

Article

Intelligent Analysis of Data Flows for Real-Time Classification of Traffic Incidents

Gary Reyes ^{1,2,3,4,*} , Roberto Tolozano-Benites ⁴ , Cristhina Ortega-Jaramillo ² , Christian Albia-Bazurto ² , Laura Lanzarini ³ , Waldo Hasperué ³ , Dayron Rumbaut ^{1,4}  and Julio Barzola-Monteses ^{1,2} 

- ¹ Artificial Intelligence Research Group, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador; drumbautr@ube.edu.ec (D.R.); jjbarzola@ube.edu.ec (J.B.-M.)
 - ² Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Cda. Universitaria Salvador Allende, Guayaquil 090514, Ecuador; cristhina.ortegaj@ug.edu.ec (C.O.-J.); christian.albiab@ug.edu.ec (C.A.-B.)
 - ³ Instituto de Investigación en Informática LIDI (Centro CICPBA), Facultad de Informática, Universidad Nacional de La Plata, Buenos Aires CP1900, Argentina; laural@lidi.info.unlp.edu.ar (L.L.); whasperue@lidi.info.unlp.edu.ar (W.H.)
 - ⁴ Carrera de Sistemas Inteligentes, Universidad Bolivariana del Ecuador, Campus Durán Km 5.5 vía Durán Yaguachi, Durán 092405, Ecuador; rtolozano@ube.edu.ec
- * Correspondence: gxreysz@ube.edu.ec

Abstract

Social media platforms have been established as relevant sources of real-time information for urban traffic analysis. This study proposes an intelligent framework for the classification and spatiotemporal analysis of traffic incidents based on semi-synthetic data streams constructed from historical geolocated seeds for controlled validation, utilizing real reports from platforms such as X and Telegram. The approach integrates adaptive machine learning and incremental density-based clustering. An Adaptive Random Forest (ARF) incremental classifier is used to identify the type of incident, allowing for continuous updating of the model in response to changes in traffic flow and concept drift. The classified events are then processed using DenStream, a clustering algorithm that incorporates a temporal decay mechanism designed to identify dynamic spatial patterns and discard older information. The evaluation is performed in a controlled streaming simulation environment that replicates the dynamics of cities such as Panama and Guayaquil. The proposed framework demonstrated robust quantitative performance, achieving a prequential accuracy of up to 86.4% and a weighted F1-score of 0.864 in the Panama scenario, maintaining high stability against semantic noise. The results suggest that this hybrid architecture is a highly viable approach for urban traffic monitoring, providing useful information for Intelligent Transportation Systems (ITS) by processing authentic social signals.

Keywords: traffic incidents; social media; data streams; incremental learning; conceptual drift; ARF; denStream; dynamic clustering; smart cities



Academic Editor: Antonio Comi

Received: 17 February 2026

Revised: 17 March 2026

Accepted: 21 March 2026

Published: 23 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

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

1. Introduction

The efficient management of urban traffic is a major challenge for contemporary cities, particularly in view of the steady increase in the number of vehicles and the burden this places on road networks. Congestion and traffic incidents, including collisions, roadworks, blockages, and unforeseen events, have far-reaching social, economic, and environmental repercussions [1]. These effects impact quality of life, productivity, and public safety, which has prompted authorities and research centers to adopt more accurate monitoring and more agile response strategies [2]. Under these circumstances, it is essential to promote

solutions that integrate different types of information and favor an operational view of the traffic phenomenon in real-time.

The traditional incident detection approach relies on fixed sensors and surveillance cameras which, despite their proven effectiveness, present structural limitations such as high implementation costs and coverage restricted mainly to primary arteries. These restrictions create information gaps in highly variable urban environments. To overcome these shortcomings and facilitate broader monitoring, social media platforms like X (formerly Twitter) and Telegram have emerged as vital complementary sources. User-generated content transforms citizens into active sensors, generating a constant flow of real-time data that directly addresses the limitations of physical infrastructure. This approach extends spatial coverage to secondary roads, significantly reduces monitoring costs, and delivers spontaneous reports with remarkable immediacy, capturing minor obstructions or specific risk situations often missed by conventional systems [3].

Effective use of these volumes of information requires tackling complex technical challenges. The speed, heterogeneity, and dynamic nature of messages on social media generate scenarios typical of stream mining, in which the phenomenon known as “concept drift”, the evolution of underlying patterns over time, is particularly critical [4]. Added to this are frequent practical obstacles: the widespread absence of georeferencing, the use of informal language, the presence of textual noise, and variability in the quality of reports. Overcoming these barriers requires the development of robust and adaptive algorithms, as well as validation frameworks that allow citizen information to be reliably integrated with instrumental and reproducible sources.

In this context, knowledge extraction from social and multidimensional data flows has been established as a critical line of research [5]. Recent studies [1] have validated the effectiveness of social networks for accident detection. However, the dynamic nature of these streams poses challenges that require real-time processing superior to static methods. In analogous domains of streaming data, ref. [6] demonstrated the operational superiority of ensemble-based online learning algorithms for handling large volumes of data. Additionally, the temporal variability of language requires robust adaptation mechanisms; the authors of [7], in their review of evolutionary strategies, noted that the mitigation of concept drift is essential to maintain model accuracy.

Although recent studies have demonstrated the viability of using social media to detect traffic incidents, highlighting work applied in specific urban contexts such as Panama City [2] and historical accident analyses [1], most of these approaches are based on static learning models. These models tend to experience rapid degradation in their predictive performance when faced with continuous data flows, due to the constant evolution of user vocabulary (concept drift) and variable urban conditions [4]. Furthermore, while some research has addressed online clustering and classification for general event detection [8], their simultaneous and adaptive application, integrating both textual semantics and spatial density specifically geared toward urban traffic management, remains an area that requires further exploration. The lack of hybrid architectures capable of autonomous updating limits the real applicability of these tools in the long term.

Based on this overview, the following research questions are formulated: How can traffic incidents be reliably detected using data generated on social media? How effective is the incorporation of adaptive algorithms in responding to flow variability and concept drift? How can clustering methods contribute to identifying spatiotemporal patterns that characterize the occurrence of traffic incidents? The answers to these questions will be left for the conclusion of this research. The main focus of this study is the development of a method to identify and classify types of traffic incidents based on a classification algorithm and a real-time clustering algorithm. This method will process traffic incident data in order

to classify and group the most affected areas accurately and efficiently. In this context, the objective of this work is not focused on maximizing traditional classification metrics in static scenarios, but rather on analyzing the behavior of the proposed approach under conditions typical of continuous data flows, characterized by semantic noise, temporal variability, and concept drift. The evaluation prioritizes performance stability over time, progressive adaptability, and the operational usefulness of the method for real-time traffic monitoring, above the individual accuracy of each classified message. Therefore, the contributions of this work can be summarized as follows:

- We propose an architecture that combines adaptive supervised classification Adaptive Random Forest (ARF) with density-based unsupervised clustering (DenStream), enabling simultaneous processing of text semantics and spatial location in real-time.
- Unlike static approaches, the method we propose incorporates a change detection algorithm based on adaptive windows (ADWIN) and temporal decay factors, ensuring that the model automatically evolves in response to changes in social media vocabulary and traffic dynamics.
- A “forgetting” mechanism was implemented in the clustering that allows the persistence and dissipation of incidents over time to be identified.
- We validate the proposal under a progressive (prequential) evaluation scheme that simulates a real production environment, overcoming the limitations of traditional cross-validation for continuous data flows.

The rest of this article is organized as follows: Section 2 delves into the relevant literature, analyzing related works and proposing various solutions; Section 3 articulates the details of the proposed method; Section 4 shows the results obtained; Section 5 presents a discussion of the results; and, finally, Section 6 presents the conclusions and outlines future work to be done.

2. Related Works

Initial research has explored the capabilities of traffic data derived from digital platforms and social networks [1–3,9–11], recognizing them as highly dynamic information flows [1,8,12] and useful for the early detection of traffic incidents [2,13–15]. In this field, different studies have applied machine learning methodologies [10–12,16,17], including neural networks and deep learning models in combination with ensemble techniques [18], to classify posts in real-time [10,11,13,14] and extract relevant signals for traffic management, taking advantage of the immediacy and considerable variability of user-generated content [1,2]. These approaches have proven effective in improving the accuracy of incident detection and prediction [10], overcoming the limitations of traditional techniques by better adapting to the complexity and speed of information flows on social media. In addition, the integration of these methodologies allows large volumes of unstructured data to be processed in real-time [8,9], which increases the effectiveness of urban traffic monitoring and management systems [2,10], providing better forecasting and response to critical events.

These approaches have widely used conventional classification techniques, such as Support Vector Machine (SVM) and Naive Bayes [8,11,19], to detect incidents based on short and erratic texts [8,11]. The results indicate that, although these methodologies allow significant patterns to be identified, their effectiveness may be limited by the rapid evolution of linguistic expressions and the variable nature of online data flows. At the same time, density-based clustering algorithms [20–25], such as DBSCAN, have made it possible to identify critical areas and spatial patterns of incidents [21], making them particularly effective for recognizing spontaneously arising traffic clusters [26,27].

The most recent research efforts have focused on approaches designed specifically for continuous data streams, addressing the need to process information at the speed at

which it is generated [8,18,28–30]. In this regard, frameworks have been proposed that combine incremental clustering algorithms with online classifiers to detect events in real-time [8,18,31] in a scalable and adaptable manner [8]. Specifically, certain architectures incorporate a Random Forest classifier as a supervised component [1,2,32,33], which takes advantage of its resilience to data noise and its ability to handle the different attributes of social media texts [28], thus reinforcing the synergy between classification models and dynamic clustering processes [8,18].

However, limitations associated with the nature of social data flows continue to prevail [2,8,19,34]. One of the main challenges is related to conceptual drift [2,4,18,28,34–37], as models need to be continuously adjusted to changes in incident reporting behaviors [4,28,34,35,38]. In addition, only a small portion of posts contain explicit geolocation metadata, which limits spatial accuracy and requires the use of textual geocoding processes [2] based on fuzzy matching and address normalization techniques to approximate the location of the reported incident [1,2]. Adding to this problem is the linguistic complexity typical of social media, characterized by abbreviations, errors, colloquialisms, hashtags, and emojis, which introduces an additional layer of noise into natural language processing (NLP) tasks [1,2,39]. Consequently, these inherent limitations translate directly into measurable performance gaps in existing static models. Specifically, unaddressed concept drift causes a rapid and sustained degradation in predictive accuracy and F1-scores over time, as the models fail to recognize new terminology. Simultaneously, extreme linguistic noise and the scarcity of explicit geolocation metadata significantly increase false-positive rates and spatial latency, severely limiting the operational reliability of these conventional architectures for real-time traffic monitoring.

The literature shows an evolution from static approaches based on supervised learning towards more adaptive and flow-oriented architectures, where the combination of classification, clustering, and geocoding techniques allows for a more comprehensive approach to the detection and analysis of traffic incidents. These advances pave the way for more robust and responsive systems, aligned with intelligent transportation system architectures and applications geared toward urban mobility decision-making, capable of operating continuously and adjusting to changing social data conditions in real-time. This study introduces an approach aimed at identifying and characterizing road incidents by combining adaptive classification and incremental clustering. Reports generated on social media are classified to determine the type of incident described and then geo-referenced or geo-inferred events are grouped to identify areas where spatio-temporal patterns of impact emerge. The dynamic updating of these groups ensures that the processed information consistently reflects the actual state of the system at any given moment, avoiding the accumulation of obsolete data and allowing for continuous and coherent analysis of urban traffic evolution. By integrating classification and clustering mechanisms with periodic updating processes, the proposed method effectively adapts to fluctuations in the flow of information and addresses the challenges identified in previous work.

