A Bayesian multi-source data fusion approach to mitigate spatiotemporal uncertainties in traffic accident records

Mahsa Movaghar*[1,2], Neil Yorke-Smith[2], Hans van Lint[1], and Serge Hoogendoorn[1]

[1]Transport and Planning, Delft University of Technology (TU Delft), Netherlands
[2]STAR Lab, Delft University of Technology (TU Delft), Netherlands

Short summary

Reliable and precise accident data are a crucial prerequisite for mitigating accident-related costs and improving the safety of the road traffic network. However, in practice, reported accident data often suffer from temporal and spatial uncertainties, creating a gap between when and where accidents are recorded and when and where their effects are observable in traffic flow and congestion. To address this challenge, we propose a Bayesian multi-source data fusion framework that reconstructs latent accident attributes by integrating heterogeneous evidence while explicitly quantifying uncertainty. **Key contributions** of this work are: (i) a latent-variable formulation under Bayesian inference that frames accident records as uncertain observations rather than deterministic inputs, (ii) a scalable multi-source data fusion approach that integrates matrix signaling information data and traffic measurements as complementary probabilistic evidence, and (iii) the resulting uncertainty-aware accident dataset can be directly used in data-driven models, such as Bayesian networks, leading to more reliable robust inference in safety and accident management applications.
Keywords: Traffic accident, Bayesian inference, Data fusion, Spatiotemporal uncertainty, Accident detection

1 Introduction

In the era of AI and data-driven methods, recent developments have opened new opportunities for analyzing and understanding real-world problems. However, these methods are highly dependent on the quality of the input data (Lint & Hoogendoorn (2010)). The importance of data quality is often overlooked, even though it directly affects the reliability of model outputs. This challenge becomes even more critical in sensitive domains such as safety and policy-making.

In transportation, many important applications, such as crash analysis and prevention, accident detection, causal inference, surrogate safety indicators, and infrastructure planning, are mainly built on real-world datasets. Crash analysis, in particular, relies heavily on reported accident data for tasks ranging from incident detection to understanding spatial and temporal patterns of crashes, identifying hazardous locations, and developing strategies to reduce accident-related costs (Wang et al. (2021)). Among all variables, the occurrence time and location of accidents are the most fundamental elements of these datasets.

Despite advances in data collection and modelling techniques, the reliability of accident datasets is rarely discussed in the literature. Due to their inherent stochastic nature, these datasets are often uncertain and incomplete in practice, which can undermine trust in data-driven approaches among practitioners and policymakers (Peng et al. (2024); Liu et al. (2022)). Reported crash data, although widely used, is one of the most unreliable sources due to its dependence on manual recording and reporting processes.

Accident data is typically collected by multiple organizations, such as police, emergency services, and traffic control centres. This human involvement introduces uncertainty in both time and location. For example, the recorded time of an accident often corresponds to when it is reported, not when it actually occurred. Similarly, the recorded location may differ from the true location due to post-incident actions such as moving vehicles. These inconsistencies create a gap between reported data and actual traffic conditions.

Although some studies have attempted to address these issues, the problem remains largely unexplored. For example, Wang & Jiang (2019) proposed a method to correct time and location

bias using speed drop information and integer programming. While this approach provides useful insights, it relies on a single source of information and does not fully exploit the potential of integrating multiple data sources.

Contributions

Despite the rapid development of AI techniques in traffic safety, a robust framework for mitigating uncertainty in accident records is still missing. To address this gap, this study aims to answer the following research question: *"How can the uncertainties and biases associated with the recorded time and location of accidents be mitigated by integrating multi-source heterogeneous traffic information?"*
The main contributions of this study are:

- A Bayesian framework is proposed to fuse multiple heterogeneous datasets and estimate a more accurate time and location of reported accidents.
- The proposed approach is flexible to any missing data source and can be extended to incorporate additional data sources in the future (e.g., social media or crowd-sourced data).
- The resulting uncertainty-aware accident dataset can be used as a reliable input for future data-driven models, leading to more accurate and robust analysis.

2 Methodology

In this study, three distinct data sources are integrated within a Bayesian framework: accident reports, matrix signal information, and traffic measurements.

