

A comparative assessment of spatial regression methods for predicting maritime events in the Caribbean Sea

Authors

Amrika Maharaj¹, Keith Miller¹, Dexter Davis¹ and Michael Sutherland¹

Abstract

Predicting maritime events requires spatially aware techniques that account for complex geographic and operational dynamics. By applying Geographic Weighted Regression, Generalized Linear Regression, Forest-Based Classification and Regression, and Empirical Bayesian Kriging, this research models the spatial distribution of vessel events across the Caribbean Sea. Key influencing variables include vessel traffic density, charted zone confidence, flag state, and vessel age. Results highlight regional hotspots such as the Panama Canal and Gulf of Paria. Comparative analysis demonstrates the strengths and limitations of each technique, informing the development of adaptive, location-specific maritime risk mitigation strategies.

Keywords

maritime accidents · geographic information systems · Caribbean Sea · regression techniques · vessel traffic density · automatic identification system

Resumé

La prévision des événements maritimes nécessite des techniques sensibles à la dimension spatiale, capables de prendre en compte la complexité des dynamiques géographiques et opérationnelles. En appliquant la régression pondérée géographique, la régression linéaire généralisée, la classification et la régression basées sur les forêts et le krigeage bayésien empirique, cette recherche modélise la diffusion spatiale des événements maritimes dans la mer des Caraïbes. Les principales variables influentes comprennent la densité du trafic maritime, la fiabilité des zones cartographiées, l'État du pavillon et l'âge des navires. Les résultats mettent en évidence des zones sensibles telles que le canal de Panama et le golfe de Paria. L'analyse comparative révèle les forces et les limites de chaque technique, contribuant à l'élaboration de stratégies adaptatives de réduction des risques maritimes, spécifiques à chaque localisation.

Resumen

La predicción de eventos marítimos requiere técnicas de conciencia espacial que tengan en cuenta las complejas dinámicas geográficas y operativas. Aplicando la Regresión Ponderada Geográficamente, la Regresión Lineal Generalizada, la Clasificación y Regresión Basadas en Bosques y el Kriging Bayesiano Empírico, esta investigación modela la distribución espacial de los eventos relacionados con buques en el Mar Caribe. Las variables clave que influyen incluyen la densidad del tráfico marítimo, la confianza en la zona cartografiada, el estado de abanderamiento y la edad del buque. Los resultados destacan puntos críticos regionales como el Canal de Panamá y el Golfo de Paria. El análisis comparativo demuestra las fortalezas y limitaciones de cada técnica, contribuyendo al desarrollo de estrategias adaptativas para mitigar los riesgos marítimos de cada localización específica.

✉ Amrika Maharaj · amrika.maharaj@hotmail.com

¹ Department of Geomatics Engineering and Land Management, The University of the West Indies, St. Augustine, The Republic of Trinidad and Tobago

1 Introduction

The Caribbean Sea, a major corridor for international shipping and regional trade, faces persistent challenges related to maritime safety. With increasing vessel traffic, outdated navigational charts, and limited Aids to Navigation (AtoNs), the region is particularly vulnerable to navigation related accidents. Over the past two decades, maritime incident reports have highlighted recurring patterns influenced by spatial and operational risk factors, including high vessel density, inadequate AtoNs, and poorly charted areas (Maharaj et al., 2025). Understanding where and why maritime accidents occur is essential for effective risk mitigation and maritime domain awareness. Given the inherently spatial nature of these events, conventional analytical tools fall short of capturing the complex relationships between vessel traffic patterns, operational practices, and environmental hazards. Spatial statistical techniques, on the other hand, allow researchers and maritime authorities to quantify these relationships, identify incident-prone zones, and develop predictive insights that can inform proactive safety measures (Waller & Gotway, 2004; Pfeiffer, 1996; Fotheringham et al., 2002).

Despite the importance of maritime safety in the Caribbean, there remains a clear research gap: limited studies have systematically compared multiple spatial regression and machine learning approaches for predicting maritime events in this region. Much of the existing literature has focused on either global maritime accident trends or localized port-level analyses, leaving a lack of comparative, region wide evaluations of model performance and predictive capacity. This absence constrains the ability of decision makers to identify the most suitable modelling techniques for operational use in complex, heterogeneous maritime environments.

To address this gap, this paper analyses various regression methods that best capture spatial heterogeneity in maritime events across the Caribbean Sea, and how consistent hotspot patterns are across different modelling techniques. This is central for determining both the explanatory and predictive strengths of different approaches, and to guiding their practical application in regional maritime risk management. Accordingly, the objective of this study is to undertake a comparative assessment of four spatial prediction techniques, Geographically Weighted Regression (GWR), Generalized Linear Regression (GLR), Forest-Based Classification and Regression (FBCR), and Empirical Bayesian Kriging (EBK). By applying these methods to a database of maritime events and associated spatial covariates including traffic density, AtoNs, current velocity, navigational hazards, bathymetry, chart confidence (CATZOC), survey age, vessel flag state, and vessel age, this research evaluates their ability to identify high-risk zones and to explain underlying causal factors.

By benchmarking the performance of statistical, spatial, and machine learning approaches, this paper

provides new insights into the suitability of different modelling frameworks for maritime accident prediction in geographically heterogeneous environments. The contribution of the study is threefold: (i) it advances understanding of how different spatial modelling techniques perform when applied to real-world maritime accident data; (ii) it highlights consistent and divergent hotspot patterns, thereby informing risk-based prioritization of maritime safety interventions; and (iii) it proposes a framework that can guide adaptive, geographically targeted strategies for enhancing navigational safety and reducing future maritime accidents. Ultimately, this work supports the development of evidence-based tools for maritime authorities to better anticipate, prepare for, and mitigate navigational risks in the Caribbean Sea.

