methods (Random Forest and Extra Trees Regression) offer lower runtime while maintaining competitive effectiveness in some scenarios, making them the best trade-off between effectiveness and efficiency for large-scale execution and frequent updates. Moreover, including calendar variables (i.e., description of whether a specific date is a holiday or not) tends to improve model fit, especially under monthly granularity, reinforcing the relevance of holiday and non-working-day effects in explaining part of the observed variability.

Therefore, the main contributions of our paper are: (i) the integration and enrichment of PRF occurrence data with socioeconomic information from cities along BR-040, allowing the highway partitioning in spatial segmentation better aligned with the territorial economic activities; (ii) a complete experimental framework that allows comparative analysis of spatiotemporal predictive models under different scenarios (temporal granularities and characteristics configurations) and perspectivities (effectiveness and efficiency); and (iii) an empirical discussion of how segmentation and granularity choices affect performance, stability, and operational trade-off between effectiveness and efficiency, with implications for intervention prioritization and risk monitoring.

2 Related Work

We observed two main approaches in the literature regarding solutions for reducing traffic accidents. The first is the spatiotemporal prediction models themselves, which aim to anticipate the risk and/or the expected number of occurrences. Another complementary approach concerns spatial representation, that is, strategies for defining spatial units (segmentation) of analysis so that predictions can be translated into planning and intervention. The main challenges lie in spatial heterogeneity, which reflects structural differences across highway segments (e.g., pavement quality, dominant economic activity per section, etc.), and in temporal dependencies, such as seasonality, routines, and long-term trends. Furthermore, maintaining interpretable results that align with management units, such as administrative boundaries and operational coverage areas, should also be considered.

Regarding modeling, spatiotemporal prediction strategies that use deep learning architectures encoding space as grids or latent structures and time as sequences have shown the best effectiveness. In [20], the authors present ConvLSTM, which has been used to incorporate spatiotemporal autocorrelation in accident prediction. More recent variants [10] combine recurrent blocks with attention mechanisms and transformers to improve temporal fusion and capture nonlinearities. Given the importance of interpretability of results from a management perspective, many studies continue to favor tabular machine learning. Along these lines, models such as Random Forest [5] and Extremely Randomized

Trees [9] are common choices for prediction/counting and risk estimation. Statistical time series models with explicit seasonal structure and exogenous covariates (e.g., SARIMAX) also remain relevant when the objective is to isolate seasonal effects and quantify the impact of external variables.

Equally important to prediction models are spatial aggregation strategies (segmentation) and the unit of analysis, both of which are often underestimated in the literature. In road safety, this debate is often framed as the modifiable areal unit problem (MAUP), in which different scales and forms of zoning can yield distinct conclusions [18]. Studies also show that results vary with segment length and the segmentation strategy adopted, making it difficult to define a universally optimal spatial unit [4]. Literature reviews also indicate that the choice of the unit of analysis, such as road segments, census areas, or functional regions, influences both the inference and the practical application of the results [21]. In this context, our contribution is not to propose a new spatial unit, but to adopt a segmentation guided by the socioeconomic context of the cities along the BR-040 highway, instead of using artificial divisions of fixed length. This brings the analysis closer to the dynamics of territorial planning and resource allocation.

Geographic information systems integrated with spatial analysis also face recurring challenges, such as heterogeneity in data quality, high computational cost, and inconsistencies in structural variables [7]. In [3], the authors combined predictive and network analysis, using models such as random forests and spatial joining to associate accidents with road network segments and to support decisions such as safer routing and the prioritization of interventions. In this study, we also used georeferenced representations, evaluated at different spatiotemporal granularities.

In Brazil, the availability of open data from the Federal Highway Police (PRF) has boosted machine learning applications for analysis and operational planning on federal highways. Studies on the BR-101/PE, for example, explore machine learning techniques for analysis, visualization, prediction, and clustering [15]. Nevertheless, there remains room for experimental protocols that consider, in an integrated way, the quantitative prediction of occurrences, the systematic comparison of models at multiple spatiotemporal granularities, and spatial segmentation compatible with the territorial and socioeconomic context. This work seeks to advance in this direction by combining the comparative evaluation of spatiotemporal models for predicting accident counts on the BR-040 highway with spatial segmentation guided by the economic activities of the cities along the corridor, thereby generating more interpretable and useful results for prioritization and mitigation.

Table 1 summarizes the main differences between our proposal and the related literature. While many studies prioritize more complex deep architectures or fixed-length spatial units, our work adds a systematic comparison at different temporal granularities and a segmentation strategy aligned with the territorial and socioeconomic context.