Data sets

(1) **Accident reports:** The first and foremost available evidence of accident occurrence is accident reports recorded and documented by the police and/or road inspectors from Rijkswaterstaat. This data source, known as BRON, compiles and maintains traffic accident records from the police and Rijkswaterstaat. Despite years of collaborative work among different organizations and institutions, this dataset still includes multiple limitations:

- Although the exact locations of registered accidents are known, the time of occurrence remains uncertain, and there is inconsistency with other traffic evidence.
- Many accidents are missing in the dataset, as the police are not called for all accidents, and in accordance with instructions, minor accidents are not recorded.
- Despite all the improvements in the quantity and quality of the dataset, Rijkswaterstaat relies on the quality of the dataset suppliers. Each dataset is inherently uncertain due to its characteristics, and it can never be complete.

(2) **Matrix signal information:** Matrix Signal Information (MSI), also known as Variable Message Sign (VMS), is an electronic display that is used to show variable messages such as warnings about congestion, lane closures, or advisory speed limits. Unlike static signs, these can change in real-time to respond to evolving road conditions. When an accident on the road is detected via traffic cameras or roadside inspectors, operators manually place lane closure symbols to close specific lanes. MSI is an important and informative dataset, specifically lane closures, as this is our only human-verified, labelled, and available data. Moreover, traffic control centres using cameras are among the first authorities to be aware of accident occurrences around the network.
However, there are multiple concerns with using only this dataset:

- This data is only available in specific locations, with an average spacing of 700 metres (not even at every loop detector). As a signal applies to the entire segment between two consecutive boards, localization error is expected.
- The time that speed changes and lane closures appear on a matrix sign is not the actual accident time. There are multiple errors associated with this dataset. First of all, there is a physical delay from data measurements from loop detectors to local preprocessing units, to central processors, and finally out to the matrix boards. In addition, the AID system triggers based on detected traffic slowdowns. If traffic is light, it may take several minutes for a queue to form and propagate to trigger the mechanism.

- Critical signals like lane closures are activated manually by a human operator in a traffic control centre. This process of human in the loop adds significant uncertainty in detecting the exact time and location of accidents.

It is important to highlight that most speed symbols are a response to a traffic jam, which is an effect of an accident, not the accident occurrence itself. Finding the "cause" in this pipeline requires a complex back-propagation of shockwave speeds.
(3) **Traffic measurements:** The most common traffic data used to detect incidents are speed and flow, measured by loop detectors. Traffic loop detector data consists of real-time, high-frequency measurements from thousands of inductive loops throughout highways, mainly collected by Nationale Databank Wegverkeergegevens (NDW). Despite the extensive use of this dataset in incident detection and management, multiple technical and operational challenges limit its effectiveness (Lin et al. (2020); ElSahly & Abdelfatah (2024)):

- Relying only on loop detector data for incident detection increases false alarm rates. Environmental factors (like heavy rain and fog), lane changes, or sensor malfunctioning mimic the data patterns of an accident. These false detections create accident alerts for situations that do not actually exist.
- As the locations of loops are fixed, the incident can only be detected once the resulting traffic queue physically reaches the adjacent upstream sensors. This creates a time and space lag between the actual crash and the observable evidence in speed and flow.
- Like any other data source, equipment malfunction can result in significant missing data and uncertainty.

As discussed, each data source is subject to uncertainty and inherent limitations; therefore, integrating these sources yields more reliable evidence, for which a Bayesian network provides a principled fusion framework.

Bayesian Network

To integrate heterogeneous data sources and explicitly account for their dependencies, a hierarchical Bayesian network is adopted. The objective is to estimate the posterior probability of an accident occurring at a given spatiotemporal location (x, t) by combining multiple sources of evidence, including accident reports, lane-closure signals, speed patterns, flow measurements, and matrix signal information.
The Bayesian framework is particularly suitable for this problem, as it allows the systematic combination of uncertain and noisy observations through probabilistic reasoning. Each data source provides partial and uncertain evidence about the true accident state, and Bayesian inference enables updating prior beliefs based on observed data to obtain a consistent posterior distribution (Sander & Beyerer (2013)). In this study, we continue with a non-informative prior ($P(A) = 1$) as we do not have any further information regarding the probability of accident occurrence. An important advantage of the Bayesian network framework is its robustness to missing data. Even when one or more data sources are unavailable, the method can still operate effectively and infer the likelihood of accidents based on the remaining available evidence.
However, a key challenge in this setting is that not all data sources are independent. In particular, speed and flow measurements originate from loop detectors, and matrix signal information, especially speed limits, is partially derived from these measurements. Ignoring these dependencies and applying a naïve Bayesian formulation would lead to double-counting of evidence and an overconfident posterior.
To address this issue, a hierarchical Bayesian structure is employed, as illustrated in Figure 1. In this formulation, the accident variable P_{accident} is treated as the root node, which directly influences independent observations such as accident reports P_{report} and lane-closure signals $P_{\text{MSI_LC}}$.
Traffic state variables, represented by speed P_{speed}, are modelled as intermediate latent observations influenced by the accident. In turn, flow P_{flow} and dependent components of matrix signal information $P_{\text{MSI_SL}}$ are conditioned on speed. Formally, the joint likelihood can be factorized according to the graph structure as:

$$P\big(A(x,t) \mid \text{data}\big) \propto P\big(A(x,t)\big) \prod_{data \in \{r,f,v,msi,lc\}} L_d(x,t) \tag{1}$$

$$\underbrace{P\big(A(x,t) \mid \text{data}\big)}_{\text{Posterior}} \propto \underbrace{P(A)}_{\text{Prior}} \underbrace{P(r \mid A)\, P(lc \mid A)\, P(v \mid A)\, P(msi \mid v)\, P(f \mid v)}_{\text{Likelihoods}} \tag{2}$$

$$P(A(x,t) \mid \text{data}) \propto P(A(x,t)) \cdot \underbrace{\mathcal{L}_{\text{r}} \cdot \mathcal{L}_{\text{lc}} \cdot \mathcal{L}_{\text{v}}}_{\text{Primary Evidence}} \cdot \underbrace{\mathcal{L}_{\text{f}} \cdot \mathcal{L}_{\text{msi}}}_{\text{Secondary Evidence (via Speed)}} \tag{3}$$

where $A(x,t)$ denotes the accident variable at location x and time t, r represents accident reports, v corresponds to speed measurements, f denotes traffic flow, lc indicates lane-closure signals, and msi refers to matrix signal information.

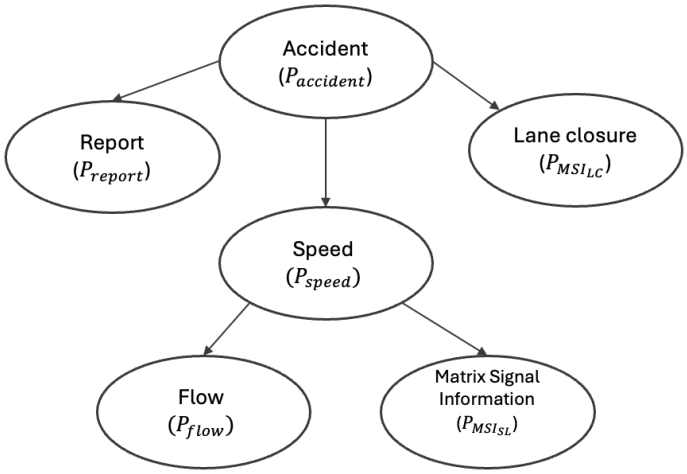


Figure 1: Hierarchical Bayesian network representing dependencies between accident occurrence and observed data sources.

Likelihoods estimation

In this section, the approach for statistical incident detection for each dataset is explained:

(1) Accident reports ($L_{\text{report}}(t,s)$): To incorporate the uncertainty in the report dataset, a probability surface showing where and when an accident most likely occurred is defined based on a reported time and location. A bivariate Gaussian distribution centred at the reported time and location is used to represent the uncertainty in accident reports. The distribution attains its maximum at the reported point and decays smoothly as the temporal and spatial distance from this point increases.

$$L_{\text{report}}(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{4}$$

where:

$$x = \begin{bmatrix} s \\ t \end{bmatrix} \qquad \text{vector of variables (space } s \text{ and time } t\text{)},$$

$$\mu = \begin{bmatrix} \mu_s \\ \mu_t \end{bmatrix} \qquad \text{vector of means (reported location and time)},$$

$$\Sigma = \begin{bmatrix} \sigma_s^2 & \text{Cov}(\sigma_s,\sigma_t) \\ \text{Cov}(\sigma_s,\sigma_t) & \sigma_t^2 \end{bmatrix} \qquad \text{covariance matrix.}$$

where:

- σ_s: space error (standard deviation in metres),
- σ_t: time error (standard deviation in minutes).