2 Literature review: Spatial approaches to maritime incident modelling

Understanding maritime risk requires methods that go beyond simple attribute analysis. Spatial statistical models incorporate location and context, enabling the identification of incident clusters and causal relationships. These models handle both spatial and aspatial data, allowing for localized exploration of how risk factors vary geographically. The prediction and analysis of maritime incidents require modelling techniques that can capture both the spatial structure of data and the complex interactions between geographic and operational risk factors. Traditional statistical models often assume independence among observations and disregard the influence of location, rendering them insufficient for spatially dependent phenomena such as vessel incidents (Fotheringham et al., 2002). As a result, spatial statistical methods have emerged as essential tools for understanding maritime risk, particularly in geographically heterogeneous environments like the Caribbean Sea.

Spatial statistical techniques focus on identifying patterns, relationships, and dependencies among spatially distributed events and their associated attributes. These methods assume that "location matters" that the spatial distribution of events is not random but influenced by underlying environmental and operational conditions (Waller & Gotway, 2004). Maritime incidents, by their nature, are spatial events influenced by locational factors such as depth (bathymetry), traffic density, charting accuracy, proximity to AtoNs, proximity to navigational hazards and current velocity. Therefore, spatial analysis becomes a critical component in determining where incidents are more likely to occur and why. Spatial data consist of two components: geometry (location) and attributes (characteristics or measurements at that location). In contrast, aspatial data include only the latter (Pfeiffer, 1996). Methods developed for aspatial data analysis often fail when applied to spatial data due to spatial autocorrelation and non-stationarity the tendency for relationships between variables to vary across space

(Fotheringham et al., 2002).

To explore and quantify these relationships, a series of regression and machine learning techniques have been developed and applied to real world maritime accident data. The following subsections review the theoretical foundations and prior applications of four key methods used in this study: GWR, GLR, FBCR, and EBK. Collectively, these approaches provide the methodological backbone for our comparative analysis of maritime accident prediction, demonstrating effectiveness in revealing hidden spatial patterns, identifying high-risk zones, and generating actionable insights for navigation safety and risk mitigation strategies.

2.1 Regression techniques

To explore and quantify the relationship between maritime events and their contributing factors, a series of spatial and non-spatial regression techniques were employed. These methods are essential for understanding how risk variables such as traffic density, chart accuracy, vessel characteristics, and environmental conditions influence the likelihood of vessel events across the Caribbean Sea. Each technique offers a unique analytical lens: some allow for localized parameter estimation, while others emphasize statistical robustness or predictive accuracy. This section explains the theoretical foundations, assumptions, and implementation procedures of each technique as applied to the maritime accident dataset.

2.1.1 Geographic weighted regression

GWR is a spatial statistical technique designed to explore and model spatially varying relationships between a dependent variable and a set of independent variables (Charlton & Fotheringham, 2009). Unlike global regression models that assume a uniform relationship across an entire study area, GWR estimates local models at each spatial location, thereby capturing the heterogeneity in relationships that may exist across geographic space. In GWR, each local model is calibrated using a subset of data points that are geographically proximate to the location of interest. Observations closer to the target location are assigned higher weights in the regression analysis, typically using a spatial weighting function based on distance decay (Lin & Wen, 2011). This allows for the production of spatially adaptive regression coefficients that can reveal how explanatory variables influence the dependent variable differently from one location to another. For this study, GWR was applied to investigate the relationship between maritime incidents and their contributing factors across the Caribbean Sea. Given the region's diverse maritime geography and operational contexts, GWR provided an initial exploratory framework to assess whether these relationships varied spatially. This was particularly important in understanding how localized features such as traffic density, AtoNs, bathymetry, and charting accuracy contribute to the likelihood of

maritime incidents (Weisent et al., 2012).

Mathematically, GWR extends the standard linear regression model by allowing the coefficients to vary at each spatial location. The GWR equation is expressed as:

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i)x_{ik} + \varepsilon_i \quad (1)$$

where:

- y_i is the dependent variable at location i (e.g., maritime incident occurrence),
- $\beta_k(u_i, v_i)$ are the location-specific coefficients for the k^{th} explanatory variable,
- (u_i, v_i) are the spatial coordinates of location i ,
- x_{ik} are the values of the independent variables at location i , and
- ε_i is the residual or random error term at location i .

The estimated β coefficients describe how each explanatory factor influences the dependent variable at a specific location. The intercept term β_0 represents the expected value of y_i when all explanatory variables are zero, while the residual ε_i captures the portion of variation not explained by the model either due to unmeasured factors or random variation. By generating a distinct regression equation for each location within the study area, GWR reveals spatial variations in model structure and predictive relationships that global models often obscure. This makes it particularly effective for identifying regional risk patterns and informing location-specific interventions. In the context of this research, GWR results were instrumental in mapping incident-prone areas and highlighting the spatial variability in how different covariates, such as vessel flag state or survey age of charts, contribute to risk.

2.1.2 Generalized linear regression

GLR is a flexible extension of classical linear regression designed to model relationships where the assumptions of normality and linearity do not hold. In the context of maritime risk assessment where the dependent variable (e.g., incident occurrence) is often non-normally distributed GLR offers a robust statistical framework for capturing complex, non-linear associations between maritime incidents and their causal factors (Awal & Hasegawa, 2017). At its core, GLR links the expected value (mean) of the dependent variable to a linear combination of independent variables through a *link function*. This function governs how the mean of the response variable relates to the linear predictor, allowing the model to accommodate a wide range of data types, including binary outcomes, count data, and skewed continuous variables (Jetz et al., 2005). The model is estimated using maximum likelihood estimation, which identifies the parameter values that best fit the observed data under the assumed probability distribution (Gao & Li, 2011). The GLR model assumes that the dependent variable follows a probability distribution from the exponential family, such as the normal,

Poisson, or binomial distribution. The appropriate distribution is selected based on the characteristics of the response variable. For example, Poisson distribution is often used for count data, while binomial distribution applies to binary outcomes (e.g., incident occurred vs. did not occur). Once the link function and distribution are specified, GLR can be used to make predictions for new observations, interpret the strength and direction of predictor variables, and test hypotheses regarding statistical significance.

