Sustainability And Resilience for Infrastructure and Logistics networks

D4.1 Validation Technical Report

Deliverable Number	D4.1
Author(s)	Gonzalo Durán Piñeiro, Pablo Jose Vallhonrat Blanco
Due/delivered Date	31.12.2025
Reviewed by	Corinna Köpke, Jan Tore Pederesen, Ana Maia, Benjamin Lickert
Dissemination Level	PU
Version of template	1.2

Start Date Project: 2023-06-01 Duration: 36 months Project ID: 101103978

This project has received funding from the Horizon Europe program of the European Union under grant agreement ID 101103978.

Document contributors

NO.	Name	Organisation	Role (content contributor, reviewer, other)
1	Gonzalo Durán Piñeiro	UVIGO	Main author
2	Pablo Jose Vallhonrat Blanco	CEM	Main author
3	Raquel Ortega Hita	UVIGO	Content contributor
4	Fereshteh Jafari Shahdani	UMINHO	Content contributor
5	Sahar Babri	SINTEF	Content contributor
6	Fatemeh Fadaei	POLIMI	Content contributor
7	Tobias Rinnert	FRAUNHOFER	Content contributor
8	Krzysztof Matysiak	L-PIT	Content contributor
9	Saimir Osmani	RINA-C	Content contributor
10	Concepción Toribio Diaz	CEM	Content contributor
11	Corinna Köpke	FRAUNHOFER	Reviewer
12	Jan Tore Pedersen	MARLO	Reviewer
13	Ana Maia	RANGEL	Security officer
14	Benjamin Lickert	FRAUNHOFER	Quality officer

Document revisions

Revision	Date	Comment	Author
00	01/07/2025	Draft	Gonzalo Durán Piñeiro, Pablo Jose Vallhonrat Blanco
01	24/09/2025	Review	Gonzalo Durán Piñeiro
02	29/10/2025	Review	Concepción Toribio Diaz
03	10/11/2025	Review	Corinna Köpke
04	10/11/2025	Review	Jan Tore Pedersen
05	27/11/2025	Review	Ana Maia
06	28/11/2025	Review	Benjamin Lickert

Executive summary

Climate change impact becomes more and more evident, as it increases the frequency and the magnitude of natural hazards, while man-made hazards (like war, pandemics, cyber-attacks) become more intense and present in our increasingly interconnected world. They pose a significant threat to the resilience of critical infrastructure, including transportation networks. Disruptions to these networks can severely impact the movement of people, goods, and services. The SARIL project (Sustainability and Resilience for Infrastructure and Logistics networks) aims for an encompassing resilience understanding of logistical operations incorporating green aspects. It is part of the European Union's Horizon Europe research and innovation programme under grant agreement No 101103978, included in the call "HORIZON-CL5-2022-D6-02-07[1] - New concepts and approaches for resilient and green freight transport and logistics networks against disruptive events (including pandemics)".

Based on the simulations and the level of development presented in document D3.3 'Results of the green resilience assessment', the research in this deliverable focuses on conducting an in-depth analysis of the tools developed in the project. The objective is to verify and validate their performance from a quantitative perspective, evaluating the dispersion and internal error of each one.

In addition, the performance of these tools is analysed, based on a sensitivity check of the parameters or against data from real disruption situations, in order to determine their level of precision, accuracy and potential under operational conditions.

The results of the seven evaluated tools demonstrated coherent behaviour and overall solid reliability following the verification and validation processes. The natural hazard and fire propagation models (Forest Fire Risk Map and FlamMap) showed high seasonal consistency and an 83% overlap with real fire events, confirming their predictive value. The Scour Monitoring system produced results consistent with reference models and InSAR data, evidencing structural robustness under normal conditions. The Disruption Information Interface verified the correct operation of its prediction and classification models, showing strong capability to identify and characterize disruptions, though with room for improvement in temporal accuracy and false positive reduction. The Energy Module and the Vulnerability and Traffic tools responded logically to variations in vehicle weight, slope, and traffic flow, although more detailed data are needed for refined calibration. The Traffic Simulation and ASTROIT tools effectively reproduced congestion patterns and delivery times, maintaining acceptable error margins despite data limitations and short-route biases. Overall, the tools demonstrate **methodological consistency, predictive capacity**, **and operational applicability** for resilience assessment and risk management in transport infrastructures, with improvement potential in **accuracy, calibration, and input data quality** to optimize performance under real-world conditions.

The results of this report will be considered in the work of Task 5.1 (that considers recommendations to be derived from the project work) as well as the work of Task 5.2 (that considers business models).

Table of Content

List of Figures

List of Tables

List of Acronyms

Acronym	Definition
MODIS	Moderate Resolution Imaging Spectroradiometer
SHM	Structural Health Monitoring
PS	Persistent Scatterers
InSAR	Interferometric Synthetic Aperture Radar
EGMS	European Ground Motion Service
SSE%	Percentage sum of squared errors
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
SMAPE	Symmetric Mean Absolute Percentage Error
RMSE	Root Mean Square Error
MCC	Matthews Correlation Coefficient
NMRSE	Normalised Root Mean Square Error
GPS	Global Position System
VOC	Volume-to-Capacity
GDP	Gross Domestic Product
WMAPE	Weighting Mean Absolute Percentage Error

1. Introduction

The transport sector plays a vital role in our society, facilitating the movement of people, goods, and services. However, the growing frequency and intensity of natural and man-made hazards, e.g., due to climate change, pose a significant threat to the resilience of transport infrastructure networks. These, when disrupted, can cause cascading effects impacting economic activity, emergency response capabilities, and overall societal well-being.

Work Package 1 contains the foundations that established the original SARIL scenarios and a common framework for assessing the sustainability and resilience of logistics networks. It also included early engagement with relevant stakeholders to align the framework with real-world priorities and challenges. Work Package 2 identified well-suited methodologies for managing resilience and sustainability, Work Package 3 applied these methodologies in a practical way with simulations and implementations of all the tools. The simulation results of Task 3.3 provided the necessary basis for quantitatively evaluating the precision and accuracy of the developed resilience models and tools during Task 4.1.

This deliverable presents a validation analysis of the tools developed in the project, thus obtaining a quantitative result of the precision and accuracy for each of the tools.

To this end, the report is divided into three sections:

1. The first section (chapter 2) describes the methodology used in the verification and validation of the tools.
2. The second section (chapters 3 to 9) uses historical data, sensitivity analysis, stakeholder data and literature to test and evaluate the precision and accuracy of each tool's actual operation.
3. The document comes to closure (chapter 10) with the conclusions drawn from this report.

This report provides the results of Task 4.1 (part of WP4) and describes a specific assessment based on the verification and validation of each tool developed in this project. This allows to estimate the scope and effectiveness of the tools in real or historical situations, which, together with the other tasks of WP4, provides the necessary information for possible development or implementation, thus visualising the potential of each tool to meet the final objectives of the SARIL project.

2. Methodology for Verification and Validation of Tools

The precision and accuracy of tools have been assessed through verification and validation processes. It is worth noting that due to the differences between certain tools and the partial lack of real data not all tools have been evaluated using the same approach or process. Subsequent chapters specifically describe the application of the evaluation methodology for each of the tools.

2.1 Verification Process (Precision)

Precision is understood as whether a model is stable and produces similar results under similar conditions. Consequently, if a model estimates two different outputs when using the same input, it is considered to have deficient precision. An example of a precision-deficient tool is shown in Figure 2-1 where a supposed tool estimates different inundation extents for the same rainfall (with the same input, different final outputs are obtained).

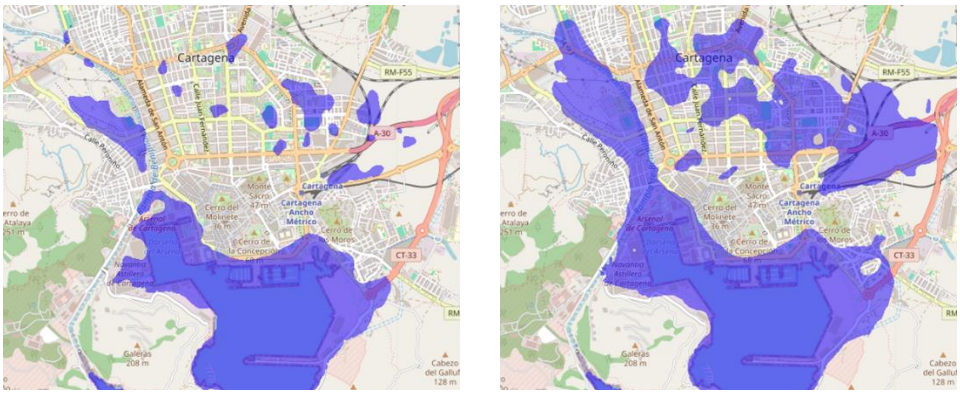


Figure 2-1: Estimated inundation extent for the same rainfall (example for illustrative purposes).

Precision has been assessed by means of verification processes, where two different approaches have been implemented:

Variation by changing input data.

Input data was modified. The outcomes of the models were evaluated to determine whether they behaved as expected in response to these variations.

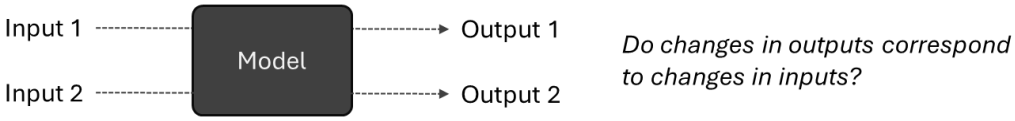


Figure 2-2: Verification via variation in input data.

Variation in modelling parameters.

Modelling parameters were modified and changes in outputs, using the same input data, were evaluated whether they behaved as expected.

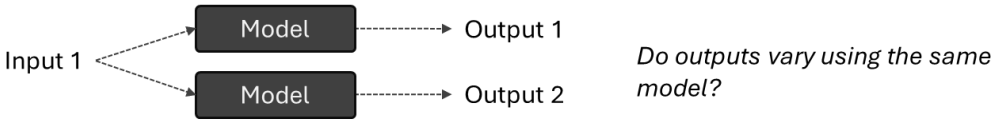


Figure 2-3: Verification via variation in modelling parameters.

2.2 Validation Process (Accuracy)

Accuracy is understood as the degree of correctness of the simulation outputs. High accuracy would therefore be related to models whose measured values are close to the true value. Accuracy has been evaluated through validation processes, where two different approaches have been implemented:

Comparison with real data.

A tool is assumed valid when its outputs reproduce reality. This is evaluated by simulating real events and ensuring that the model's outputs match those observed.

Figure 2-4 illustrates an example of a validation where it is observed that simulated data could predict the flooding in the southern area but was limited in predicting flooding in the northern region correctly.

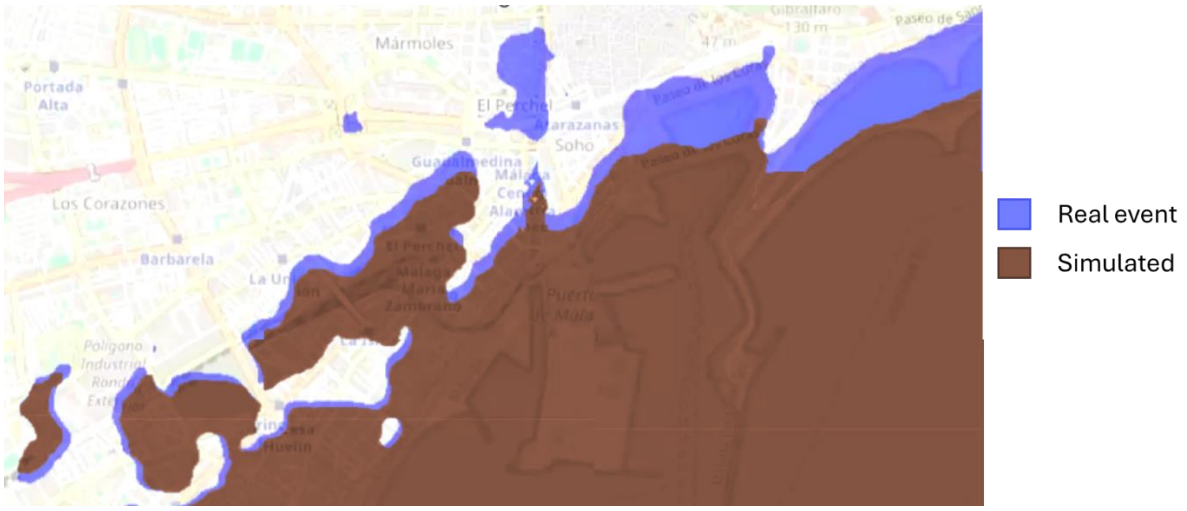


Figure 2-4: Example of validation process of flood prediction (only for illustrative purposes).

Comparison with other models.

When real data was unavailable, the validation was done by comparing outputs with those from other models that had been previously validated.

3. Natural Hazard Maps

Natural Hazard Maps is a tool that uses open-access sources to address infrastructure disruptions caused by forest fires and adverse weather events. This tool provides:

- A Forest Fire Risk Map, which uses a scale from 1 (very low risk) to 5 (very high risk) and assesses the risk of ignition in the vicinity of infrastructure based on multiple factors, such as vegetation, climate and historical data.
- Hourly maps showing potential alerts for extreme weather events, such as heavy rainfall, snow, or strong winds in a given region.
- Forest fire spread maps, indicating the Minimum Arrival Time of the fire to a specified infrastructure. To generate each of these maps, data sources such as satellite imagery, vegetation, topography, infrastructure, and meteorological conditions are essential.

This section presents the methods developed to verify and validate the Forest Fire Risk Map and the Minimum Arrival Time (propagation model of a forest fire) obtained via Flammap.

3.1 Verification (Including Sensitivity Analyses)

The verification of the Forest Fire Risk Map was conducted through a sensitivity analysis given that the module is built on analytical equations. The sensitivity analysis was carried out by substituting the NDVI (Normalized Difference Vegetation Index) and FWI (Fire Weather Index) layers with data from different dates. The main objective was to evaluate the model's responsiveness to temporal variations in vegetation condition and weather to assess the stability and robustness of the forest fire risk map outputs under different environmental scenarios.

Two maps were generated using NDVI and FWI layers characteristic of the summer and autumn seasons. Figure 3-1 presents the Forest Fire Risk Maps corresponding to the August (left) and Autumn (right) seasons, classified into five levels of risk: very low (1), low (2), moderate (3), high (4), and very high (5). In the image representing the summer season, there is a broad concentration of areas with very high fire risk (category 5), particularly in the central part of the analysed area. These zones, shown in dark red, decrease significantly in the autumn season, being largely replaced by areas of high risk (level 4).

In autumn, there is also a noticeable increase in areas with moderate (category 3) and low (category 2) fire risk, especially towards the northern and south-western parts of the map. These regions, represented in light green and light blue respectively, were less prevalent in the map corresponding to summer.

A relevant aspect is the behaviour of fire risk around road networks (depicted as blue lines on the maps). These areas predominantly show high fire risk levels during summer, although a slight reduction to moderate levels is observed along certain sections in autumn. On the other hand, some areas classified as high risk (level 4) remain constant between both seasons. Variations in fire risk maps between August and Autumn behave as expected, as they align with risk levels characteristic of each season.

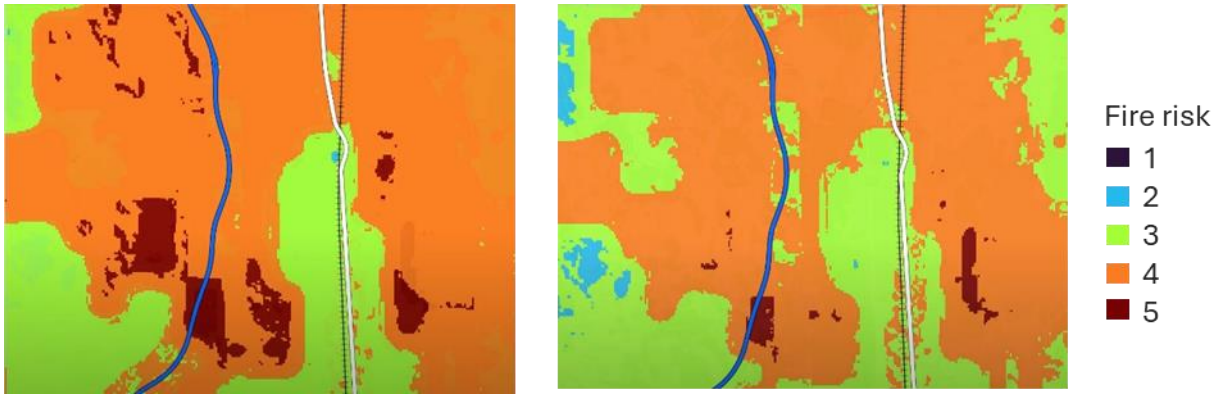


Figure 3-1: Fire risk map for the August (left) and Autumn (right) seasons.

The Minimum Arrival Time from the Flammap software was verified by adjusting internal simulation parameters. To verify this propagation model, the same fire was simulated while keeping all the necessary parameters constant, modifying only Flammap's internal settings. Specifically, the spatial resolution (pixel size) used to recalculate fire spread was varied. Subsequently, the burned area in each simulation was compared as a function of pixel size, and the percentage error was calculated with respect to the area obtained using the simulator's default resolution (30 m per pixel).

Figure 3-2 presents the errors obtained when simulating fire behaviour using different spatial resolutions, expressed in metres per pixel. The aim was to assess how the model's resolution affects the accuracy of the results, particularly in terms of the simulated burned area. The default resolution in Flammap, 30 m/pixel, is used as the reference point for calculating relative errors. When the resolution is decreased to 25 m/pixel, thereby enhancing the level of detail, the error rises sharply to 85%, indicating that higher resolution does not necessarily lead to improved model accuracy. Conversely, as the pixel size increases, a general trend of rising error is observed, although not in a strictly linear fashion. Table 3-1 displays the results for the tested resolutions. At the extreme end, with a resolution of 80 m/pixel, no ignition occurs, meaning the model fails to detect sufficient conditions to initiate fire spread. This results in a 100% error, as no burned area is generated. For improved visual clarity, the extreme error values at 25 m/pixel (85%) and 80 m/pixel (100%) have been excluded from Figure 3-2, allowing for a more detailed comparison of the intermediate resolutions.

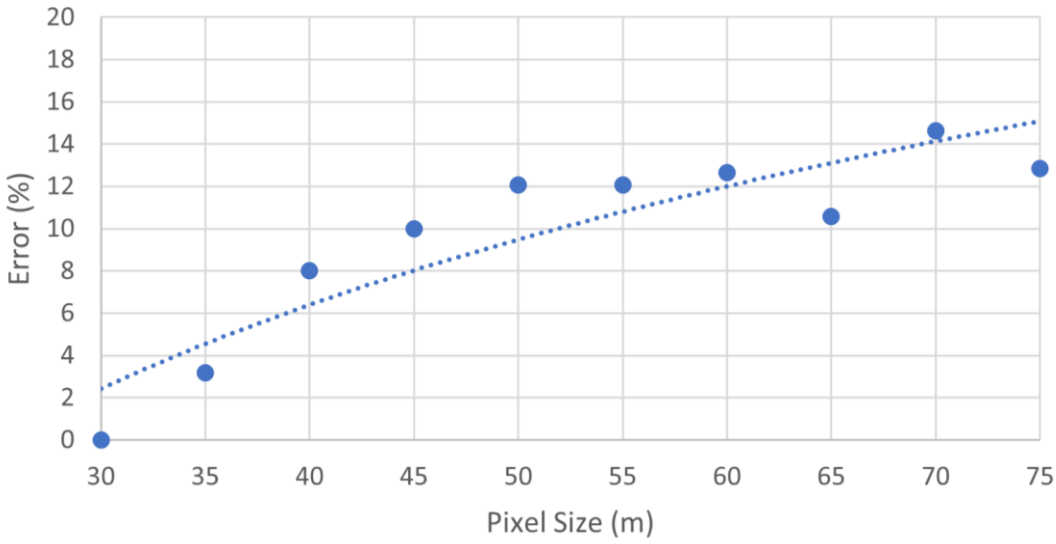


Figure 3-2: Verification results for different pixel sizes.

Table 3-1: Simulated resolutions and resulting errors.

Pixel Resolution (m)	Error (%)
25	85
30 (default)	0
35	3,16
40	7,99
45	9,98
50	12,08
60	12,07
65	10,57
70	14,64
75	12,84
80	100

The aim of this analysis has been to select a resolution that provides accurate results without significantly increasing resource consumption or computation time. The SARIL project adopts FlamMap's default resolution of 30 m/pixel, as evaluations using both higher and lower resolutions exhibited increased deviations from the default output without offering methodological advantages.

3.2 Validation

To validate the Forest Fire Risk Map, the spatial distribution of moderate-risk (level 3), high-risk (level 4) and very high-risk (level 5) pixels were compared with historical forest fires, obtained from MODIS satellite observations. This approach has been qualitatively applied in previous studies (1), (2).

The hotspots detected by the MODIS sensor correspond to active fires identified by NASA's Terra and Aqua satellites. Each point has a spatial resolution of 1 kilometre, meaning it represents an area of 1 km × 1 km where thermal activity consistent with fire has been recorded. To validate the Forest Fire Risk Map, the historical record of these detections has been compiled from 2017 to August 2025. From this point forward, these records will be referred to as *real events*.

The methodology employed involved comparing the real events with the Forest Fire Risk Map. To do this, from all the real events falling within the study region, only those that coincided with areas classified as moderate, high, or very high risk were selected. Subsequently, the total area corresponding to these selected events was calculated and divided by the total area classified under moderate, high, and very high risk levels. This metric is referred to as the *Detection Rate*, and its expression is given by Equation 3-1.

Figure 3-3 displays simulated risk levels 3, 4, and 5 in orange, areas where real fire events intersect with simulated are shown in yellow, while the remaining fire events are marked in black.