3. Materials and Methods

Before detailing the specific components of the proposed architecture, it is essential to establish the theoretical justification for selecting the ARF as the core classification algorithm. In the context of traffic incident detection via social media, data streams present unique challenges: continuous high-speed flow, fixed incident categories, and severe concept drift due to the evolving nature of informal language. Table 1 presents a theoretical comparison between our selected approach and two highly relevant incremental learning paradigms: the Hoeffding Adaptive Tree (HAT), a standard single-model algorithm for

data streams [37], and iCaRL, a state-of-the-art ensemble framework for Class-Incremental Learning [40].

Table 1. Comparison of incremental learning approaches and concept drift management.

Algorithm/ Paradigm	Learning Style	Concept Drift Handling	Architecture Type	Robustness to Textual Noise
iCaRL	Class Incremental	Low (Focuses on avoiding catastrophic forgetting of old classes)	Neural Network/Exemplar	High
Hoeffding Adaptive Tree (HAT)	Data Incremental	High (Uses early drift detection)	Single Tree	Low (Vulnerable to unstructured text)
ARF	Data Incremental	High (Incorporates ADWIN sliding window algorithm)	Ensemble (Multiple Trees)	High (Reduces variance and noise)

As illustrated in Table 1, while all three algorithms operate incrementally, their design objectives differ significantly. iCaRL is an exceptional architecture for scenarios requiring the continuous learning of new categories over time [40]. However, for Intelligent Transportation Systems (ITS), the categories of incidents are typically static (e.g., accident, traffic, obstacle). The primary obstacle is not the emergence of new labels, but the concept drift within the existing ones [41]. On the other hand, while the Hoeffding Adaptive Tree (HAT) manages concept drift efficiently, its single-tree architecture lacks the necessary resilience to handle the sparsity and noise inherent in social media text. Consequently, ARF emerges as the most suitable algorithm for this framework. By combining the ADWIN sliding window algorithm with an ensemble architecture, it natively adapts to language variations while maintaining high predictive stability against informal textual noise, all within the constraints of real-time processing.

This study proposes an integrated method for the detection, classification, and spatiotemporal analysis of traffic incidents based on continuous flows of textual data from social networks, whose general architecture is presented in Figure 1. Given the dynamic, brief, and unstructured nature of this type of publication, the method is conceived as a continuous processing flow oriented toward real-time analysis and adaptation to language evolution. The process begins with the collection and preprocessing of messages, including text cleaning and normalization. The data is then analyzed using an ARF incremental classifier, which identifies reports associated with traffic incidents and adapts to the concept drift present in the stream. Events classified as incidents feed into a dynamic spatial clustering component based on DenStream, which incorporates a temporal decay mechanism to identify, maintain, and update spatial patterns of incident concentration. Finally, the results of the classification and clustering processes are integrated into a geospatial analysis and visualization module, providing an up-to-date and consistent representation of urban traffic conditions.

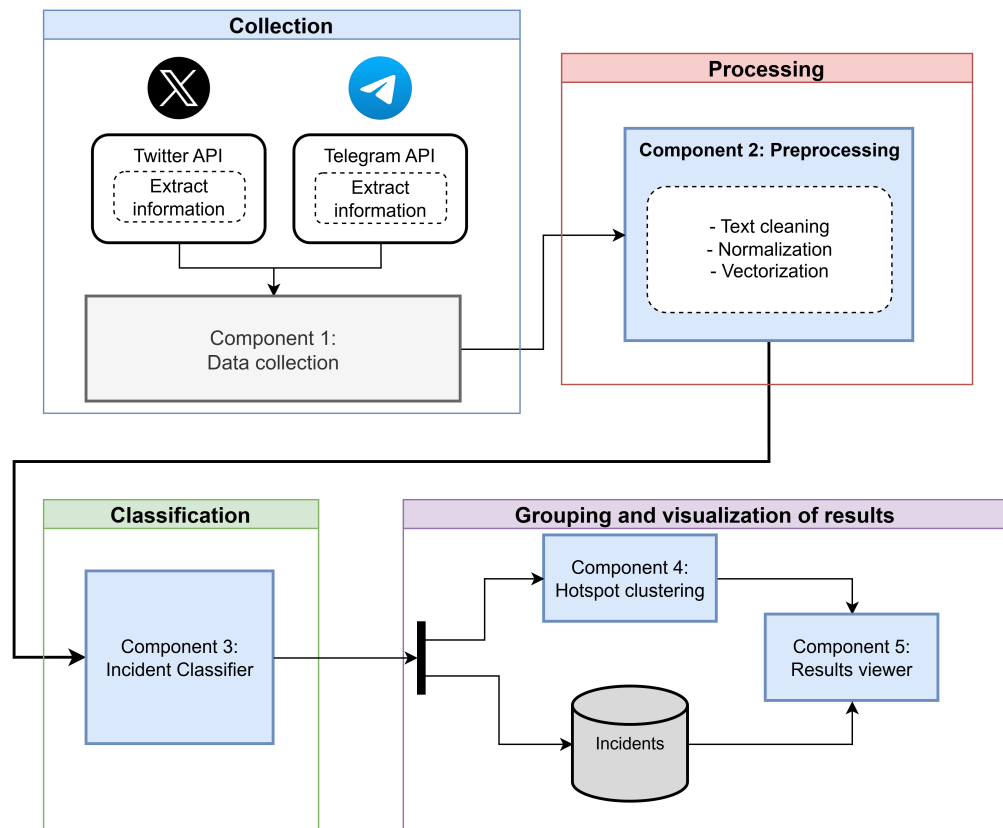


Figure 1. Components of the proposed method.

3.1. Component 1: Data Collection

The data collection is designed as a process aimed at capturing textual flows generated on social networks with high citizen participation around urban mobility, such as X and specialized Telegram channels or groups, where users spontaneously report traffic incidents, congestion, and relevant road events. Although the proposed approach is fully applicable to real data obtained directly from these platforms through continuous monitoring and keyword filtering mechanisms, this work uses controlled experimental datasets that reproduce the structural and dynamic characteristics of such real flows. These hybrid data flows preserve key aspects such as the temporal nature of the stream, the linguistic variability of the messages, the distribution of incident categories, and the presence of spatial information, allowing the behavior of the methodology to be evaluated in a controlled and reproducible environment. In this way, the collection component establishes a representative and consistent basis for the subsequent stages of preprocessing, incremental classification, and spatiotemporal analysis of traffic incidents.

3.2. Component 2: Preprocessing

The purpose of the preprocessing component is to adapt the textual flow from social networks for incremental analysis in real-time. At this stage, each message is processed sequentially using a set of basic operations aimed at reducing noise and preserving information relevant to the detection of traffic incidents.

Initially, the textual content is cleaned up, which includes removing web links, user mentions, hashtag symbols, special characters, and other non-alphabetic elements that do not provide direct semantic information about traffic events. This process mitigates the noise characteristic of the informal language used on social media.

Subsequently, the text is converted to lowercase in order to reduce lexical variability associated with spelling differences between messages. Next, simple tokenization is applied, in which each message is segmented into individual lexical units. During this stage, empty tokens generated after text cleaning are discarded, as well as fragments with no informational meaning that do not contribute to the characterization of traffic incidents.

The resulting textual representation is integrated into an Online Bag-of-Words scheme, where the vocabulary is not predefined, but is progressively constructed and updated as new messages enter the data stream. This approach allows each publication to be represented incrementally, without requiring prior knowledge of the entire corpus or recalculating the overall representation of the dataset.

As a result, each message is converted into a vector representation that can be immediately processed by the incremental classifier, preserving the temporal nature of the flow and allowing the method to continuously adapt to the evolution of language and citizen reporting patterns.

To ensure full experimental reproducibility, this Online Bag-of-Words scheme is configured to extract unigram frequencies (Term Frequency, TF) strictly incrementally, avoiding global weighting techniques like TF-IDF that would improperly assume prior knowledge of the dataset's overall distribution.

3.3. Component 3: Incident Classifier

The classification process is structured around a continuous flow of messages derived from real social media traces, organized into discrete time cycles for experimental purposes. In this work, a 1 h time window was considered, within which messages are progressively entered into 10 min sub-periods, with an approximate average of 500 posts per interval. This organization allows us to model the dynamic behavior of the flow and evaluate the performance of the classifier in conditions close to a real-time environment.

An incremental supervised classifier based on ARF was implemented on this flow, responsible for identifying messages corresponding to traffic incidents from those that do not report road events. Based on the vector representation generated in the preprocessing stage, each message is evaluated individually and sequentially by the classifier, which issues a multi-class prediction that assigns each message to one of the defined categories: accident, traffic, obstacle, or no incident.

While other incremental learning models, such as iCaRL [40], are highly effective for incorporating new classes over time (Class-Incremental Learning), our approach deals with a fixed set of traffic categories. The primary challenge in this context is the continuous concept drift in social media vocabulary, rather than the emergence of new labels. Consequently, ARF was selected because its architecture, combined with drift detection mechanisms, is specifically designed to adapt to rapid distributional changes in continuous text streams. Although ensemble methods are often considered 'black boxes' and can theoretically increase prediction times, interpretability at the individual message level is secondary to predictive robustness in our framework. To mitigate any potential latency in real-time processing without sacrificing performance, the ensemble parameters were carefully bounded.

The ARF classifier was configured as an ensemble of 15 base models, corresponding to incremental decision trees trained in parallel. This structure allows complementary linguistic patterns present in citizen reports to be captured and improves the stability of the classification in the face of noise inherent in informal language. The maximum depth of the trees was limited in order to reduce overfitting and maintain a balance between predictive power and computational efficiency.

To address changes in flow distribution over time, the ADWIN (ADaptive WINdowing) sliding window algorithm was incorporated for both early warning and conceptual drift detection. These mechanisms allow for the identification of statistically significant variations in language and reporting patterns observed over time cycles, facilitating the progressive adaptation of the model without interrupting the classification process.

As a result of this phase, only messages classified as traffic incidents are preserved and sent to the next component, while the rest of the flow is discarded. This strategy reduces the volume of data processed in subsequent stages and ensures that spatio-temporal analysis is performed exclusively on information that is relevant and up-to-date for urban traffic management.

To explicitly clarify the data flow and dependencies between the components: the ARF classification module acts as a strict semantic filter for the subsequent clustering phase. Once the ARF model predicts an incoming message as a valid road incident (i.e., accident, obstacle, or traffic), the system extracts its geolocation metadata (latitude and longitude) and timestamp. The original textual payload and vector representation are then discarded to optimize memory. Finally, these extracted spatial-temporal coordinates are immediately forwarded as a continuous, noise-free numerical stream to the DenStream algorithm. This sequential pipeline ensures that the unsupervised clustering component is completely dependent on, and fed exclusively by, the semantically validated output of the supervised classifier.

3.4. Component 4: Hotspot Clustering

The clustering component aims to identify dynamic spatial concentrations of traffic incidents from previously classified messages. Each incident report contributes its geographic coordinates to the spatial analysis, serving as the basis for the incremental detection of clusters that reflect recent developments in urban dynamics. Once a message is classified as an incident, its latitude and longitude coordinates are incrementally incorporated into the clustering model.

The process is implemented using the DenStream algorithm, which allows continuous spatial data flows to be managed without requiring global reprocessing or complete storage of the event history. It is important to note that DenStream retains only summary descriptors (micro-clusters) rather than tracking discrete entities (individual messages) over time. This intentional abstraction is highly suitable for Intelligent Transportation Systems (ITS), as it prioritizes spatial and temporal event density, ensuring the algorithm remains computationally lightweight while inherently protecting citizen data privacy. In the first stage, each new point feeds into the structure of potential micro-clusters, where its spatial proximity to existing micro-clusters is evaluated based on a predefined neighborhood radius. If the point is within the region of influence of a micro-cluster, the latter is updated, reinforcing its density; otherwise, a new atypical micro-cluster (outlier) is created, representing a possible emerging concentration of incidents.