In this study, GLR was implemented to assess the global relationships between maritime accidents across the Caribbean Sea and a set of explanatory variables, including traffic density, navigational hazards, CATZOC, survey age, vessel flag, and vessel age. The technique proved particularly useful given the sparsity and variability of the incident data. It allowed for robust statistical inference by estimating coefficients, odds ratios, and confidence intervals enabling a clearer understanding of how each factor influences incident likelihood. Although GLR does not account for spatial heterogeneity treating the relationships as constant across the entire study area, its interpretability and adaptability make it a valuable component of the analytical framework. In this research, it provided a complementary perspective to localized models such as GWR and informed the identification of dominant risk factors at a global scale.

2.1.3 Forest-based classification and regression

FBCR is a supervised machine learning technique designed to model complex relationships between a set of explanatory variables and a dependent variable. Based on Leo Breiman's Random Forest algorithm, FBCR combines two approaches: Random Forest (RF) for classification problems and Regression Forest for regression problems (Scott & Bennett, 2012). In both cases, the method builds an ensemble of decision trees each trained on a randomly selected subset of the input data and variables and aggregates their outputs to produce a final prediction (ESRI, 2016). For classification tasks, the final prediction is typically derived through majority voting, whereas for regression problems, the predictions from all trees are averaged (Wilmsmeier et al., 2006). This ensemble approach allows FBCR to capture non-linear and high-dimensional relationships that traditional regression techniques may overlook. Moreover, the method is robust to outliers and missing data, making it particularly useful for working with real-world maritime datasets where data completeness and distribution may vary (Zhang et al., 2011). In this study, FBCR was implemented to generate predictive surfaces for maritime incidents across the Caribbean Sea. The training phase involved constructing a forest that models the relationship between the dependent variable (incident occurrence) and the independent variables, which included maritime traffic density, AtoNs, bathymetry,

current velocity, navigational hazards, CATZOC, vessel flag, vessel age, and survey age of navigational charts (Jetz et al., 2005). During this phase, the model automatically excluded 10 % of the training data for validation purposes. Predictions generated for this excluded portion were then compared against observed values to assess model accuracy. A key output of the FBCR analysis was the Variable Importance table, which ranks explanatory variables based on their contribution to the model's predictive power. This importance is calculated using Gini coefficients, a measure reflecting how often a variable is used to split decision nodes and the relative impact of those splits across all trees in the forest (Su et al., 2012). Variables that appear more frequently in high-impact splits are considered more influential in predicting the outcome. FBCR's ability to handle complex interactions, high-dimensional data, and missing values while also producing interpretable outputs such as variable importance makes it a powerful tool for maritime risk modelling. Despite its advantages, however, the technique may be susceptible to overfitting, particularly in cases with sparse or highly imbalanced data. Therefore, model tuning and validation remain critical to ensuring reliable predictions. In this research, FBCR provided not only spatial predictions of incident likelihood but also insights into which factors contributed most significantly to those predictions, enhancing our understanding of maritime risk dynamics across the Caribbean Sea.

2.1.4 Empirical Bayesian Kriging

EBK is a geostatistical interpolation method designed to estimate values at unsampled locations by incorporating spatial autocorrelation and uncertainty into its predictions. It is an enhancement of traditional kriging, which estimates unknown values as weighted averages of observed data points. EBK improves upon this by automatically accounting for spatial heterogeneity and uncertainty in the model parameters through a Bayesian framework (Mainuri & Owino, 2017; Scott & Bennett, 2012). The core innovation of EBK lies in its use of Bayesian simulation to repeatedly estimate semivariogram models from subsets of the observed data. These simulated models reflect different plausible spatial structures and are averaged to produce a final predictive surface. This approach allows EBK to incorporate prior knowledge about spatial patterns and produce more robust predictions, especially in areas where the spatial structure varies significantly across the study region. In EBK, predictions at unsampled locations are based on observed data within a specified local neighborhood. Weights are assigned based on both geographic distance and the strength of spatial autocorrelation derived from the local semivariogram (Wang et al., 2013). Unlike traditional kriging, which typically fits a single global semivariogram for the entire dataset, EBK fits a unique local semivariogram for each neighborhood, making it highly responsive to spatial

heterogeneity. In the context of this study, EBK was applied to generate predictive surfaces of maritime incident likelihood across the Caribbean Sea. Its capacity to model local variation made it particularly well suited for interpolating across regions with sparse data, variable maritime infrastructure, and diverse navigational conditions. While EBK does not produce explicit regression coefficients or quantify the impact of individual variables, its strength lies in producing highly localized, accurate predictions that complement the explanatory power of other models such as GWR, GLR, and FBCR. As part of the broader modelling framework, EBK offered a valuable spatial perspective, especially in identifying incident-prone areas that may be underrepresented in the observed dataset. Its minimal reliance on user defined parameters and robust handling of nonstationary data further supports its application in maritime risk modelling.

3 Approach to spatial prediction

For the purpose of this study, the dependent variable was defined as the spatial occurrence of maritime events across the Caribbean Sea. The independent variables were identified through a quantitative analysis of maritime accident records from 2002 to 2021 obtained from the GISIS IMO database (Maharaj et al., 2025). Each accident report was examined to determine contributing causes, which informed the selection of the risk factors most relevant to navigation safety. These variables, vessel traffic density (transits), current velocity, proximity to navigational hazards, bathymetry, the location of AtoNs, and the quality and survey age of Admiralty nautical charts provided by the UK Hydrographic Office (UKHO), along with vessel flag state and vessel age represent a comprehensive set of physical, environmental, and operational elements influencing the likelihood of maritime events. The study area was delineated using the Caribbean Maritime Boundary Exclusive Economic Zone (EEZ), which provided a legally recognized and geographically bounded framework for analysis. Within this EEZ, a spatial grid was developed in ArcGIS Pro at a 1-kilometre resolution. This grid served as the foundational spatial unit for storing both the dependent and independent variable data in a standardized format. The choice of 1 km as the grid dimension was informed by an initial investigation into the distribution of maritime accident data. Given the relatively low number of recorded events in certain areas of the Caribbean Sea, a finer resolution was avoided to prevent sparse data issues and minimize the risk of multicollinearity among the explanatory variables.