Table 1. Comparison of related work on highway crash prediction and spatiotemporal analysis.

Study	Models	Spatial Unit	Temporal Res.	Socioecon. Context
Yuan et al. [20]	ConvLSTM	Grids/Segments	High	No
Li et al. [10]	Deep Learning	Latent Structures	Multiple	No
Bisht et al. [4]	Statistical	Fixed Segments	Varies	No
Taveira et al. [15]	ML (Clustering)	Road Points	N/A	No
Berhanu et al. [3]	Random Forest	Network-based	Static	Partial
This Work	**SARIMAX/RF/ET**	**Socioecon. Segm.**	**Weekly/Monthly**	**Yes**

3 Methodology

3.1 Dataset

In this study, we used the open database of traffic accidents on Brazilian federal highways, made available by the PRF (Federal Highway Police) in monthly CSV files [14]. We considered the period from 2017 to 2024 because, starting in 2017, the PRF adopted a more comprehensive data collection process, with additional variables relevant for spatio-temporal analyses [13]. In total, the analysis includes 26,540 accident records on the BR-040 highway during this period. The data acquisition, integration, cleaning, segmentation, and evaluation steps are presented below.

The monthly files were obtained from the PRF open data portal [14] and compiled into a single longitudinal table covering 2017 to 2024. Table 2 summarizes the main attributes of the dataset used in the filtering, enrichment, and aggregation steps.

Table 2. Main attributes from the PRF dataset used in this study (2017–2024).

Attribute	Description
id	Unique identifier of the crash/occurrence.
data_inversa	Crash date (YYYY-MM-DD).
dia_semana, *horario*	Day of week and time of the crash.
uf, *br*, *km*, *municipio*	State (UF), highway, km marker, and municipality.
latitude, *longitude*	Geographic coordinates of the crash.
causa_principal, *causa_acidente*	Primary cause and detailed cause (PRF).
tipo_acidente, *classificacao_acidente*	Crash type and severity (no victim/injured/fatal).
fase_dia, *condicao_meteorologica*	Time-of-day period and weather condition.
tipo_pista, *tracado_via*, *uso_solo*	Road and surrounding-environment characteristics.
regional, *delegacia*, *uop*	PRF operational structure for the segment.
pessoas, *veiculos*, *feridos*	Counts associated with the occurrence.

First, we removed records with missing values in fields required for spatiotemporal aggregation and eliminated duplicates based on id. For continuous numeric variables, we standardized decimal separators (comma → dot) to ensure consistent parsing. For categorical variables, we performed numeric encoding (mapping categories to integers) to enable their use in algorithms that require numeric inputs. Table 3 presents the mapping for the categorical variables effectively used by the models.

Table 3. Numeric encoding of categorical variables used by the models.

Variable	Category	Code
fase_dia	Dawn	0
	Daytime	1
	Dusk	2
	Nighttime	3
sentido_via	Increasing	0
	Decreasing	1
condicao_meteorologica	Sun	0
	Clear sky	1
	Cloudy	2
	Rain	3
	Drizzle	4
	Fog/Mist	5
	Wind	6
	Hail	7
	Unknown	8
tipo_pista	Single carriageway	0
	Dual carriageway	1
	Multiple	2

3.2 Feature Engineering and Selection

We applied feature construction to incorporate signals from existing variables and subsequently applied feature selection procedures to reduce redundancy and mitigate multicollinearity. In particular, `tracado_via` has high cardinality (174 values) and was decomposed into more compact structural attributes (Table 4), as suggested in [12].

Additionally, we constructed temporal attributes to capture calendar effects (holidays and weekends), derived from `data_inversa`. These signals are relevant for modeling routines and seasonality and were evaluated in scenarios with and without such variables (Table 5).

Table 4. Attributes derived from *tracado_via*.

Derived attribute	Label	Code
condicaoPista	Normal	0
	Under construction	1
tipoInclinacao	Flat	0
	Uphill	1
	Downhill	2
tipoSuperficie	Straight	0
	Curve	1
tipoManobra	None	0
	Intersection	1
	Regulated U-turn	2
	Roundabout	3
tipoEstrutura	None	0
	Bridge	1
	Overpass	2
	Tunnel	3

Table 5. Derived temporal attributes (holidays and non-working days).

Attribute	Description
feriado	1 if the date matches a national holiday; otherwise 0.
pre_feriado	1 if the date is the day before a national holiday; otherwise 0.
pos_feriado	1 if the date is the day after a national holiday; otherwise 0.
fim_de_semana	1 if the date is Saturday or Sunday; otherwise 0.