Progressively, atypical micro-clusters that reach a minimum density level and receive continuous reinforcement over time are promoted to consolidated micro-clusters, which constitute the active spatial clusters of the method. At the same time, DenStream incorporates a temporal decay mechanism that reduces the influence of old events, allowing micro-clusters that no longer receive new reports to lose relevance and eventually be discarded as spatial noise.

The DenStream configuration was adjusted to capture relevant spatial concentrations incrementally, controlling both spatial proximity and the minimum density required for cluster formation. Among the most significant parameters are the neighborhood radius (ϵ), the minimum density threshold (μ), and the decay factor, which determined the consolida-

tion of micro-clusters and the dynamic updating of clusters, ensuring that they reflected the recent evolution of the incident flow and progressively attenuated the influence of older events.

As shown in Figure 2, the clustering process is applied sequentially to the classified incident flows, allowing for the analysis of spatial patterns associated with different types of road events, such as traffic, accidents, and obstacles. As a result, dynamic spatio-temporal clusters are obtained, characterized by their representative location, density of reports, and persistence over time, which form the basis for geospatial analysis and visualization of urban traffic hotspots, where the scale bar indicates the distance in meters to provide a detailed view of the localized micro-clusters.

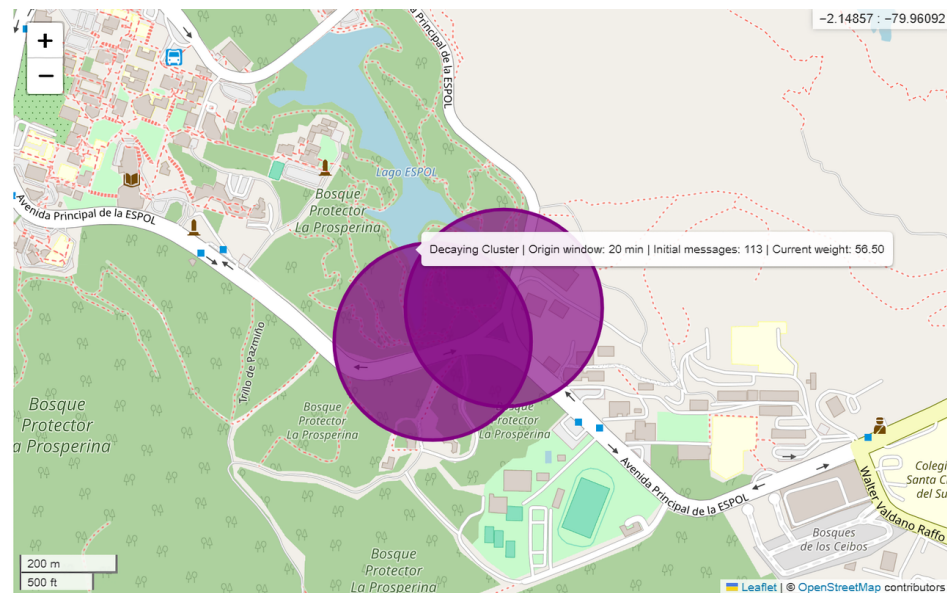


Figure 2. Formed Clusters (Scale in meters to show a localized view).

Temporary Decay Mechanism

Given that the approach operates on continuous flows of geolocated reports, a temporal decay mechanism is incorporated into the spatial clustering process using DenStream. This mechanism regulates the influence of incidents based on elapsed time and the continuity of new events in their spatial vicinity. In the implemented model, each incident contributes to the density of a micro-cluster with a weight that decreases progressively according to the defined temporal decay factor $\lambda = 0.01$. If a micro-cluster stops receiving new reports nearby within the established neighborhood radius, the cumulative density associated with its incidents gradually attenuates.

As a result, micro-clusters that are not reinforced by recent events lose relevance and may fall below the minimum density thresholds, being treated as spatial noise or automatically discarded by the algorithm. In this context, information considered obsolete corresponds to old incidents whose contribution to the density of the cluster has decreased due to the absence of recent activity in the same geographical area.

This behavior allows active spatial clusters to dynamically reflect the current state of urban traffic, strengthening those areas where incidents persist over time and progressively weakening areas where events have ceased to occur. In this way, the method prioritizes recent and relevant information, continuously adapting to the evolution of urban dynamics and facilitating the identification of current hot spots for analysis and real-time decision-making.

3.5. Component 5: Results Viewer

The visualization component is the final component of the method and aims to integrate the results of classification and spatial clustering to provide a clear and up-to-date representation of urban traffic conditions. In this part, the clusters generated by incremental analysis are displayed on a geospatial map, allowing the observation of concentrations of events that reflect critical patterns of road impact. The visualization is updated periodically as the data flow progresses, showing the temporal evolution of the clusters, their persistence or decay, and the emergence of new areas of interest. These clusters, originating from reported incidents (traffic, obstacle, or accident), are represented by graphic elements that indicate their location, spatial extent, and relative activity level. In addition, by consolidating the observation time, the interface provides information labels for each cluster visible on the map. These labels detail key attributes, such as the status of the cluster, the time window of origin, the number of reports contained, and their weight. This approach allows for intuitive analysis of the relationship between time, space, and incident type, acting as a bridge between automated processing and decision-making in traffic management and intelligent transportation systems contexts. As shown in Figure 3, the scale bar indicates the distance in kilometers to show the large-scale distribution of micro clusters throughout the city.

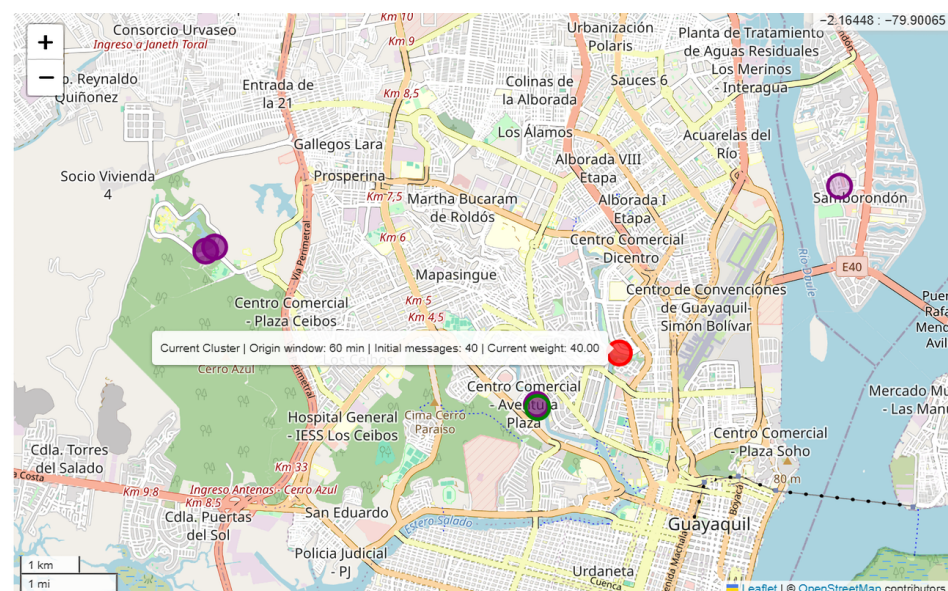


Figure 3. Clusters formed on the map (Scale in km). Status: decay (purple), previous (green), current (red).

4. Results

This section presents the findings obtained after implementing the proposed architecture. The evaluation is structured in two main phases: first, the composition of the datasets used for controlled validation is detailed; and second, the performance of the selected algorithms is analyzed. The performance of the ARF incremental classifier is examined using accuracy metrics and confusion matrices, while the effectiveness of spatial clustering with DenStream is validated through the consistency of the detected clusters and their temporal evolution in the Panama and Guayaquil scenarios.

4.1. Data Used

Two datasets were used, totaling 6000 processed records distributed equally between the two study scenarios (3000 for Panama and 3000 for Guayaquil). Both datasets share a standardized structure to ensure reproducibility of the experiments and consistency in flow

intake. Each record includes the following key metadata for processing: unique identifier (tweet_id), timestamp associated with the simulated flow (tweet_created), actual geospatial coordinates (latitude, longitude), textual content of the message or report (text), and the class label (category) that serves as ground truth for supervised validation (classified as: Accident, Traffic, Obstacle, No Incident).

Both sets were generated using a simulation strategy based on real traces, preserving statistical properties observed in historical data, such as the temporal distribution of events, spatial density on the main road network, and the relative proportion between categories. The semi-synthetic datasets are publicly available through the GitHub repository mentioned in the Data Availability Statement.

4.1.1. Panama Dataset

A semi-synthetic data flow designed to replicate the urban mobility dynamics of Panama City was used. Unlike purely artificial datasets, this dataset was constructed from historical seeds derived from a corpus of traffic reported in previous studies, based on georeferenced publications from the X platform. These seeds served as a reference for modeling and preserving real patterns of temporal distribution, spatial density, and linguistic variability observed in citizen reports.

From these seeds, 3000 semi-synthetic records were generated, simulating scenarios of high information density and allowing the algorithm's capacity to be evaluated in the face of load peaks and semantic noise.

4.1.2. Guayaquil Dataset

For the Guayaquil, Ecuador scenario, a semi-synthetic dataset derived from real seeds was constructed, integrating the city's authentic road topology with semantic content from real citizen reports.

The generation procedure was based on collecting seeds extracted directly from platform X (formerly Twitter) and from traffic monitoring groups during the period 2025, capturing regional linguistic variability, local idioms, and non-standard abbreviations. From these seeds, 3000 semi-synthetic records were generated, preserving geographical consistency and semantic fidelity, and ensuring experimental control and reproducibility.

4.2. Method Parameter Selection

The selection of the ARF classifier hyperparameters was carried out through preliminary experiments (warm-up phase) using the first 10% of the data flow, with the aim of identifying a configuration that offered an adequate balance between accuracy, stability, and adaptability to conceptual drift. Given the incremental nature of the problem, traditional cross-validation schemes were not applied; instead, a prequential evaluation was used, in which each instance is first used for evaluation and then for model updating.

The number of base models (n models) was set at 15, following preliminary tests that allowed us to find an appropriate balance between prediction stability and computational efficiency. This configuration enabled the set of models to capture the diversity of linguistic patterns present in the message flow, while maintaining a responsive response to textual variability.

The maximum depth of the trees (max depth) was adjusted to control overfitting in the face of the high dimensionality and sparsity of the Bag-of-Words representation. Reduced depths limited the model's discrimination capacity, while excessive values tended to capture patterns that were not very general. Based on preliminary experiments, a maximum depth of 30 levels was set, which allowed relevant patterns to be preserved without amplifying noise.

The sampling hyperparameter (λ) was set to 6, a value that allowed diversity among the trees in the set to be maintained and ensured stability in the incremental updating of the classifier. It was found that lower values reduced the models' ability to capture different patterns in the text flow, while higher values could generate instability in early adaptation.

To detect changes in flow distribution, the ADWIN sliding window algorithm was used for both warning and drift detection, selected for their ability to identify statistically significant variations without requiring predefined thresholds. This configuration allowed for timely reactions to changes in language and reporting patterns observed throughout the time windows.

Finally, a random seed (seed = 42) was set to ensure the reproducibility of the experiments. The final configuration of the ARF classifier remained fixed for all experiments reported in Section 4, being applied uniformly to both experimental datasets used in the study. Table 2 summarizes the final configuration of the ARF classifier used in all reported experiments.