The final dataset consisted of 302 grid cells, each representing a unique spatial unit across the Caribbean Sea (Fig. 1). These grid cells aggregated the aspatial attributes of the explanatory variables, enabling consistent input for all subsequent modelling techniques. Prior to model calibration, a multicollinearity diagnostic was conducted to assess

the interdependencies among the explanatory variables. Multicollinearity, a condition in which two or more independent variables are highly correlated, can significantly inflate the standard errors of regression coefficients, leading to unreliable parameter estimates (Callaghan & Chen, 2008). Following established guidelines (Wheeler & Tiefelsdorf, 2005), variables were evaluated and transformed as necessary to ensure all predictor fields were numeric, continuous, and exhibited acceptable levels of variance. Binary variables, which typically exhibit weak collinearity, were retained where applicable, while highly collinear continuous variables were excluded or restructured. Given that some regression techniques used in this study, such as GWR, are not suited for binary outcome prediction, care was taken to ensure that the dependent variable remained continuous and appropriately distributed for the chosen modelling approaches. This rigorous data preparation process ensured the analytical integrity and statistical robustness of all subsequent regression and machine learning models.

4 Results of regression techniques

The results of the GWR model provided spatial insights into where maritime incidents are most likely to occur across the Caribbean Sea. Fig. 2 illustrates the spatial distribution of predicted incident probabilities, highlighting key maritime zones with statistically significant co-movements between the dependent and independent variables. The fitted values derived from the regression equation are presented as p-values, with darker shades indicating areas of high incident likelihood.

The model revealed that strong correlations both positive and negative exist across different regions. Notably, areas with the highest probability of incident occurrence included:

- The Panama Canal region
- Waters north of Trinidad
- The vicinity of Grenada, St. Lucia, the British Virgin Islands, and Jamaica

These regions are characterized by “under-predictions” locations where actual incident counts exceed model estimates. Conversely, lighter shaded areas represent “over-predictions”, where the model estimates exceed the observed incident volumes.

The GWR model produces several spatial diagnostics to evaluate the performance and reliability of the regression results. The diagnostics derived from the GWR model provide valuable insights into the reliability and performance of the regression analysis. The condition number values indicated that multicollinearity was not present in the model, as all values remained below the threshold commonly associated with unstable estimates. This suggests that the explanatory variables were sufficiently independent to produce robust parameter estimates. The local R^2 values, which quantify how well the model explains

variation in the dependent variable across space, varied throughout the study area. Higher values were observed in key maritime zones, indicating strong model performance and good predictive fit in these regions, while lower values in other areas suggest reduced explanatory power. Standardized residuals, representing the difference between observed and predicted values, were generally low in regions with a high local R^2 , indicating accurate model predictions. However, isolated areas showed higher residuals, reflecting local deviations that the model could not fully explain. Finally, the spatial variation in regression coefficients highlighted the heterogeneity in the strength and direction of relationships between the explanatory variables and the occurrence of maritime incidents. These locally varying coefficients underscore the importance of accounting for geographic context when assessing maritime risk, as the influence of specific variables changes significantly across the Caribbean Sea.



Fig. 1 302 grid cells within the Caribbean Sea EEZ.

The GWR model yielded significant parameter estimates for nine explanatory variables, each characterized by a direction (sign) and strength (magnitude). A positive coefficient indicates that an increase in the respective variable is associated with a corresponding increase in the likelihood of a maritime incident, while a negative coefficient suggests a mitigating effect. The magnitude of the parameter reflects the degree of influence that a one-unit change in the variable has on the dependent outcome (Chang et al., 2008). Among the most influential factors, the analysis revealed that regions surrounding the Panama Canal exhibited the strongest associations, particularly for vessel flag state, survey age of nautical charts (CATZOC), and vessel traffic density (Table 1). These same variables were also prominent in other high-risk locations such as northwest Trinidad and Jamaica. Interestingly, current velocity and CATZOC values in the Panama Canal area contributed less to incident prediction relative to other factors. GWR produces a surface of parameter estimates, allowing the model to capture spatial changes in the magnitude of influence for each covariate. This ability to reflect the locally varying strength of relationships is central to the concept of spatial heterogeneity. In some locations, certain variables exert a much stronger influence on the dependent variable than in others. Conversely, parameter estimates close to zero are often spatially clustered, indicating that in those regions, changes in the respective variables have little to no effect on

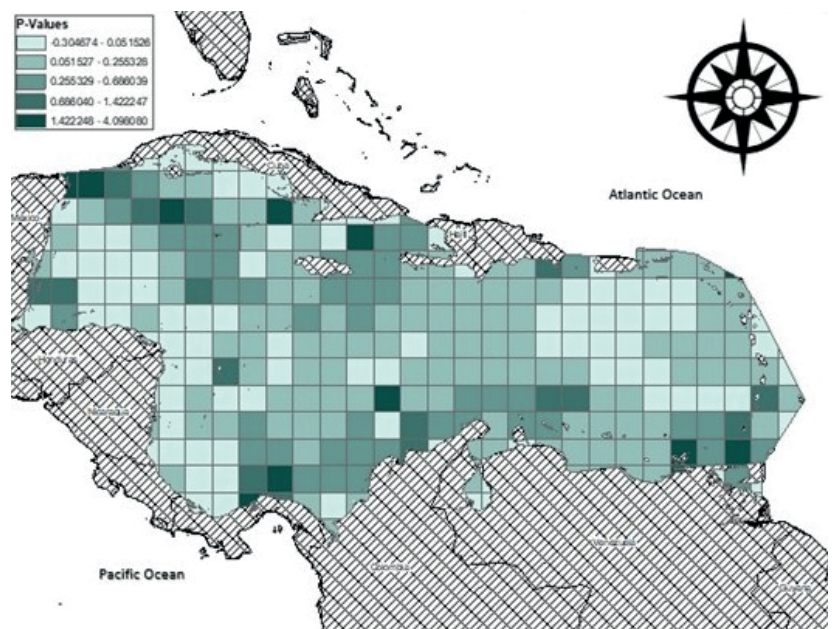


Fig. 2 GWR P-values.

incident likelihood. These clusters may result from uniform local conditions, weak variable variance, or the absence of relevant interactions. Such spatial dynamics invite deeper exploration into the processes shaping maritime risk and underscore the importance of localized modelling in revealing nuanced patterns that global models may obscure. Together, these findings enrich the understanding of the region's risk

Table 1 Parameter estimates of hotspot locations.