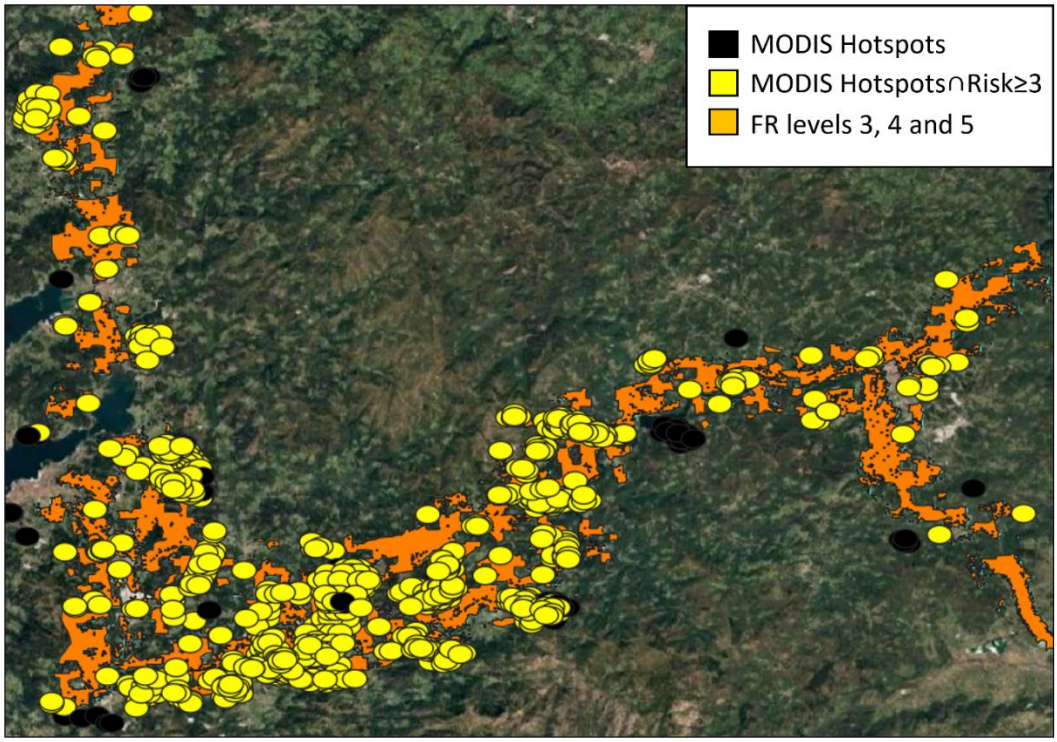


Figure 3-3: Comparison of moderate, high and very high risk areas with historic fire events.

Validation was conducted using the Detection Rate metric, which represents the percentage of fires that occur in areas classified as risk levels 3, 4, and 5, relative to the total area of the map assigned to these levels. The numerator corresponds to the area of overlap between the area of the real events and the areas of the Forest Fire Risk Map classified as level 3 or higher, while the denominator represents the total area of the map covered by levels 3, 4, and 5.

$$Detection\ Rate = \frac{Area(Hotspots) \cap Area(Risk \geq 3)}{Area(Risk \geq 3)} \qquad \textit{Equation 3-1}$$

Table 3-2 shows the relationship between real events and areas classified as moderate risk or higher in the Forest Fire Risk Map. The area affected by fires or areal events (MODIS hotspot) that coincides with moderate, high and very high risk levels (3, 4 and 5) is 109,900 hectares, while the total area estimated by the Fire Risk Map classified at these levels amounts to 131,619.48 hectares. This corresponds to a detection rate of 83 %. The results show that, between 2017 and August 2025, 83 % of fires in the study region in Galicia took place in regions classified as moderate risk or higher, which means that the Forest Fire Risk map successfully identified the majority of the areas where fires effectively occurred. An 83% detection rate can be considered a strong performance given the complexity and variability of fire behaviour. The ambition of the analysis was to assess and verify the reliability of the risk map in anticipating the spatial distribution of fire events. Therefore, achieving more than 80% alignment between predicted risk levels and actual fire locations suggests that the map fulfils its objective satisfactorily, although there is still room for improvement in capturing the remaining 17% of events that occurred outside the predicted risk zones.

Table 3-2: MODIS Fire Hotspots (real events) overlapping moderate and high risk areas (levels 3–5).

Affected Area (Hotspots∩Risk≥3) (Ha)	Total Area (Ha)	Detection Rate
109.900,00	131.619,48	83 %

To validate the Minimum Arrival Time given by Flammap, the results of a simulated forest fire event were compared with a real forest fire for which both the duration and the final burned area are known. This data has been provided by the public administration of the Autonomous Community of Galicia. The simulation was run using the same ignition point and weather as the real event, including wind, temperature, relative humidity, precipitations and cloud cover conditions. The simulated total area affected was then compared to the real data, according to the following equation, which represents the relative error,

$$\varepsilon = \frac{|Area_{real} - Area_{simulated}|}{Area_{real}} \cdot 100 \qquad \textit{Equation 3-2}$$

Table 3-3 presents the distribution of errors obtained from the simulation of various forest fires. On average, an error of 48% was recorded. Figure 3-4 illustrates the analysed fires, where it can be seen that, qualitatively, simulations reasonably replicate the shape and extent of real fires. However, the quantitative analysis reveals a significant discrepancy, with an error of 40.5% for the 1st forest fire, 45% for the 2nd one and 8,28% for the 4th one. Although the results show considerable variability in the errors this does not mean that the tool is unreliable or unsuitable for practical use. It is important to highlight that the simulations were carried out attempting to reproduce real fires based on their known final area and the estimated time at which they were brought under control. However, these times do not necessarily correspond to the moment when the fire would have evolved freely, but rather to operational decisions made by suppression commanders, which can vary between incidents and do not always reflect a natural spread dynamic. This subjectivity in estimating the "final time" introduces an unavoidable source of discrepancy in the quantitative comparison. The aim of these simulations is not to replicate the burnt area with absolute accuracy, but rather to assess the model's ability to reasonably reproduce the general patterns of fire spread. From this perspective, the qualitative agreement observed in the simulated shapes and extents indicates that the tool is useful as a support for planning and analysing fire behaviour, although it does present limitations when used for high-precision quantitative reconstructions. Table 3-3 summarises the data used for the simulations, along with the error obtained for each case.

Table 3-3: Data on simulated forest fires.

NO.	Total Minutes	Real Burnt Area (Ha)	Simulated Burnt Area (Ha)	ε (%)
1	1200	75,89	106,65	40,5
2	2000	63,60	92,43	45
3	330	22,7	0	100
4	1470	39,73	43,02	8,28

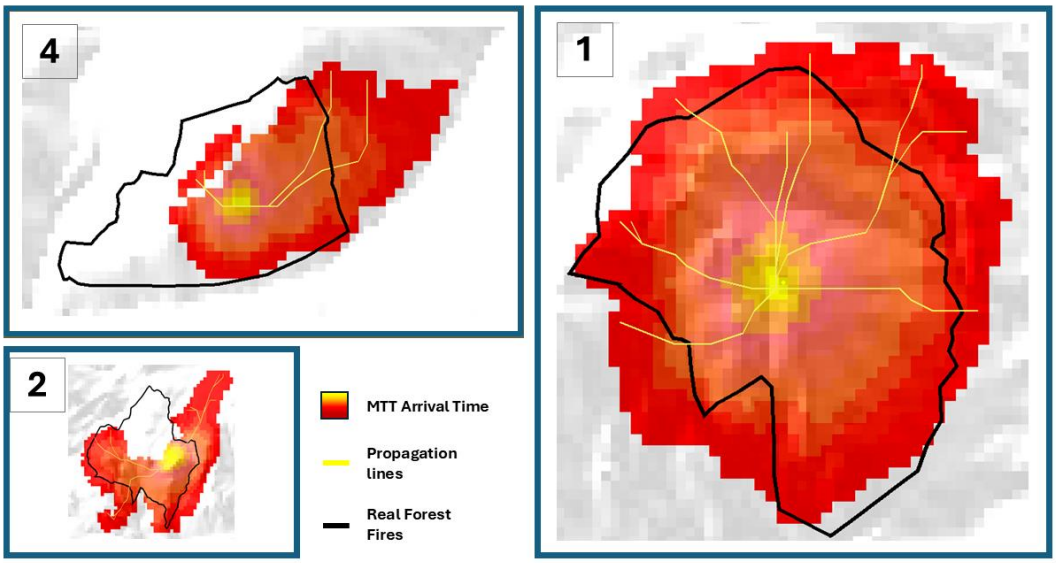


Figure 3-4: Comparison between simulated burnt areas and real burnt areas.

FlamMap is a software tool that models fire spread based on meteorological data, fuel characteristics, and terrain topography. While it allows for the definition of containment zones where fire does not spread, it **does not incorporate active suppression measures**. In this study, such zones were not considered due to the lack of detailed information regarding fire management in each case.

Given this limitation, only real forest fires with simple characteristics were selected: small areas and short development periods (no more than 1.5 days). However, these timings are not objective; they depend on the decision made by the command post responsible for coordinating the firefighting resources to determine whether the fire is stabilised, controlled, or extinguished. According to the *PLADIGA* and the *Manual de Prevención e Defensa Contra os Incendios Forestais de Galiza*, developed by the public administration, a fire is considered active when it is in the process of spreading, with visible flames and a progressing fire front, which requires direct intervention by firefighting teams (3).

Once the spread of the fire has been halted, but there is still a risk of reactivation, the fire enters the stabilised phase. During this stage, monitoring continues, and firefighting resources remain on the ground to prevent the fire from reigniting or spreading again.

A fire is declared controlled when there is no possibility of further spread, although hotspots may remain within the affected perimeter. The perimeter is secured, and the fire is under operational control. This declaration is made by the technical director of the operation, based on the fire's progression and the prevailing weather and terrain conditions.

Finally, a fire is considered extinguished when no flames or hotspots remain, and it has been verified that there is no risk of reactivation. At this point, the intervention is deemed complete (3), (4).

In the simulations, the timings used correspond to the official declaration of the fire as stabilised. However, as noted in the text, these timings may be conservative, meaning they are estimated from a precautionary standpoint. This implies that they do not necessarily reflect the exact moment the fire stops advancing but are adopted as a safety measure to ensure that the fire does not reignite or spread

again. Under these conditions, the average error observed in the simulated fire propagation times is 48%.

Additional discrepancies between the simulated and actual fire perimeters can be explained by several key factors. One of the primary limitations of our FlamMap simulations is the inability to incorporate active fire suppression efforts. In real wildfire events, ground and aerial firefighting operations play a decisive role in limiting fire spread. Due to the lack of data on firefighting operations, such as the use of preventive burning to slow the advance of the fire (introduced in FlamMap as containment zones or firebreaks), these interventions could not be reflected in the simulations, likely resulting in an overestimation of the burned area.

Furthermore, the selection of cases focused exclusively on small, short-duration fires, under the assumption that their behaviour would be easier to replicate. However, even in these simplified scenarios, the accuracy of the simulations is highly dependent on the quality of the input data. The use of generalised fuel models, averaged weather conditions, and limited topographic detail, combined with the spatial resolution and currency of the input layers, introduces a degree of uncertainty into the results.

In summary, while the model is capable of approximating the general shape and extent of a fire under certain conditions, notable quantitative differences remain. This analysis underscores that such discrepancies stem from a combination of modelling limitations, data constraints, and the inherent unpredictability of real-world fire behaviour. FlamMap is particularly suited for analysing potential fire behaviour in small to medium-sized events, especially under relatively homogeneous terrain and fuel conditions, and for applications where the objective is to identify likely spread directions, assess risk, or support preventative planning. For complex fire behaviour or scenarios involving extensive suppression interventions, outputs should be interpreted with caution. Enhancing data availability, refining model calibration, and fostering closer collaboration with relevant authorities will be essential to improve the reliability and practical utility of future simulations.

4. Scour Monitoring for Decision Support

For effective emergency management of a bridge network during flood events, accurately estimating the scour hole at bridge piers and abutments in real-time is critical. The Scour Monitoring for Decision Support tool is a tool which leverages real-time SHM sensor data from critical bridges to infer scour depth at unmonitored locations. It then integrates the reliability estimates with an analysis of both direct and indirect consequences of different management strategies and based on the Principle of Maximum Utility, gives the optimal strategy as the management plan that maximises expected utility.

4.1 Verification (Including Sensitivity Analyses)

To verify that the tool performs as intended, a sensitivity analysis was conducted on the risk outcomes. The objective was to assess whether varying the input values leads to reasonable and consistent results. The analysis focused on key inputs influencing flood intensity across the four bridges outlined in D3.3: the discharge level (q) and the correction factor for the flow's angle of attack (k_2).

As shown in Figure 4-1, an increase in discharge resulted in a shift of the optimal management plan—the plan associated with the minimum risk. A management plan represents the combination of actions applied to the four bridges—either closure or remaining open—yielding 16 possible configurations for the network. For instance, at a discharge of 100 m^3/s (measured upstream of the bridges), the optimal strategy was to keep the first and third bridges open while closing the second and fourth. However, once the discharge exceeded 150 m^3/s, the optimal action changed to closing all four bridges. This outcome is logical, as higher discharge levels correspond to more severe flood events, which increase both the probability of deeper scour and the likelihood of bridge failure.

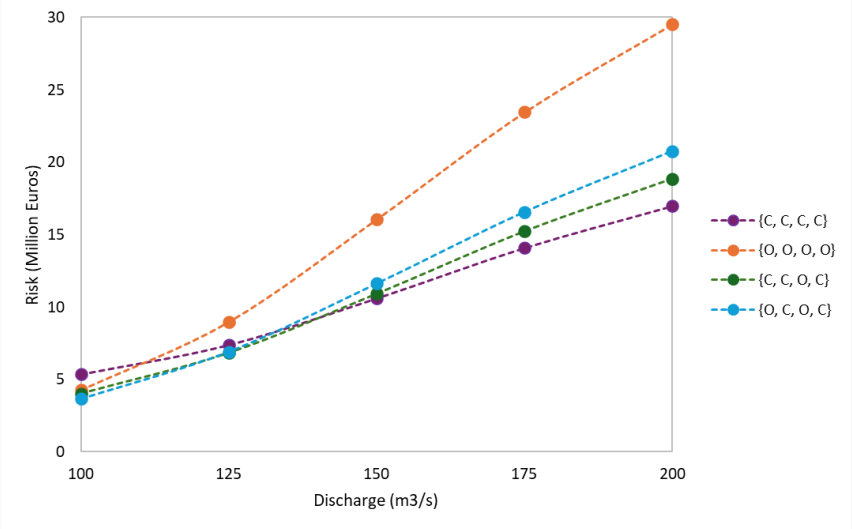


Figure 4-1: Risk versus Discharge plot for four different decision alternatives. The plot legend represents the decision alternatives, where C indicates that the bridge is closed, while O indicates that the bridge remains open.

According to (5), the correction factor for the angle of attack of flow depends on both the ratio of a pier's length to its width and the skew angle of the flow. As illustrated in Figure 4-2, an increase in the angle of attack—representing a more severe condition—leads to a shift in the optimal management plan. Specifically, the strategy changes from keeping all bridges open to closing all bridges, which is a logical outcome given the heightened likelihood of having a deeper scour depth associated with a more skewed flow.

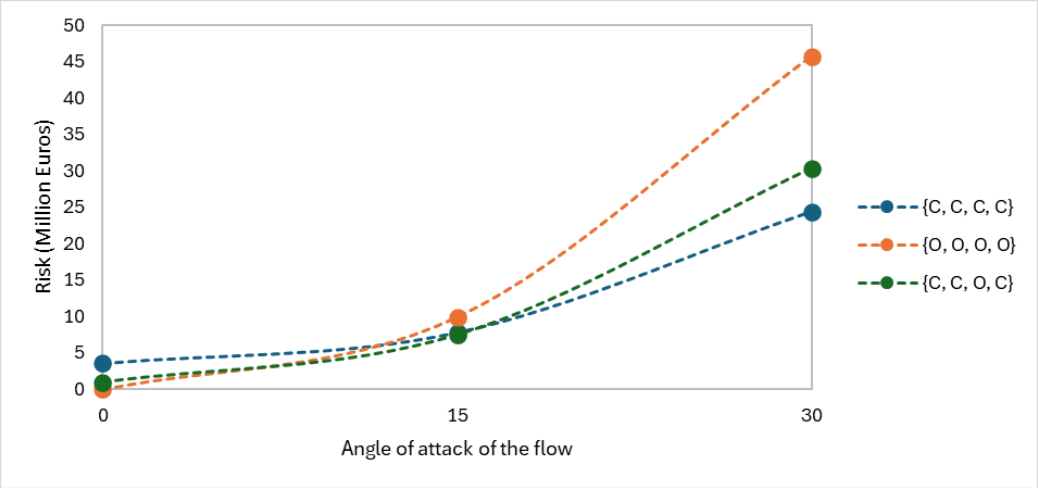


Figure 4-2: Risk versus Angle of attack of the flow for three different decision alternatives

4.2 Validation

Unlike predictive models, which can be validated by comparing their outputs to observed real-world data, a decision support tool cannot be validated in the same way because its outcome is a decision rather than a measurable quantity. In the case of bridge management during floods, there is no record of a true decision that can be used as a benchmark. Historically, decisions were made heuristically, such as closing all bridges once a certain water level was reached or expected, which does not provide a reliable ground truth for comparison. Moreover, a true decision does not objectively exist. Instead, the Scour Monitoring for Decision Support tool supports decision-making by identifying the decision alternative that minimises risk. While one cannot claim that a given decision is the true or only correct choice, the tool enables the selection of a decision that balances safety and losses in functionality in the most effective way.

As a result, the tool was validated by comparing the predicted scour depth outcomes from the Scour Monitoring for Decision Support tool with those obtained from a well-established model in the literature. For this purpose, Johnson's model was selected, as it uses the same input parameters as the Scour Monitoring for Decision Support tool (6). The prediction errors were then quantified using the following equation:

$$SSE\% = \frac{\sum(y_s^{Johnson} - y_s^{SMDST})^2}{\sum(y_s^{Johnson})^2} \times 100 \qquad \textit{Equation 4-1}$$

where SSE% represents the percentage sum of squared errors, $y_s^{Johnson}$ denotes the scour depth predicted by Johnson's model, and y_s^{SMDST} denotes the scour depth predicted by the Scour Monitoring for Decision Support tool.

To perform this validation, scour depth predictions from the Scour Monitoring for Decision Support tool at the second pier of the first bridge (an unmonitored pier) were compared with Johnson's model predictions across a range of discharges. The resulting SSE values are presented in Table 4-1.

Table 4-1: Sum of squared errors (%) for scour depth.

	Q = 100 m³/s	Q = 125 m³/s	Q = 150 m³/s	Q = 175 m³/s	Q = 200 m³/s
$y_s^{Johnson}$	8.7708	9.4824	10.1137	10.6795	11.199
y_s^{SMDST}	11.5014	12.1524	12.7249	13.2414	13.7074
SSE (%)	11.40	9.28	7.81	6.76	5.92

The results in Table 4-1 show that the percentage sum of squared errors (SSE%) between the scour depths predicted by the Scour Monitoring for Decision Support tool and Johnson's model decreases consistently as the discharge increases. At lower flows (Q = 100 m³/s), the error is 11.40%, but it reduces steadily to 5.92% at Q = 200 m³/s. This trend indicates that the predictions of the Scour Monitoring for Decision Support tool become increasingly consistent with those of the established Johnson's model under higher flow conditions. Although some discrepancy is observed at lower discharges, the overall errors remain within a moderate range, suggesting that the tool provides reasonably accurate scour depth estimates. These results support the reliability of the Scour Monitoring for Decision Support tool as a basis for decision-making during flood events.

Usage of InSAR

Among the structural parameters used to detect anomalies in civil structures and infrastructure, displacements are widely employed in Structural Health Monitoring (SHM) of existing assets. More recently, the increasing availability of satellite remote sensing data, particularly Synthetic Aperture Radar (SAR) imagery, has enabled large-scale, non-contact monitoring with medium-to-high temporal and spatial resolution. In particular Interferometric Synthetic Aperture Radar (InSAR) techniques provide displacement time series with millimetric precision for specific reflective elements referred to as Persistent Scatterers (PS), i.e., buildings, rocks, and corner reflectors.

The potential of InSAR for scour detection was demonstrated, for example, by (7) focusing on the case of Tadcaster Bridge in England. The bridge suffered a partial collapse on the evening of December 29, 2015, following severe rainfall and flooding. Using the Small Baseline Subset (SBAS) InSAR method, 48 high-resolution TerraSAR-X (X-band) images collected over the two years preceding the incident were analyzed. The results showed that, although the bridge as a whole remained stable (with deformation within 2 mm per year), a specific distributed scatterer near the failed pier displayed distinct non-linear, millimetre-scale deformation. This anomalous behavior was detected in the final two image acquisitions, approximately one month before the collapse, coinciding with a period of intense rainfall and elevated river flow.

Building on this approach, one way to further validate the Scour Monitoring for Decision Support tool, is to check the consistency between its estimated probability of failure for the piers of a specific bridge in the network and displacement data derived from InSAR observations for the corresponding area of that bridge.

Unfortunately, since data for periods of heavy rainfall and flooding are unavailable, we will conduct the analysis under normal conditions to support the verification of our tool's precision and accuracy.

The measured displacements are expected to remain within the range associated with normal structural performance. This same procedure could be applied when adequate data for periods of heavy rainfall and flooding are available.

In this section, the validation procedure is demonstrated under normal conditions. The displacement data from the European Ground Motion Service (EGMS) were analysed. EGMS products are made available for free by the Copernicus Land Monitoring Service and provide standardized, validated ground motion measurements derived from Sentinel-1 imagery. Two datasets were used: EGMS_L2a_117_0269_IW1_VV_2019_2023_1 (ascending orbit, track 117) and EGMS_L2a_168_0804_IW2_VV_2019_2023_1 (descending orbit, track 168), spanning the period 2019-2024.

Of the four bridges examined in document D3.3, Bridge 2 was selected for validation for two main reasons. First, the EGMS datasets provided the best PS coverage on this structure compared to the other bridges in the network. Second, its longitudinal axis is almost aligned with the east-west direction, which makes it particularly suitable for InSAR analysis, as displacements along this orientation are more effectively detected by the radar (8).

To investigate longitudinal displacements, the bridge deck was subdivided into 7 distinct zones with a length of 30 meter each, spanning from west to east, as shown in Figure 4-3. Negative velocities (in red) indicate displacements away from the satellite, and positive velocities (in blue) indicate displacements toward the satellite.