Table 2. ARF classifier parameters.

Hyperparameters	Value
n_models	15
max_depth	30
lambda_value	6
drift_detector	ADWIN
warning_detector	ADWIN
seed	42

The DenStream parameters were defined considering the incremental nature of the incident flow, the geographical scale of the urban area analyzed, and the need to identify dynamic spatial concentrations in real-time. The configuration adopted seeks to balance spatial sensitivity, cluster stability, and adaptability to changes in the distribution of events.

The spatial proximity parameter $\epsilon = 0.002$ was selected to reflect the geographical proximity between reports within the urban context of the cities analyzed. This value allows incidents occurring at short distances to be grouped together, avoiding the merging of unrelated events and favoring the detection of localized hot spots.

The minimum density threshold $\mu = 4$ was defined with the aim of ensuring that a cluster represents a significant concentration of incidents, discarding clusters formed by a few isolated events. This value was determined after preliminary tests, observing that lower values tended to generate spurious clusters, while higher values delayed the detection of emerging critical areas.

The parameter $\beta = 0.75$ was used to determine the weight threshold necessary to promote an atypical micro-cluster to the category of consolidated (or potential) micro-cluster. This configuration allowed only those clusters that receive consistent reinforcement over time to remain active, facilitating the elimination of spatial noise.

The number of initial samples $n_{\text{init}} = 50$ was set to allow the formation of sufficiently stable initial micro-clusters before starting the complete incremental process, reducing sensitivity to early fluctuations in the flow. Likewise, the flow rate $v = 100$ was defined to match the model update rate to the incident arrival rate observed in the simulated data.

The final DenStream configuration was kept fixed for all experiments reported in Section 4, being applied uniformly to both test datasets. Table 3 summarizes the final configuration of the DenStream algorithm used in all reported experiments.

Table 3. DenStream algorithm parameters.

Parameter	Value
epsilon (ϵ)	0.002
mu (μ)	4
beta β	0.75
decaying_factor	0.01
n_samples_init	50
stream_speed	100

4.3. Comparison Method

The proposed method was evaluated by comparing the predictions generated by the model with the actual labels present in the data streams, using a prequential evaluation scheme suitable for streaming learning scenarios. In the incremental classification approach, each incoming message was processed sequentially using ARF to predict the event category based on its semantic content. Each prediction was then immediately compared with the corresponding actual label before updating the model, allowing its performance to be measured continuously under realistic operating conditions.

The labeled datasets from Panama and Guayaquil were used as a reference for this comparison, ensuring that the evaluation was based on direct agreement between the model's predictions and the reality represented in the data. Based on this sequential comparison, standard incremental classification metrics were calculated, including precision, recall, and F1-score, as well as confusion matrices, which allowed the model's ability to correctly identify different types of incidents in the flow to be evaluated.

In addition, spatial clustering was performed using DenStream, which allowed dynamic micro-clusters to be identified based on the geographical distribution of events. The spatial coherence of the detected clusters was analyzed in relation to the concentration of incidents present in the data, allowing us to observe the formation, evolution, and persistence of high-density areas over time. This analysis complements the evaluation of the classifier by providing a spatial representation of the patterns detected in the data flow.

The application of this evaluation scheme to the Guayaquil and Panama datasets made it possible to analyze the ability of the proposed method to maintain consistent behavior in different urban contexts, verifying its adaptability and consistency with the reality represented in the data flows without requiring structural adjustments or modifications to the model configuration.

4.4. Method Validation

In order to validate the robustness and adaptability of the proposed method, a sequential experimental scheme was designed based on the incremental processing of two simulated data streams corresponding to the cities of Panama and Guayaquil. This approach allowed the behavior of the proposal to be analyzed in relation to different spatial configurations, event densities, and reporting patterns, while maintaining the model configuration unchanged.

In the first stage, the method was evaluated using the continuous data stream from the semi-synthetic Panama dataset, which allowed for analysis of the baseline performance of the incremental classifier and the behavior of the DenStream algorithm in the formation and evolution of spatial micro-clusters. During this phase, the approach's ability to identify coherent spatial concentrations of incidents and to discard isolated reports as noise was observed, validating the expected performance of the method in a controlled environment.

Subsequently, validation was extended to the data flow corresponding to the city of Guayaquil, with the aim of analyzing the response of the proposal to a change in geographical context and a different spatial distribution of incidents. This phase allowed us to

evaluate the adaptability of the method without introducing modifications to the parameters or the structure of the architecture, demonstrating that the model is not overfitted to a single urban topology.

Throughout both experiments, the temporal evolution of spatial clusters was analyzed, observing their formation, persistence, and dissolution as new reports entered the flow. The results obtained in both scenarios confirm that the proposed method maintains stable and consistent behavior in the face of changes in the distribution of events, showing promising performance in adapting to traffic dynamics in controlled experimental environments.

4.5. Results Obtained

This section presents the results obtained from applying the proposed method, which combines supervised classification and unsupervised clustering techniques for the analysis of urban incidents. The experiments were conducted using different datasets in order to evaluate the model's performance in both controlled scenarios and real-world contexts.

In particular, two datasets were analyzed: a dataset based on real traces corresponding to the city of Guayaquil, constructed from geolocated reports representing urban mobility events, and a control scenario associated with the city of Panama, designed to validate the performance of the proposal under controlled and reproducible conditions. For both datasets, the results of the classification phase are presented, including confusion matrices and performance metrics, as well as the results of the spatial clustering process through the visualization of the generated clusters. The datasets used in the study are detailed below and summarized in Table 4.

Table 4. Summary of synthetic datasets used.

Dataset	Source	Records	Categories
Panama Semi-synthetic	X/Telegram	3000	Accident, Non-incident, Obstacle, and traffic
Guayaquil Semi-synthetic	X/Telegram	3000	Accident, Non-incident, Obstacle, and traffic

4.5.1. Results with the Guayaquil Dataset

The first set of experiments was conducted using a synthetic hybrid corresponding to the city of Guayaquil, generated to simulate urban incident reports with associated textual and geographic information. This dataset incorporates controlled variability, allowing the model's performance to be evaluated in a scenario designed to emulate the city's information dynamics. The unstructured nature of the data introduces specific challenges, such as the use of colloquial language, abbreviations, and ambiguous descriptions, as well as a high density of messages that must be effectively filtered by the classifier.

The classification of incidents was carried out using the ARF model. The results of this stage are presented through the confusion matrix shown in Figure 4, which shows the distribution of the model's predictions against the actual classes of incidents generated for the Guayaquil environment. This multi-class classification phase is essential for identifying specific mobility patterns, differentiating between accidents, obstacles, traffic, and other road events.

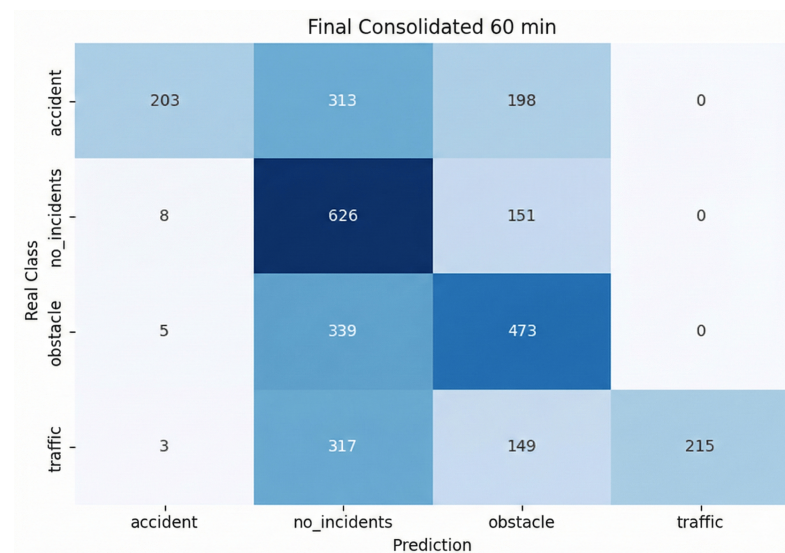


Figure 4. Confusion matrix of the Guayaquil ARF model.

In Figure 4, we can see that the classifier correctly identifies 626 reports corresponding to the “no incidents” category, which is the class with the highest number of correct predictions. This result demonstrates the model’s solid ability to filter out irrelevant posts from the information flow. Similarly, there are 473 correct classifications in the “obstacle” category, followed by 215 correct classifications in the “traffic” category and 203 in the “accident” category.

However, the matrix reveals a considerable level of confusion between categories, due to the semantic complexity of the evaluated dataset. In particular, it is observed that 313 actual reports of “accident” and 339 of “obstacle” were misclassified as “no incidents.” This behavior suggests that, when faced with descriptions with a low level of detail or intentional ambiguity in the processed texts, the model tends to make conservative predictions toward the filtering class.

Likewise, cross-confusion is evident, especially in the “traffic” category, where 317 actual reports were assigned to “no incidents” and 149 to “obstacle.” This pattern is consistent with the nature of the language used on social media, even in simulated scenarios, where congestion, partial blockages, or delays are often described using generic terms that make it difficult to accurately distinguish between routine traffic and physical events on the road. Despite these confusions, the model maintains sufficient accuracy to feed into the subsequent stages of spatial analysis.

To complement the confusion matrix-based analysis, Table 5 presents the overall performance metrics of the ARF classifier obtained for the Guayaquil hybrid dataset, considering precision, recall, F1-score, and support per category.

Table 5. Classification model results for the Guayaquil dataset.

Category	Precision	Recall	F1-Score	Support
Accident	0.927	0.284	0.435	714
No incidents	0.392	0.797	0.526	785
Obstacle	0.487	0.579	0.529	817
Traffic	1.000	0.314	0.478	684
Accuracy		0.506		3000
Macro avg	0.702	0.494	0.492	3000
Weighted avg	0.684	0.506	0.494	3000

As shown in Table 5, the model achieves an overall accuracy of 50.6% over a total of 3000 processed instances. At the class level, the traffic category shows perfect accuracy (1.000), indicating that all instances classified under this label correspond to congestion events; however, its low recall value (0.314) reflects the model's limited ability to retrieve all simulated traffic reports, evidencing the classifier's restrictive behavior.

A similar pattern is observed in the accident class, which achieves high precision (0.927) but low recall (0.284), suggesting that the model prioritizes minimizing false positives for critical events at the expense of omitting a significant fraction of these incidents. On the other hand, the non-incident category has the highest recall value (0.797), confirming the effectiveness of the model in the initial filtering phase, although with lower accuracy (0.392) due to the displacement of other categories towards this class. The obstacle class presents the most balanced performance with an F1-score of 0.529, reflecting a moderate ability to identify this type of event in a complex evaluation environment. Overall, the macro averages (0.702 in precision) highlight the difficulties inherent in classification in highly variable urban contexts, where semantic similarity between categories limits the balanced performance of the ARF model.

It is important to note that the accuracy value observed in this controlled scenario should not be interpreted as a limitation of the proposed approach, but rather as a direct consequence of the incremental nature of the data flow and the conservative strategy adopted by the model. In highly dynamic urban contexts characterized by semantic ambiguity, the classifier prioritizes the reduction of false positives and the stability of the filtering process, even at the expense of sacrificing overall accuracy. This behavior is consistent with the objective of the method, whose purpose is not to maximize the individual classification of messages, but to ensure that the events used in the later stages of spatial analysis are relevant, reliable, and stable over time.