Coefficients current velocity	Coefficients traffic density	Coefficients CATZOC	Coefficients survey age	Coefficients flag state	Location
-0.089040303	0.0000839	-0.030459982	0.02242219	0.644299132	Panama Canal
-0.115263682	0.0000818	-0.037947247	0.022592015	0.630958731	NW, Trinidad
-0.092792259	0.0000860	-0.031691154	0.02369738	0.649360068	Jamaica
-0.136586501	0.0000819	-0.044245746	0.024081995	0.625691053	British Virgin Islands
-0.154913036	0.0000913	-0.04870322	0.03102929	0.653840374	Grenada

landscape and highlight the practical utility of GWR in informing maritime safety interventions across diverse operational environments.

The GLR approach provided a global view of the relationships between the dependent variable and its predictors, offering valuable insights into which factors most strongly influence the likelihood of incidents across the Caribbean Sea. Among the variables assessed, vessel flag state, vessel age, and the CATZOC emerged as the most influential contributors to the dependent variable. Notably, traffic density was found to be statistically significant, indicating a robust and direct association with incident likelihood. The model also generated odds ratios to interpret the strength of association for each variable. An odds ratio greater than one suggests that increased exposure to a given variable elevates the probability of an incident occurring. In this study, variables such as traffic density, AtoNs, current velocity, CATZOC, vessel age, and flag state exhibited odds ratios above one, underscoring their importance in predicting risk (Table 2). Conversely, variables with odds ratios close to or below one, such as survey age and bathymetry, were associated with either minimal or inverse effects. Further diagnostics were used to assess model reliability. The Wald's Chi-Squared test confirmed the statistical significance of the explanatory variables included, validating their relevance within the GLR framework. Additionally, the Variance Inflation Factor (VIF) was calculated to evaluate multicollinearity among predictors. All VIF values were well below the threshold of 10, indicating that collinearity did not undermine the model's stability. For example, the survey age variable showed a VIF of approximately 1.8, suggesting moderate inflation but not to a degree that would compromise the results. The spatial prediction output of the GLR model revealed high-risk areas across the Caribbean, with particular concentration around maritime corridors such as the Panama Canal, the Gulf of Paria (Trinidad), Grenada,

St. Vincent and the Grenadines, St. Lucia, Antigua, Montserrat, the British Virgin Islands, Jamaica, and Belize (Fig. 3). These locations were identified as zones with a high probability of incident occurrence based on the values of the explanatory variables. In contrast, areas classified with low, or zero probability were primarily situated in open sea regions, where maritime activity and associated risk factors are comparatively minimal.

The FBCR model provided critical insight into the relative influence of each explanatory variable in predicting maritime incident risk across the Caribbean Sea. Variable importance was determined using the sum of Gini coefficients generated from all decision trees in the model. These values indicate how frequently and effectively each variable contributed to data splits within the forest, offering a robust diagnostic for understanding which inputs drive the model's predictions. According to the analysis, traffic density was identified as the most influential variable, accounting for 15 % of the model's total predictive power. This was followed closely by the location of Aids to Navigation (13 %), navigational hazards (12 %), and CATZOC (12 %). It is important to note that variable importance measures do not reflect model accuracy unlike metrics such as R^2 used in regression analysis but rather indicate the contribution of each variable to the predictive decision-making process. The FBCR model's prediction results aligned closely with the outputs of GWR and GLR, identifying a consistent set of high-risk locations. These included the Panama Canal, Gulf of Paria, Trinidad, Grenada, St. Vincent and the Grenadines, St. Lucia, Montserrat, Antigua, British Virgin Islands, Jamaica, and Belize. The similarity in prediction trends across models strengthens the reliability of the identified hotspots and highlights the spatial coherence of incident-prone areas in the region (Fig. 4). FBCR thus reinforces the value of ensemble decision-tree approaches in revealing dominant maritime risk factors and supporting targeted mitigation efforts across the Caribbean Sea.

The EBK model produced a predictive surface that highlights spatial variations in the likelihood of maritime incidents across the Caribbean Sea. The model identified several high-risk areas, characterized by a higher probability of incident occurrence, particularly in regions surrounding the Panama Canal, Gulf of Paria, Trinidad, Grenada, St. Vincent and the Grenadines, St. Lucia, Montserrat, Antigua, British Virgin Islands, Jamaica, and Belize (Fig. 5). These areas exhibited elevated prediction values compared to surrounding open sea regions, which were associated with lower risk levels. The EBK output demonstrated strong alignment with the prediction patterns generated by the other regression techniques applied in this study, namely GWR, GLR, and FBCR. This consistency reinforces the validity of the identified hotspots and confirms the robustness of spatial interpolation methods in capturing underlying spatial patterns of maritime risk. EBK's ability to model local

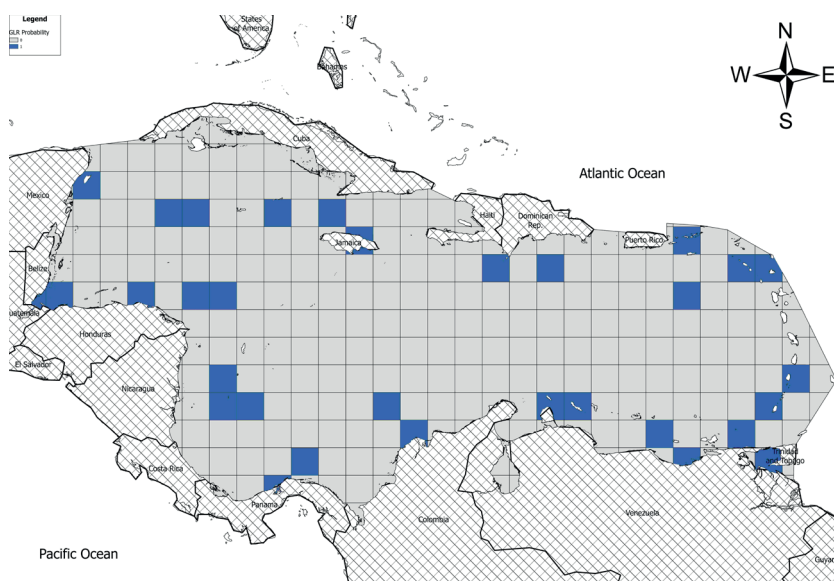


Fig. 3 Prediction surface generated using GLR.