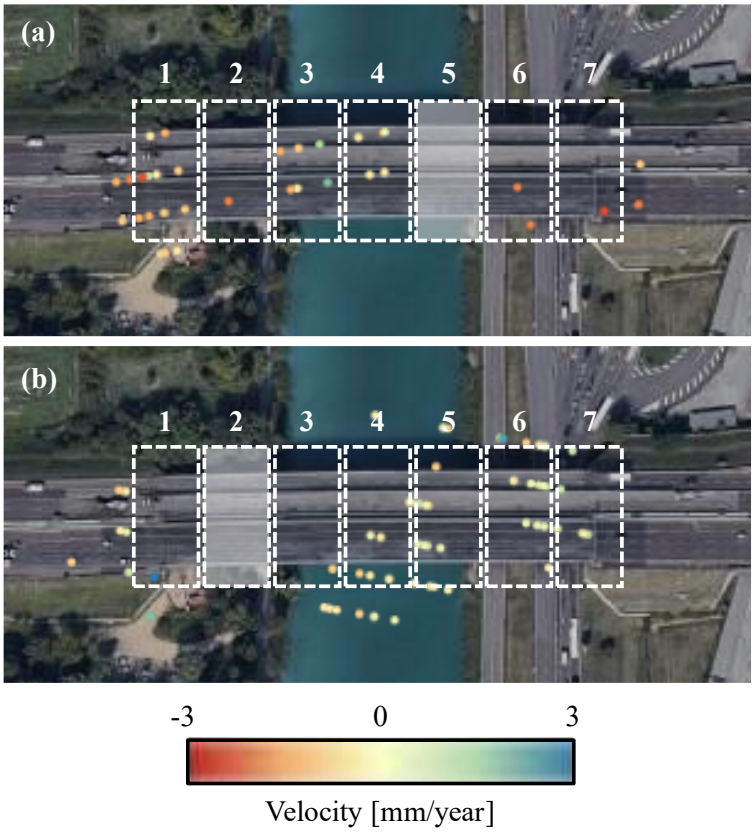


Figure 4-3: Clustering of Persistent Scatterers (PSs): (a) ascending dataset; (b) descending dataset.

Displacements trends along the longitudinal direction were estimated for zones 1, 3, 4, 6 and 7 by the combination of ascending and descending orbit datasets, according to the methodology described in (9). Figure 4-4 presents these displacements.

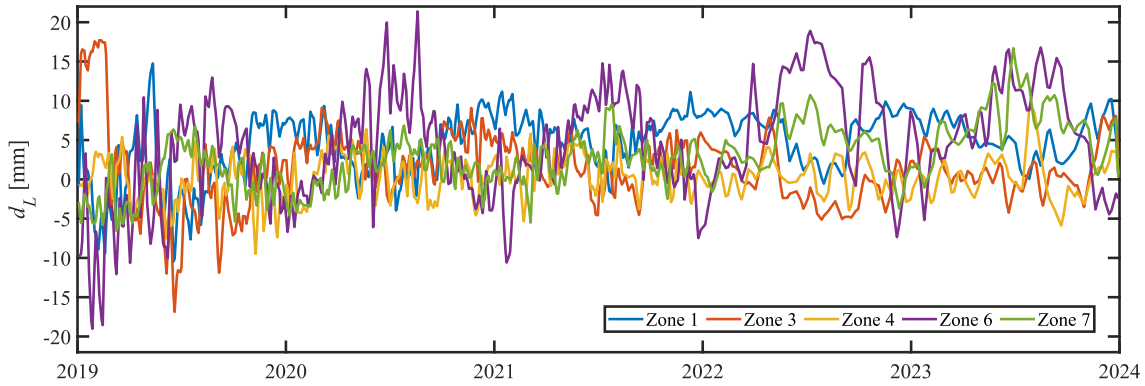


Figure 4-4: Longitudinal displacement of the influence zones

The analysis aimed to evaluate whether the longitudinal displacements in zone 4 (the zone where piers are located), after removing the thermal effects—since longitudinal displacements are strongly influenced by environmental temperature—indicated normal structural behaviour. Because the bridge did not experience any major flooding or extreme water flow events during this period—conditions that could have caused significant scour and led to anomalous displacements—the observed displacements were expected to remain within the range associated with normal structural performance.

Figure 4-5 presents the scatter plots of longitudinal displacements versus environmental temperature for some of the zones, where each point represents a paired observation of displacement and temperature. The position of each point along the x-axis corresponds to the temperature value, while its position along the y-axis reflects the associated displacement. The scatter plots in Figure 4-5 reveal that the majority of the displacement time histories exhibit a high linear correlation with temperature, either positive or negative, depending on the relative position with respect to the piers. The correlation between the longitudinal displacements of zone 4 and the temperature is minor due to the location of the fixed supports (piers).

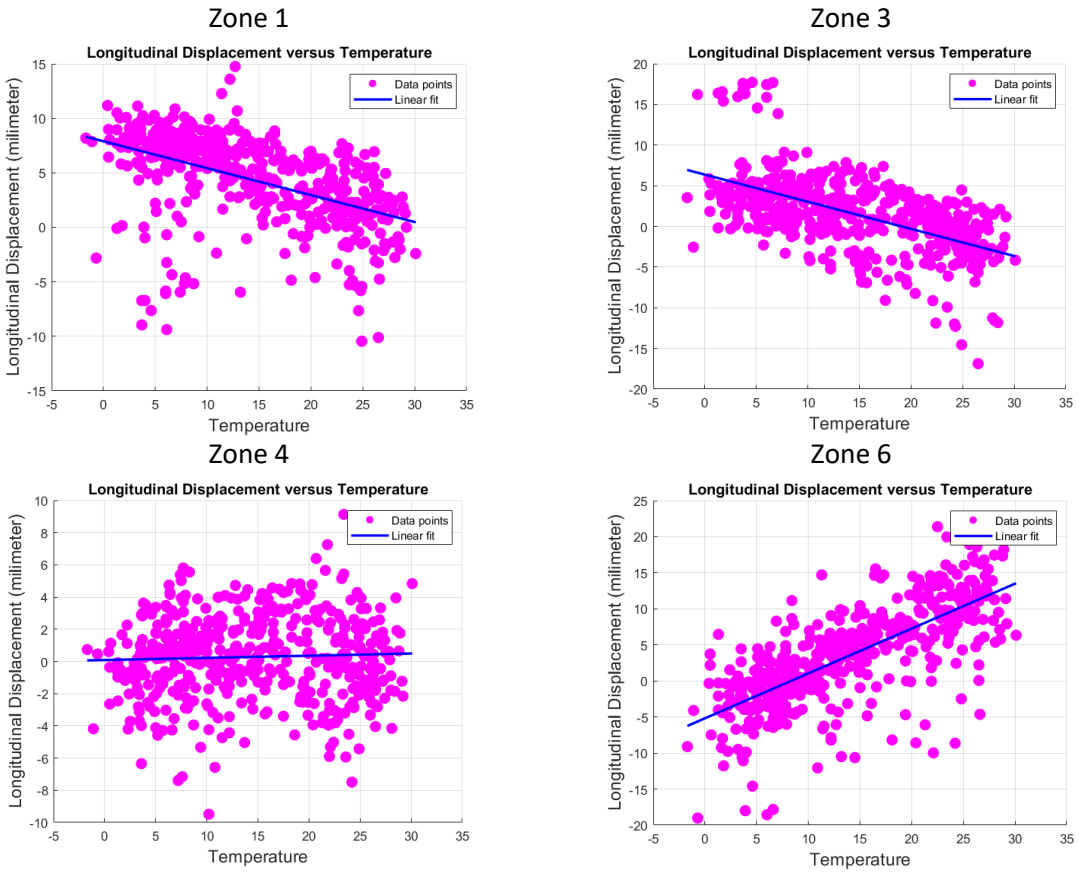


Figure 4-5: Plots of longitudinal displacements and temperature datasets.

These results in all zones suggest that the longitudinal displacements are likely to be caused by the thermal expansion of the bridge due to the variation of the environmental temperature.

To quantify the relationship between temperature and longitudinal displacements, a linear regression analysis was performed:

$$\text{Longitudinal displacement}_{\text{Zone i}} = \beta_{1,\text{Zone i}} \times \text{Temperature} + \beta_{0,\text{Zone i}} \qquad \textit{Equation 4-2}$$

where $\beta_{0,\text{Zone i}}$ is the longitudinal displacement of the zone i, i = 1,…, 7, at 0°C and $\beta_{1,\text{Zone i}}$ is the slope of the fitting line. Table 4-2 contains the coefficients of the linear model for each zone.

Using Equation 4-2, we can remove the component of the longitudinal displacements attributed to thermal effects and then assess whether the remaining (residual) displacements fall within an acceptable range.

To achieve this, the bridge is assumed to have been in a healthy condition during the first two years of recorded data (2019–2021). The coefficients of Equation 4-2 derived from this training period (Table 4-2) are then used to mimic the displacements caused by thermal effects in the subsequent three years. Figure 4-6 presents the residuals, defined as the difference between the observed and simulated displacements.

Table 4-2: Coefficients of the linear model between longitudinal displacement and temperature data for the training period.

	β_1	β_0
Zone 1	-0.2256	5.9474
Zone 3	-0.4503	8.4747
Zone 4	0.0028	-0.1605
Zone 6	0.6165	-6.9577

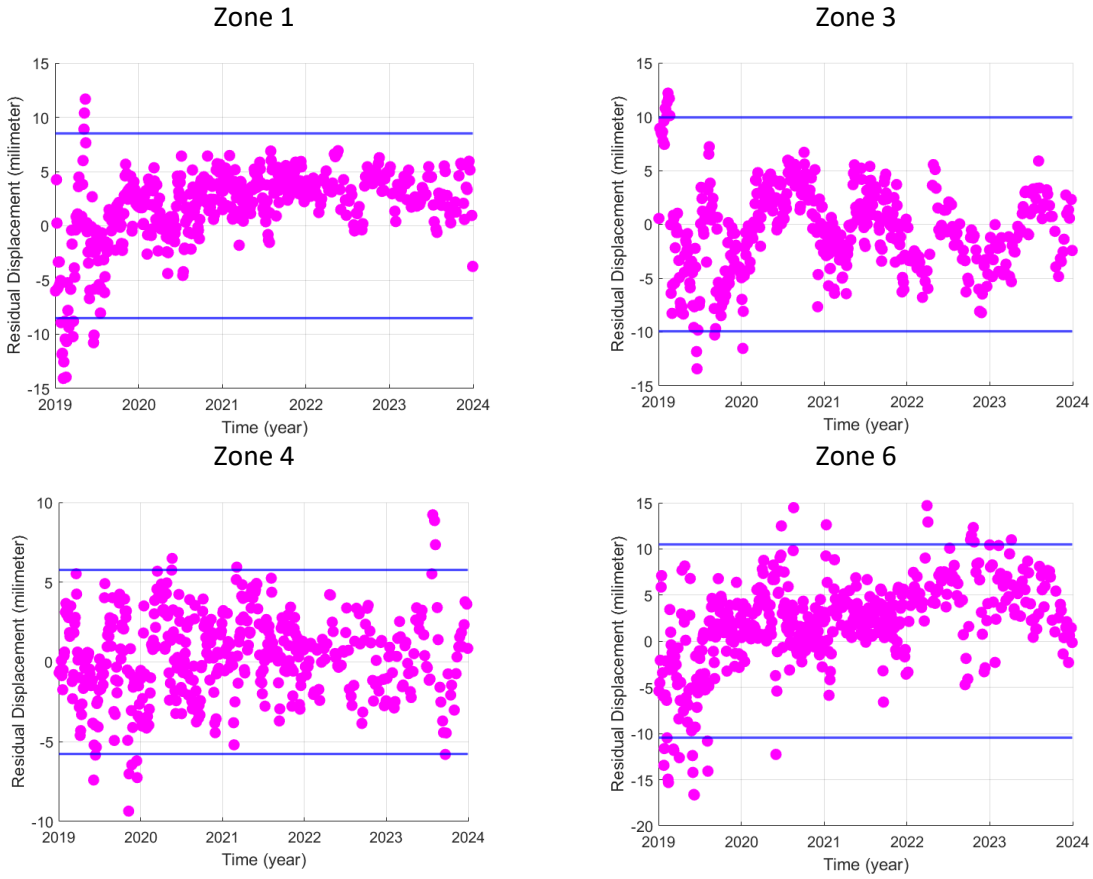


Figure 4-6: Residuals related to the simulated longitudinal displacements.

Figure 4-6 shows that, for Zone 4 (the zone containing the piers), most residuals lie within the band defined by the mean residual of the training period ± 2σ (indicated by the blue lines), suggesting that the bridge's longitudinal displacement has not changed significantly over time. This observation aligns with expectations, as no extreme flow events occurred during this period that could have induced significant scour affecting the bridge's displacement.

As previously stated, validation of the Scour Monitoring for Decision Support tool would require flood events or periods of heavy rainfall to fully assess the performance of the tool. Therefore, the current approach illustrates the methodology and workflow for validation, which can be applied once relevant satellite or observational data become available.

5. Disruption Information Interface

The Disruption Information Interface informs the transport network user about ongoing disruptions, their type (e.g., flood, accident), and provides a brief description of the problem. The tool also collects hazard information from transportation services or project partners tools (e.g. output files with fire risk level from Natural Hazard Maps tool [Figure 3-1]). A significant emphasis has been placed on supporting the user in resolving operational issues by providing an end time prediction for disruptions. Furthermore, to support the end user, the tool highlights messages that indicate disruptions, saving time that would otherwise be required by the end user to sort through hundreds of aggregated messages. The presented requirements were directly translated into validation and verification issues, which are analysed below. Both verification and validation will be performed in the context of two features: “correctness of the function for predicting the end of road disruptions” and “classifier functionality for messages containing disruption information”.

5.1 Evaluation of the Model Used for Predicting the End of Road Disruptions (Survival Analysis)

To assess the predictive reliability of the model created for predicting the end of road disruptions, an out-of-sample validation process was designed Figure 5-1). The model is based on Kaplan-Meier survival estimates (10) (11), a statistical method traditionally used in medical survival analysis. Survival analysis is a set of various statistical methods primarily used in medicine to predict the time of death of a patient from the time of diagnosis. Survival analysis methods allow us to detect correlations between covariates (e.g., the concentration of administered medications, whether the patient was treated), and the estimated time to death. Traffic disruptions can be modelled in a similar way, where the end time indicates the elimination of the traffic problem. The difference with a similar application in medicine lies in the interpretation of the duration – in the case of traffic disruptions, the shorter the duration, the better for road users. Historical data on traffic disruptions, including precise timestamps for event start and end, were collected from information published in government traffic databases. The training period was limited to events recorded from July 24, 2024 up to March 21, 2025, ensuring that future disruptions could serve as unseen validation cases. For all events after this cutoff date, only the start time and categorical attributes (e.g., disruption type) were kept, while the actual end times were deliberately hidden. This data will be referred to as the *sample* in the following sections.

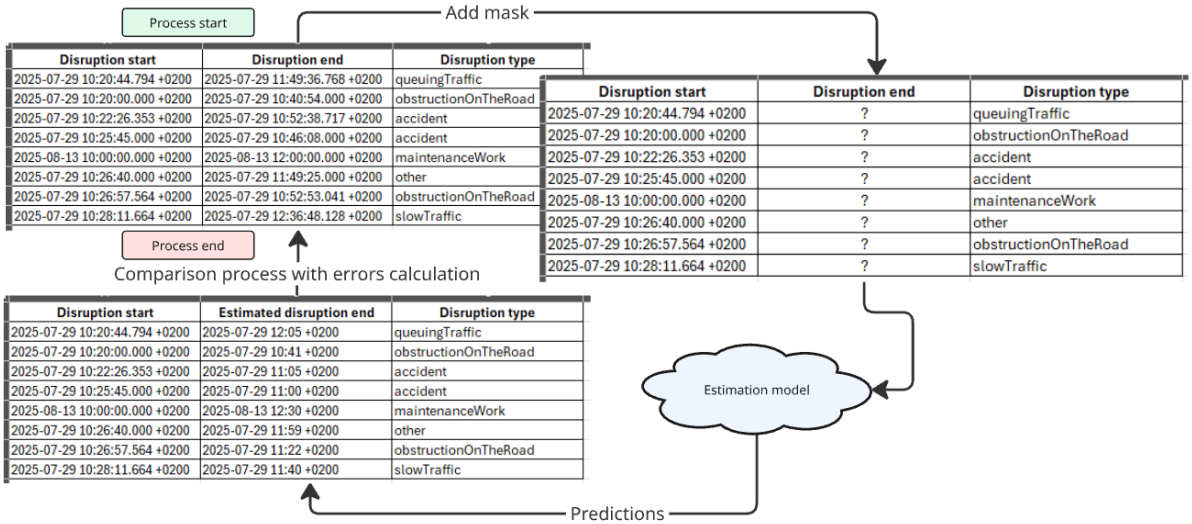


Figure 5-1: Evaluation process flow.

This approach allowed the survival model to generate predictions under realistic conditions, without access to future outcomes.

The masked end times were then used as a ground truth benchmark against the Kaplan-Meier survival functions. Specifically, for each disruption type, the model's estimated probability of event continuation at fixed time horizons (e.g. 2h, 4h, 6h, etc.) was compared to the observed termination time of the disruptions. Performance was quantified using established error metrics:

- **MAE** (Mean Absolute Error) – tells us the average size of the errors in hours, showing by how much predictions differ from the actual disruption duration, regardless of direction:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Actual_i - Predicted_i| \qquad \text{Equation 5-1}$$

- **MAPE** (Mean Absolute Percentage Error) – expresses the average error as a percentage of the actual duration, highlighting how large the errors are compared to the true values:

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{Actual_i - Predicted_i}{Actual_i}\right| \qquad \text{Equation 5-2}$$

- **SMAPE** (Symmetric Mean Absolute Percentage Error) – a percentage-based error similar to MAPE, but normalized to avoid extreme values when actual durations are very small. It balances over- and under-estimations:

$$SMAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|Predicted_i - Actual_i|}{(|Actual_i| + |Predicted_i|)/2}$$ *Equation 5-3*

- **RMSE** (Root Mean Square Error) – gives more weight to large errors by squaring them before averaging. This highlights situations where the model makes a few very large mistakes:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Actual_i - Predicted_i)^2}$$ *Equation 5-4*

This framework provided a robust mechanism to assess both the accuracy and precision of survival-based predictions across disruption categories, highlighting where the survival curves aligned with or diverged from real-word outcomes (Figure 5-1).

For each category of disruption, the evaluation incorporated a systematic choice of the horizon. This means in practice:

- Identification of the smallest horizon *t* at which the survival curve indicated that more than 60% of disruption were already ended (i.e. *S(t) < 0.4*). To determine the most appropriate horizon, multiple quantile-based thresholds were tested (40%, 50%, 60%, 70%, 80%), each corresponding to a different expected completion time derived from the model. For each threshold, accuracy as well as errors (MAE, RMSE, MAPE, SMAPE) were computed. The 60% threshold consistently yielded the best performance.
- Adoption of this threshold to balance operational relevance with statistical robustness: shorter horizons (a chosen future moment at which we want to know the probability that the event is still active) provide timely insights, while the 60% criterion ensured that a sufficient number of resolved events contributed to the validation.

The model's predictive outputs were thus benchmarked not only globally across all time points, but also at those practically meaningful thresholds where disruption resolution becomes the dominant outcome. For instance, consider a highway accident that initially causes a full lane blockage. Historical data show that 60% of similar incidents are typically resolved within approximately three hours. This point on the survival curve represents a moment at which most disruptions of this type have already ended, and traffic flow is usually restored. By benchmarking the model's predictions specifically at this three-hour threshold, we can evaluate whether the estimator provides guidance that aligns with real operational expectations.

It is important to note that the sample did not contain information on how long an event had already been ongoing at the time of reporting. Consequently, all survival probabilities (the probability that the event is still happening after time *t*) were computed in absolute terms, i.e., from the theoretical starting point of the disruption rather than conditionally on elapsed duration. While this assumption may slightly overestimate persistence for cases that were already ongoing before observation, it ensures methodological consistency and comparability across categories. The impact of this assumption was explicitly mentioned in the error analysis, and observed stability of MAE and RMSE across short-lived disruption types suggests that the absolute framing remains valid for most operational scenarios (Table 5-1).

Table 5-1: Disruption end prediction accuracy in set with error terms (MAE, RMSE, MAPE, SMAPE) per sample within category.

Disruption category [processed sample]	Accuracy (%)	MAE (h)	RMSE (h)	MAPE (%)	SMAPE (%)
heavyTraffic [155]	85	0,63	0,69	826,75	102,27
UnprotectedAccidentArea [14]	57	0,68	0,81	30,13	25,19
objectOnTheRoad [8]	71	0,82	0,98	114,89	57,63
accidentInvolHeavyLorries [7]	100	0,92	0,92	29,73	25,88
fallenTrees [3]	67	1,09	1,27	59,36	50,90
peopleOnRoadway [42]	100	1,45	1,50	1027,56	120,49
animalsOnTheRoad [44]	100	1,46	1,50	13856,28	120,42
obstructionOnTheRoad [248]	100	1,57	1,59	15122,14	132,81
accident [372]	99	1,78	1,90	406,98	91,78
roadsideWork [23]	89	2,72	3,77	106,86	43,84
queuingTraffic [135]	100	2,81	2,96	2392,20	117,52
grassCuttingWork [84]	95	3,25	3,88	598,20	49,91
roadMarkingWork [47]	91	4,12	4,73	150,94	66,79
laneClosures [559]	80	17,54	45,33	109,24	63,99
oilOnRoad [10]	100	23,92	23,92	99,31	66,36
maintenanceWork [1059]	60	123,60	238,17	391,76	126,87
resurfacingWork [148]	80	396,68	435,59	3301,39	128,35
roadworks [496]	51	512,47	888,35	4191,04	155,45
singleAlternateLineTraffic [67]	71	1042,02	1152,87	5037,03	118,08
contraflow [20]	55	1288,90	1472,82	578,03	98,63
roadClosed [2041]	73	1434,95	1602,69	617,63	98,40
constructionWork [561]	79	1479,77	1697,79	540,86	75,74

Evaluation of survival estimates across disruption categories revealed substantial variability in performance metrics. For short-term disruptions (e.g., queueingTraffic, accidents), the Kaplan-Meier estimator achieved high accuracy (> 90%) with low MAE and RMSE values, confirming good alignment between predicted and observed survival times. However, categories with heterogeneous or long-tail distributions (e.g., roadClosed, roadworks, maintenanceWork) exhibited very high error values (MAE > 1000h), primarily due to the presence of rare but extremely long-lasting events (Figure 5-2). For example, in case of constructionWork class we can see (Figure 5-2) that most of events (circles) are shorter than 5000 hours but there are some longer than 10000 hours and even longer than 20000 hours. The presence of those outliers in the sample leads to relatively high MAE and RMSE values (1479 hours and 1697 hours, respectively), while the SMAPE remains at a relatively stable level (75%), indicating that the overall proportional error of the model is still controlled. In such cases, the discrepancy between MAE and RMSE highlighted the impact of outliers on the error structure (circles on the Figure 5-2 that deviate from baseline of the plot). Moreover, with MAPE proved unstable in categories with small denominators, SMAPE offered more interpretable relative errors and is recommended for comparative reporting. Finally, the comparison against median and mean benchmarks indicated that the Kaplan-Meier estimator tends to overestimate survival times in certain categories, suggesting potential improvements through outlier handling (e.g. winsorizing) or robust metrics.