A maximum time error of ± 15 minutes is assumed for police reports, and the maximum space error is assumed as the average distance between two consecutive sensor locations. Since time and space errors are assumed independent, the covariance matrix simplifies to:

$$\Sigma = \begin{bmatrix} 15^2 & 0 \\ 0 & \left(\frac{\sum_i d_i}{n}\right)^2 \end{bmatrix} \tag{5}$$

These bandwidth parameters were set based on reasonable assumptions about temporal and spatial uncertainty in accident reports. A systematic sensitivity analysis and optimization of these parameters will be addressed in future work.

(2) Matrix Signal Information ($L_{\mathbf{MSI}}$): Estimating the most probable time and location of accidents using only the MSI dataset is challenging, as this dataset is partially derived from loop detector measurements. Integrating these two data sources therefore introduces a classical Bayesian challenge, since they are both spatially and temporally correlated. As discussed earlier, the signaling system is largely a product of loop detector observations and thus cannot be treated as independent evidence from speed data. To address this issue, a Bayesian fusion framework that explicitly accounts for the functional dependency between loop detector data and signaling symbols is developed. In this approach, the independent and dependent components of MSI are modeled separately. Lane closures are treated as independent signals, while speed-limit signals are modeled as dependent observations.
For the independent component, a Gaussian kernel is constructed based on lane-closure signals. For the dependent components, instead of adopting a simple Naïve Bayes assumption, a hierarchical Bayesian model is employed. In this formulation, the signaling system (MSI) is interpreted as a noisy observation of loop detector data, which itself is a noisy observation of the underlying traffic state influenced by an accident.

Lane Closure Likelihood ($L_{\mathbf{MSI\text{-}LC}}$)

To estimate the likelihood $L_{\text{MSI_LC}}(\text{MSI} \mid x, t)$, we focus exclusively on how well a hypothesized accident location x and time t explain the observed signaling data. The likelihood is formulated as the product of individual signal probabilities, which is computed in the logarithmic domain as a summation:

$$\ln L(\text{MSI} \mid x, t) = \sum_{s \in S} \ln P(s \mid x, t) \tag{6}$$

where $S = \{s_1, s_2, \ldots, s_n\}$ denotes the set of observed signals.
Each individual signal contributes to the total likelihood based on three factors: its type, its spatial distance, and its temporal distance from the hypothesized accident.
The core of the likelihood formulation is a Gaussian distance kernel. This model assumes that signals closer to the accident in both space and time are more likely to have been caused by that accident. The probability of observing a single signal s is defined as:

$$L_{\text{MSI}}(s \mid x, t) \propto w_s \exp\left(-\frac{(t_s - t)^2}{2\tau_t^2} - \frac{(x_s - x)^2}{2\tau_x^2}\right) \tag{7}$$

where:

- (x_s, t_s) represent the spatial location (hectometre) and timestamp of the signal,
- (x, t) denote the candidate accident location and time,
- τ_x and τ_t are bandwidth parametres controlling the spatial and temporal decay (set to 1 km and 15 minutes, respectively),
- w_s is a weighting factor reflecting the informativeness of each signal type.

Lane-closure signals are assigned a weight of $w_{\text{LC}} = 1.0$, as they are strong indicators of localized incidents. In contrast, speed-limit signals are assigned a low or zero weight, since reductions in speed can arise from various factors unrelated to accidents.
(3) **Traffic measurements ($L_{\mathbf{speed}}$ and $L_{\mathbf{flow}}$):**
To incorporate traffic dynamics into the Bayesian fusion framework, we derive a likelihood function based on speed measurements obtained from loop detectors. Instead of directly using raw speed values, we quantify anomalies in speed using a rolling Z-score formulation. This allows us to capture deviations from normal traffic conditions while accounting for local temporal variations. Using the rolling statistics, the standardized anomaly (Z-score) is computed as:

$$Z_v(x, t) = \frac{v(x, t) - \mu_v(x, t)}{\sigma_v(x, t)} \tag{8}$$

Negative Z-scores correspond to speed drops, which are indicative of potential traffic disturbances such as accidents. To transform the Z-score into a likelihood value bounded between 0 and 1 (as

in the normalized probability p_{data}), the cumulative distribution function (CDF) of the standard normal distribution can be used. Specifically, the probability of an anomaly is defined as:

$$L_{\text{speed}}(x,t) = 1 - \Phi(z) \tag{9}$$

where $\Phi(z)$ denotes the standard normal cumulative distribution function, given by:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}}\, dt \tag{10}$$

This formulation ensures that:

- Large negative Z-scores (significant speed drops) yield high likelihood values,
- Near-zero or positive Z-scores (normal or free-flow conditions) contribute minimally.