After feature construction, we employed backward elimination from the full set, iteratively removing variables whose exclusion did not compromise the adopted performance criterion. The backward selection approach is conservative and consistent with the fact that the derived attributes were defined based on domain knowledge; thus, we initially considered only redundancies for elimination.

3.3 Models

The predictive task consists of estimating the volume of accidents in each segment of the BR-040 highway at two temporal aggregation levels: weekly and monthly. To this end, we evaluated three families of models with different modeling characteristics and operational uses: Random Forest, Extra Trees Regression, and SARIMAX. Random Forest is suitable for capturing non-linear interactions among heterogeneous covariates and also provides variable importance estimates, which facilitate the interpretation and identification of risk factors [5,12]. Extra Trees Regression introduces more randomness in tree construction, which can

reduce variance and maintain stable performance at a lower computational cost, a relevant aspect when testing multiple segmentations and many segments [9]. SARIMAX explicitly models temporal dependence, trend, and seasonality, and allows the use of exogenous regressors, such as calendar indicators, making it a suitable reference for aggregated series in weekly and monthly windows [1].

The evaluated models require defining hyperparameters that affect the learning process and the complexity of the adjustment. In this work, these parameters were defined based on configurations recurrent in the literature and in preliminary validation experiments. Table 6 presents the main hyperparameters adopted for each model.

Table 6. Best hyperparameter combinations for the evaluated algorithms.

Model	Best Hyperparameters
Random Forest	n_estimators = 100, max_depth = None, min_samples_leaf = 2, max_features = 0.5
Extra Tree Regression	n_estimators = 100, max_depth = 15, min_samples_split = 10, min_samples_leaf = 4, max_features = 'sqrt'
SARIMAX	Monthly: (p, d, q) = (2, 1, 1), (P, D, Q, s) = (2, 1, 1, 12) Weekly: (p, d, q) = (2, 1, 1), (P, D, Q, s) = (2, 1, 1, 52)

Spatial segmentation is a central decision in the analysis of extensive highways. Uniform kilometer-based partitions are simple and reproducible, but tend to ignore territorial differences, such as economic dynamics, logistics function, and population density, which affect exposure and accident risk. In this work, we adopted a segmentation oriented toward the socioeconomic context along the BR-040 highway, incorporating municipal variables to define units that are more closely aligned with the highway's territorial transitions. To analyze the sensitivity to spatial granularity, we tested four configurations, with 18, 24, 30, and 36 segments. This allows us to observe the balance between greater statistical aggregation, with more events per unit, and greater sensitivity to local patterns, with smaller units.

We also evaluated two temporal resolutions, monthly and weekly. Monthly aggregation highlights trends and annual seasonality, reducing short-term oscillations. Weekly aggregation preserves faster variations, linked to routines, long weekends, and local events, and may be more useful for short-term operational decisions.

To avoid information leaks and maintain adherence to operational use, we adopted temporal validation, dividing the data into training (2017–2022), vali-

dation (2023), and testing (2024). This protocol preserves the temporal order of the data and enables a prospective evaluation of generalization capacity.

We evaluated predictive performance using MAE, RMSE, coefficient of determination (R^2), and total execution time. MAE and RMSE measure the average error, with RMSE penalizing larger errors more heavily [6]. R^2 indicates the proportion of variance explained by the model in the evaluated dataset [8]. Execution time was used as a feasibility measure in monitoring and decision support scenarios, where frequent updates may be necessary.

For pairwise comparisons between models and configurations, we used the nonparametric Wilcoxon signed-rank test, which is suitable for paired samples and makes no normality assumption [16]. In the results section, the best performance values are highlighted in bold.

4 Results

In this section, we compare the SARIMAX, Random Forest, and Extra Trees Regression models in predicting accidents on the BR-040 highway at two temporal granularities, monthly and weekly, and in two attribute configurations, with and without holiday and non-working day variables, totaling four experimental scenarios. The models were evaluated in different segmentations based on R^2, MAE, RMSE, and execution time, in order to jointly analyze predictive performance and computational cost. The experiments were performed on a machine with 16 GB of RAM, an AMD Ryzen 5 5600G processor, and an AMD Radeon Vega 7 GPU.