For the spatial analysis of the Guayaquil dataset, the geographic coordinates associated with reports classified as incidents by the ARF model were used. In order to analyze the temporal evolution of the reported events, the data were processed in an incremental input stream, where each 10 min interval incorporates approximately 500 new messages, simulating a continuous real-time monitoring scenario. In the first observation interval (0–10 min), for example, the flow consisted of 85 accident reports, 44 obstacle reports, and 316 traffic reports, which allows us to initially observe the operational load of the approach and the early distribution of the types of incidents detected. Figure 5 shows the evolution of this flow.

During the initial phase, shown in Figure 5a,b, there is a dispersion of reports that begins to concentrate in an incipient manner. In these first 20 min, the method begins to outline the location of events on road corridors, where traffic reports (blue) clearly predominate, followed by a few isolated accidents (red) and a minimal presence of obstacles (orange), which mark the beginning of detected activity on the road network.

Subsequently, in the central intervals of Figure 5c,d, it is evident that the accumulation of incidents reaches its maximum saturation point. In this phase, the density of points clearly outlines the city's main avenues and roadways, where the overlap of accident and congestion reports suggests that the initial events have escalated, severely affecting the flow of traffic on the busiest arteries.

Finally, when observing Figure 5e,f, a notable reduction in point density can be seen. This decrease in the arrival of new messages in the last 20 min indicates a transition in the intensity of the information flow within the monitoring window. However, the persistence of certain nodes on the main roads allows us to identify areas of recurrence that remain active, demonstrating that certain critical incidents have a prolonged impact on the road

network, which the algorithm continues to track effectively despite the decrease in the volume of new data.

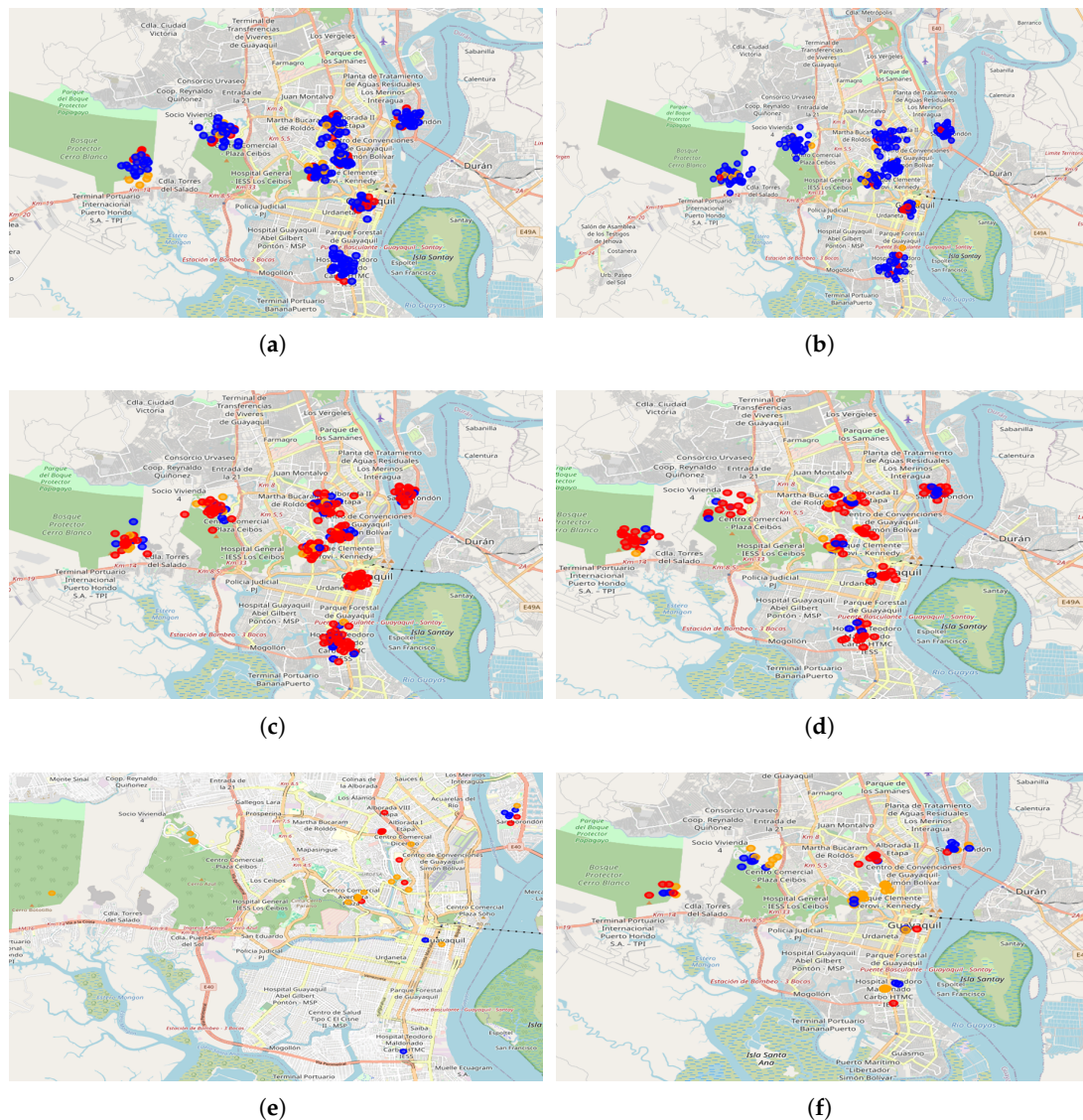


Figure 5. Spatial evolution of traffic incidents in 10 min intervals: (a) 0–10 min. (b) 10–20 min. (c) 20–30 min. (d) 30–40 min. (e) 40–50 min. (f) 50–60 min. The colors represent: traffic (blue), accidents (red), and obstacles (orange).

These visualizations show how incremental report management allows for the identification of areas with a higher recurrence of incidents in real-time, reflecting significant variations in both type and density throughout the entire observation hour.

In order to complement the spatial analysis based on the point distribution of reports, the evolution of the clusters generated by the DenStream algorithm is analyzed below. While Figure 5 shows the dispersion and progressive concentration of individual incidents, Figure 6 summarizes this behavior by identifying dense clusters that represent areas of persistent traffic conflict. These clusters emerge dynamically as the density of events in the flow increases and allow the abstraction of specific information into more stable and representative spatial structures for real-time monitoring.

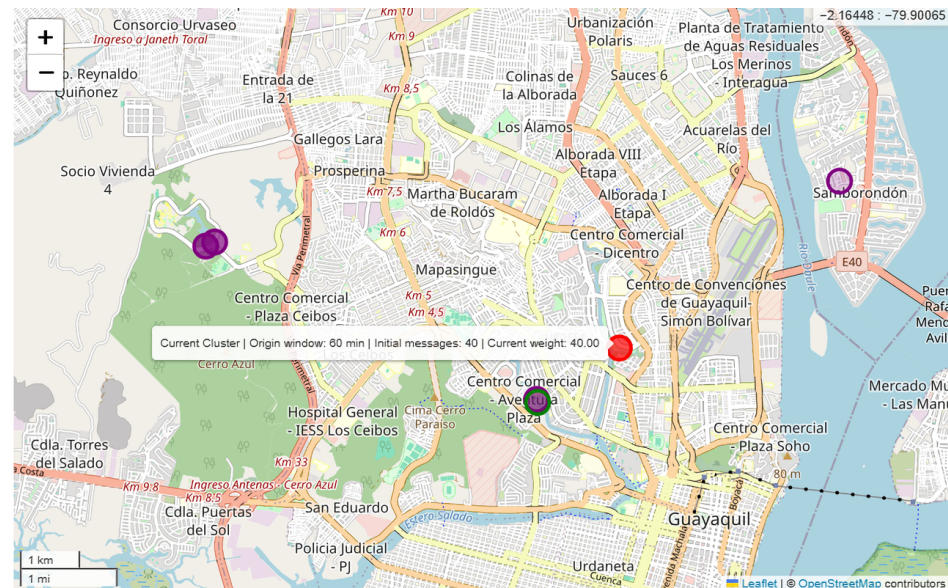


Figure 6. Evolution of clusters in Guayaquil (Scale in km). Status: decay (purple), previous (green), current (red).

As shown in Figure 6, a global visualization of the spatial clustering status after 60 min of continuous processing is presented, where the scale bar indicates the distance in kilometers. This figure simultaneously shows the clusters detected from the first 10 min to the final snapshot, differentiated by their internal status within the model. The clusters represented in purple correspond to dense centers in the decay phase, while the cluster in green represents the center detected in the previous snapshot (50 min). The cluster highlighted in red corresponds to the current dense center, generated after 60 min of processing, which concentrates 40 reports within the neighborhood radius ϵ and has a weight $w = 40.00$. This behavior demonstrates DenStream's ability to maintain memory of past events and progressively update spatial concentrations as new reports are incorporated. Although the visualization mainly emphasizes the state of the most recent cluster, the model incrementally identifies multiple dense centers throughout the flow, approximately one every 10 min, reflecting changes in both the geographic location and semantic composition of the incidents.

The values mentioned above are summarized quantitatively in Table 6, which shows, for each temporal snapshot of the flow (in minutes), the dense center detected, the status of the cluster within the model, the number of cluster groups formed, the number of reports contained within the neighborhood radius (ϵ), the weight associated with the cluster, the distribution of reports by category, and the dominant category of each cluster. This table clearly shows the temporal evolution of dense centers, complementing the information presented in Figure 6 and facilitating the interpretation of incident concentration patterns in the city.

Table 6. Dense centers detected by DenStream for the Guayaquil dataset.

Snapshot (min)	State	Group	Reports ϵ	Weight (w)	Category
10	Decay	1	87	36.58	Traffic
20	Decay	1	113	56.50	Obstacle
30	Decay	1	61	36.27	Obstacle
40	Decay	1	16	11.31	Traffic
50	Decay	1	6	5.05	Obstacle
60	Actual	1	40	40.00	Obstacle

Table 6 allows for a detailed analysis of the temporal evolution of dense centers detected by DenStream throughout the data flow corresponding to the city of Guayaquil. In the initial 10 min snapshot, a dense cluster with 87 reports within the neighborhood radius ϵ and a weight $w = 36.58$ is identified, dominated by the traffic category (72 reports), reflecting an early phase characterized by recurring traffic congestion. At 20 min, the cluster significantly increases its density to 113 reports within ϵ and reaches the highest weight observed ($w = 56.50$), with a clear transition to the obstacle category as dominant, a trend that consolidates at 30 min with a dense cluster composed exclusively of obstacle reports. In the 40 and 50 min snapshots, a progressive decrease in both the number of reports and the weight of the cluster can be observed, indicating a process of temporary dissipation of events. Finally, after 60 min, the model detects a dense center in its current state with 40 reports within ϵ and a weight $w = 40.00$, again dominated by the obstacle category, evidencing the consolidation of a persistent incident that continues to affect urban mobility. Taken together, these results confirm that DenStream not only identifies relevant spatial concentrations, but also captures the temporal transition in the nature of traffic incidents within a continuous flow.

4.5.2. Results with the Panama Dataset

This subsection presents the results obtained from the validation dataset corresponding to Panama City, which was designed to simulate georeferenced traffic reports published on social media under controlled conditions. Thanks to its design based on historical seeds, the dataset incorporates semantic variability, category diversity, and realistic spatial distribution, allowing the behavior of the proposal to be evaluated in a representative and reproducible urban scenario.

The classification process was carried out using the ARF model, which operated incrementally to simulate a continuous flow of incoming messages. The performance of the classifier was evaluated using the confusion matrix presented in Figure 7, which allows analyzing the correspondence between the predicted classes and the actual labels of the Panama dataset.