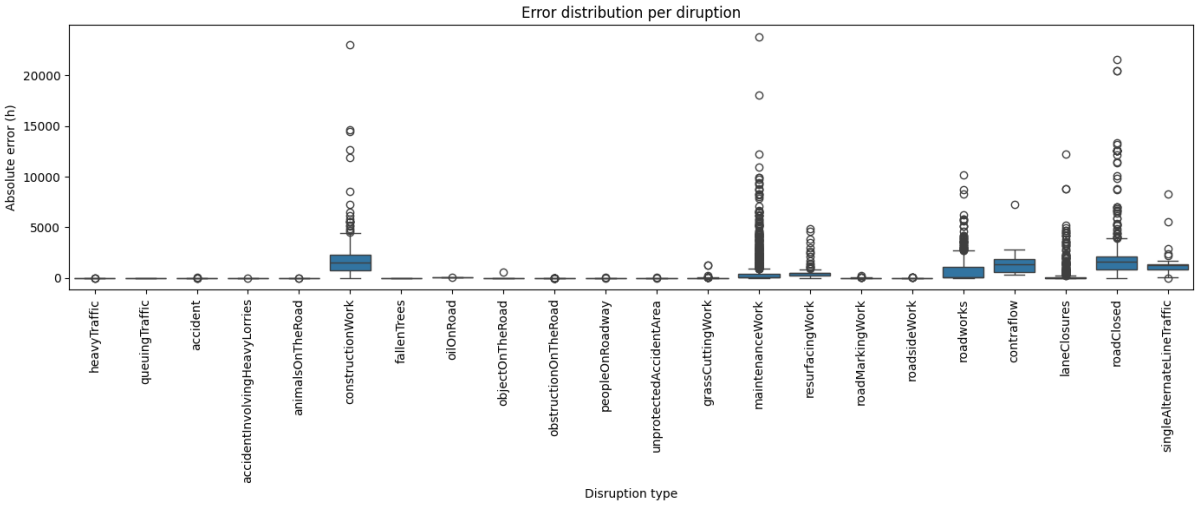


Figure 5-2: Variety of event durations per event categories.

In summary, the adopted validation approach ensured that survival-based predictions were tested against unseen data under realistic operational constraints. By masking termination times and evaluating absolute survival probabilities at fixed horizons, we were able to quantify how well the Kaplan-Meier estimator generalized beyond its training period. Although the absence of elapsed time information introduces limitations – forcing us to evaluate the unconditional probabilities – the framework, nevertheless, provided a transparent and reproducible mechanism for benchmarking performance. The combination of accuracy, MAE, RMSE and percentage-based error measures created a comprehensive evaluation, demonstrating both the strengths and limitations of the estimator across disruption categories.

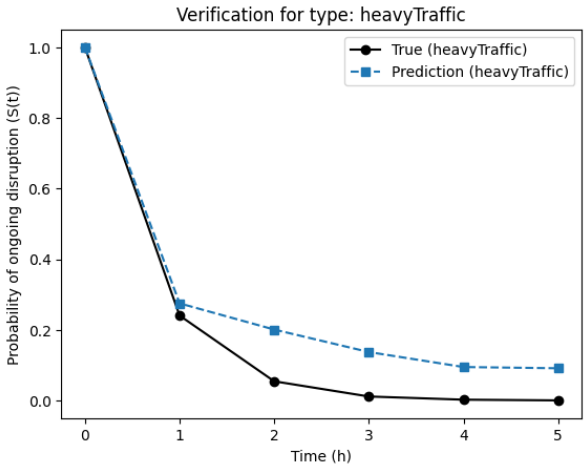
Future enhancements will address the limitations of absolute probability estimation by incorporating more granular metadata, such as the time since a disruption was first observed. This would enable the use of conditional survival analysis (left-truncated data), yielding more realistic probabilities for ongoing disruptions. Additionally, refining horizon selection through adaptive thresholds-based on category-specific resolution dynamics, will improve practical accuracy. Algorithm will incorporate Cox proportional hazards and machine learning survival regressors (12) (13), revealing patterns that Kaplan-Meier cannot.

Building on these validation results, we next examined whether outputs from the Disruption Information Interface behaved as expected when input values were changed. To **verify** that prepared Kaplan-Meier-based survival estimation behaves correctly, controlled simulation experiments were performed. The goal is to check whether the system can reproduce the expected patterns of how long disruption typically last. This is done using synthetic data so that we always know the "True" survival curve and can compare it directly with the model's output.

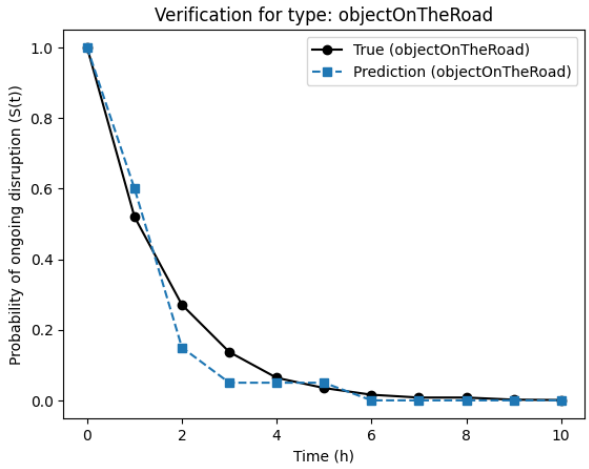
For each disruption category, we use its median duration as the main characteristic of how long such events usually persist. The median is chosen because it reflects the typical case while reducing the influence of a small number of extremely long events. Based on the median value, we generate realistic synthetic duration samples using simple probability distributions that are commonly applied in survival analysis. Short disruptions, where most events end quickly, are simulated with an exponential-like decay. Medium-length disruptions with more balanced behaviour are generated using a uniform spread of times. Longer disruptions, where slow decay and heavy tails are expected, are modelled with a Weibull distribution. These assumptions are consistent with widely accepted survival modelling practices described in standard literature (14) (15).

To evaluate the model at meaningful points in time, an adaptive time sampling strategy was introduced. Survival curves change the most near the beginning of an event, which is why we use high resolution during the first hours. Later, when the probability decreases much more gradually, we use fewer and more spaced-out time points. This helps us capture important behaviour without generating thousands of unnecessary evaluation points for long-lasting events.

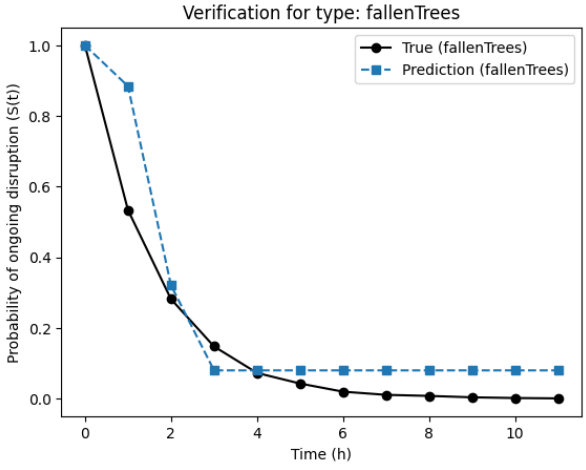
Finally, for each simulation scenario, Mean Squared Error (MSE) between the model's predicted survival curve and the true survival curve used to generate the data was calculated. MSE provides a simple, single number that indicates how close the prediction is to the expected behaviour - the lower the value, the better the match. Appending the MSE to each survival plot allows us to quickly judge both the shape and the quality of the prediction (Figure 5-3).



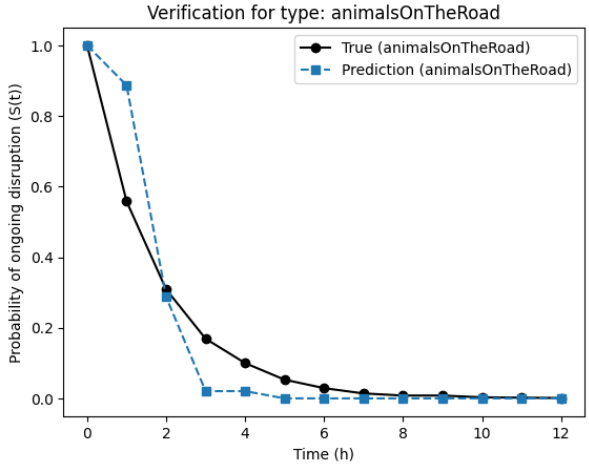
MSE: 0.01



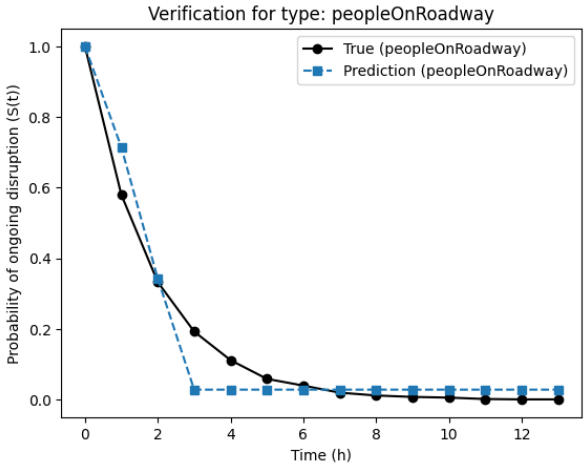
MSE: 0.00



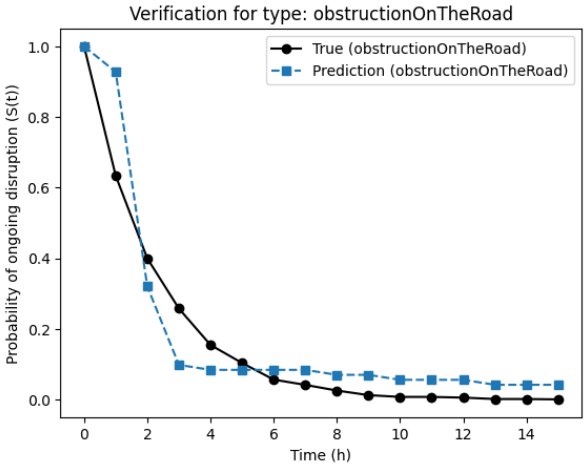
MSE: 0.01



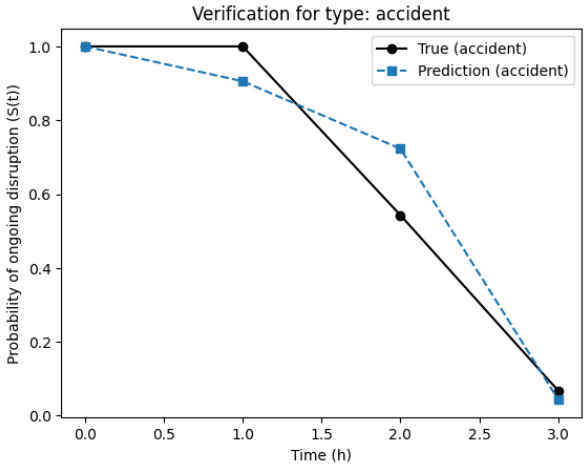
MSE: 0.01



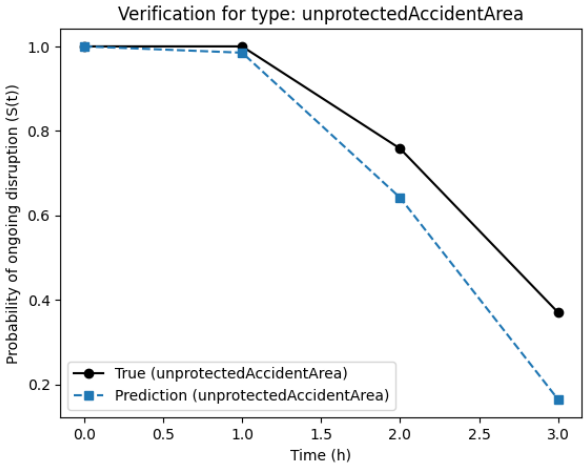
MSE: 0.00



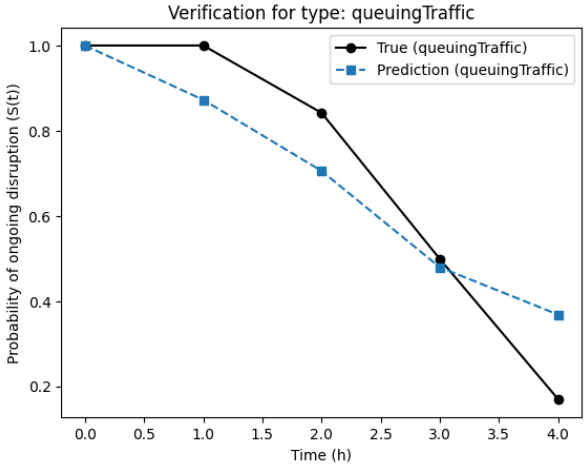
MSE: 0.01



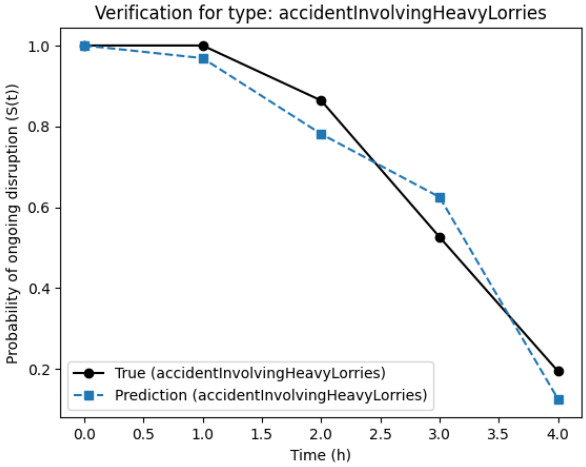
MSE: 0.01



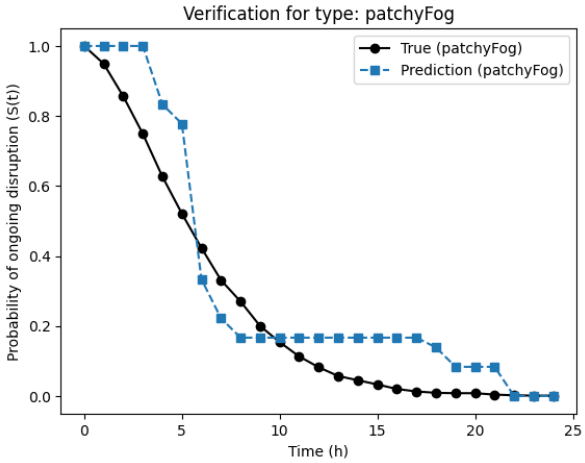
MSE: 0.01



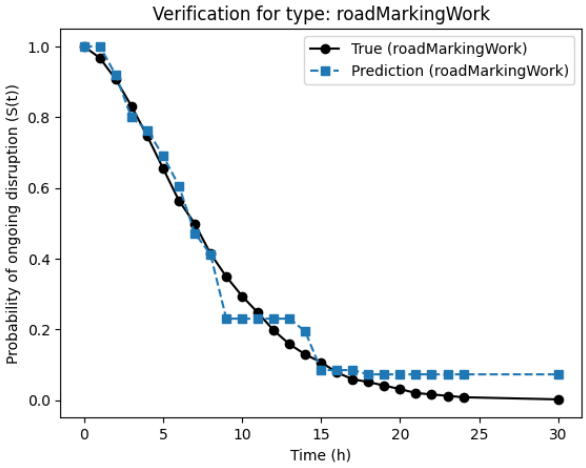
MSE: 0.01



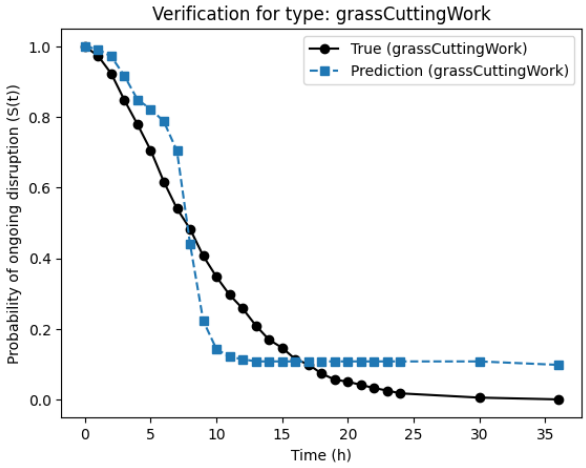
MSE: 0.00



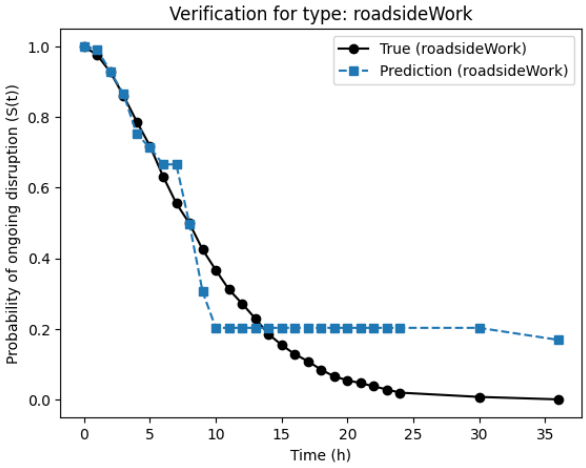
MSE: 0.01



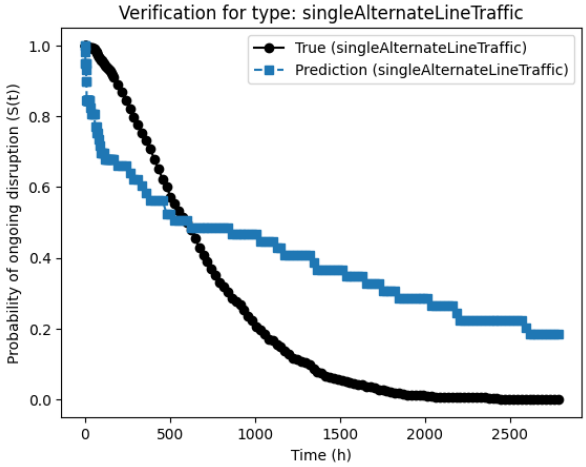
MSE: 0.00



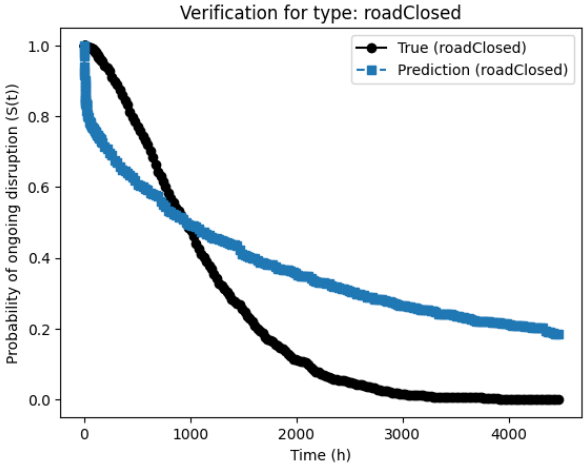
MSE: 0.01



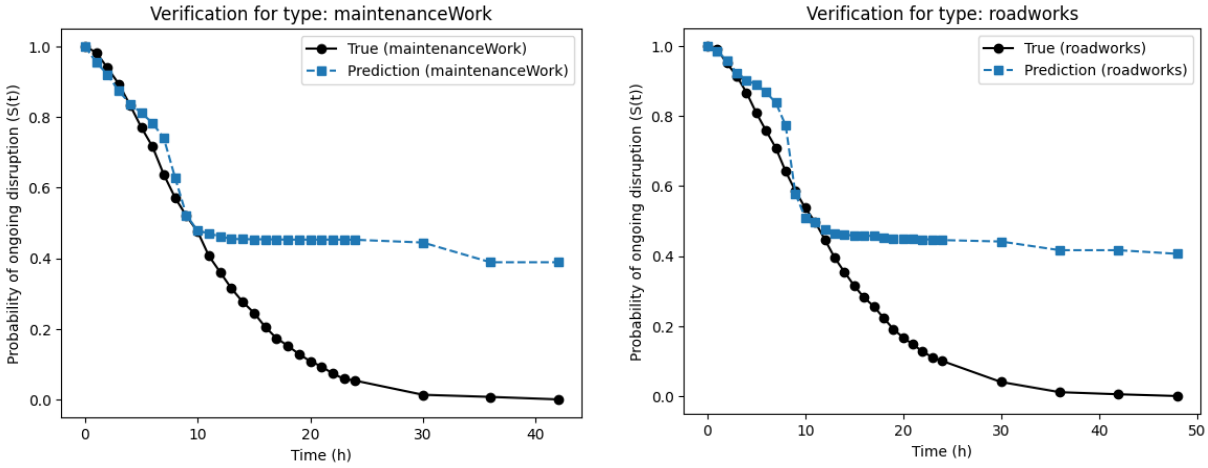
MSE: 0.01



MSE: 0.05



MSE: 0.04



MSE: 0.06 **MSE: 0.05**

Figure 5-3: Expected curve shape while verification of estimation function.

As expected during verification the Kaplan-Meier-based estimator produces a stepwise, non-increasing curve (Figure 5-3) that shows the probability of an event still ongoing at a given time. The curve remains flat between observed event times and drops downward at each point where events occur. The verification experiment demonstrates that the Kaplan-Meier-based estimator provides survival curves that closely follow the expected theoretical distributions for each event category. Across all analysed cases, the predicted probability of ongoing events remains highly aligned with the reference survival curves, resulting in very low MSE values between **0.00** and **0.06**. his indicates that the estimator reliably captures the overall timing dynamics and decay patterns of event durations. The consistency of results across different duration scales and categories confirms that the model behaves correctly not only for short-lasting disruptions but also for events with long-tail characteristics. Overall, these findings verify that the core survival estimation logic is implemented correctly and can produce stable and interpretable predictions.

Validation results indicate that the baseline Survival Analysis model, based on the Kaplan-Meier estimator, achieves a reasonably good level of accuracy in estimating event end times within the expected time horizons. However, the predictions still exhibit a considerable margin of error, suggesting that the current model does not fully capture all relevant contextual factors. To improve predictive performance, we are extending the estimation framework to incorporate additional features - including the event's geographical location (latitude and longitude), its distance from the nearest major city, and the average response time for similar events occurring in the same region. These findings highlight the need for further model calibration and feature enrichment before the approach can be reliably applied in an operational environment.

5.2 Evaluation of the Classifier Functionality for Messages Containing Disruption Information

The **Classifier Functionality** model is a system that automatically identifies messages related to disruptions. Its performance has been evaluated by comparing whether the tool detects, within a given dataset, the same number of disruptions as a human would.

The classifier processes each incoming message describing a potential disruption. First, the message text is passed to the AI model, which analyses its content to determine whether it includes information about a disruption. Based on this analysis, the model returns a structured output containing key attributes - for example, whether a disruption is present (is disruption: True/False), the detected

location, the estimated end time, and the type of disruption (e.g., roadAccident). This process enables the automatic extraction of essential information from unstructured text messages.