4.1 Model Evaluation with Monthly Granularity and Holiday/Non-Working-Day Variables

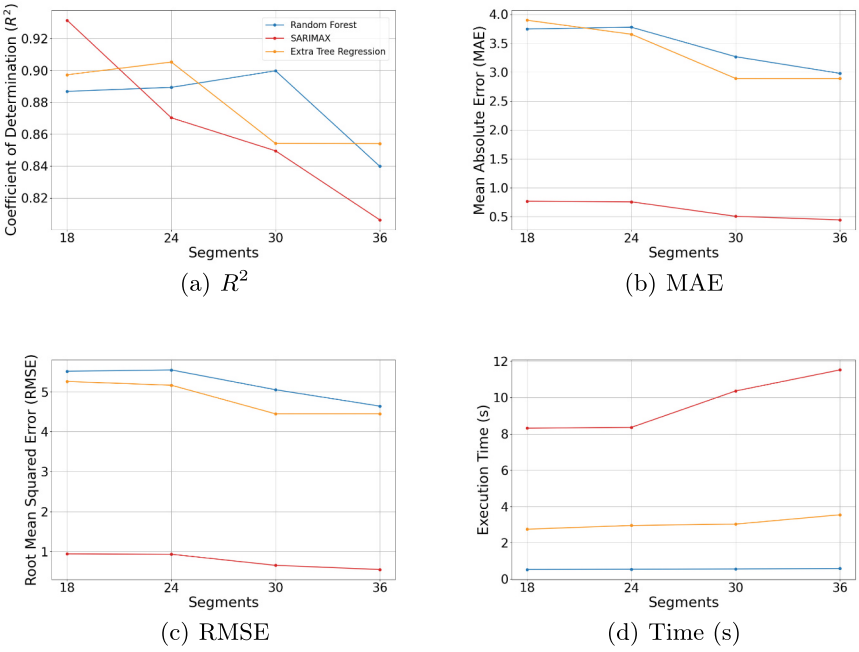
Table 7 shows a consistent relationship between accuracy and computational cost at the monthly granularity with calendar variables, also observed in Fig. 1. SARIMAX presented the lowest MAE and RMSE values in all segmentations, indicating greater accuracy and a better ability to capture temporal patterns when calendar effects are included. Conversely, this gain came with a longer execution time, which increases with the number of segments and may limit its use in scenarios of frequent retraining or larger scale. Random Forest, on the other hand, maintained execution times below 0.6 s, being the most efficient alternative. Furthermore, along with Extra Trees Regression, it presented R^2 values higher than those of SARIMAX in most scenarios, especially with 24, 30, and 36 segments, suggesting that ensemble methods can explain a relevant part of the variability with operational advantage, although with higher absolute errors.

4.2 Model Evaluation with Monthly Granularity and *without* Holiday/Non-Working-Day Variables

Under the monthly configuration without holiday and non-working-day variables (Table 8), we observe an overall reduction in explanatory power compared to the

Table 7. Model performance under monthly *with* holiday and non-working-day variables.

Segments	Model	R^2	MAE	RMSE	Time (s)
18	Random Forest	0.8869	3.7472	5.5110	**0.5310**
	SARIMAX	**0.9315**	**0.7655**	**0.9408**	8.3178
	Extra Tree Regression	0.8973	3.8983	5.2553	2.7524
24	Random Forest	0.8894	3.7779	5.5443	**0.5473**
	SARIMAX	0.8703	**0.7554**	**0.9311**	8.3626
	Extra Tree Regression	**0.9053**	3.6563	5.1618	2.9576
30	Random Forest	**0.8998**	3.2676	5.0489	**0.5599**
	SARIMAX	0.8496	**0.5048**	**0.6535**	10.3678
	Extra Tree Regression	0.8542	2.8912	4.4451	3.0312
36	Random Forest	0.8399	2.9795	4.6387	**0.5889**
	SARIMAX	0.8063	**0.4437**	**0.5491**	11.5266
	Extra Tree Regression	**0.8541**	2.8913	4.4456	3.5443



(a) R^2 (b) MAE

(c) RMSE (d) Time (s)

Fig. 1. Performance under monthly granularity with calendar variables (holidays and non-working days).