Table 2 GLR coefficient diagnostic.

Variable	Coefficient [a]	StdError	Z-statistic	Probability [b]	Odds ratio [c]	Wald's low (95 %) [d]	Wald's high (95 %) [d]	VIF [e]
Intercept	-2.302016	3.670614	-0.627147	0.530563	0.100057	0.000075	133.271415	-
Survey age	-0.914001	0.732144	-1.24839	0.211888	0.400917	0.095464	1.683713	1.867827
Navigational hazards	-0.004042	0.006317	-0.639797	0.522305	0.995967	0.983711	1.008375	1.78854
Bathymetry	-0.000004	0.000276	-0.013114	0.989537	0.999996	0.999456	1.000537	1.145088
Traffic density	0.000348	0.000129	2.695553	0.007027*	1.000348	1.000095	1.000602	1.077603
Atons	0.021859	0.029867	0.731878	0.464243	1.022099	0.963985	1.083717	1.740065
Current velocity	0.090804	0.882044	0.102947	0.918005	1.095055	0.194368	6.169459	1.347341
CATZOC	0.593922	0.530677	1.119178	0.263064	1.811077	0.640053	5.124575	1.046801
Age of vessel	13.825304	4075.428641	0.003392	0.997293	1009841.863	0.941245	1.000602	1.502486
Flag of vessel	25.26346	5005.806328	0.005047	0.995973	93708958281	0.847512	1.002513	1.509231

variations in spatial autocorrelation allowed for a nuanced depiction of incident-prone zones, particularly in areas where observed incident data were sparse or unevenly distributed. As a result, the model effectively contributes to a more comprehensive understanding of regional maritime safety dynamics and supports the development of targeted risk mitigation strategies.

6 Discussion

Maritime accidents, regardless of their nature, remain one of the most significant threats to seafarers and shipping operations (Maharaj, 2017). Whether occurring in confined or open waters, the risks are heightened in regions characterized by dense vessel traffic, poorly charted waters, inadequate AtoNs, and the presence of navigational hazards. Despite advancements in satellite-based navigation and precision technology, maritime casualties continue to occur (Calle & Alves, 2015). As such, the ability to model and predict areas vulnerable to vessel incidents is essential to reducing risk and improving maritime safety. This research demonstrated that specific causal factors when concentrated geographically can significantly increase the probability of incidents, thereby creating identifiable hotspots across the Caribbean Sea. To assess these relationships and predict incident-prone areas, four spatial modelling techniques were applied: GWR, GLR, FBCR, and EBK. GWR, GLR, and FBCR produced localized parameter estimates, enabling the capture of spatial heterogeneity and revealing complex relationships between maritime incidents and their explanatory variables (Gao & Li, 2011). Key variables such as traffic density, flag state, vessel age, CATZOC, and survey age emerged as dominant contributors across all models. GWR highlighted the influence of current velocity, GLR identified flag state as the most significant statistical contributor, and FBCR emphasized traffic density as the most impactful variable. Although EBK did not produce explicit parameter estimates, it delivered reliable spatial predictions that closely mirrored the outputs of the other models.

All four techniques identified high-risk zones, particularly around the Panama Canal, Gulf of Paria,

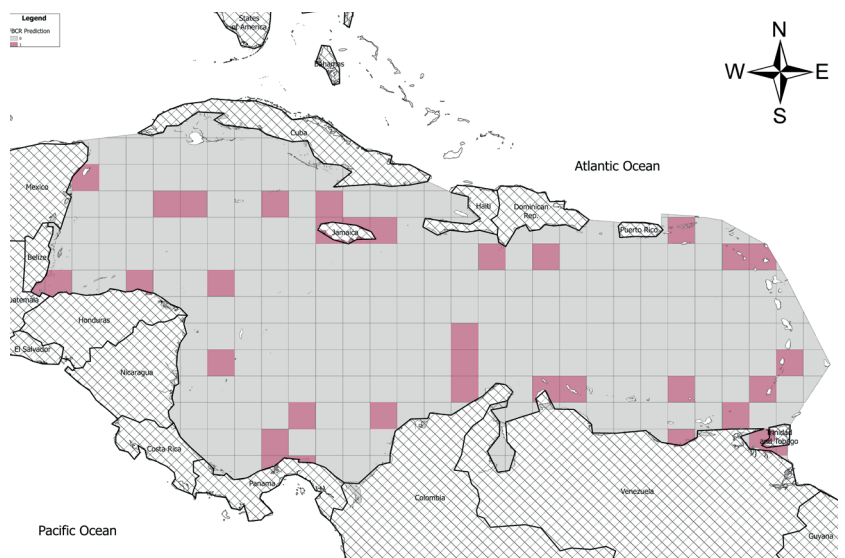


Fig. 4 Prediction surface generated using FBCR regression.