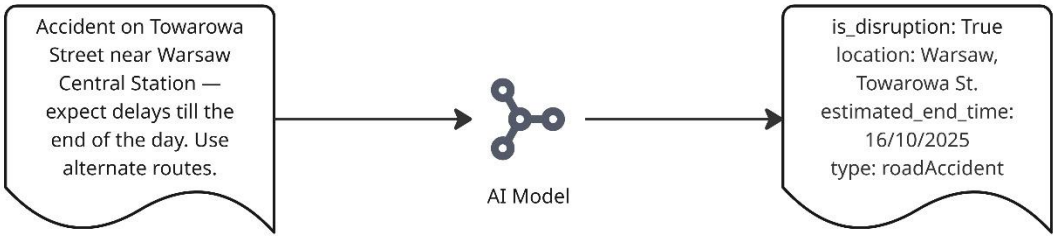


Figure 5-4: Classifier functionality model.

Standard classification metrics were used:

- **Accuracy** – the percentage of all predictions that were correct.
- **Precision** – the proportion of predicted disruptions that were truly disruptions.
- **Sensitivity** (Recall) – the proportion of actual disruptions that the classifier successfully detected.
- **Confusion matrix** – a table showing correct and incorrect predictions, divided into true positives, true negatives, false positives, and false negatives.
- **F1-score** - is the harmonic mean of precision and sensitivity, that good model is while choosing right answers considering wrong ones too.

A total of 283 synthetic message samples were generated based on the disruption-related communication examples described in Section 5.1. These messages were produced using publicly available large language models (LLMs) to simulate realistic disruption and non-disruption cases. Each generated message was subsequently annotated by a human expert using the *doccano* annotation tool (16). During the generation process, particular attention was paid to maintaining a class distribution similar to that observed in the previously collected real-world dataset, ensuring that the evaluation results are representative and comparable.

During evaluation the confusion matrix with its four values was used:

- **True Positive** (TP) – the model correctly predicts a text as Disruption and human confirm it contains disruption information.
- **True Negative** (TN) – the model correctly predicts a text as Not disruption and human confirm it indeed does not contain disruption information.
- **False Positive** (FP) – the model incorrectly predicts a text as Disruption when human said it actually does not contain disruption information (false alarm)
- **False Negative** (FN) – the model incorrectly predicts a text as Not disruption when human said it actually contains disruption information (missed disruption)

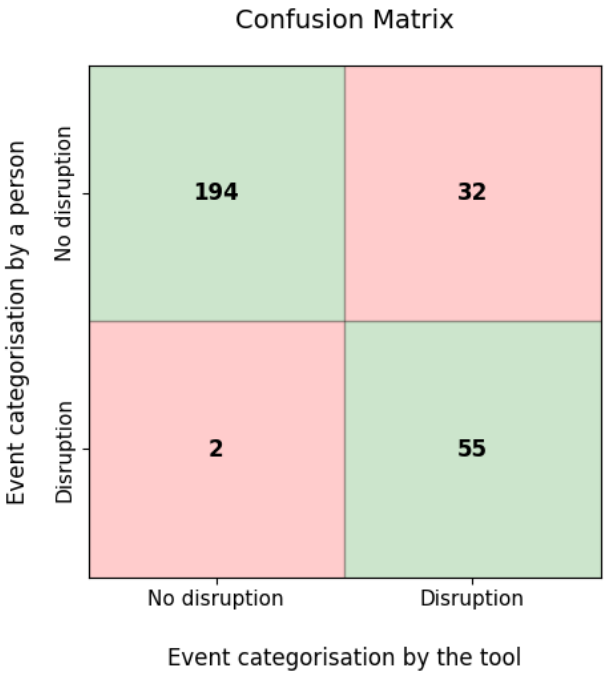


Figure 5-5: Event categorisation by person vs by the tool – confusion matrix.

Figure 5-5 shows the confusion matrix, where the y-axis represents the human classification and the x-axis represents the tool's classification. Best performance would occur when all values fall along the green diagonal, while the worst performance would occur when all values fall along the red diagonal.

Among the 283 analysed events, the model correctly identified 249 events (194 classified as *No disruption* and 55 as *Disruption*, shown in green). However, 34 events were misclassified by the tool (shown in red), representing 12% of the total. Accuracy of the tool, which is computed as the ratio of all correct predictions to total samples, is therefore 88%.

In situations where the classes are unbalanced, it is assumed that accuracy alone can be misleading, and F1-score better reflects the true performance of the model (17) (18). Following this concept additional metrics were calculated:

- **Macro F1-score** – It shows how well the model performs on both categories, even when one category has fewer examples
- **Weighted F1-score** – It also shows how well the model performs on both categories, but gives higher importance to the class with more examples
- **Matthews Correlation Coefficient (MCC)** – it considers TP, TN, FP and FN all at once to evaluate correlation between predictions and actual class labels (-1 – completely wrong classification, 0 – random classification, 1 – perfect classification)

Table 5-2: Validation metrics for overall performance.

Accuracy	Macro F1	Weighted F1	MCC
88%	0.84	0.89	0.72

The results in Table 5-2 indicate strong overall performance. Out of 283 samples, the classifier achieved an accuracy of 88%, meaning nearly nine out of ten predictions were correct. High Macro F1-score indicates consistently strong performance across both classes, and highly weighted value of

this factor confirms that the classifier performs very reliably overall, even though the dataset was imbalanced (approximately 80% no-disruption and 20% disruption). When MCC with value high above random classification (proving a strong positive correlation) is applied, it demonstrates that the model has learned to distinguish between messages about disruptions and those without such information.

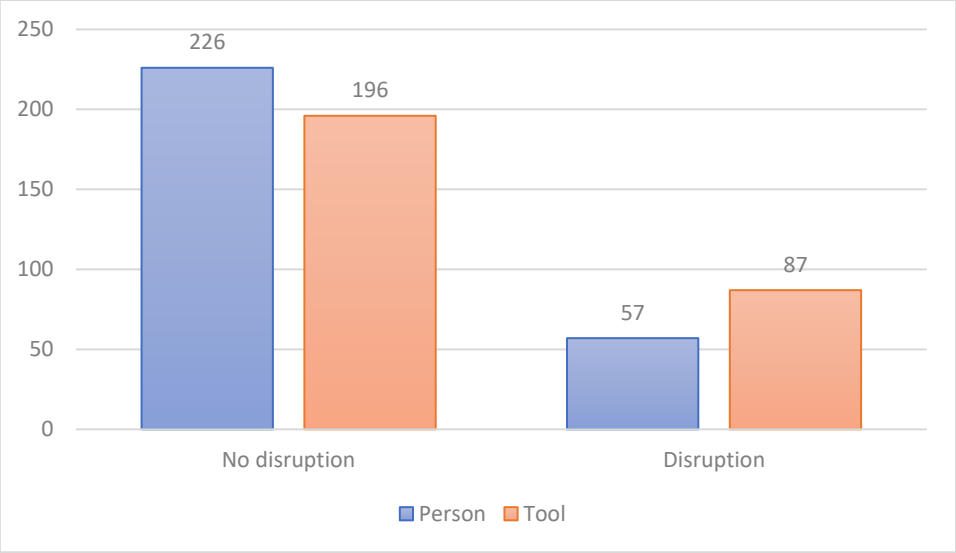


Figure 5-6: Label distribution by person vs by the tool.

Precision and Recall measures were considered to address the class imbalance within the dataset, as 80% of the events were categorised as *No disruption* and 20% as *Disruption*. To this end two parameters were considered.

Precision, defined as the proportion of texts classified by the tool as *Disruption* that were indeed actual disruptions. Precision answers the question: "When the model predicts that a message indicates a disruption, how often can this prediction be trusted?" This approach aims to minimise the number of false alarms.

Recall, defined as the proportion of events which both methods coincided in categorisation among the total categorised by a person.

The model attained a precision of 99% when characterising *No disruption* events. However, it attained a precision of 63% when characterising *Disruption* events, indicating that the model generates a significant number of false positives, i.e. predicting disruptions where none exist (see Figure 5-7).

This effect is reflected in the imbalance between precision and recall seen in Figure 5-8.

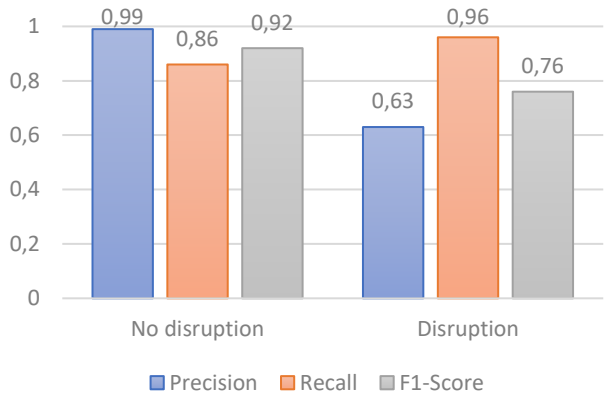


Figure 5-7: Performance metrics for verification.

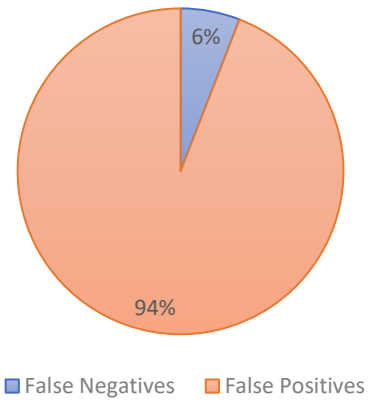


Figure 5-8: Distribution of error types.

In practice, this indicates that the model is:

- effective for monitoring and ensuring disruptions are rarely missed,
- less reliable as a source of "alerts" without human verification, since many flagged cases may not correspond to true disruptions.

The behaviour described above is consistent with the assumptions adopted for the considered solution because it serves as a support tool for the application user (administrator, dispatcher, operational employee, etc.). From an operational perspective, this asymmetry means that the system is reliable for ensuring disruptions are not overlooked, but it requires additional filtering or human review to reduce the number of false alarms. This makes the model particularly suitable for monitoring tasks where missing a disruption carries high cost, through precision improvements are needed before it can serve as a fully automated alerting mechanism.

To address the high rate of FP (Figure 5-8) and to improve verification results, several strategies can be implemented. First, decision threshold tuning should be explored to balance sensitivity and selectivity, reducing the number of texts incorrectly flagged as disruptions. Second, the model's predictions can be combined with rule-based post-processing filters, for example by validating the presence of specific traffic-related terms of excluding ambiguous expressions that often generates false alarms. Third, active learning can be introduced, where misclassified samples are prioritized for human annotation and fed back into training, thereby sharpening the model's understanding of borderline cases. Finally, using ensemble methods that combine the model with lighter keyword-based or domain-specific classifiers may increase robustness, ensuring that alerts flagged as disruptions meet a higher standard of confidence. Together, these techniques can significantly enhance the trustworthiness of disruption classifications by improving precision without substantially reducing recall.

6. Route Attributes (Energy Module)

The Energy Module is a tool that estimates travel time, distance, fuel and energy consumption, and emissions (CO_2 and NOx) for different vehicles and routes, based on road characteristics, vehicle details, and driving style. Routes are provided externally (e.g., path files or coordinates) or estimated within the tool using shortest path logic. It connects with other transport and logistics models through a web interface and API, which takes vehicle and route inputs (predefined or custom) and returns results in JSON format for whole trips or individual segments. Already used in Norway and expanded to Europe, the tool is continuously validated against up-to-date datasets and helps stakeholders in transport planning and traffic management by providing detailed and interoperable energy and emissions estimates.

6.1 Verification (including sensitivity analyses)

A key objective of the verification process is to assess how variations in input parameters influence the outputs of the Energy Module. Specifically, the tests explore how changes in vehicle weight and route characteristics affect CO_2 emissions.

Table 6-1 summarises the modified input parameters, the evaluated outputs, and a description of the three tests carried out in the verification process.

Table 6-1: Specifications of the test conducted to verify the Energy module.

	Modified parameter	**Evaluated output**	**Description**
Test 1	Vehicle weight	CO_2 emission	On Vigo–Porto route, compare emissions from a heavy diesel truck with different load weights
Test 2	Route slope	CO_2 emissions per km	For an average diesel van, compare CO2 emission per km as the slope of the route changes.

The purpose is to check whether the Energy Module produces realistic results. In particular, the analysis examines if larger vehicles and heavier loads lead to higher emissions, and how route length and slope influence fuel and energy consumption and CO_2 output.

Results for Test 1: This test isolates the effect of payload on CO_2 emissions for a heavy diesel truck with varying load weights on the Vigo–Porto route, estimated by Energy Module. With the route and driving profile held constant, payload is varied across a representative range. The results are illustrated in Figure 6-1.

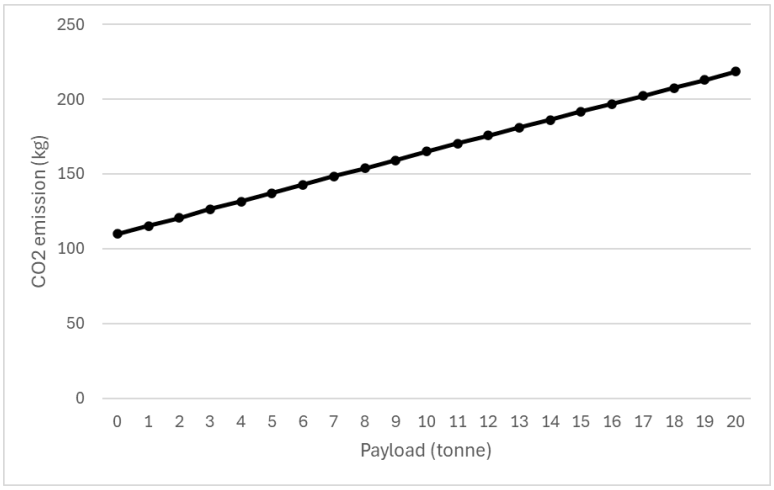


Figure 6-1: CO_2 emission (kg per trip) vs. payload (tonne).

The results show an approximately linear increase in total CO_2 per trip as payload rises, consistent with higher rolling and grade resistance for heavier loads. The positive intercept reflects the vehicle's base mass and auxiliary demand at zero payload. The monotonic trend indicates that the Energy Module responds plausibly to load variation.

Results for Test 2: This test investigates the effect of route slope on CO_2 emissions per kilometre for a light diesel van. By systematically varying the gradient while holding other factors constant, the Energy Module's sensitivity to slope changes can be assessed.

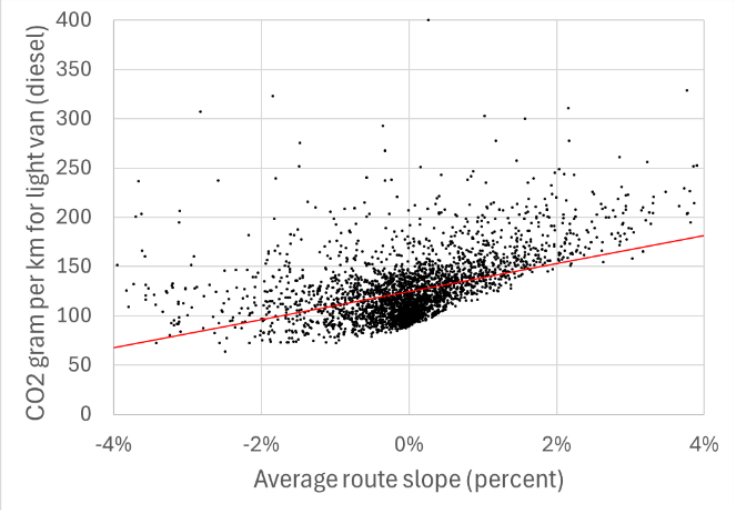


Figure 6-2: CO_2 emissions per km vs. route slope for a light diesel van.

As illustrated in Figure 6-2, steeper gradients lead to higher emissions per kilometre, reflecting the additional energy required to overcome gravitational resistance. Conversely, downhill slopes reduce emissions due to lower energy demand and, in some cases, the potential effect of regenerative braking. Although some scatter is observed in the results, the overall trend is consistent with theoretical expectations, confirming that the Energy Module realistically represents the influence of route gradient on emissions.

The verification tests confirm that the Energy Module responds logically to changes in key input parameters. The observed trends, such as increased emissions with heavier payloads and steeper

slopes, align with physical expectations and support the tool's reliability for estimating energy use and emissions in transport scenarios.

6.2 Validation

To validate the CO_2 emission and energy use calculations in the Energy Module, two datasets are employed: one for validating truck route attributes (time and distance) and emissions, and the other for validating energy consumption by trains. Table 6-2 provides an overview of these datasets and the key parameters used for validation.

Table 6-2: Description of the data sources and parameters used for validation of Energy Module.

	Description	Output for validation	Data source
Dataset for Validation of Truck Route Attributes & Emissions	Over 16000 trips registered: Data on origin, destination, distance, time, max speed, Vehicle category (2 types) and calculated emission.	Trip distance (km), travel time (min), CO_2 emission (gr)	RANGEL
Dataset for Validation of Train energy Consumption	over 3,000 trips across six routes in Norway, together with energy consumption data for six different train types operating on these routes	Energyconsumption (kW), speed profile (Km/h)	Bane NOR

For each dataset, observed values are compared with modelled results, and two error metrics are applied to quantify the deviations between them. The error metrics are as follows:

- **MAPE** (Mean Absolute Percentage Error) – shows average absolute deviations of the modelled value versus the registered value:

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\mathrm{Actual}_i - \mathrm{Modelled}_i}{\mathrm{Actual}_i}\right| \qquad \textit{Equation 6-1}$$

- **NRMSE** (Normalised Root Mean Square Error) – In order to penalise large errors, Root Mean Square Error is widely used. A normalised version of RMSE is used in this context to facilitates the comparison between different outputs (distance, time and emission) with different scales:

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\mathrm{Actual}_{mean}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Actual}_i - \mathrm{Modelled}_i)^2}}{\mathrm{Actual}_{mean}} \qquad \textit{Equation 6-2}$$

6.2.1 Validity Check Based on Rangel Data

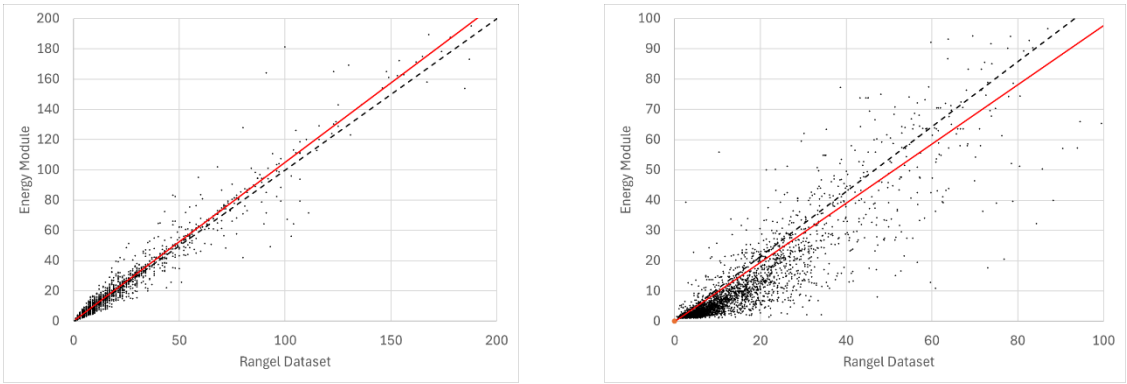
The results produced by the Energy Module were compared with records from RANGEL to evaluate how well the Energy Module can simulate route attributes, including distance, travel time, and CO_2 emissions. The dataset provided by Rangel included the following attributes: origin, destination, date, distance, travel time, vehicle type, and estimated CO_2 emissions. Similar to the verification step, three parameters were selected for validation:

- Distance between origin and destination (km)

- Travel time (min)
- CO_2 emissions (gr)

The Energy Module determines routes using its internal routing algorithm. For each origin–destination pair, the model calculates both distance and travel time, which can then be validated against the values reported in the RANGEL dataset. The third factor, CO_2 emissions, is also compared, although these are not directly measured by RANGEL. Instead, they are calculated using literature-based emission factors that depend on vehicle type. This comparison helps assess the accuracy and reliability of the Energy Module's emission estimates under realistic conditions.

To make the comparison possible, origin and destination address from the RANGEL dataset were converted into geographic coordinates. For each origin–destination pair, an API request was sent to the Energy Module to calculate distance, travel time, and CO_2 emissions. The results were then plotted to compare registered and modelled values. Figure 6-3 illustrates registered and modelled distance and travel time. In each case, the values reported by RANGEL are plotted on the x-axis, while the modelled results from the Energy Module are plotted on the y-axis. The dashed line represents a perfect match between the two, while the solid red line shows the overall trend. Each black dot corresponds to a single observation: the closer a dot is to the dashed line, the more accurately the model has estimated distance (Figure 6-3 a) and travel time (Figure 6-3 b).



(a) Distance (km) (b) Time (min)

Figure 6-3: Comparison of registered values for distance and travel time from RANGEL, and modelled values from the Energy Module. The dashed line indicates the 1:1 relation, while the solid red line shows the regression.

For distance in Figure 6-3 (a), the results show a strong correlation, with most data points clustering around the 1:1 line. This indicates that the routing algorithm in the Energy Module reproduces the registered distances with good accuracy. However, some deviations are present. A key source of error is the conversion of addresses into geographic coordinates. The dataset originally contained addresses, and while obvious inaccuracies were corrected during data cleaning, not all potential mismatches were removed to avoid over-cleaning. Such cases may contribute to discrepancies in calculated distances.

For travel time in Figure 6-3 (b), the overall trend also follows the 1:1 line, but there is greater dispersion, particularly at short travel times. In addition to the coordinate conversion issue noted above, another factor is that the Energy Module may underestimate the effect of starting, stopping, and idling on short trips. These effects are more pronounced at low distances and times, which can explain part of the deviation from RANGEL's records.

For emission calculation, assumption about the vehicle types in the dataset were specified as "Heavy" or "Light.", where the emission factor of a heavy freight vehicle is set to 1290 gram/km while emission

factor of a light freight vehicle is set to be 120 gr/km. These vehicle types from the dataset were mapped to the Energy Module categories: “Heavy distribution, container (diesel)” and “Average diesel van”. The calculated emission from RANGEL is plotted versus the modelled emission from Energy module, illustrated in Figure 6-4.