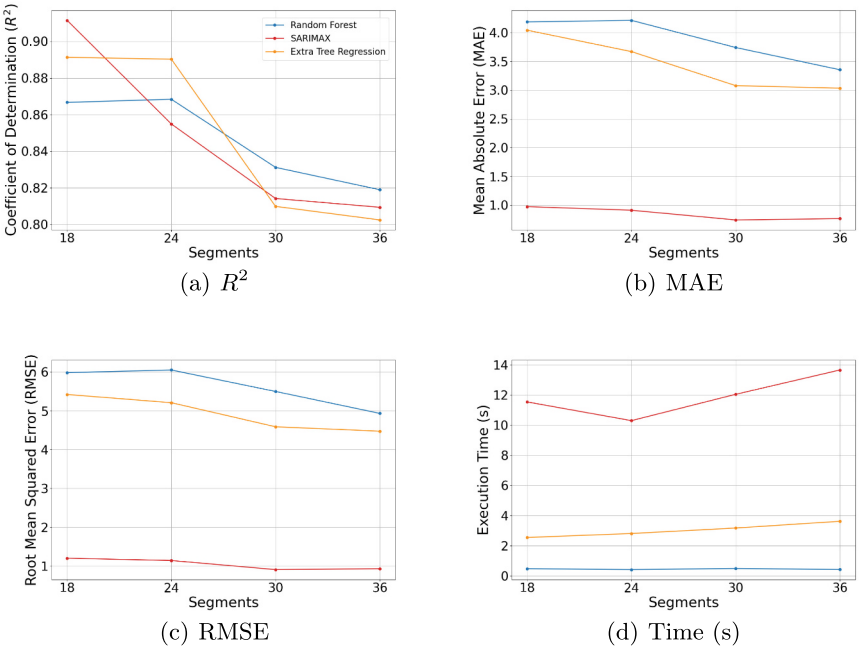
calendar-aware setting, reflected in lower R^2 across all methods, indicating that part of the previously captured variability was associated with calendar effects. Still, SARIMAX remained the best in residual minimization, yielding the lowest MAE and RMSE for all segmentations (Fig. 2), but at high computational cost (runtimes above 10 s and increasing with the number of segments). Conversely, Random Forest remained the most efficient alternative (runtimes below 0.5 s) and achieved the highest R^2 for 30 and 36 segments, while Extra Tree Regression achieved the best R^2 for 24 segments. Overall, the results suggest that, in the absence of exogenous calendar regressors, tree-based ensembles tend to preserve higher explanatory power (R^2) in some scenarios, whereas SARIMAX continues to produce lower absolute error, reinforcing the trade-off between precision and scalability.

Table 8. Model performance under monthly granularity *without* holiday and non-working-day variables.

Segments	Model	R^2	MAE	RMSE	Time (s)
18	Random Forest	0.8667	4.1883	5.9809	**0.4799**
	SARIMAX	**0.9115**	**0.9737**	**1.1991**	11.5360
	Extra Tree Regression	0.8913	4.0438	5.4191	2.5481
24	Random Forest	0.8684	4.2153	6.0496	**0.4192**
	SARIMAX	0.8549	**0.9106**	**1.1407**	10.2971
	Extra Tree Regression	**0.8903**	3.6714	5.2072	2.8131
30	Random Forest	**0.8312**	3.7428	5.4970	**0.4944**
	SARIMAX	0.8142	**0.7423**	**0.9069**	12.0413
	Extra Tree Regression	0.8099	3.0793	4.5871	3.1745
36	Random Forest	**0.8190**	3.3557	4.9324	**0.4249**
	SARIMAX	0.8094	**0.7648**	**0.9298**	13.6540
	Extra Tree Regression	0.8025	3.0343	4.4732	3.6211

4.3 Model Evaluation with Weekly Granularity and Holiday/Non-Working-Day Variables

With weekly granularity and calendar variables (Table 9), the computational-cost difference across approaches becomes dominant. As shown in Fig. 3, SARIMAX runtimes reach hundreds of seconds, making large-scale execution impractical despite its high precision for weekly dynamics. SARIMAX achieved the lowest errors (MAE and RMSE) across all segmentations, with values below 0.4, indicating high precision in modeling weekly dynamics with statistical terms and exogenous regressors. However, this precision comes at the expense of extremely high runtimes (447–695 s) that increase with the number of segments. In terms of overall fit, SARIMAX was superior in R^2 for 18 and 24 segments, but was outperformed by the ensembles for 30 and 36 segments, indicating that as the spatial partitioning becomes more fine-grained, tree-based models maintain more



(a) R^2 (b) MAE

(c) RMSE (d) Time (s)

Fig. 2. Performance under monthly granularity without calendar variables.