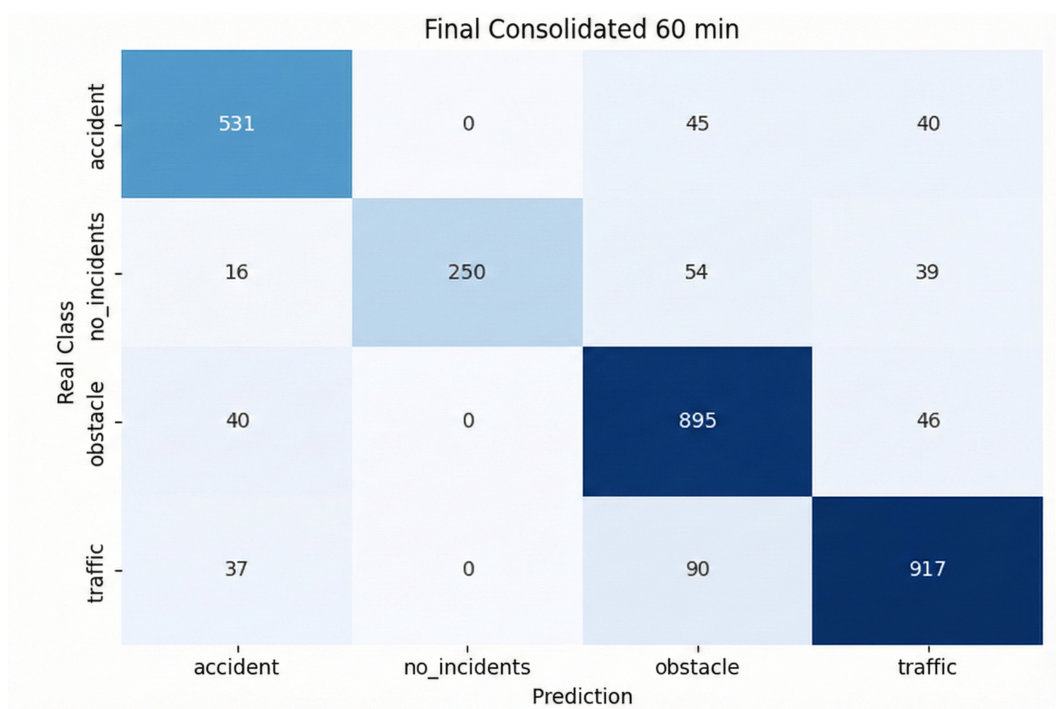


Figure 7. Confusion matrix of the ARF model for incident classification in the Panama dataset.

In Figure 7, we can see that the model correctly classified 917 real reports in the “traffic” category, consolidating itself as the class with the highest number of correct predictions in this scenario. This is followed by the “obstacle” category with 895 correct predictions, while the classifier accurately identified 531 instances of ‘accident’ and 250 corresponding to “no incidents.” These results confirm the robustness of the approach for operating in simulated urban environments with high information density.

However, the confusion matrix reveals ambiguities arising from the operational correlation of incidents. There were 90 cases in which actual reports of “traffic” were classified as “obstacles,” as well as 54 instances of “non-incidents” mistakenly assigned to this same category. Similarly, 40 reports of “accidents” were predicted as “traffic” and 45 as “obstacles,” reflecting the close conceptual relationship between physical events on the road and their immediate effects on traffic congestion. Despite these specific confusions, the low cross-error rate between critical categories confirms the effectiveness of the model in identifying the general nature of incidents in a continuous data stream.

To complement the confusion matrix analysis, additional performance metrics were calculated, including precision, recall, and F1-score for each category, as well as the corresponding support. Table 7 summarizes these results, providing a quantitative view of the classifier’s performance by class.

Table 7. Classification model results for the Panama dataset.

Category	Precision	Recall	F1-Score	Support
Accident	0.851	0.862	0.856	616
No incidents	1.000	0.696	0.821	359
Obstacle	0.826	0.912	0.867	981
Traffic	0.880	0.878	0.879	1044
Accuracy		0.864		0.864
Macro avg	0.889	0.837	0.856	3000
Weighted avg	0.871	0.864	0.864	3000

In the case of the Panama dataset, the model’s performance shows greater stability in the classification metrics, reflecting less semantic variability in the validation data flow. However, as in the Guayaquil scenario, the main objective of the classification process is to serve as a reliable filtering stage for incremental spatial analysis. From this perspective, the model’s usefulness is evaluated based on its ability to consistently feed the clustering process, rather than on the isolated optimization of traditional classification metrics.

As shown in Table 7, the model achieved an overall accuracy of 86.4%, accompanied by a weighted F1-score of 0.864, calculated over a total of 3000 instances. These values confirm that the ARF classifier maintains solid and consistent performance under controlled conditions, being able to handle continuous data streams with multiple categories and high event density.

The analysis by category shows balanced performance. The “No incidents” class has perfect accuracy (1.000), indicating that the method effectively filters out irrelevant messages, eliminating false positives in this category. Meanwhile, the “Obstacle” category achieves the highest recall value (0.912), demonstrating the model’s ability to capture the vast majority of events related to road obstructions. The “Traffic” and “Accident” classes have F1-scores of 0.879 and 0.856, respectively, ensuring reliable discrimination between routine traffic congestion and events of greater operational criticality.

For the spatial analysis of the Panama dataset, the geographic coordinates associated with reports classified as incidents by the ARF model within a 60 min time window were processed. The data were analyzed incrementally in 10 min intervals, with each window incorporating approximately 500 messages. In the first observation interval (0 to 10 min), for example, the flow consisted of 37 accident reports, 18 obstacle reports, and 405 traffic-related reports, allowing us to visualize the operational load of the algorithm and the predominance of events associated with traffic congestion from early stages. Figure 8 shows the spatial distribution of individual incidents in each time interval, allowing us to observe how the progressive accumulation of events reflects the urban mobility dynamics modeled in Panama City.

In Figure 8a, we can see a dispersion that aligns with the main road corridors and avenues, where traffic reports (blue) predominate massively, with a very limited presence of accidents and obstacles. However, just 10 min later in Figure 8b, the flow dynamics change dramatically; there is a sudden increase in orange and red dots, which gives greater visibility to critical events. This trend is accentuated in the central intervals in Figure 8c,d. In these phases, the method reaches its highest saturation density, showing a strong prevalence of accidents (red) and obstacles (orange) that almost completely overshadow traffic reports. This saturation of high-priority incidents clearly delineates the arteries with the highest flow, suggesting that physical events are conditioning mobility throughout the monitored area. As the observation window closes, the composition of the flow changes again. In Figure 8e, a transition can be seen where obstacles (orange) become the almost exclusive category on the map, while in the final interval in Figure 8f, the algorithm records a return to the predominance of traffic reports (blue). This final evolution is key, as it demonstrates the model's ability to track how a situation of multiple accidents and obstacles leads back to a condition of widespread traffic congestion before the monitoring hour is complete. While the spatial distribution of individual incidents allows for the identification of general mobility patterns and areas with high operational load, an analysis based solely on points limits the ability to distinguish persistent concentrations of events in space. In order to complement this spatial analysis and capture road impact areas in a more structured way, an incremental clustering approach using the DenStream algorithm is incorporated below, which allows the identification and tracking of the temporal evolution of incident clusters as the data flow progresses.

Figure 9 shows the overall status of spatial clustering after 60 min of continuous processing. The visualization integrates the clusters identified from the first 10 min to the final snapshot, allowing us to observe their temporal evolution and their status within the model. The clusters represented in purple correspond to dense centers that are in a decay phase, while the cluster in green represents the center detected in the previous snapshot (50 min). The cluster highlighted in red identifies the current dense center, generated at 60 min, which groups 162 reports within the neighborhood radius ϵ , reaching a weight of $w = 162.00$. This result demonstrates the ability of the DenStream algorithm to preserve relevant historical information and dynamically update spatial concentrations as new data is incorporated, facilitating the detection of persistent clusters of high relevance in complex urban scenarios. The values mentioned above are detailed in Table 8, which shows, for each temporal snapshot of the flow (in minutes), the dense center detected, the status of the cluster within the model, the number of cluster groups formed, the number of reports contained within the neighborhood radius (ϵ), the weight associated with the cluster, the distribution of reports by category, and the dominant category in each observation interval.

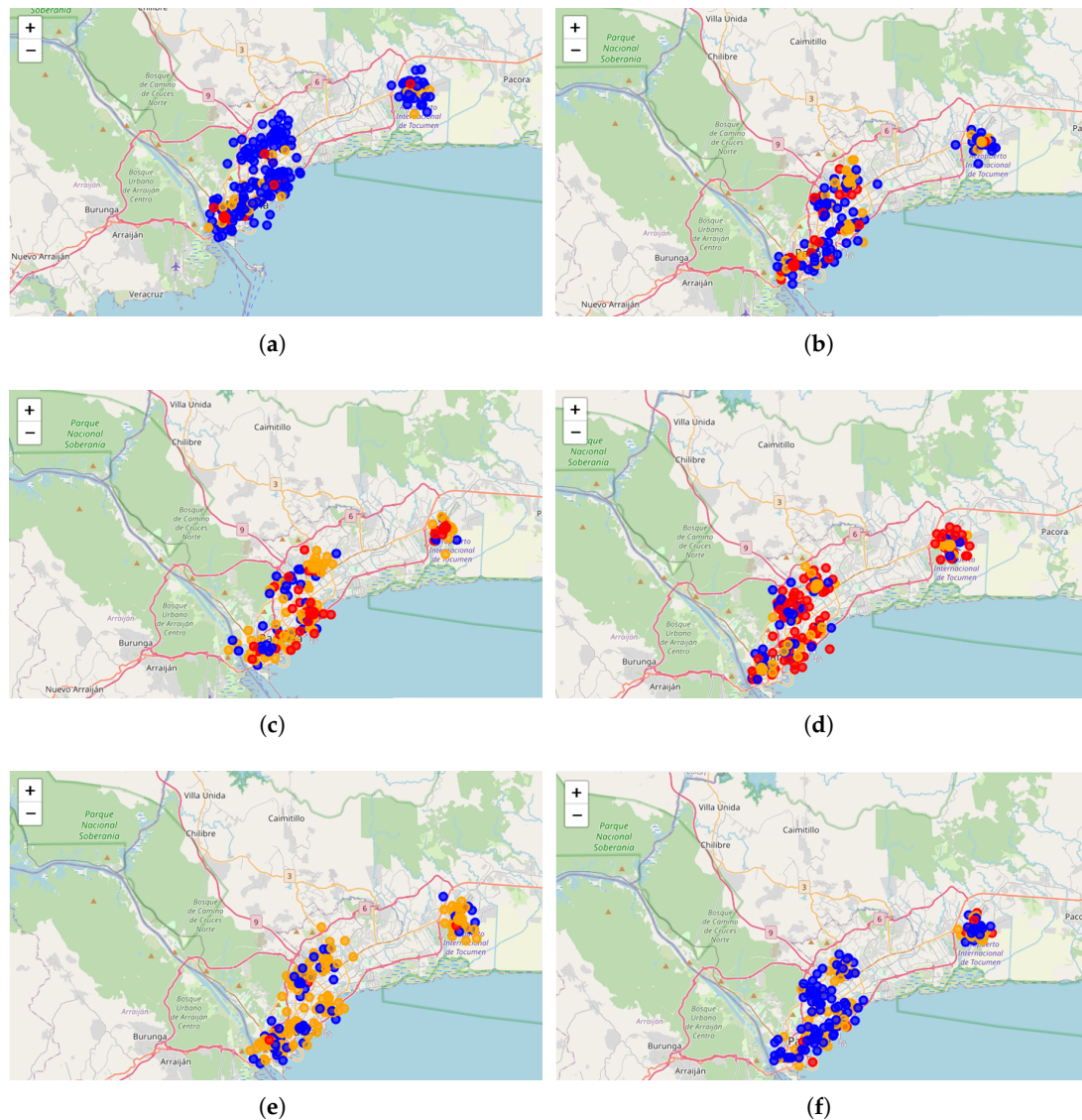


Figure 8. Spatial evolution of incidents in Panama City at 10 min intervals: (a) 0–10 min. (b) 10–20 min. (c) 20–30 min. (d) 30–40 min. (e) 40–50 min. (f) 50–60 min. The colors represent: accidents (red), traffic (blue), and obstacles (orange).