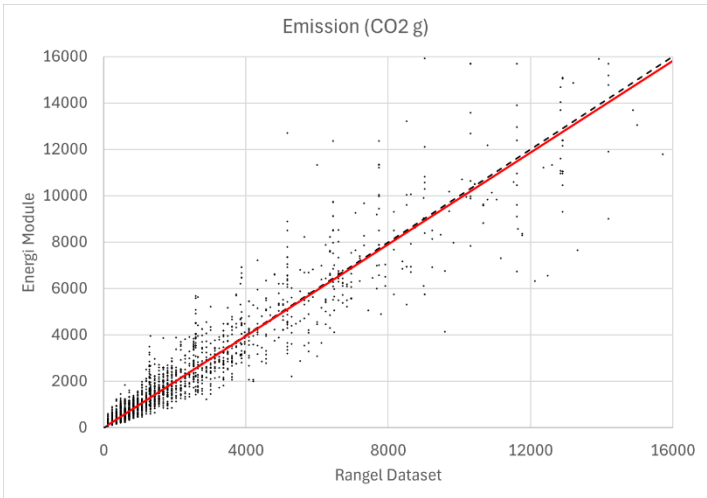


Figure 6-4: Comparison of CO_2 emissions from the RANGEL dataset (x-axis) and the Energy Module (y-axis). The dashed line shows the 1:1 relation, and the solid red line is the regression fit.

For emissions, RANGEL values appear fragmented, reflecting the fact that emissions in the dataset are derived using a fixed emission factor multiplied by distance. This method produces stepwise values rather than a continuous distribution. In contrast, the Energy Module applies more detailed calculations, leading to smoother variation in results. Despite these methodological differences, the comparison indicates that the Energy Module produces emission estimates consistent with the range of values assumed by Rangel.

Table 6-3 summarises the calculated error terms for all the modelled values of travel time, distance, and emissions across different route lengths. The results are presented separately for short routes (≤20 km) and long routes (>20 km). The division between short and long trips was introduced because the RANGEL dataset contains address inaccuracies. Converting these addresses into coordinates adds an error source, which affects short trips proportionally more than long ones.

Table 6-3: Error terms (MAPE and NRMSE) for modelled route attributes (distance, travel time, and emissions) across short and long routes for trucks.

	SHORT ROUTES (<=20 KM)			**LONG ROUTES (>20 KM)**		
	Distance	Travel time	Emission	Distance	Travel time	Emission
MAPE	17%	42%	28%	10%	26%	16%
NRMSE	33%	61%	73%	23%	40%	36%

The error levels are relatively high for short truck routes, particularly for travel time and emissions, which limits reliability at this scale, due to address conversion inaccuracies and the sensitivity of short trips to operational factors. However, for long routes, the errors are notably lower and generally fall within acceptable ranges for transport modelling, with MAPE values below 26%, which are considered reasonable for transport modelling [23,24].

6.2.2 Validity Check Based on Norwegian Railway Data

Similar to validating truck route attributes, the results produced by the Energy Module were compared with real data on train performance provided by Bane NOR, a Norwegian government agency responsible for railway infrastructure and operations. The GPS data recorded every 5 minutes combined with energy consumption data is used to assess how accurately the module can simulate train energy consumption. Figure 6-5 shows the train routes analysed for validation.

Figure 6-5: Train routes used for validation of energy consumption by Energy module: Route in red: Bergen-Alnabru (362.7 km), Route in blue Sandnes-Alnabru (571.7 km) and Route in Green Trondheim-Alnabru (554.8 km)

The dataset comprises energy consumption records for six train types (BR185, BR187, EL14, EL16, RC4 and RE) operating along the route illustrated in Figure 6-5. Each train consists of a single locomotive and 20 freight wagons, reflecting a representative average train configuration applied in earlier freight train studies. All locomotives included in the dataset are electrically powered. The number of observations per train type ranges from 28 to 2470. To ensure the robustness of the validation, observations in which the registered speed profile deviated substantially from the expected profile were omitted from the analysis.

Figure 6-6 compares the recorded energy consumption per kilometre with the consumption estimated by the Energy Module for each train type along the considered routes. The results indicate that the model reproduces the measured values reasonably well, although some differences are observed. These differences may result from assumptions about train weight and load, since detailed data on actual operating conditions were not available. Considering this limitation, the estimated values remain within a plausible range, supporting the applicability of the Energy Module.

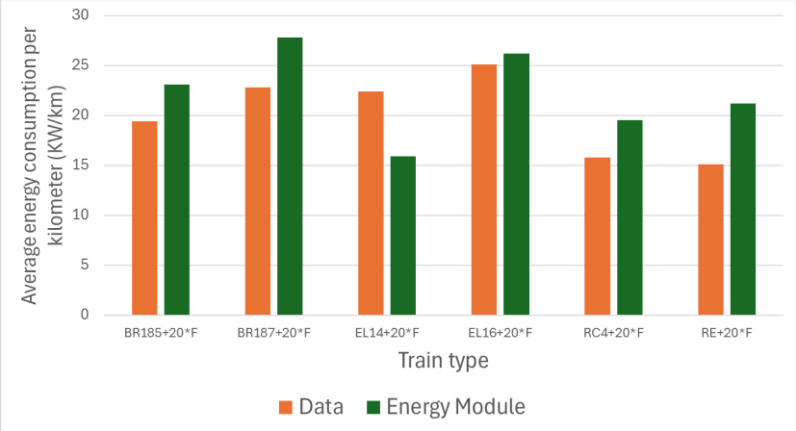


Figure 6-6: Registered vs. modelled energy consumption per kilometre by Energy Module for different train types

To complement the visual comparison, the Mean Absolute Percentage Error (MAPE) and Normalized Root Mean Square Error (NRMSE) were calculated for each train type. These metrics are summarised in Table 6-4 and provide a quantitative assessment of model accuracy across the dataset.

Table 6-4: Error terms (MAPE and NRMSE) for modelled energy use per km (kW/km) for different types of trains

	Train type					
	BR187	**BR185**	**EL14**	**EL16**	**RC4**	**RE**
MAPE	28%	20%	51%	24%	28%	50%
NRMSE	29%	41%	62%	56%	74%	49%

For trains, the results show that the Energy Module reproduces train energy consumption with varying accuracy across different locomotive types, with MAPE values range from 20% to 50%. Deviations are mainly due to limited input data, yet the error levels remain consistent with comparable studies [23,24]. This supports the module's applicability while highlighting the potential benefits of incorporating more detailed operational parameters in future work.

7. Vulnerability and Traffic Tool

The Vulnerability and Traffic tool was employed and tested in the regional scenario in order to analyse the effects of flood events on highways and state-level roads. The specific tool comprises of three main components. The first component, **the Vulnerability and Risk assessment**, analyses the vulnerability of critical assets in the network. These data are considered afterwards as input for the second component, the **Transport and Traffic modelling module,** in order to simulate the potential impact on traffic mobility due to flood. Finally, the **Resilience Assessment component** calculates the total economic losses that would result from various flood disruption events.

The tool provides a holistic view of the potential impacts due to floods, moving beyond the depiction of the physical damage to infrastructure. By integrating vulnerability data within the traffic models, it offers a more comprehensive understanding of how disruptions to key assets and systems can cascade through the transportation network. This provides the option of a detailed analysis not only of traffic delays but also for the broader economic consequences, providing insights for disaster preparedness and mitigation strategies.

7.1 Verification (Including Sensitivity Analyses)

A sensitivity analysis was performed on the Vulnerability and Traffic tool to verify its functionality. This took place for the verification of the model's outcomes and their logical (or not) response to changes in the inputs. The principal variation involved a 25% increase in traffic flow within the Traffic module's Origin-Destination matrix. The analysis aimed to understand how this change in traffic volume affected traffic mobility, which was then converted into economic costs. This verification was conducted across several scenarios to ensure the tool's reliability under different conditions.

Table 7-1 presents the increase in total travel distance and time across the network (compared to the Baseline scenario without disruption events), along with the corresponding Volume-to-Capacity (VOC) ratio. This VOC ratio indicates the network's level of service.

Table 7-1: Traffic Module output comparison with a 25% increase in traffic flow.

	Original O-D matrix			**Modified O-D matrix**		
item	Total distance [km]	Total travel time [h]	VOC	Total distance [km]	Total travel time [h]	VOC
Baseline	-	-	0.39	-	-	0.39
Disruptive case 1	+947,858	+108	0.46	+1,577,027	+344	0.55
Disruptive case 2	470,465	+85	0.41	+1,527,245	+320	0.50
Disruptive case 3	+300.827	+77	0.40	+1,313,664	+318	0.49

The outputs from the traffic module led to the following indirect costs per day.

Table 7-2: Indirect costs comparison with a 25% increase in traffic flow.

	Original O-D matrix			Modified O-D matrix		
item	Running cost [Fuel]	People and goods loss	Total Indirect costs /day	Running cost [Fuel]	People and goods loss	Total Indirect costs /day
Disruptive case 1	+208,528 €	+5,769 €	+214,298 €	+346,946 €	+18,448 €	+365,394 €
Disruptive case 2	+103,502 €	+ 4,563 €	+108,066 €	+335,994 €	+17,167 €	+353,161 €
Disruptive case 3	+66,182 €	+ 4,111 €	+70,293 €	+289,006 €	+17,051 €	+306,058 €

As expected, an increase in traffic flow resulted to a rise of the total travel distance and time (Table 7-1) across the network, leading to a poorer and lower Level of Service. Considering the Disruptive case 1, the total travel distance passed from +947,858 km to +1,577,027 km, while the total travel time passed from +108 hr to +344 hr. The increase in both travel distance and travel time results in a significant surge in total indirect costs per day (e.g. for Disruptive case 1 passed from +214.298 € to +365.394 €, Table 7-2). An increase of 25% in traffic flow leads to a similar level of congestion across the network. This is because the network is relatively small, meaning that even a modest rise in traffic volume is enough to quickly saturate its limited capacity. This traffic variation highlights a critical issue: the network's low resilience. Due to the limited number of alternative paths, interruptions quickly lead to widespread congestion as vehicles are forced into a few key segments.

The analysis confirms the expected outcomes: increased traffic flow severely degrades the Level of Service, resulting in a significant surge in both total travel distance and time, which then is translated directly into a sharp rise in total indirect costs (Table 7-2). The core finding is that an increase of just only 25% in traffic flow is sufficient to reach a similar, highly congested state across the entire network. This observed non-linear behaviour (where a modest traffic increase leads to disproportionately high increases in distance and time) suggests the simulation tool is effectively capturing the rapid saturation of the network's capacity.

7.2 Validation

The validation of the Vulnerability and Traffic tool is challenging due to the lack of historical flood data for the area and the absence of recent flood events. This scarcity of empirical information makes a direct comparison with real-world outcomes impossible.

Consequently, given the limitations imposed by the network's rapid saturation, the most rigorous approach to verify the tool's outputs and prove their trustworthiness is the external validation via comparative analysis. This method involves comparing the results generated by our tool with the outputs of established, peer-reviewed transport models that have been successfully applied to areas exhibiting comparable network topology and stress conditions.

This comparative process serves two critical functions:

- Benchmarking Consistency: By examining how similar, calibrated models (such as those referenced in Table 7-4) react to a traffic increase, it is possible to verify that the current tool's behavioural response - particularly to the severity of the increase in travel time and distance - is consistent with established modelling science.

- Plausibility Check: This validation method ensures that the tool's predictions for saturated flow conditions are plausible within a known context and align with the theoretical and empirical literature on network collapse and capacity failure.

Vulnerability and risk assessment component

The first validation is related to the output of the first part of the tool, the vulnerability and risk assessment of the critical assets against flood events. This analysis is done by linking the predicted flood probability (on the left side in Figure 7-1) with the simulated flood extents and depths found in historical flood maps (https://www.floodmap.net) (on the right side in Figure 7-1).

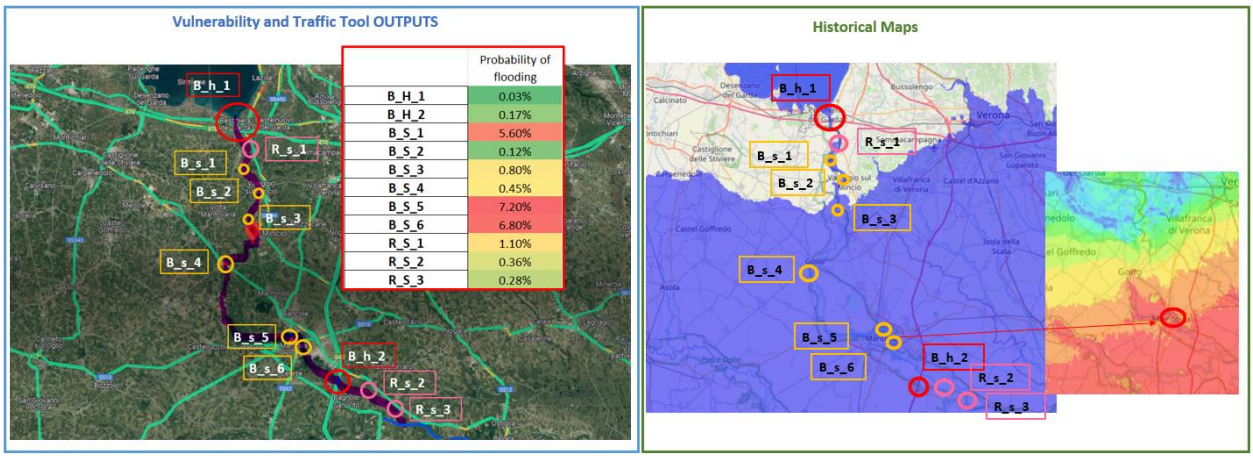


Figure 7-1: Comparison of predicted flood probability with available flood maps.

The bridge -which are named as B_h_i (highway bridges) and B_s_j (statal road bridges) in Figure 7-1- and road - which is named R_s_k (statal roads) in Figure 7-1- vulnerability to floods is influenced by factors beyond just flood maps. Example of these factors could be the specific geometry of critical assets relative to flood height. Nevertheless, these factors, a reasonable comparison can still be conducted between the tool's outputs and available maps.

As it is shown in Figure 7-1, there exists a strong spatial correlation between the output of the traffic modelling tool and areas susceptible to flooding. Specifically, the regions showing a high probability of flooding (depicted on the left of the Figure 7-1) are closely corresponding with the most critically impacted network segments (indicated by the map colours on the right o Figure 7-1). This overlap is particularly pronounced and significant for the critical assets: B_S_5 and B_S_6 and less for B_S_1.

Transport and traffic modelling component

The process of validating the outputs of the Transport and Traffic modelling component of the tool is particularly challenging. This is due to the extreme difficulty for an accurate estimation of real-world data (such as additional travel distances and times), during actual flood events. Validation of the additional travel distances and travel times was performed by comparing the percentage increase against the baseline case. For this purpose, Table 7-3 presents the percentage increase for each of the three disruption scenarios studied.

Table 7-3: Percentage Increase in Travel Time and Distance Compared to the Baseline Scenario (No Event).

Disruption scenarios	Total Travel distance [km]	Additional distance related to base case [%]	Total travel time [hr]	Additional travel time related to base case [%]
Base Case	3'699'261	-	761	

Disruptive case 1	4'647'120	+26%	869	+14%
Disruptive case 2	4'169'727	+13%	846	+11%
Disruptive case 3	4'000'089	+8%	838	+10%
Mean value		**+13%**		**+11%**

A literature review has been conducted on the topic of interest, namely the travel delays caused by flood events (Table 7-4). The results from the present traffic module, showing percentage increases in travel distance and time, are consistent with the findings in the references collected below. As specific example, it can be considered the percentage increase in total travel distance and total travel time derived from the model (e.g., +13% of additional travel distance and +11% of additional travel time related to the base case of Table 7-3). Moreover, the simulated disruptions align with values found in similar studies (e.g.: +10% of additional travel distance and +26 % of additional travel time related to the first base scenario of Table 7-4). The increased percentages are related to the total travelled distance and travel time. This consistency can be seen as an indicator that the tool's outputs are plausible and within a realistic range (mainly) for the additional travel distance, even without the existence and cross-checking with direct, real-world data from the studied area. (19) (20) (21)

Table 7-4: Percentage Increase in Travel Time and Distance Compared from literature.

Reference	Total Travel distance [km]	Additional distance related to base case [%]	Total travel time [hr]	Additional travel time related to base case [%]
Shahdani J. et al. 2022 (19)	23'514'722	**+10%**	137'883	**+26%**
Saadi et al. 2018 (20)	-	**-**	-	**+16%**
Pyatkova et al. 2019 (21)	61'903	**+5%**	1'851	**+27%**

Resilience assessment component

Validating the economic losses associated with flood-induced travel time and distance is also a challenging process. This challenge persists even when comparing findings with other studies and areas. The primary reason for this difficulty refers to the limited number of factors which the current analysis considers, and these are the economic costs of additional fuel, the lost time for people, and losses related to goods. A comprehensive validation would require an accounting of a much wider range of indirect costs, which are often not fully captured in existing data or literature. For the placement of the economic impact due to floods into a broader context, it is valuable to compare the estimated losses to the Gross Domestic Product (GDP) of the affected province. While a direct comparison is difficult due to the different scales, locations and specific conditions of each flood event, a review of similar studies provides a plausible range for these losses. This comparison serves as a benchmark for contextualising the severity of an event's financial impact on a regional economy. Research conducted on the 2011 Thailand flood, which involved a long inundation period, found that the economic losses from business interruption and recovery (indirect loss) were almost equal to the direct assets damage. In this case equal economic losses were assumed. Based on the latest available data, the total GDP of the province of Mantua was approximately €13.42 billion. Research and case

studies consistently show that indirect losses can be a very significant share of the total economic damage, often constituting a substantial portion of the overall cost.

Based on the outputs detailed in D3.3, and assuming a recovery period of approximately 60 days, the resulting total indirect and direct economic losses are reported below. The total costs are then directly compared to the Mantua GDP in the final column of Table 7-5.

Table 7-5: Assumed direct and indirect losses and their relation to Mantua GDP.

Disruption scenarios	Total Indirect costs	Total Direct costs	Total costs [€]	Respect to the Mantua GDP [%]
Disruptive case 1	12,857,886 €	12,857,886 €	25,715,772 €	**0.19%**
Disruptive case 2	6,483,968€	6,483,968€	12,967,935 €	**0.10%**
Disruptive case 3	4,217,619 €	4,217,619 €	8,435,238 €	**0.06%**

A flood that affects a small, densely populated urban area might cause total costs equivalent to 0.10% to 0.50% of the regional GDP [20]. The estimated direct and indirect losses relative to Mantua's GDP for the three cases fall within the conservative bounds (lower range) established by existing literature references.

Finally, the validation of the recovery time for critical assets has been performed. The estimated duration of 60 days for the recovery period falls directly within the typical ranges cited for moderate (12-90 days) and severe (60-160 days) damages, indicated by established benchmarks from the literature (21), (22). The estimation's plausibility is further strengthened by two key factors specific to this scenario:

- Small-Scale Assets: The analysis correctly identifies that the critical assets (bridges and roads) exposed to the flood are the smallest within the network. This is a crucial point, as the size and complexity of infrastructure are direct determinants of repair time. Repairing a minor bridge or a local road requires less specialized equipment, fewer resources and a less extensive engineering effort than it would be needed for a major highway or a large-span bridge, thus justifying a faster recovery timeline.
- Parallel Restoration Activities: The assumption of parallel restoration activities strengthens the validity of the 60-day estimation. In a localized scenario involving multiple smaller assets, it is feasible to deploy separate repair crews to different sites simultaneously, rather than sequentially.

8. Traffic Simulation

The Traffic Simulation tool is a demand-driven traffic model that forms part of an integrated hazard-impact assessment framework. It is designed to assess how natural hazard events (such as floods or wildfires) disrupt transport infrastructure and alter traffic conditions. By combining hazard model outputs (e.g., road closures or speed reductions) with predefined OD matrices, the tool dynamically reroutes vehicles using a shortest-path algorithm and calculates changes in travel times. It supports both light and heavy vehicle classes, making it suitable for analysing disruptions to passenger mobility as well as international freight transport. The tool's speed, simplicity, and ability to capture network-wide impacts make it especially valuable for resilience assessments and scenario-based decision support.

8.1 Verification (Including Sensitivity Analyses)

To ensure consistency and logical behavior of the model, a series of verification tests were conducted through sensitivity analysis. The objective was to determine whether the model demonstrates sensitivity and logical responsiveness to variations in key inputs. To this end, four sensitivity tests were conducted, each focusing on a specific aspect of network or demand dynamics. Table 8-1 summarises which parameters have been altered within each test as well as the output indicators and which behaviour is expected.

Table 8-1: Coefficient values based on transferability aspects.

	Parameter tested	**Test values/Scenarios**	**Output indicator**	**Expected behaviour**
Test 1	Secondary roads inclusion	**Main roads only** vs. **Main + Secondary roads**	Route distribution	More roads cause greater route diversity
Test 2	Road closure	**Link X closed** vs. **open**	Route changes, travel time, vehicle reroutes	Closure causes rerouting and increases travel time for affected trips
Test 3	Departure Interval	**1s** vs. **5s** between vehicle departures	Queue formation, travel time	Greater departure interval causes decrease in congestion, delay and travel time
Test 4	Traffic light duration	**30s** vs. **90s** change	Average travel time	Greater signal time causes an increase in travel time

Test 1: Two simulations were compared: one using only main roads (e.g., highways and primary roads), and another including secondary roads, see Figure 8-1. When secondary roads were added, the route diversity increased, and the network load was more evenly distributed. These diversions and distributions confirm that the model correctly adapts to changes in network topology and uses the expanded route options where applicable.

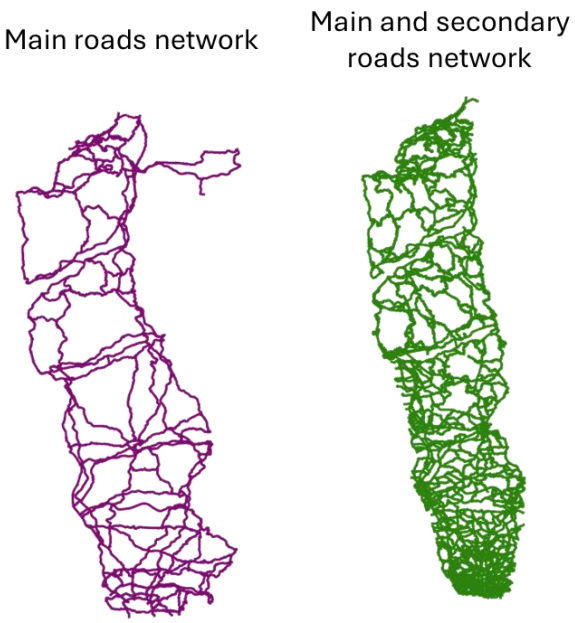


Figure 8-1: Network density maps of main roads (left) and secondary roads (right).