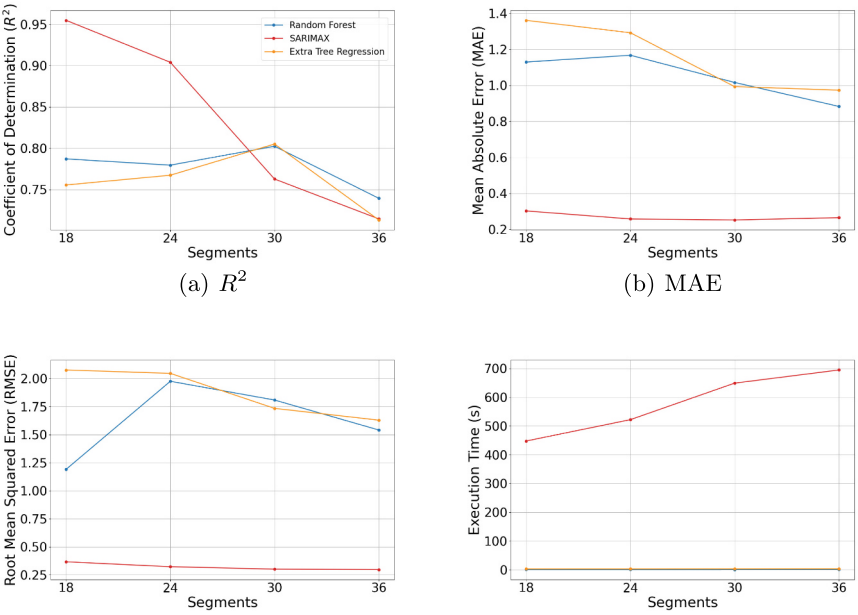
Table 9. Model performance under weekly granularity *with* holiday and non-working-day variables.

Segments	Model	R^2	MAE	RMSE	Time (s)
18	Random Forest	0.7871	1.1295	1.1915	**0.6615**
	SARIMAX	**0.9548**	**0.3028**	**0.3672**	447.9933
	Extra Tree Regression	0.7555	1.3611	2.0760	2.9343
24	Random Forest	0.7795	1.1672	1.9757	**0.6552**
	SARIMAX	**0.9040**	**0.2582**	**0.3234**	522.2830
	Extra Tree Regression	0.7673	1.2922	2.0461	2.9548
30	Random Forest	0.8024	1.0162	1.8085	**0.5980**
	SARIMAX	0.7627	**0.2529**	**0.3019**	649.3398
	Extra Tree Regression	**0.8051**	0.9931	1.7343	3.1126
36	Random Forest	**0.7394**	0.8832	1.5417	**0.8074**
	SARIMAX	0.7145	**0.2653**	**0.2978**	694.9754
	Extra Tree Regression	0.7132	0.9733	1.6293	3.2913

stable explanatory performance even with larger residuals. Thus, under weekly granularity, SARIMAX maximizes accuracy, whereas Random Forest and Extra Tree Regression offer a better cost–benefit trade-off for operational applications.

4.4 Model Evaluation with Weekly Granularity and *without* Holiday/Non-Working-Day Variables

Table 10 reinforces SARIMAX's superiority in predictive accuracy under the weekly configuration without calendar variables, with consistently lower MAE and RMSE and high R^2 across all segmentations. However, this performance is accompanied by a substantial computational cost (744–1106 s), hundreds of times larger than the alternatives. Among the ensembles, Extra Tree Regression provides the best compromise, preserving R^2 and reducing errors compared to Random Forest, which shows marked degradation for 36 segments ($R^2 = 0.4841$), suggesting higher sensitivity to spatial heterogeneity when combined with weekly variability (Fig. 4). In summary, although SARIMAX maximizes precision, its scalability is limited, and tree-based methods, especially Extra Tree Regression, are more viable alternatives for large-scale execution.



(a) R^2 (b) MAE

(c) RMSE (d) Time (s)

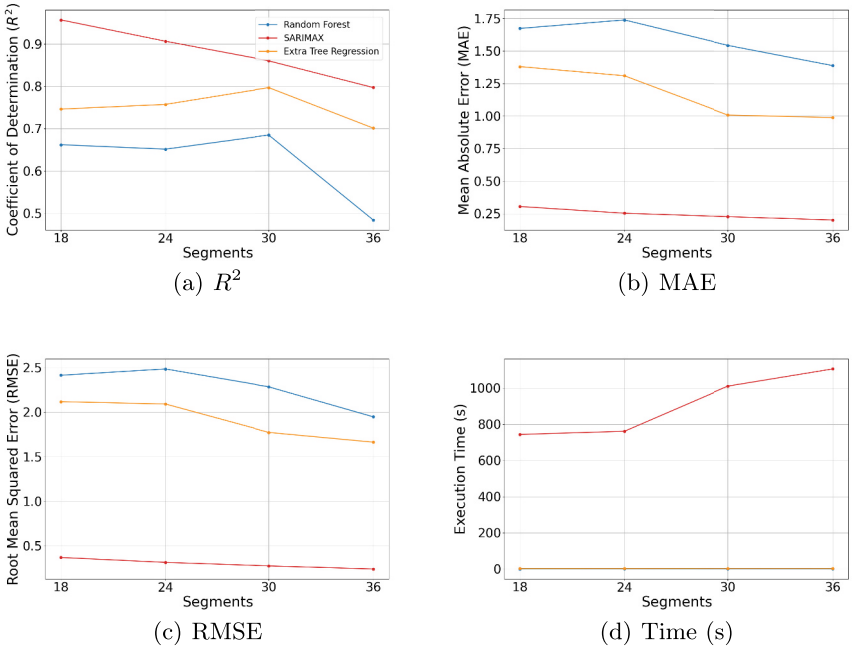
Fig. 3. Performance under weekly granularity with calendar variables.