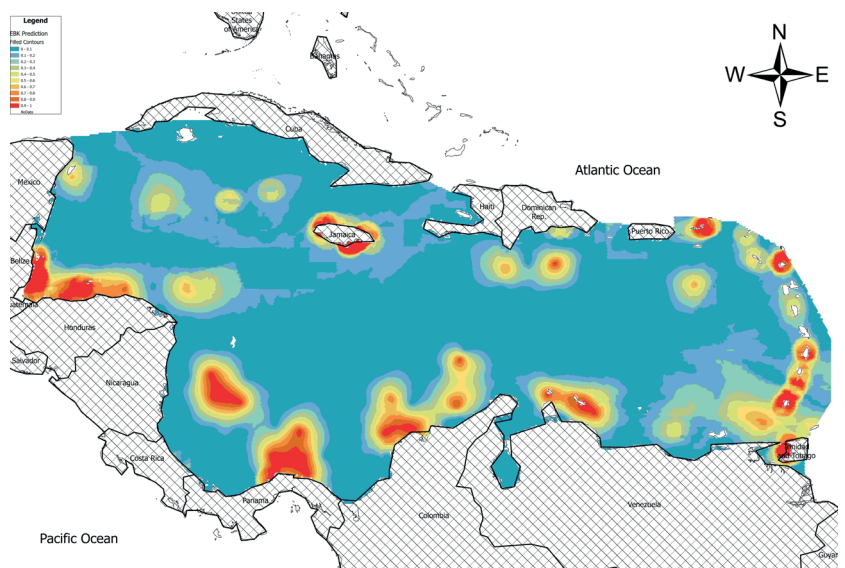


Fig. 5 Prediction surface generated using EBK regression.

Trinidad, Grenada, St. Lucia, Jamaica, and the British Virgin Islands. While GWR, GLR, and FBCR were effective in revealing variable-specific contributions

and spatial trends, their ability to account for highly localized dynamics was limited. EBK, on the other hand, excelled in generating spatially adaptive predictions without requiring complex model configurations, making it especially useful when incident data are sparse or exhibit moderate nonstationarity. However, each method showed notable limitations. GLR assumes a linear relationship and a normally distributed dependent variable assumptions not satisfied by the maritime incident data. While FBCR was more suited to non-linear relationships, its performance can be hindered by data sparsity and overfitting. EBK, although spatially flexible, does not estimate the strength or direction of explanatory variable effects and therefore lacks interpretability for causal analysis. GWR, while spatially explicit, is sensitive to multicollinearity and may produce unstable coefficients in areas with limited data. The analysis further revealed that conventional regression models face challenges when applied to real-world maritime risk data. These models often require the aggregation of point-based incident data into spatial units, which can reduce precision. Moreover, their statistical assumptions and sensitivity to data quality can limit their applicability in operational decision-making. It was also observed that regression-based approaches are not well suited for complex, spatially variable phenomena, such as maritime incidents influenced by interacting environmental and operational factors.

More critically, the study confirmed several key limitations of the regression techniques:

1. The GIS-derived covariates, though essential for spatial modelling, were not always suitable for statistically robust regression outcomes.
2. Spatial heterogeneity was evident but poorly captured by traditional methods.
3. The models lacked capacity to make causal inferences or to identify interactions between variables.
4. Spatial autocorrelation introduced bias in traditional models, leading to potential overcounting effects.
5. Improvements in data quality, frequency, and granularity are needed to support more accurate spatial modelling of maritime incidents.

The findings of this research underscore the limitations of traditional regression models in capturing the full complexity of maritime incident risk. While techniques such as GWR, GLR, FBCR, and EBK are valuable for exploratory spatial analysis and identifying general trends, they fall short when it comes to modelling the intricate, non-linear, and multi-dimensional interactions that characterize real-world maritime environments. These methods are constrained by assumptions of linearity, normality, and spatial stationarity assumptions rarely satisfied in dynamic coastal and open-sea environment. As such, they should be regarded as supportive analytical tools rather than definitive predictors of risk. To address these

shortcomings, more sophisticated approaches and advanced modelling techniques can offer a powerful alternative (Maharaj et al., 2025). These alternative techniques are not bound by rigid statistical assumptions and are capable of learning complex, hidden relationships within the data. Their ability to model non-linear behaviours and adapt to spatial and operational variability makes them especially well-suited for risk prediction in heterogeneous maritime regions like the Caribbean Sea. Potential techniques include artificial intelligence and machine learning approaches such as gradient boosted trees, support vector machines, deep neural networks, and ensemble learning; spatio-temporal models such as Bayesian hierarchical frameworks and spatial panel regression; and hybrid approaches that combine AIS trajectory mining with environmental and operational data. Incorporating such advanced methods allows for a more robust, data-driven modelling framework, one that aligns more closely with the realities of maritime navigation and safety planning.

7 Conclusion

Understanding where and why maritime incidents occur is essential to improving navigational safety and risk mitigation across the Caribbean Sea. The application of spatial regression techniques, GWR, GLR, FBCR, and EBK offered valuable insights into the geographic distribution of risk and the influence of key operational and environmental variables. Factors such as traffic density, vessel flag state, age of navigational charts, and proximity to hazards consistently emerged as significant across the models. However, while these methods proved useful for exploratory spatial analysis and surface-level prediction, they fell short in capturing the deeper, non-linear, and interactive dynamics that characterize real-world maritime risk. Their reliance on assumptions of normality, linearity, and fixed spatial relationships limits their effectiveness, especially in complex and data-sparse environments like the Caribbean. The inability to model variable interactions or adapt to rapidly changing maritime conditions further constrains their operational utility. These findings highlight the importance of adopting more advanced modelling frameworks that go beyond traditional regression. Unlike regression models, advanced modelling techniques can accommodate non-linear, high dimensional relationships and uncover latent patterns within data, offering a more robust and adaptive tool for maritime risk prediction. In moving toward a hybrid spatial modelling approach one that combines the interpretability of regression techniques with the predictive capabilities of machine learning researchers and maritime authorities can develop more precise, targeted, and data informed strategies to enhance safety at sea. The lessons learned from this comparative analysis serve as a stepping stone for advancing predictive analytics in the maritime domain, with the ultimate goal of fostering safer, smarter navigation throughout the region.