Uncertainty Quantification

With the posterior distribution obtained as the product of the estimated likelihoods from the Bayesian network, the core of Bayesian uncertainty quantification (UQ) is already established. The posterior itself serves as the main tool for UQ, as it naturally represents the relative plausibility of different time and location values (Melev & Kauermann (2025); Peixoto (2025)). In this study, the following uncertainty quantification measures are considered: Credible intervals (50% and 95%), Posterior standard deviation, Posterior modality analysis (mode-based UQ), and Highest Posterior Density (HPD).

3 Results and discussion

To demonstrate the applicability of the proposed method, it has been implemented on a case study in the Netherlands. An accident was reported on the A13 motorway at 08:04:01. Three groups of results are particularly relevant for discussion: (1) likelihood estimations, (2) posterior estimation, and (3) uncertainty quantification.

Likelihood estimations

Based on the methods explained earlier, the likelihood probabilities have been estimated for available evidence.
The likelihood derived from speed data and accident reports reflects the consistency between observed traffic patterns and the reported accident (Figure 2). The rolling Z-score formulation captures abnormal speed drops, while the report likelihood provides a prior spatial and temporal focus. The combined likelihood highlights regions where both sources agree, resulting in a more concentrated probability mass around the true accident location.
The likelihood based on lane-closure signals captures the direct response of the traffic management system to a localized accident (Figure 3). Since lane closures are strong indicators of disruptions, this likelihood tends to be more spatially focused and less noisy compared to speed-based likelihoods. The Gaussian kernel formulation ensures that signals closer in time and space to the hypothesized accident contribute more significantly.

Posterior estimation

The posterior distribution is obtained by combining all likelihood components within the Bayesian framework (Figure 4). This integration balances the strengths and limitations of each data source, resulting in a refined estimation of the accident's time and location. Compared to individual likelihoods, the posterior is more stable and exhibits reduced uncertainty, as it reflects the consensus across multiple sources of evidence. The posterior distribution is summarized using the Maximum A Posteriori (MAP) estimate, which corresponds to the peak of the posterior surface. The MAP is identified at time $t = 07:57:30$ and location $x = 11.60$, representing the most probable accident occurrence given all observed data.
This result demonstrates how the Bayesian fusion framework refines the initial report by incorporating additional evidence from traffic and signaling data. The resulting estimate provides a more reliable and physically consistent representation of the accident location in both space and time.

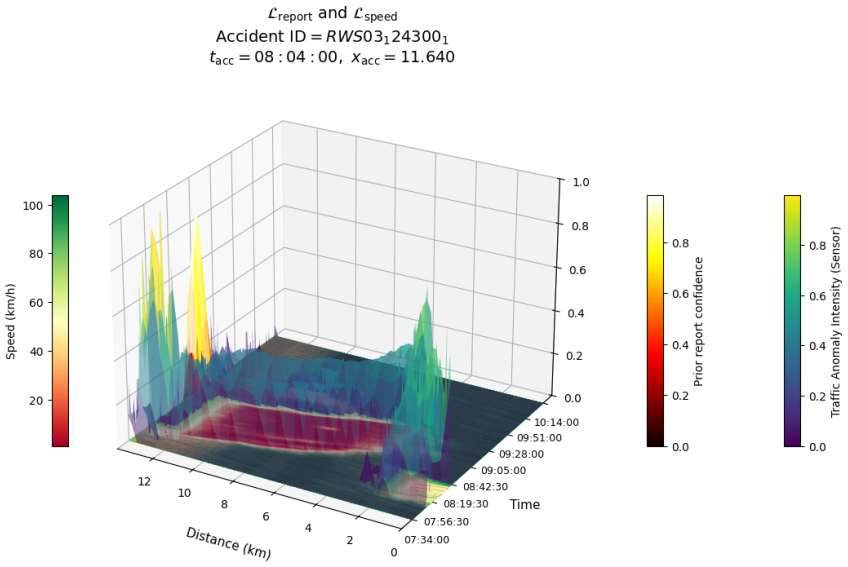


Figure 2: 3D representation of the likelihood based on speed data and accident reports.