5 Discussion

The results obtained demonstrate that the forecasting models analyzed can contribute to supporting road safety policies, particularly in highway planning and management. Overall, none of the models outperformed the others across all scenarios or metrics, highlighting a trade-off between predictive accuracy and computational cost.

Table 10. Model performance under weekly granularity *without* holiday and non-working-day variables.

Segments	Model	R^2	MAE	RMSE	Time (s)
18	Random Forest	0.6620	1.6727	2.4136	**0.4644**
	SARIMAX	**0.9570**	**0.3054**	**0.3698**	744.0230
	Extra Tree Regression	0.7462	1.3794	2.1172	3.0037
24	Random Forest	0.6514	1.7378	2.4843	**0.5764**
	SARIMAX	**0.9066**	**0.2537**	**0.3149**	761.4545
	Extra Tree Regression	0.7572	1.3090	2.0912	2.9859
30	Random Forest	0.6850	1.5425	2.2836	**0.5050**
	SARIMAX	**0.8602**	**0.2272**	**0.2753**	1010.9557
	Extra Tree Regression	0.7970	1.0074	1.7722	3.1855
36	Random Forest	0.4841	1.3872	1.9476	**0.6327**
	SARIMAX	**0.7972**	**0.2023**	**0.2414**	1106.0378
	Extra Tree Regression	0.7012	0.9884	1.6630	3.7680



(a) R^2 (b) MAE

(c) RMSE (d) Time (s)

Fig. 4. Performance under weekly granularity without calendar variables.

Thus, the results indicate that the different models can be used in a complementary manner for future highway planning. Models such as SARIMAX, due to their higher accuracy, are more suitable for supporting strategic medium-

and long-term policy decisions, helping to identify periods with a higher risk of accidents and thereby defining preventive actions. On the other hand, ensemble-based models, with lower computational cost, are more appropriate for operational applications and continuous monitoring systems, where fast predictions are essential to support real-time decision-making.

Therefore, this study reinforces the potential of predictive models as a tool to support the formulation of more efficient public policies, especially in the context of road safety, enabling both strategic planning and operational management.

Although the results demonstrate the ability of the models to support decision-making, particularly in the context of highway planning and management, the study is limited to a single Brazilian highway, which may restrict the generalization of the conclusions to other regions. For future work, it is suggested to extend the analysis to different highways and geographic contexts in order to validate the robustness and generalizability of the proposed models.

6 Conclusion and Future Work

This work developed and comparatively evaluated spatiotemporal models for crash-count forecasting on the BR-040 highway using PRF open data (2017–2024). The experimental protocol combines a socioeconomic-guided spatial segmentation with alternative temporal granularities (monthly and weekly), and assesses setups both with and without calendar variables (holidays and non-working days). Across the four experimental scenarios, the results consistently exposed a practical accuracy-scalability trade-off. SARIMAX tended to achieve the lowest MAE/RMSE and a strong fit across multiple settings; however, this advantage came at a substantial computational cost, and the model became impractical at weekly resolution as runtimes increased to the order of hundreds of seconds. In contrast, tree-based ensembles-particularly Extra Tree Regression—ran orders of magnitude faster while remaining competitive in several scenarios, which makes them a better match for large-scale monitoring or frequently updated operational pipelines.

As future work, we will proceed in three directions. First, as data availability improves, we will incorporate richer roadway and infrastructure variables (e.g., pavement condition and signage) to better capture structural risk factors. Second, we will expand the benchmark to include additional spatiotemporal methods, including those that may require greater computational resources, to further characterize the performance frontier under realistic constraints. Third, we will investigate finer temporal resolutions (e.g., daily aggregation) to capture short-term patterns relevant to rapid response. Finally, to facilitate adoption by road-safety stakeholders, we plan to translate the modeling outputs into an interactive geospatial dashboard that turns forecasts into segment-level risk maps and alerting mechanisms for prioritization and intervention.