The Table 8 allows for a detailed analysis of the temporal evolution of the dense centers detected by DenStream throughout the data flow corresponding to Panama City. In the snapshot of the first 10 min, a dense center is identified with 234 reports within the neighborhood radius ϵ and a weight $w = 234.00$, composed exclusively of the traffic category, reflecting an initial phase characterized by massive and highly concentrated traffic congestion. In the 20 and 30 min snapshots, a geographical shift of the dense center is observed, accompanied by a marked semantic transition, where the clusters become dominated by accident reports, concentrating 129 and 135 events respectively, which evidences a phase of high criticality associated with physical incidents that exceed routine traffic. Then, in the 40 and 50 min snapshots, the model detects dense centers dominated by the obstacle category, with 102 and 146 reports within ϵ , suggesting the persistence of prolonged blockages or interruptions in the road network. Finally, in the 60 min snapshot, DenStream identifies a dense center in its current state with 162 reports and a weight $w = 162.00$, where traffic is once again the dominant category, indicating a reconfiguration of urban dynamics toward sustained vehicle saturation as a result of previous incidents.

Taken together, these results demonstrate the algorithm's ability to incrementally capture not only the location of critical areas, but also the temporal transformation in the nature of road events within a real urban environment.

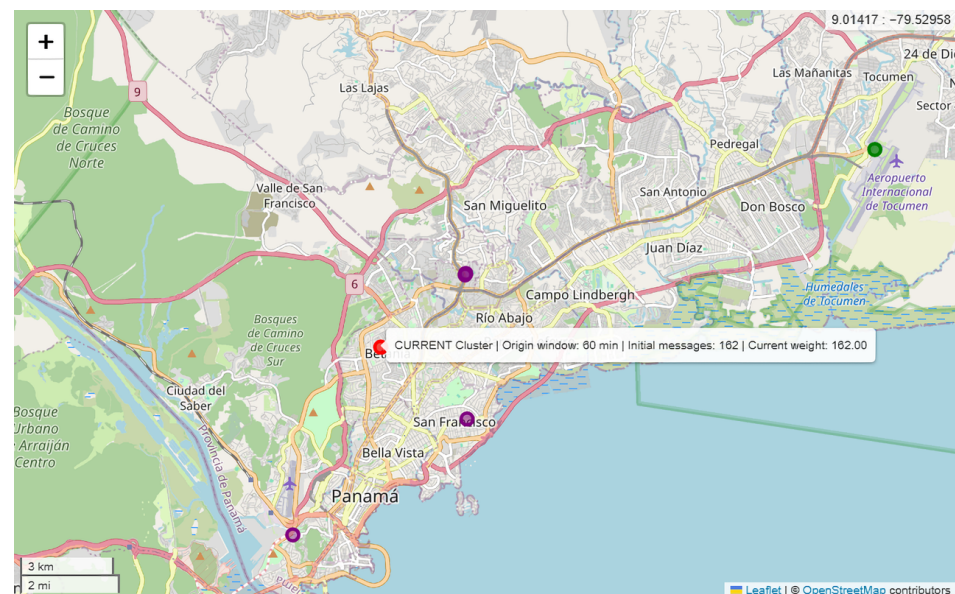


Figure 9. Evolution of clusters in Panama (Scale in km). Status: decay (purple), previous (green), current (red).

Table 8. Dense centers detected by DenStream for the Panama dataset.

Snapshot (min)	State	Group	Reports ϵ	Weight (w)	Category
10	Decay	1	234	234.00	Traffic
20	Decay	1	129	129.00	Accident
30	Decay	1	135	135.00	Accident
40	Decay	1	102	102.00	Obstacle
50	Decay	1	146	146.00	Obstacle
60	Actual	1	162	162.00	Traffic

5. Discussion

The proposed approach for traffic incident detection combines incremental classification using ARF with DenStream's dynamic spatial clustering, allowing for the changing nature of social media data flows in Guayaquil and Panama City to be managed. Unlike traditional batch learning-based methodologies, this proposed method is based on the collection and analysis of streaming data, allowing for adaptive evaluation of traffic dynamics through the use of incremental learning algorithms, responding efficiently to changes in the underlying patterns (concept drift) inherent in social sensors.

When comparing the performance of the proposal with the existing literature, significant advantages in terms of operational immediacy are evident. Compared to [1], which focuses its analysis on accident extraction using mining techniques on static and historical datasets, our approach prioritizes real-time processing. Although the methodology presented in [1] is effective for retrospective analysis, it has its own latency that limits its applicability for live traffic management. In contrast, our method overcomes this restriction by processing the flow message by message, applying incremental filtering that reduces informational noise before the spatial analysis phase, thus optimizing reliability without the delays observed in [1].

To put the performance of our proposal into perspective, it is useful to compare our results with the specialized literature applied to similar scenarios. In the context of

Panama City, previous studies such as [2] have used static machine learning models (e.g., traditional Random Forest), achieving high-precision metrics on closed historical datasets. However, these static approaches often suffer from severe performance degradation when implemented on continuous real-time streams. In contrast, our ARF-based method achieved a sustained pre-decision accuracy of 86.4% and a weighted F1 score of 0.864 on the Panama stream. While a static model may show slightly higher absolute metrics in an offline batch evaluation, the results presented here are operationally superior because they remain stable over time. This superiority is due to the integration of change detectors within the adaptive architecture, which allows the algorithm to autonomously adjust its decision trees in response to conceptual drift and linguistic variability in social media, ensuring consistently high accuracy without the need for periodic manual retraining.

On the other hand, validation in the Panama City scenario allows for a direct comparison with previous work carried out in the same geographical context. Unlike [2], where incident identification is based on static training models that tend to degrade in accuracy when faced with changes in social media vocabulary, our proposal incorporates mechanisms to handle conceptual drift by autonomously updating the classifier structure. While in [2] the validity of the model depends on periodic manual retraining, our approach maintains its effectiveness in the face of linguistic variability over time, ensuring superior scalability to the solution proposed in [2].

The spatial dimension managed by the clustering algorithm complements this adaptability by allowing clusters to indicate not only the location but also the life cycle of the incident (escalation and dissipation). However, it is crucial to recognize certain limitations inherent in the experimental design. First, reliance on social data sources introduces coverage bias, as incident detection is contingent on the density of active users and technology penetration in each urban area. Second, although data streams based on real traces were used to ensure semantic fidelity, validation was performed in a controlled environment. While this allows for scientific reproducibility, implementation in a real production environment could face additional challenges of network latency and unstructured noise that are not present in test scenarios. Finally, the accuracy of geolocation inferred from text remains an open challenge that could affect the granularity of clusters detected in areas of high urban density.

Despite promising results obtained in simulated environments, this study has certain limitations and implementation risks that must be taken into account. The proposed approach is based on social media data, which may introduce spatial and temporal biases related to user activity and reporting behavior. Furthermore, although the datasets are derived from real historical seeds, their semi-synthetic nature—which are indispensable for controlled pre-sequential validation—may not fully capture the extreme data bursts and unpredictable anomalies of a live deployment. From a performance standpoint, the model could experience degradation in highly ambiguous scenarios, such as widespread use of sarcasm. Likewise, when previously unseen local slang suddenly appears, there is an inherent latency in adaptation: predictive performance may momentarily decline as the ADWIN algorithm accumulates the statistical evidence needed to update the Online Bag-of-Words scheme and readjust the classifier. Finally, bringing this architecture into a real operating environment (ITS) carries structural risks: monitoring depends on the availability and limitations of third-party APIs (such as X and Telegram). A massive urban event could trigger a data avalanche that exceeds these quotas or overwhelms computational resources, making dynamic cloud scaling and secondary physical fallback mechanisms indispensable.

6. Conclusions

The results obtained indicate that the proposed analysis method, based on incremental learning and dynamic clustering, demonstrates promising performance in controlled data-flow environments for the identification of traffic incidents. Returning to the research questions posed at the beginning of this study, the conclusions are explicitly structured as follows:

First, regarding how to reliably detect traffic incidents from social media, the research demonstrates that the ability to process these text streams using the ARF algorithm enables effective classification of citizen reports, successfully distinguishing between informational noise and critical traffic events. This approach offers a viable and superior alternative to traditional static approaches, adapting to the continuous nature of the information.

Second, regarding the effectiveness of adaptive algorithms in the face of traffic variability and concept drift, it has been demonstrated that the integration of streaming data mining methods constitutes a highly effective strategy. These methods are able to adapt to constantly evolving traffic patterns and changes in vocabulary, autonomously identifying shifting behaviors and reports without requiring manual retraining.

Third, regarding how clustering methods contribute to identifying spatiotemporal patterns, the study concludes that the introduction of decay functions (forgetting mechanism) via DenStream enables much more efficient spatial information management. This ensures that structures are constantly updated by prioritizing recent incidents and filtering out older ones, facilitating a dynamic representation that enables the early detection of forming clusters and reflects traffic scaling processes in great detail.

Based on the experimental results, it is worth noting that the method demonstrated stable performance and a high degree of resilience to semantic ambiguity, enabling the majority of detected events to be correctly located within the neighborhood radius (ϵ) of the identified dense clusters. The metrics achieved underscore the positive impact of the methodology, highlighting it as a dynamic tool that contributes to a deep understanding of constantly evolving urban dynamics. For future work, we propose improving the algorithm's adaptability in larger-scale urban environments and incorporating massive streams of real-world data to overcome the limitations of current semi-synthetic datasets, thereby validating the model's robustness in the face of extreme volumes of live data. Additionally, we will prioritize optimizing the method's sensitivity through automatic parameter tuning based on circadian rhythms. We also aim to investigate the integration of large language models (LLMs) adapted to continuous data streams, analyzing factors such as sarcasm and irony to improve accuracy in message retrieval in areas with high linguistic variability. Experiments in multimodal scenarios, incorporating data from video surveillance cameras and IoT sensors, will allow us to evaluate the algorithm's scalability and robustness when dealing with heterogeneous information sources. To strengthen the relational understanding of the congestion problem, we propose exploring the influence of road network topology and social media user density on cluster formation, using advanced spatial analysis and data fusion tools. Finally, the implementation of AI-based predictive models, supported by the generated micro-cluster history, is considered a key way to anticipate and prevent congestion patterns before they become established on the road network.

Author Contributions: Conceptualization, G.R.; methodology, G.R., C.O.-J. and C.A.-B.; validation, L.L. and W.H.; formal analysis, R.T.-B. and J.B.-M.; investigation, G.R., C.O.-J. and C.A.-B.; data curation, G.R.; writing—original draft preparation, G.R.; writing—review and editing, R.T.-B., L.L., W.H., D.R. and J.B.-M.; supervision, R.T.-B. and J.B.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Universidad Bolivariana del Ecuador (UBE) through the research project FCI (PROY-UBE-2023-026), entitled “Intelligent System for Automatic Vehicular traffic Flow Control”, including financial support for the publication of the research results.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting the reported results are available at the GitHub repository: <https://github.com/gary-reyes-zambrano/Incident-Detection-River> (accessed on 4 February 2026).