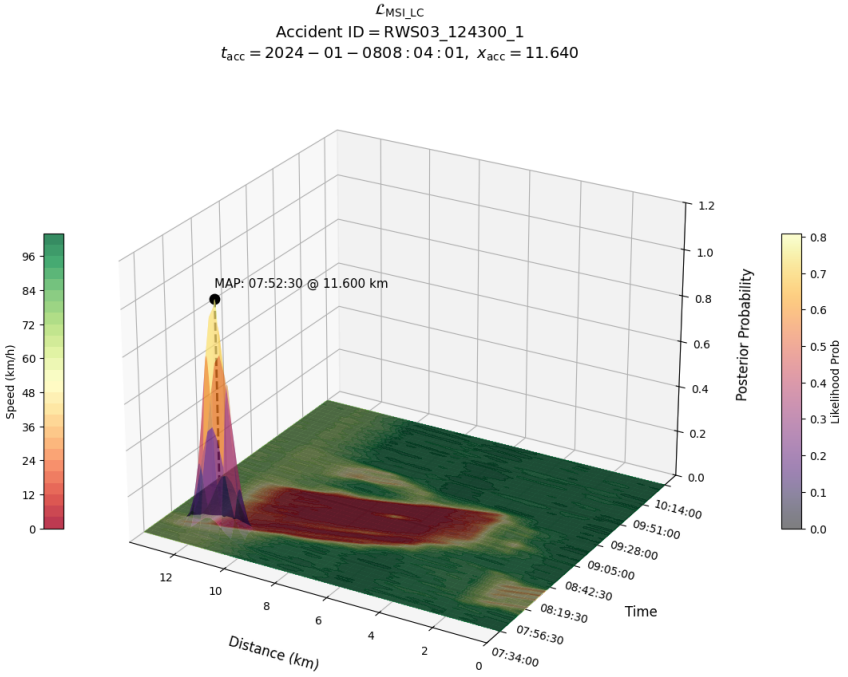


Figure 3: 3D representation of the likelihood based on lane-closure signals.

Uncertainty quantification

The uncertainty analysis in Table 1 reveals that the fused posterior shifts the estimated accident occurrence earlier than the reported time, with the 95% credible interval ending before the recorded timestamp. This suggests that disturbance patterns in traffic and signaling data emerge prior to the official report, highlighting temporal uncertainty in accident records. In terms of source contributions, signaling and report data have the largest global influence on the posterior structure (35%), while speed contributes relatively less (27%). However, at the Maximum A Posteriori (MAP) point, the estimate is primarily supported by speed and signaling, with negligible contribution from the report prior. This indicates a discrepancy between the reported record and the data-driven evidence. The posterior modality analysis revealed two distinct peaks in the inferred accident space–time distribution. However, the primary mode at 07:57 and 11.6 km was approximately 19.5 times stronger than the secondary mode at 07:52 and 12.0 km, indicating a

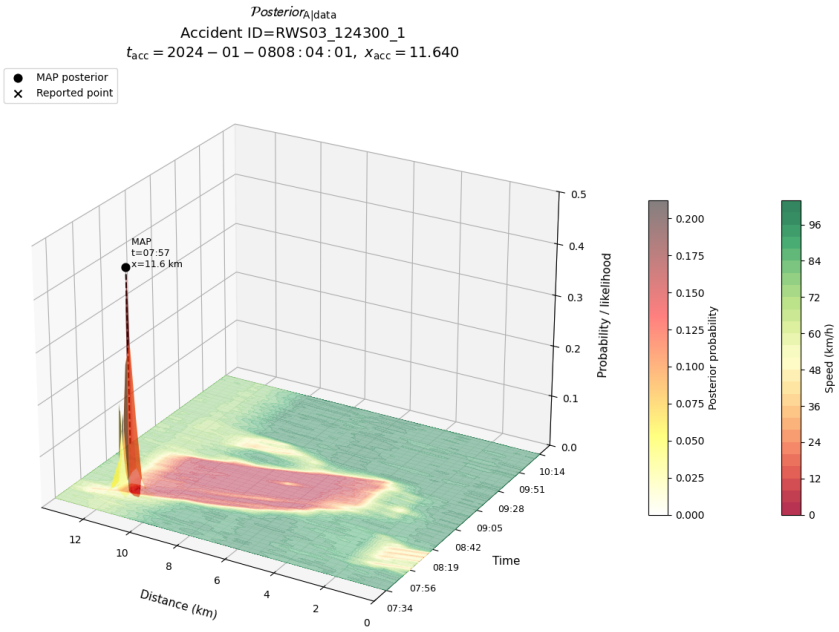


Figure 4: 3D posterior distribution showing the estimated accident location and time after Bayesian fusion.