References

- Awal, Z. I. and Hasegawa, K. (2017). A study on accident theories and application to maritime accidents. *Procedia Engineering*, 194, pp.298–306. <https://doi.org/10.1016/j.proeng.2017.08.149>
- Callaghan, K. and Chen, J. (2008). Revisiting the collinear data problem: An assessment of estimator 'ill-conditioning' in linear regression. *Practical Assessment, Research and Evaluation*, 13(5), pp. 1–6.
- Calle, M. A. G. and Alves, M. (2015). A review-analysis on material failure modeling in ship collision. *Ocean Engineering*, 106, pp.20–38. <https://doi.org/10.1016/j.oceaneng.2015.06.032>
- Charlton, M. and Fotheringham, A. S. (2009). Geographically weighted regression: White paper. *White Paper*, pp. 1–17. <https://doi.org/10.1111/1467-9884.00145>
- ESRI (2016). Esri Training – Getting started with ArcGIS Pro. <http://training.esri.com/gateway/index.cfm?fa=catalog.webCourseDetail&courseid=2889> (last accessed 26 March 2025).
- Fotheringham, A. S., Brunsdon, C. and Charlton, M. (2002). *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester: John Wiley & Sons.
- Gao, J. and Li, S. (2011). Detecting spatially non-stationary and scale-dependent relationships between urban landscape fragmentation and related factors using geographically weighted regression. *Applied Geography*, 31(1), pp. 292–302. <https://doi.org/10.1016/j.apgeog.2010.06.003>
- Jetz, W., Rahbek, C. and Lichstein, J. W. (2005). Local and global approaches to spatial data analysis in ecology. *Global Ecology and Biogeography*, 14(1), pp. 97–98. <https://doi.org/10.1111/j.1466-822X.2004.00129.x>
- Lin, C. H. and Wen, T. H. (2011). Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. *International Journal of Environmental Research and Public Health*, 8(7), pp. 2798–2815. <https://doi.org/10.3390/ijerph8072798>
- Mainuri, G. and Owino, J. O. (2017). Spatial variability of soil aggregate stability in a disturbed river watershed. *European Journal of Economics and Business Studies*, 3(3), pp. 278–290.
- Maharaj, A. (2017). *Geo-statistical analysis of marine traffic in the Gulf of Paria* [MSc thesis, Department of Geomatics Engineering and Land Management, The University of the West Indies].
- Maharaj, A., Miller, K., Davis, D. and Sutherland, M. (2025). Geostatistical analysis of maritime accidents: Identifying contributory factors and risk patterns in maritime navigation. *The International Hydrographic Review*, 31(2), pp. 102–121. <https://doi.org/10.58440/ihr-31-2-a10>
- Pfeiffer, D. U. (1996). Issues related to handling of spatial data. *Proceedings of the Epidemiology and State Veterinary Programmes, New Zealand Veterinary Association / Australian Veterinary Association Second Pan-Pacific Veterinary Conference*. Christchurch, New Zealand. ftp://131.252.97.79/Transfer/ES_Pubs/ESVal/spatial_statistics/issues_related_to_spatial_data.pdf (last accessed 10 November 2015).
- Scott, L. M. and Bennett, L. R. (2012). Modeling spatial relationships using regression analysis. *ESRI International User Conference*.
- Su, C. M., Chang, K. Y. and Cheng, C. Y. (2012). Fuzzy decision on optimal collision avoidance measures for ships in vessel traffic service. *Journal of Marine Science and Technology*, 20(1), pp. 38–48.
- Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*. New Jersey, NY: Wiley & Sons.
- Wang, H., Jiang, H. and Yin, L. (2013). Cause mechanism study to human factors in maritime accidents: Towards a complex system brittleness analysis approach. *Procedia - Social and Behavioral Sciences*, 96, pp. 723–727. <https://doi.org/10.1016/j.sbspro.2013.08.083>
- Weisent, J., Rohrbach, B. and Dunn, J. R. (2012). Socioeconomic determinants of geographic disparities in campylobacteriosis risk: A comparison of global and local modeling approaches. *International Journal of Health Geographics*, 11(1), p. 45.
- Wilmsmeier, G., Hoffmann, J. and Sanchez, R. J. (2006). The impact of port characteristics on international maritime transport costs. *Research in Transportation Economics*, 16, pp. 117–140. [https://doi.org/10.1016/S0739-8859\(06\)16006-0](https://doi.org/10.1016/S0739-8859(06)16006-0)
- Zhang, C., Tang, Y., Xu, X. and Kiely, G. (2011). Towards spatial geochemical modelling: Use of geographically weighted regression for mapping soil organic carbon contents in Ireland. *Applied Geochemistry*, 26(7), pp. 1239–1248. <https://doi.org/10.1016/j.apgeochem.2011.04.014>

Authors' biographies

Dr. Amrika Maharaj is a Lecturer in the Department of Geomatics Engineering & Land Management, Faculty of Engineering, The University of the West Indies. Her research focuses on spatial data analysis, geostatistics, and risk assessment, with emphasis on maritime navigation in the Caribbean Sea. Some of her current areas of research expertise include numerical modelling, GIS, remote sensing, hydrography, and maritime navigation studies.



Amrika Maharaj



Keith Miller

Dr. Keith Miller is a former lecturer at the University of the West Indies where he specialized in Geodesy and Hydrography, now retired and living in Trinidad and Tobago. He has worked as a lecturer at universities in the UK, Egypt and Australia. In a professional capacity, he was a member of the Chartered Institution of Civil Engineering Surveyors and served for eight years on the IBSC.



Dexter Davis

Dr. Dexter Davis is a Lecturer in Geomatics Engineering and Head Department of Geomatics Engineering & Land Management. His research is focused on engineering surveying, digital photogrammetry and digital mapping, geodesy and geodetic applications, as well as GNSS and hydrographic surveying. Some of his current areas of research expertise include sea level monitoring and low cost and disaster relief mapping.



Michael Sutherland

Dr. Michael Sutherland is currently an Adjunct Professor at the Department of Geodesy and Geomatics Engineering, University of New Brunswick. He was also a Professor of Land and Marine Administration Systems at the Department of Geomatics Engineering and Land Management, The University of the West Indies, St. Augustine Campus, Trinidad and Tobago. Dr. Sutherland has many years of direct and indirect geomatics-related research experience in the fields of GIS, multidimensional geospatial modelling, land and marine administration systems, climate change, and ocean governance.