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (GPT-5.2, OpenAI) for the purposes of language editing, text organization, and improvement in clarity and readability. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Suat-Rojas, N.; Gutierrez-Osorio, C.; Pedraza, C. Extraction and Analysis of Social Networks Data to Detect Traffic Accidents. *Information* **2022**, *13*, 26. [CrossRef]
2. Liu, L.; Guevara, A.; Sanchez-Galan, J.E. Identification and Classification of Road Traffic Incidents in Panama City through the Analysis of a Social Media Stream and Machine Learning. *Intell. Syst. Appl.* **2022**, *16*, 200158. [CrossRef]
3. Casa Lumbi, J.P. Análisis de Tweets como Fuente de Información para Mejorar las Estrategias de Prevención en Siniestros de Tránsito en Quito. Bachelor’s Thesis, Universidad Politécnica Salesiana, Sede Quito, Quito, Ecuador, 2024. Available online: <http://dspace.ups.edu.ec/handle/123456789/26783> (accessed on 15 February 2026).
4. Sakurai, G.Y.; Lopes, J.F.; Zarpelão, B.B.; Barbon Junior, S. Benchmarking Change Detector Algorithms from Different Concept Drift Perspectives. *Future Internet* **2023**, *15*, 145. [CrossRef]
5. Reyes, G.; Tolozano-Benites, R.; Lanzarini, L.; Hasperué, W.; Barzola-Monteses, J. Intelligent Learning on Multidimensional Data Streams: A Bibliometric Analysis of Research Evolution and Future Directions. *Information* **2025**, *16*, 1067. [CrossRef]
6. Martindale, N.; Ismail, M.; Talbert, D.A. Ensemble-based Online Machine Learning Algorithms for Network Intrusion Detection Systems Using Streaming Data. *Information* **2020**, *11*, 315. [CrossRef]
7. Hovakimyan, G.; Bravo, J.M. Evolving Strategies in Machine Learning: A Systematic Review of Concept Drift Detection. *Information* **2024**, *15*, 786. [CrossRef]
8. Angaramo, F.; Rossi, C. Online Clustering and Classification for Real-Time Event Detection in Twitter. In Proceedings of the International Conference on Information Systems for Crisis Response and Management (ISCRAM), Rochester, NY, USA, 20–23 May 2018. Available online: https://www.idl.iscram.org/files/federicoangaramo/2018/2182_FedericoAngaramo+ClaudioRossi2018.pdf (accessed on 15 February 2026).
9. Vargas Montero, F. Análisis de Datos de Accidentalidad Vial de la Ciudad de Bogotá a Partir de Datos Abiertos y Datos Obtenidos Desde Redes Sociales. Trabajo de Grado (Maestría en Ingeniería de Sistemas y Computación), Universidad Nacional de Colombia, Bogotá, Colombia, 2022. Available online: <https://repositorio.unal.edu.co/handle/unal/81571> (accessed on 15 February 2026).
10. Gutierrez-Osorio, C.; Gonzalez, F.A.; Pedraza, C.A. Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data. *Computers* **2022**, *11*, 126. [CrossRef]
11. Vallejos, S.; Caimmi, B.; Alonso, D.; Soria, Á.; Berdun, L. Detectando Incidentes de Tránsito en Redes Sociales: Un Enfoque Inteligente basado en Twitter vs. Waze. In *Proceedings of the Jornadas Argentinas de Informática*, 2017. Available online: <https://clei.org/clei2017/sites/default/files/Mem/ASAI/asai-02.pdf> (accessed on 15 February 2026).
12. Nallaperuma, D.; Nawaratne, R.; Bandaragoda, T.; Adikari, A.; Nguyen, S.; Kempitiya, T.; de Silva, D.; Alahakoon, D.; Pothuhera, D. Online Incremental Machine Learning Platform for Big Data-Driven Smart Traffic Management. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 4679–4690. [CrossRef]
13. Wan, X.; Lucic, M.C.; Ghazzai, H.; Massoud, Y. Empowering Real-Time Traffic Reporting Systems with NLP-Processed Social Media Data. *IEEE Open J. Intell. Transp. Syst.* **2020**, *1*, 159–175. [CrossRef]
14. Alomari, E.; Katib, I.; Albeshri, A.; Yigitcanlar, T.; Mehmood, R. Iktishaf+: A Big Data Tool with Automatic Labeling for Road Traffic Social Sensing and Event Detection Using Distributed Machine Learning. *Sensors* **2021**, *21*, 2993. [CrossRef]
15. He, M.; Meng, G.; Wu, X.; Han, X.; Fan, J. Road Traffic Accident Prediction Based on Multi-Source Data—A Systematic Review. *Promet-Traffic Transp.* **2025**, *37*, 499–522. [CrossRef]
16. Jiber, M.; Mbarek, A.; Yahyaouy, A.; Sabri, M.A.; Boumhidi, J. Road Traffic Prediction Model Using Extreme Learning Machine: The Case Study of Tangier, Morocco. *Information* **2020**, *11*, 542. [CrossRef]

17. Xu, C.; Mao, Y. An Improved Traffic Congestion Monitoring System Based on Federated Learning. *Information* **2020**, *11*, 542. [\[CrossRef\]](#)
18. Liu, N.; Zhao, J. Streaming Data Classification Based on Hierarchical Concept Drift and Online Ensemble. *IEEE Access* **2023**, *11*, 126040–126051. [\[CrossRef\]](#)
19. Gaber, M.M.; Zaslavsky, A.; Krishnaswamy, S. Mining Data Streams: A Review. *SIGMOD Rec.* **2005**, *34*, 18–26. [\[CrossRef\]](#)
20. Salvatierra, D.; Laborde, J.; León-Granizo, O. Aplicación de Algoritmos de Clasificación, Agrupación y Predicción en la Detección de Patrones Asociados a la Movilidad Usando Datos de Trayectorias Vehiculares. *Ecuadorian Sci. J.* **2023**, *7*, 10–18. [\[CrossRef\]](#)
21. Reyes, G.; Tolozano-Benites, R.; Lanzarini, L.; Estrebou, C.; Bariviera, A.F.; Barzola-Monteses, J. Methodology for the Identification of Vehicle Congestion Based on Dynamic Clustering. *Sustainability* **2023**, *15*, 16575. [\[CrossRef\]](#)
22. Reyes, G.; Lanzarini, L.; Estrebou, C.; Bariviera, A. Dynamic Grouping of Vehicle Trajectories. *J. Comput. Sci. Technol.* **2022**, *22*, 141–150. [\[CrossRef\]](#)
23. Reyes, G.; Lanzarini, L.; Hasperué, W.; Bariviera, A.F. Proposal for a Pivot-Based Vehicle Trajectory Clustering Method. *Transp. Res. Rec.* **2022**, *2676*, 281–295. [\[CrossRef\]](#)
24. Qaiyum, S.; Aziz, I.; Hasan, M.H.; Khan, A.I.; Almalawi, A. Incremental Interval Type-2 Fuzzy Clustering of Data Streams Using Single-Pass Method. *Sensors* **2020**, *20*, 3210. [\[CrossRef\]](#)
25. Mutambik, I. An Entropy-Based Clustering Algorithm for Real-Time High-Dimensional IoT Data Streams. *Sensors* **2024**, *24*, 7412. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ferjani, I.; Alsaif, S.A. Dynamic Road Anomaly Detection: Harnessing Smartphone Accelerometer Data with Incremental Concept Drift Detection and Classification. *Sensors* **2024**, *24*, 8112. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Wei, Y.; Zeng, Z.; He, T.; Yu, S.; Du, Y.; Zhao, C. An Adaptive Vehicle Detection Model for Traffic Surveillance of Highway Tunnels Considering Luminance Intensity. *Sensors* **2024**, *24*, 5912. [\[CrossRef\]](#) [\[PubMed\]](#)
28. Alqabbany, A.O.; Azmi, A.M. Measuring the Effectiveness of Adaptive Random Forest for Handling Concept Drift in Big Data Streams. *Entropy* **2021**, *23*, 859. [\[CrossRef\]](#)
29. Bechini, A.; Bondielli, A.; Ducange, P.; Marcelloni, F.; Renda, A. Addressing Event-Driven Concept Drift in Twitter Stream: A Stance Detection Application. *IEEE Access* **2021**, *9*, 77758–77770. [\[CrossRef\]](#)
30. Yang, R.; Xu, S.; Feng, L. An Ensemble Extreme Learning Machine for Data Stream Classification. *Algorithms* **2018**, *11*, 107. [\[CrossRef\]](#)
31. Pesaranghader, A.; Viktor, H.; Paquet, E. Reservoir of Diverse Adaptive Learners and Stacking Fast Hoeffding Drift Detection Methods for Evolving Data Streams. *Mach. Learn.* **2018**, *107*, 1711–1743. [\[CrossRef\]](#)
32. Patiño Pérez, D.; Silva Bustillos, R.; Munive Mora, C.; Botto-Tobar, M. Predicción de COVID-19 con el Uso del Algoritmo Random Forest y Redes Neuronales Artificiales. *Ecuadorian Sci. J.* **2020**, *4*, 101–110. [\[CrossRef\]](#)
33. Gomes, H.M.; Bifet, A.; Read, J.; Barddal, J.P.; Enembreck, F.; Pfahringer, B.; Holmes, G.; Abdessalem, T. Adaptive Random Forests for Evolving Data Stream Classification. *Mach. Learn.* **2017**, *106*, 1469–1495. [\[CrossRef\]](#)
34. Sun, Y.; Pfahringer, B.; Gomes, H.M.; Bifet, A. SOKNL: A Novel Way of Integrating K-Nearest Neighbours with Adaptive Random Forest Regression for Data Streams. *Data Min. Knowl. Discov.* **2022**, *36*, 2091–2128. [\[CrossRef\]](#)
35. Halstead, B.; Koh, Y.S.; Riddle, P.; Pears, R.; Pechenizkiy, M.; Bifet, A.; Olivares, G.; Coulson, G. Analyzing and Repairing Concept Drift Adaptation in Data Stream Classification. *Mach. Learn.* **2022**, *111*, 3489–3523. [\[CrossRef\]](#)
36. Mehmood, H.; Kostakos, P.; Cortes, M.; Anagnostopoulos, T.; Pirttikangas, S.; Gilman, E. Concept Drift Adaptation Techniques in Distributed Environment for Real-World Data Streams. *Smart Cities* **2021**, *4*, 21. [\[CrossRef\]](#)
37. Gama, J.; Zliobaite, I.; Bifet, A.; Pechenizkiy, M.; Bouchachia, A. A Survey on Concept Drift Adaptation. *ACM Comput. Surv.* **2014**, *46*, 1–37. [\[CrossRef\]](#)
38. Giannini, F.; Ziffer, G.; Cossu, A.; Lomonaco, V. Streaming Continual Learning for Unified Adaptive Intelligence in Dynamic Environments. *IEEE Intell. Syst.* **2024**, *39*, 81–85. [\[CrossRef\]](#)
39. Espin-Riofrio, C.; Ortiz-Zambrano, J.; Montejo-Ráez, A. An Approach to Lexicon Filtering for Author Profiling. *Proces. Leng. Nat.* **2023**, *71*, 75–86. [\[CrossRef\]](#)
40. Rebuffi, S.A.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental Classifier and Representation Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2001–2010. [\[CrossRef\]](#)
41. Althabiti, M.S.; Abdullah, M. CDDM: Concept Drift Detection Model for Data Stream. *Int. J. Interact. Mob. Technol.* **2020**, *14*, 187–198. [\[CrossRef\]](#)

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