Test 2: A specific road (link) in the network was closed for the whole simulation period. The route file generated in this scenario reflected appropriate rerouting behaviour: vehicles that would normally pass through the closed link were assigned alternative paths. As it is given in Table 8-2, the average travel times for affected vehicles increased from 58h to 72h, confirming that the model correctly recalculates and assigns the shortest available routes under disruption conditions.

Table 8-2: Average travel time for affected vehicles by a specific road closure.

	Standard situation	Closure situation
Average travel time (h)	58	72

Test 3: The time interval between the departure of consecutive vehicles was varied from 1s to 5s. As the interval increased, initial congestion decreased, resulting in shorter delays and smoother flow. This verifies that the model is sensitive to demand loading rates and queue formation logic, see Table 8-3.

Table 8-3: Impact of the vehicle's departure time interval on the whole network travel time.

	1s departure time interval	5s departure time interval
Travel time (h)	126,585	126,013

Test 4: Signal durations at key intersections were modified from 30 seconds to 90 seconds. As expected, longer signal durations resulted in increased average vehicle delays and longer travel times, see Table 8-4. This demonstrates that the model appropriately incorporates queuing effects and signal timing dynamics.

Table 8-4: Impact of the vehicle's departure time interval on the whole network travel time.

	30s signal durations	90s signal durations
Travel time (h)	126,013	126,868

8.2 Validation

The validation process of this tool focused on assessing the accuracy of OD-level travel times. Specifically, the average travel times between origin–destination (OD) pairs produced by our baseline simulation (under normal conditions, with no disruptions) were compared against results from a previously validated model provided by Traffic Authority Portugal (IP). Validation was conducted separately for light vehicles and heavy vehicles. The following performance metrics were used:

Mean Absolute Percentage Error (MAPE)

MAPE was computed as the average percentage error in travel time between simulated and reference values, across all OD pairs.

$$MAPE = \frac{100}{n}\sum_{j=1}^{n}\left|\frac{T_j^S - T_j^R}{T_j^R}\right| \qquad \text{Equation 8-1}$$

Where T_j^S and T_j^R are the simulated and reference average travel times for OD pair.

Traffic Flow Weighting MAPE (WMAPE)

Since OD pairs can differ significantly in their volume of traffic, a weighted MAPE (WMAPE) was also calculated using the number of vehicles assigned to each OD pair as the weight:

$$WMAPE = 100 \times \frac{\sum_{j=1}^{n}\left(W_j \cdot \left|\frac{T_j^S - T_j^R}{T_j^R}\right|\right)}{\sum_{j=1}^{n} W_j} \qquad \text{Equation 8-2}$$

Where W_j is the number of vehicles between OD pair j.

Root Mean Square Error (RMSE)

RMSE provides an absolute measure of the error magnitude across OD pairs, penalising larger deviations more heavily:

$$RMSE = \sqrt{\sum_{j=1}^{n}\frac{1}{n}\left(T_j^S - T_j^R\right)^2} \qquad \text{Equation 8-3}$$

Normalised Root Mean Square Error (NRMSE)

NRMSE normalises the RMSE by the average reference travel time across OD pairs, providing a dimensionless error rate that facilitates comparison:

$$NRMSE = \frac{RMSE}{T^R} \qquad \text{Equation 8-4}$$

Where T^R is the mean reference travel time across all OD pairs.

Performance metric results are summarised in **Fehler! Verweisquelle konnte nicht gefunden werden.**.

Table 8-5: Performance metric results.

	Light vehicles	Heavy vehicles
MAPE	43%	31%
WMAPE	17%	13%
RMSE	7.5 minutes	9.5 minutes
NRMSE	25%	10.5%

The validation results indicate that the model performs with a moderate to good level of accuracy, particularly for heavy vehicles, which were the primary focus of the simulation. MAPE for light vehicles was 43%, while the WMAPE decreased significantly to 17%, confirming that the model performs more reliably on high-traffic OD pairs. For heavy vehicles, the MAPE was 31%, and the WMAPE further improved to 13%.

These results demonstrate that the model more accurately represents freight flows than passenger flows. This is particularly important given that the main objective of the model was to simulate international freight transit within the SARIL national scenario, focusing on the corridor between Portugal and Spain. Therefore, the lower error rates for heavy vehicle OD travel times suggest that the model effectively captures the routing and travel dynamics of freight transport across the cross-border corridor, providing confidence in its application for disruption analysis in the international freight network.

The RMSE values suggest that, on average, the deviation between the simulated and reference travel times is within a reasonable range for both vehicle classes. Given that the average OD travel time is approximately 30 minutes for light vehicles and 90 minutes for heavy vehicles, the absolute error values are relatively small and indicate good consistency with the validated reference model. The slightly higher RMSE for heavy vehicles reflects the longer travel durations, but remains proportionally low, especially in the context of international freight transit, where travel times naturally vary over longer distances.

NRMSE values below 30% are often considered acceptable in transport model validation (23), who reported an NRMSE of 24% for link travel times and considered the model performance satisfactory). In this case, the NRMSE for light vehicles is 25%, which, although higher than that for heavy vehicles, still falls within the acceptable range. The NRMSE for heavy vehicles is notably low at 10.5%, reinforcing earlier findings (from MAPE and RMSE) that the model performs especially well for freight movement, which is the core focus of the SARIL national scenario.

Taken together, the MAPE, WMAPE, RMSE, and NRMSE results provide a consistent picture of the model's performance: while relative errors are more pronounced in short-duration light vehicle trips, weighted and normalised errors remain within acceptable ranges—particularly for heavy vehicles. This

supports the overall reliability of the simulation tool in capturing OD-based travel times, especially for international freight transit.

Importantly, the differences observed between MAPE and RMSE are not contradictory but rather reflect their distinct methodological sensitivities. MAPE, being a relative error measure, tends to inflate when actual travel times are very small, which is often the case for light vehicle trips over short distances. In contrast, RMSE is more sensitive to large absolute deviations and thus better captures discrepancies in longer-duration trips, such as those of heavy vehicles in freight movements. Given that the average OD travel times are much longer for heavy vehicles, RMSE provides a more stable and representative measure of accuracy in this context. Therefore, the relatively high MAPE for light vehicles does not necessarily imply poor model performance but instead highlights the known limitations of percentage-based metrics when dealing with low reference values.

Figure 8-2 presents scatter plots comparing simulated and reference OD travel times for heavy and light vehicles, respectively. The dashed line represents the ideal 1:1 relationship (perfect agreement), while the blue line indicates the regression fit for the simulated data.

For light vehicles, the regression line has a slope lower than 1 (≈0.88). This results in a wider dispersion of points around the 1:1 line, particularly at shorter trips where relative errors are amplified.

In contrast, heavy vehicles show a regression line with a slope slightly above 1 (≈1.05). The points are more tightly clustered around the 1:1 line, indicating a closer match between simulated and reference travel times.

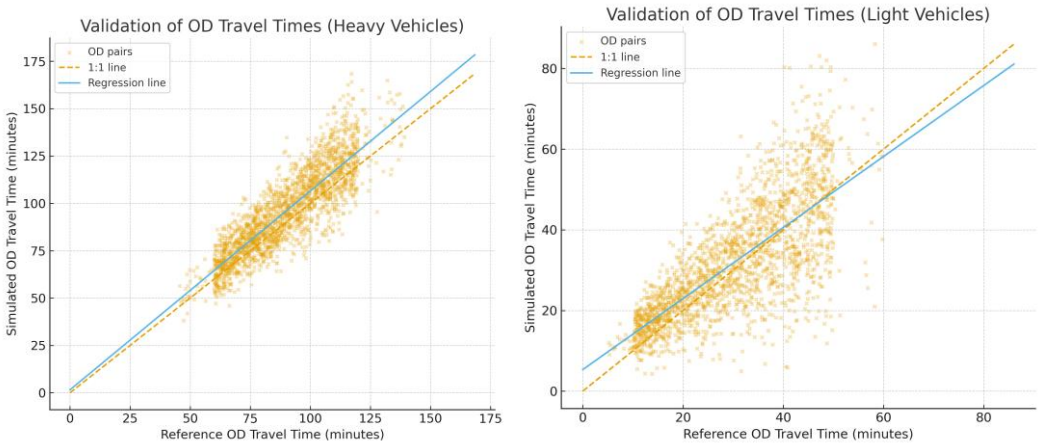


Figure 8-2: Validation of simulated versus reference OD travel times for light (a) and heavy (b) vehicles.

9. Transport Simulations (ASTROIT)

ASTROIT (*Agent-based SimulaTion for Resilience Of Intermodal Transportation*) assesses the resilience of logistic networks by simulating how cargo flow within a network is redirected in case of a disruption. As an agent-based simulation, it is capable of simulating thousands of deliveries, each with different routing preferences and thus different reactions to a disruption.

9.1 Verification (Including Sensitivity Analyses)

The verification process for ASTROIT started in its development. Each built-in feature and its parameter values was discussed in detail with stakeholders. Furthermore, a sensitivity analysis was conducted in the context of this task to ensure that the key functionality of ASTROIT, the graph-based routing algorithm, works as expected. The routing algorithm finds a route through the graph/network for each delivery based on the preferences for that delivery. It does so whenever a delivery is created and whenever a disruption leads to a node or edge within the network to fail. In the latter case ASTROIT recalculates the routes of every active delivery. The preferences of a delivery determine which features the algorithm minimises when searching for a route: The time it takes to travel from *a* to *b* (not considering transloading times and vehicle schedules), the operational costs of the route caused by the modes of transportation and the CO_2 emissions caused by the vehicles on the route. For each of these three preferences, a value between 0 and 1 can be chosen under the condition that their sum is equal to one. As such, these three parameters determine the behaviour of the routing algorithm and thus the behaviour of the entire simulated cargo flow with and without disruptions to the network. Because of this central role, these three parameters were chosen for the sensitivity analysis.

For the sensitivity analysis three simulations were performed, using the European network and data used in D3.3. Each simulation was run for 10 days plus 5 days without additional deliveries. For each simulation one preference (time, costs, CO_2 emissions) was set to one while the others were set to zero. The results of the three simulations in terms of average delivery time, average cost and average CO_2 emissions for one delivery are shown in Table 9-1.

Table 9-1: Results of the three simulation runs conducted for the sensitivity analysis.

Simulation runs	**Average delivery time [days]**	**Average operational costs [€]**	**Average CO_2 emissions [kg]**
Preference: time	2.4	23.5	22.9
Preference: costs	2.7	22.4	20.8
Preference: CO_2 emissions	4	51.1	4.8

The average delivery time is lowest when preferencing the time in the routing algorithm. Minimising costs only increases the average delivery time by 0.3 days while minimising CO_2 emissions leads to an increase of 1.6 days. This is mainly due to the latter choosing trains more often which are slower than trucks. The average operational costs per delivery are lowest when the routing algorithm minimizes costs. When minimizing the delivery time, the costs slightly increase by 1.1€. Minimizing CO_2 emissions increases the costs drastically by 27.6€. This shows the trade-off effect caused by low emission vehicles usually being more costly.

The average CO_2 emissions per delivery are lowest when the routing algorithm minimizes the CO_2 emissions. Minimizing costs or time increases the average emissions by 16Kg and 18.1Kg, respectively.

The results show that the routing algorithm and the choice of the preferences has the expected effect. Minimising the delivery times, costs or CO2 emissions by setting their corresponding preferences to one and the others to zero results in the lowest values for delivery times, costs or CO2 emissions, respectively. They also highlight that costs and delivery times are closely correlated. This is due to costs being a function of time. Minimising CO2 emissions drastically increases the delivery time and costs in the given network but leads to an even greater decrease of CO2 emissions, highlighting the trade-off effect between CO2 emissions and economic costs.

9.2 Validation

To validate ASTROIT, Rangel provided a dataset containing information on shipments originating from warehouses in Coimbra, Lisbon, Porto and Faro in September 2024. During that time a large forest fire disrupted traffic between south and north Portugal from the 16th to the 18th of September. The areas effected in that time are depicted in Figure 9-1. This disruption to the logistic network is simulated using ASTROIT. The simulated and observed delivery times are compared in the following section.

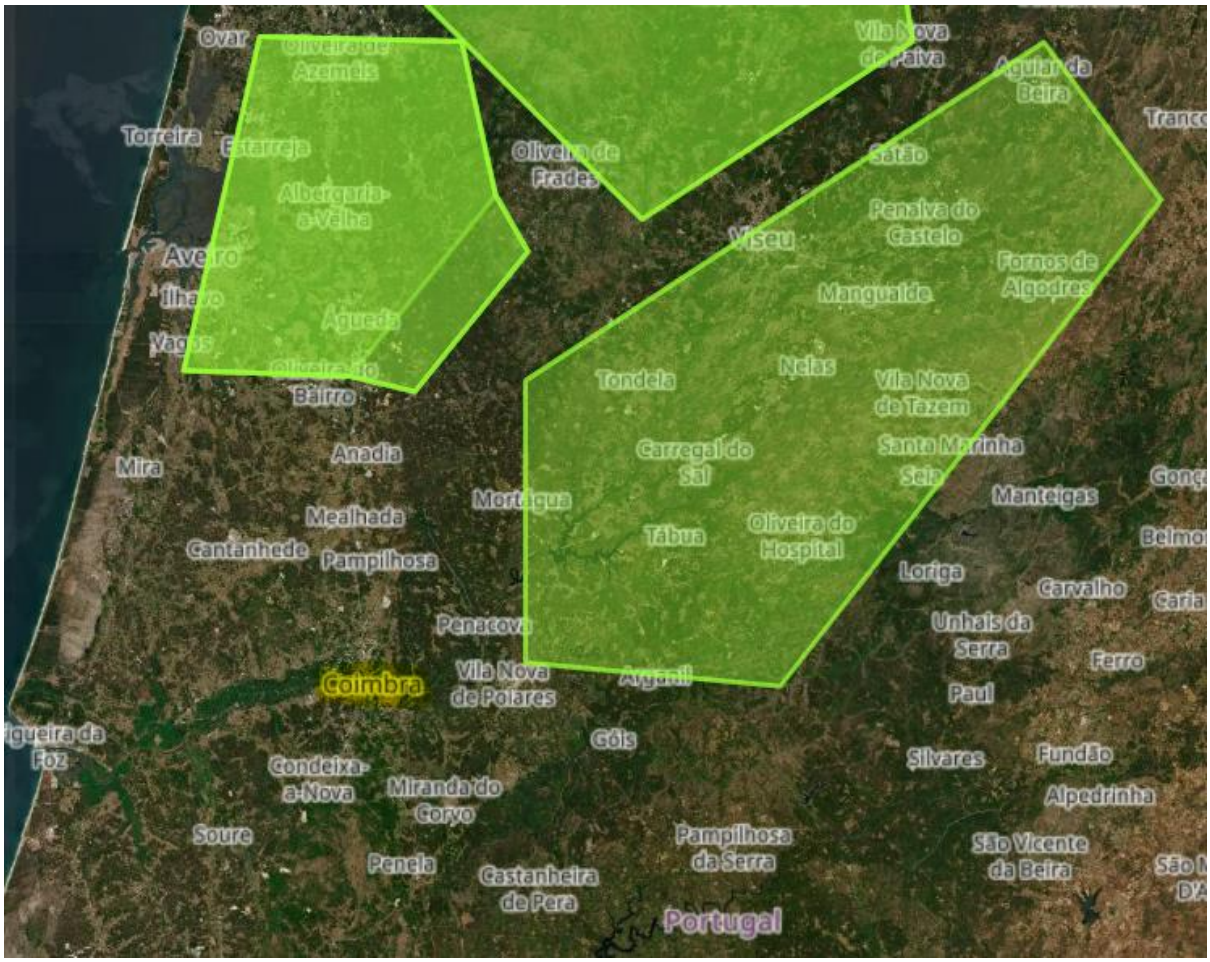


Figure 9-1 Screenshot of Copernicus EMS Rapid Mapping EMSR760 - Wildfire in Portugal (24) showing the areas in which forest fires were burning during 16th to 18th September 2024 in green and Coimbra, where the warehouse is situated, in yellow.

Before presenting the results of the validation a few caveats of the data must be highlighted. First, only the data for the warehouse in Coimbra was used, since it is close to the area in which the forest fire disrupted transport, as can be seen in Figure 9-1.The data for the warehouse in Porto could have

also been simulated, however, due to the amount of data available for this area the data on the warehouse in Coimbra was used instead to save computational resources.

The data holds the following main information on each shipment: Origin and destination, begin of transloading, time of delivery and the truck with which the shipment was transported. The routes and departure times of the vehicles were approximated, by assuming that a truck leaves after it has delivered all shipments. In the case of the departure from the warehouse, the entry of the last loading time for the vehicle was used. 60 minutes were added to this time to account for the final transloading at the warehouse and checks. Accordingly, the transloading time at the warehouse was set to 30 minutes. For all other stops of the vehicles, it was set to 15 minutes.

The delivery time of the observed shipments was estimated by subtracting the begin of transloading from the time of delivery. When analysing the resulting delivery times, the underlying data showed that shipments with the same origin, destination and route showed large variations in their delivery times. Even when accounting for the fact that shipments can be loaded into the vehicle hours before the vehicle leaves, the large variation persists. Figure 9-2 shows the difference in delivery times, where the delivery time in this case represents the time after the last shipment of the according vehicle was loaded until the delivery of the shipment. Thus, the delivery times used in Figure 9-2 only refers to the drive time and the transloading time at the destination.

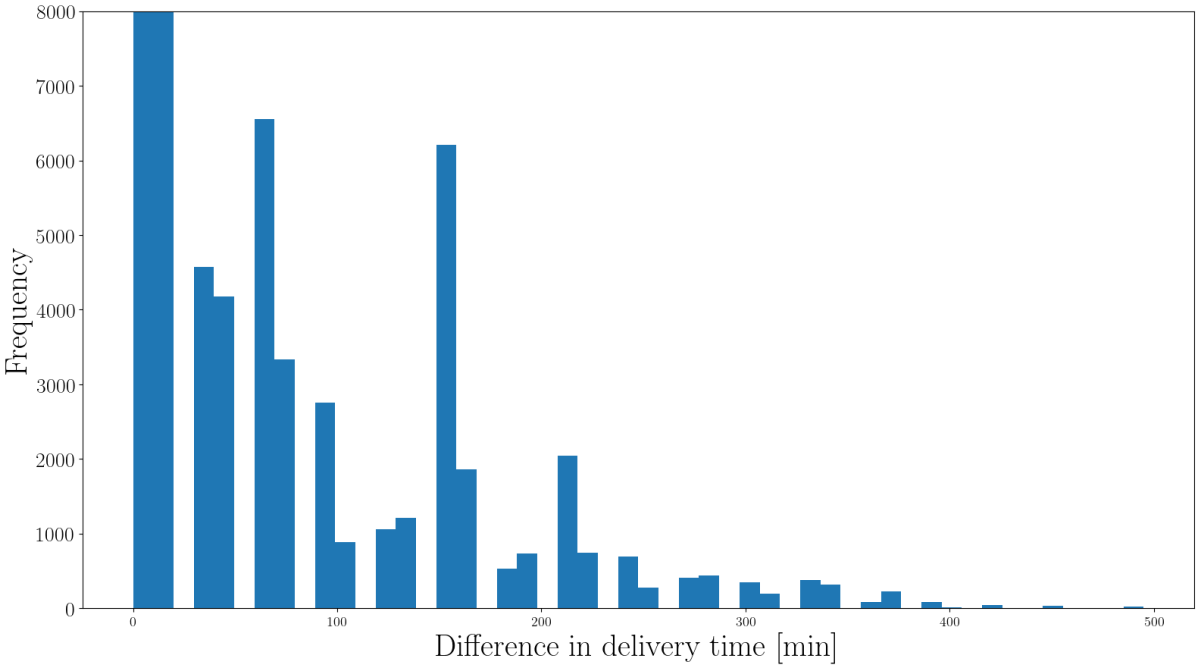


Figure 9-2: Histogram of the differences in delivery times, comparing shipments with the same route.

In total the data holds 76707 pairs of shipments with the same origin, destination and route. Figure 9-2 shows that while most of the delivery time deltas (53%) are between 0 and 30 minutes, 32% show differences in delivery times of over an hour and 3% of over four hours. The maximum difference in delivery time of shipments taking the same route in the data is 8 hours and 15 minutes. Possible explanations for these fluctuations are plentiful: Drivers could have taken a break, unknown congestions in the traffic could have occurred or the time of delivery was simply entered incorrectly. Since the delivery times are the focus of the validation, the variation of the underlying data should be considered when discussing the validation results.

The length and driving time on the final edges of the network were calculated using the Energy Module of SIN (25). The disruption was simulated by impacting all edges within the impact zones of the forest fires. The impact zones were downloaded from Copernicus EMS Rapid Mapping EMSR760 - Wildfire in Portugal (24) and can be seen in Figure 9-1.

The performance of all services within these areas was halved during the disruption, leading to all vehicles driving on these edges in that time to drive with half the velocity.

The final dataset used for the simulation holds 3819 shipments, 296 nodes, 731 edges from which 109 are disrupted by the forest fire between the 16th and 18th of September and 168 vehicles. The final network can be seen in Figure 9-3. In total 28 days were simulated.