Table 1: Uncertainty quantification summary in time and space.

	Time	**Space (km)**
CI50	07:56:30 – 07:59:00	11.6 – 11.8
CI95	07:53:30 – 08:02:00	11.4 – 12.2
Posterior std	2.05 min	0.2

clearly dominant solution with limited residual ambiguity. This suggests that, although a minor alternative hypothesis exists, the fused evidence from traffic and signaling data strongly supports a single, well-localized accident estimate. The HPD analysis indicates that 50% of the posterior probability mass is concentrated within a small region of 16 cells (1.6 min·km), reflecting a well-defined and localized core estimate. The 95% HPD region expands to 82 cells (8.2 min·km), suggesting moderate but structured spatiotemporal uncertainty around the dominant solution. Overall, the results demonstrate that while accident reports shape the global probability field, the final localized estimate is driven by traffic and signaling patterns. This emphasizes the importance of multi-source fusion in mitigating uncertainty and improving the accuracy of accident time and location estimation.

4 Conclusions

In this study, we showed that combining different traffic data sources using a Bayesian framework can improve the estimation of accident time and location. Instead of relying only on reported data, which can be uncertain or incomplete, the method uses additional information from traffic measurements and signaling systems to update and refine the results. The case study demonstrated that the fused posterior provides a more consistent and realistic estimate, often correcting the reported values. Moreover, the framework is flexible and can still work even when some data sources are missing. Overall, this approach helps reduce uncertainty in accident data and provides a more reliable basis for traffic safety analysis and decision-making. Future work will focus on systematic optimization of model parameters to improve robustness and calibration of the fusion process. In addition, the framework can be extended by incorporating richer data sources, such as congestion characteristics, capturing spatiotemporal propagation patterns.

References

ElSahly, O., & Abdelfatah, A. (2024). Developing a machine-learning-based automatic incident detection system for traffic safety: Promises and limitations. *Infrastructures*, *9*. doi: 10.3390/infrastructures9100170

Lin, Y., Li, L., Jing, H., Ran, B., & Sun, D. (2020). Automated traffic incident detection with a smaller dataset based on generative adversarial networks. *Accident Analysis and Prevention*, *144*. doi: 10.1016/j.aap.2020.105628

Lint, J. V., & Hoogendoorn, S. P. (2010). A robust and efficient method for fusing heterogeneous data from traffic sensors on freeways. *Computer-Aided Civil and Infrastructure Engineering*, *25*, 596-612. doi: 10.1111/j.1467-8667.2009.00617.x

Liu, G., Jin, H., Li, J., Hu, X., & Li, J. (2022). A bayesian deep learning method for freeway incident detection with uncertainty quantification. *Accident Analysis and Prevention*, *176*. doi: 10.1016/j.aap.2022.106796

Melev, I., & Kauermann, G. (2025). Position: There is no free bayesian uncertainty quantification.

Peixoto, T. P. (2025). Uncertainty quantification and posterior sampling for network reconstruction.

Peng, J., Shangguan, W., Peng, C., Chai, L., & Luo, R. (2024). A traffic incident detection method of highway based on spatial-temporal characteristics. *Journal of Intelligent Transportation Systems*. doi: 10.1080/15472450.2024.2416532

Sander, J., & Beyerer, J. (2013). Bayesian fusion: Modeling and application. In *2013 workshop on sensor data fusion: Trends, solutions, applications (sdf)* (p. 1-6). IEEE. doi: 10.1109/SDF.2013.6698254

Wang, Z., & Jiang, H. (2019). Simultaneous correction of the time and location bias associated with a reported crash by exploiting the spatiotemporal evolution of travel speed. *Transportation Research Part B: Methodological*, *123*, 199-223. doi: 10.1016/J.TRB.2019.03.011

Wang, Z., Liu, K., Zhu, L., & Jiang, H. (2021). Detecting the occurrence times and locations of multiple traffic crashes simultaneously with probe vehicle data. *Transportation Research Part C: Emerging Technologies*, *126*, 103014. doi: 10.1016/J.TRC.2021.103014