Acknowledgment. The authors acknowledge partial funding from the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES) - Funding Code 001; the National Council for Scientific and Technological Development - Brazil (CNPq) - Funding Codes 442344/2023–6, 444602/2024–0, and 407466/2025–8; the Espírito Santo Research and Innovation Support Foundation (FAPES) – Call No. 02/2026 (Participation in Technical-Scientific Events) – Grant No. 428/2026 and project #2024-9G1WP BPC; and the Minas Gerais Research Funding Foundation (FAPEMIG)

References

1. Alharbi, F.R., Csala, D.: A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach. Inventions **7**(4), 94 (2022). https://doi.org/10.3390/inventions7040094
2. Amorim, B.D.S.P., Firmino, A.A., Baptista, C.D.S., Braz Júnior, G., Paiva, A.C., Almeida Júnior, F.E.: A machine learning approach for classifying road accident hotspots. ISPRS Int. J. Geo Inf. **12**(6), 227 (2023). https://doi.org/10.3390/ijgi12060227
3. Berhanu, Y., Schröder, D., Wodajo, B.T., Alemayehu, E.: Machine learning for predictions of road traffic accidents and spatial network analysis for safe routing on accident and congestion-prone road networks. Results Eng. **23**, 102737 (2024). https://doi.org/10.1016/j.rineng.2024.102737
4. Bisht, L.S., Chand, S., Tiwari, G.: Proposed segment length for safety evaluation studies on rural divided highways in India. IATSS Res. **48**(3), 331–341 (2024). https://doi.org/10.1016/j.iatssr.2024.05.002
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
6. Chai, T., Draxler, R.R.: Root mean square error (RMSE) or mean absolute error (MAE)? arguments against avoiding RMSE in the literature. Geosci. Model Dev. **7**, 1247–1250 (2014). https://doi.org/10.5194/gmd-7-1247-2014
7. Deressa, B.F., et al.: A systematic review of GIS-driven road traffic accident evaluation. Vehicles **7**(4), 161 (2025). https://doi.org/10.3390/vehicles7040161
8. Draper, N.R., Smith, H.: Applied Regression Analysis. John Wiley & Sons, 3 edn. (1998). https://doi.org/10.1002/9781118625590
9. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
10. Li, K., Guo, K., Liang, Z., Xu, H., Duan, X.: Traffic accident severity prediction and analysis via spatio-temporal deep learning: a ConvLSTM-Transformer approach. Chaos Sol. Fractals **200**, 117047 (2025). https://doi.org/10.1016/j.chaos.2025.117047
11. Li, Y., Yu, R., Shahabi, C., Liu, Y.: Diffusion convolutional recurrent neural network: data-driven traffic forecasting. In: Proceedings of 6th International Conference on Learning Representations (ICLR) (2018), Conference Track Proceedings. OpenReview.net (2018). https://openreview.net/forum?id=SJiHXGWAZ
12. Mafort, E., Kappel, M.A.A.: Machine learning techniques for predicting the severity of accidents on highways in the state of Rio de Janeiro. Revista Eletrônica de Iniciação Científica em Computação **23**, 244–252 (2025). https://doi.org/10.5753/reic.2025.6274
13. Polícia Rodoviária Federal: Dados abertos. Portal gov.br (2026). https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos. Accessed 10 Feb 2026

14. Polícia Rodoviária Federal: Dados abertos da PRF. Portal Gov.br (2026). https://www.gov.br/prf/pt-br/acesso-a-informacao/dados-abertos/dados-abertos-da-prf. Accessed 10 Feb 2026
15. Taveira, J., et al.: Applying machine learning techniques to traffic accident data on BR-101 in Pernambuco. In: VI Seminário Internacional: Ciência, Tecnologia e Inovação em Segurança Pública (2024)
16. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bull. **1**(6), 80–83 (1945). https://doi.org/10.2307/3001968
17. World Health Organization: Road traffic injuries. WHO Fact sheet (2023). https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed 10 Feb 2026, published 13 Dec 2023
18. Xu, P., Huang, H., Dong, N.: The modifiable areal unit problem in traffic safety: basic issue, potential solutions and future research. J. Traffic Transp. Eng. (English Edn.) **5**(2), 73–82 (2018). https://doi.org/10.1016/j.jtte.2015.09.010
19. Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp. 3634–3640. International Joint Conferences on Artificial Intelligence Organization (2018). https://doi.org/10.24963/ijcai.2018/505
20. Yuan, Y., Zhou, J., Yang, J.: Hetero-ConvLSTM: a deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 984–992 (2018). https://doi.org/10.1145/3219819.3219922
21. Ziakopoulos, A., Yannis, G.: A review of spatial approaches in road safety. Acc. Anal. Prev. **135**, 105323 (2020). https://doi.org/10.1016/j.aap.2019.105323