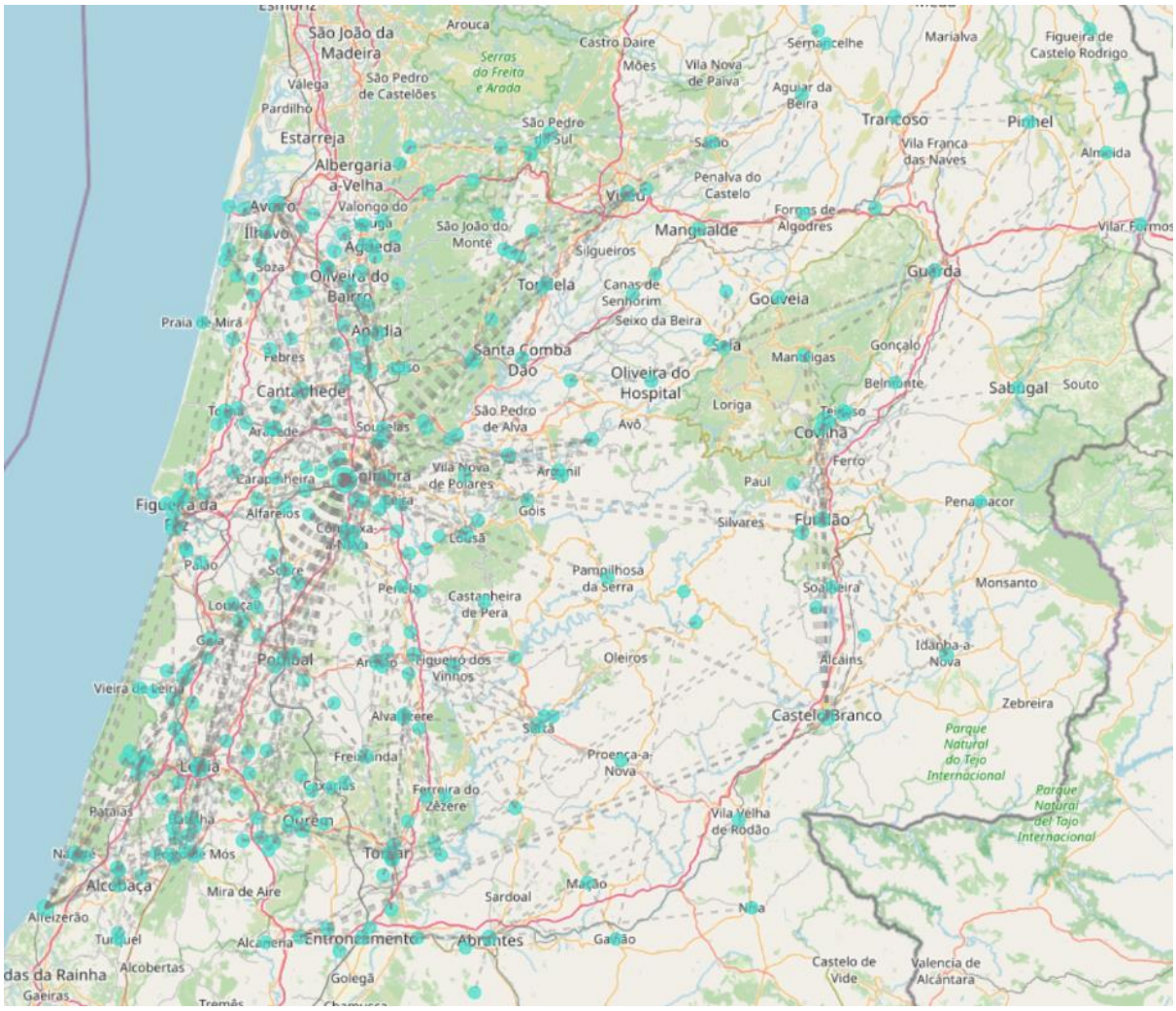


Figure 9-3: Screenshot of the ASTROIT UI, showing the simulated Coimbra logistic network.

Due to the nature of the data and simulated network, some features of ASTROIT had to be deactivated to ensure accurate results. First the pathfinding algorithm of ASTROIT is designed for networks with periodic services on each edge and thus not for last-mile logistics. Since the routes and schedules of the vehicles within the dataset are covering last-mile shipments and thus are governed by the daily fluctuations of the destinations of shipments, the vehicles drive highly irregularly. Thus, the vehicles and routes taken by each shipment was predefined for the simulation using the data. For a verification of the routing algorithm of ASTROIT, see section 9.1.

The validation was conducted by comparing the simulated delivery times with the observed delivery times for each shipment. Delivery times in this context includes the transloading times of the shipments.

Figure 9-4 shows the distributions of delivery times in the simulated and observed data.

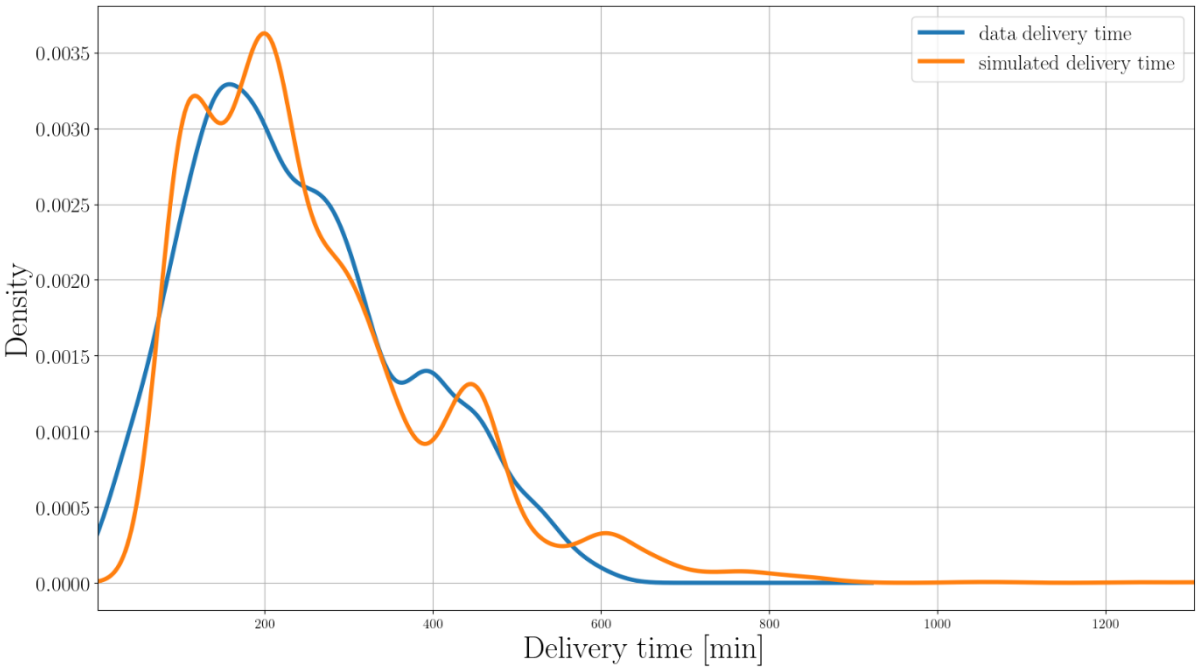


Figure 9-4: Distribution of delivery times of simulated data in orange and observed data in blue

One can clearly see that ASTROIT does replicate the shape of the distribution, with the simulated distribution fluctuating around the observed distribution. However, ASTROIT predicts delivery times larger than 600 minutes significantly more often than was observed in the real deliverable data.

To quantify the accuracy of ASTROIT the following two statistics were used:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad \textit{Equation 9-1}$$

$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{(|y_i + \hat{y}_i|)/2} \qquad \textit{Equation 9-2}$$

where y_i represents the observed and $\hat{y}_i$ the simulated delivery time of shipment i and n is equal to the total number of shipments, 3819.

Table 9-2 shows the scores of the validation for the MAPE and the sMAPE.

Table 9-2: Results of the validation of ASTROITS using MAPE and sMAPE for the delivery times.

MAPE	sMAPE
75%	45%

Both metrics clearly show variations between the simulated and observed delivery times. However, considering the property of the sMAPE to better balance over- and underestimation compared to the regular MAPE, the better performance of the sMAPE indicate that large outliers in the errors have a strong influence on the results. Figure 9-5 shows the error of the delivery times $y_i - \hat{y}_i$ in minutes for each loading start time i. Each dot represents an error of a predicted delivery time, measured by subtracting the simulated from the observed delivery. The time of the disruption is marked by red lines.

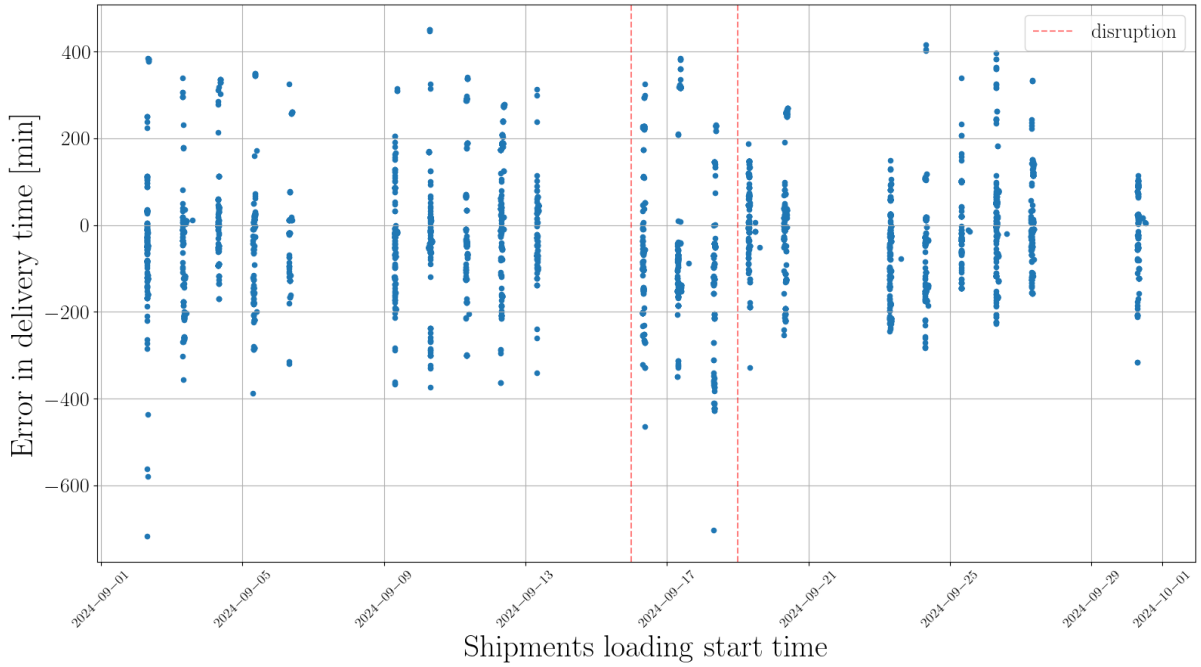


Figure 9-5: The error of the delivery times in minutes for each loading start time. The time of the disruption is marked by red lines.

Figure 9-5 shows that shipments usually get loaded in the morning of each day and only on weekdays, leading to four groups of five vertical clusters and one final vertical cluster indicating the workdays of their respected week. Overestimations of the delivery time led to the highest errors of over minus 700 minutes between the 1st of September and the 5th and during the disruption form the 16th to the 18th. Underestimations of the delivery times on average are also common, but are less extreme, with errors only reaching values of around 400 minutes. In general, the errors seem to be evenly distributed around zero, especially when disregarding the three outliers under minus 500 minutes. For the fourth week of the simulated time from days 16 to 20 the error shows no values under minus 300 minutes. This is the only week where this is the case. Furthermore, the disruption seems to have no clear effect on the error in delivery times. This indicates that the chosen effect of the disruption does fit the observed data. However, given the general spread of the errors, this conclusion does not hold with certainty.

Figure 9-6 depicts the frequency of errors in minutes.

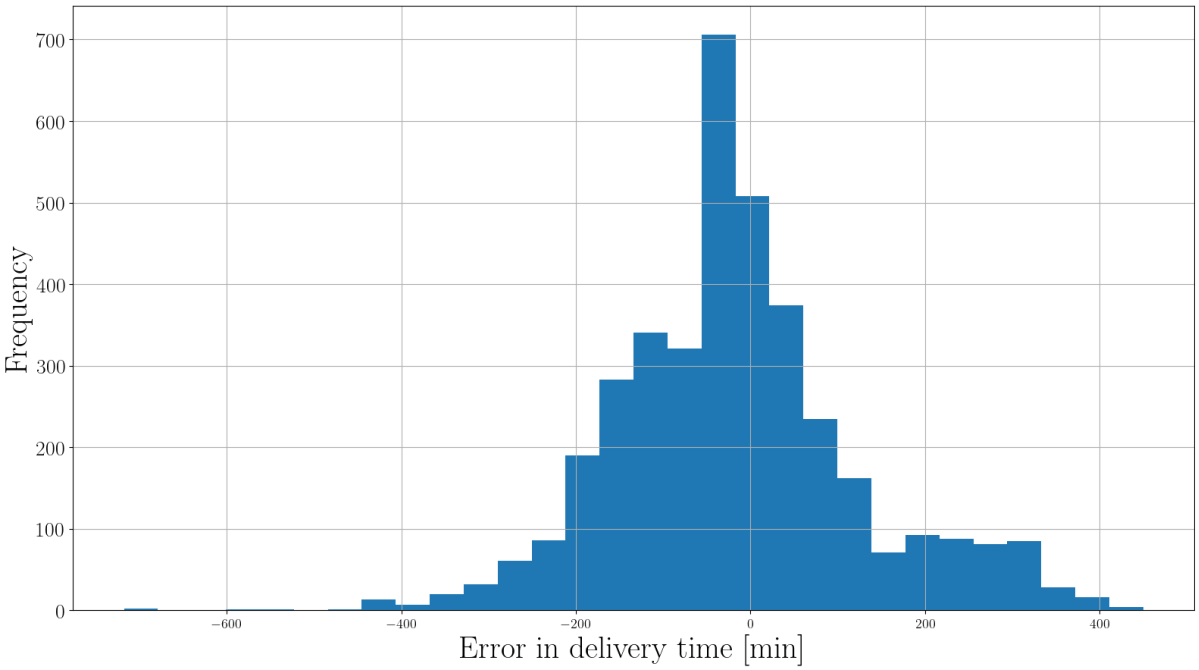


Figure 9-6: Underlying distribution of delivery time errors

The plot confirms that the errors are distributed around zero with a larger tail in the negative values. This confirms the observation that ASTROIT more often overestimates delivery times. The fluctuations in the underlying data shown in Figure 9-2 explain part of the variation Figure 9-6.

The results clearly show that many unknown variables govern the delivery times in the used dataset, leading to large fluctuations of shipments with the same route in the observed data, as shown Figure 9-2. Furthermore, the used dataset did not hold the actual departure times of the vehicles. Thus, the departure times had to be estimated. Therefore, data quality and unpredictable behaviour of drivers can have a large influence on the validation results, leading to positive and negative errors a like. ASTROIT predicted distribution of delivery times does match the shape of the observed distribution of delivery times. However, on average ASTROIT tends to overestimate delivery times and should be further developed to better account for these unknown variables.

10. Conclusion

The research shown in this report builds upon the work conducted in D3.3, where simulations of the tools were carried out and complementary mitigation measures were defined. Based on this framework, the set of tools and models for addressing the precision and accuracy of specific aspects of disruption and freight transport has been analysed.

The proposed tools have been designed to address a wide range of aspects and disruption events within the context of green resilience assessment. Key disruption events include pandemics, wartime scenarios, cyberattacks, and natural disasters of varying severity.

Building on the existing toolset, several opportunities for further development can be identified to enhance its capabilities and address emerging challenges in green resilience assessment.

The potential and limitations of the current SARIL toolkit can be analysed to support the continued advancement of resilience capabilities in the face of disruptions.

The conclusions for each tool are presented below, together with their respective results, proposed improvements, and identified limitations.

- **Natural Hazard Maps**

  The verification of the Forest Fire Risk Map involved a sensitivity analysis in which two of the layers that make up the final risk map were varied for two different seasons. These layers were the Fire Weather Index and the NDVI. The seasons compared were summer and autumn. The results showed how, for the same area, the risk levels varied from one season to another. For example, a higher proportion of high and very high-risk levels (levels 4 and 5) were observed in summer, whereas in autumn these levels decreased in extent and were replaced by moderate and low risk levels. The analysis indicates that no inconsistent risk patterns appeared and that the model responds appropriately to the sensitivity assessment.

  The Minimum Arrival Time of the fire from the FlamMap software was validated by adjusting its internal simulation parameters. To verify this fire propagation model, the same fire event was simulated while keeping all external parameters constant and modifying only FlamMap's internal settings. Specifically, the spatial resolution (pixel size) used to recalculate fire spread was varied. Results indicated that increasing the resolution to 25 m/pixel produced a significant error (85%), demonstrating that higher spatial detail does not necessarily enhance model performance. Conversely, when the pixel size exceeded 70 m, the model underestimated fire propagation, and at 80 m/pixel ignition failed completely, resulting in a 100% error. Resolutions between 30 m and 45 m showed errors ranging from 2% to 10% when compared with FlamMap's default resolution of 30 m/pixel, which has therefore been adopted for the SARIL Project. This validation approach confirmed the reliability of the propagation model and provided a set of resolution–error relationships that enable an effective balance between simulation accuracy and computational efficiency.

  The validation of the Forest Fire Risk Map was carried out by comparing the spatial distribution of areas classified as moderate (level 3), high (level 4), and very high (level 5) risk with historical forest fire occurrences detected by the MODIS satellite sensor between 2017 and August 2025. The analysis involved overlaying these historical fire events (real events) with the Forest Fire Risk Map and calculating the proportion of real fire events occurring within areas of

moderate or higher risk. The results showed that 83 % of recorded fires coincided with zones classified as level 3 or above, indicating strong spatial agreement between predicted and observed fire activity. This outcome demonstrates the high predictive reliability of the Forest Fire Risk Map, confirming its suitability as a decision-support tool for fire prevention, territorial planning, and environmental risk management.

To validate the Minimum Arrival Time results from FlamMap, simulated fire events were compared with real forest fires for which both duration and burned area were known. Comparing the simulated and real burned areas yielded relative an average error of 48 %. While the simulations reproduced the overall shape and spread of the fires reasonably well, quantitative discrepancies were significant. These differences are mainly attributed to the model's inability to include active suppression efforts and the subjectivity and conservative declaration of the fire phases by the authorities. Despite these limitations, the analysis indicates that FlamMap can effectively simulate the general behaviour of small-scale, short-duration forest fires, provided its constraints are properly considered.

- **Scour Monitoring for Decision Support**

A sensitivity analysis was conducted to verify the tool's performance by varying two key inputs: the discharge level (q) and the flow angle correction factor (k_2), both influencing flood intensity. The results showed that higher discharge levels and more severe flow angles caused a logical shift in the optimal management strategy—from keeping some or all bridges open under mild conditions to closing all bridges under more extreme flooding scenarios, reflecting the increased risk of scour and bridge failure.

The Scour Monitoring for Decision Support tool was validated in two ways. First, its predicted scour depths were compared with those obtained from Johnson's well-established model. The comparison, performed for the second pier of the first bridge across a range of discharges, showed that the percentage sum of squared errors (SSE%) between the two models decreased consistently as discharge increased—from 11.40% at 100 m^3/s to 5.92% at 200 m^3/s. This trend demonstrates that the tool's predictions become increasingly consistent with those of Johnson's model under higher flow conditions, supporting its reliability in estimating scour depth.

Second, the tool's estimated probability of failure for bridge piers was evaluated against InSAR-derived displacement data for the same bridge area. Since data from periods of heavy rainfall and flooding were unavailable, the analysis was conducted under normal conditions to verify precision and accuracy. After correcting for temperature-related effects, most displacement residuals in Zone 4 (where the piers are located) remained within the defined threshold, indicating stable longitudinal behaviour. This aligns with expectations in the absence of extreme flow events, further confirming the tool's accuracy and robustness.

- **Disruption Information Interface**

  Evaluation of the function used for predicting the end of road disruptions (Survival Analysis)

  The verification process demonstrated that the model functions correctly across different disruptions categories, producing survival curves that closely align with theoretical expectations. The low MSE values confirm that the model's internal logic precisely captures the decline in survival probability over time. This outcome verifies that the estimator effectively models both short-term and long-duration events, maintaining numerical stability and interpretability throughout. The results collectively indicate that the implemented survival analysis framework behaves as intended and provides a solid foundation for further refinement using real-world data.

  Validation using actual operation data confirmed that model provides a meaningful approximation of event duration dynamics, though with a noticeable margin of error in predicting exact end times. These discrepancies suggest that while the baseline survival model captures general temporal behaviour, it lacks certain contextual inputs necessary for precise forecasting. To address this limitation, future iterations will extend the estimation process with additional spatial and operational features. This enhancement is expected to improve both calibration and predictive accuracy, enabling the survival model to transition from a proof-of-concept tool into a reliable operational forecasting component.

  Evaluation of the Classifier Functionality for Messages Containing Disruption Information

  The verification process confirmed that the model's core classification logic operates as intended, correctly identifying text patterns associated with traffic disruptions. The analysis of confusion matrices and performance metrics showed that, while recall remains high - indicating the model's effectiveness in capturing true disruptions - the number of false positives contributes to moderate overall precision. This outcome aligns with the system's intended role as a decision-support component, where missing a disruption would be more costly than raising additional alerts. Therefore, the verification results demonstrate that the classifier behaves consistently with its design assumptions and is dependable as a first-stage filtering mechanism for identifying potential incidents in logistic chains.

  Validation conducted on real-world data confirmed the same behaviour patterns, showing that the model reliably detects the disruptions but tends to over-predict under ambiguous linguistic conditions. From an operational standpoint, this asymmetry supports system reliability by ensuring that no significant events are overlooked, though it requires additional review or filtering to reduce false alarms. This trade-off makes the model particularly suitable for monitoring and early-warning applications where high recall is prioritized. Further refinements - such as decision threshold tuning, post-processing filters or hybrid ensemble methods - will help increase precision and move the classifier closer to the level of automation required for fully autonomous disruption alerting.

- **Route Attributes (Energy Module)**

  The verification of the Energy Module involved sensitivity analyses through conducting two tests to evaluate how changes in vehicle weight and route slope affect CO_2 emissions. In Test 1, a heavy diesel truck with varying payloads was simulated on the Vigo–Porto route. The results showed a near-linear increase in CO_2 emissions with increasing payload, consistent

with physical expectations. In Test 2, the effect of route slope was assessed using a light diesel van. Steeper gradients led to higher emissions per kilometre, while downhill slopes reduced emissions. These results confirm that the Energy Module responds logically to input variations and behaves consistently under controlled conditions.

The validity of the Energy Module was assessed using two datasets: one for truck route attributes and emissions (RANGEL), and one for train fuel consumption (Bane NOR). For trucks, modelled results were compared with observed data across short (≤20 km) and long (>20 km) routes. Error metrics showed that short routes had relatively high deviations, MAPE of 42% for travel time and 28% for emissions, limiting reliability at this scale. For long routes, errors were notably lower, with MAPE values of 10% for distance, 26% for travel time, and 16% for emissions, generally within acceptable ranges for transport modelling.

For trains, the Energy Module was validated against GPS and fuel consumption data for six locomotive types. MAPE values ranged from 20% to 51%, and NRMSE values from 29% to 74%, depending on the locomotive type. The analysis also highlights the need for improved input parameters, particularly more detailed vehicle specifications and operational data, to further enhance model accuracy and reduce uncertainty in future applications.

- **Vulnerability and Traffic tool**

To verify the functionality of the 'Vulnerability and Traffic' tool and ascertain the logical coherence of the model's outcomes relative to input changes, a sensitivity analysis was performed. The key variation applied was a 25% increase in traffic flow across the Traffic module's Origin-Destination matrix. Consistent with expectations, the verification shows that elevated traffic volumes drastically degrade the Level of Service, resulting in marked increases in total travel distance and time. These metrics directly correlate to a substantial escalation in total indirect costs. Notably, a minimal 25% increase in traffic flow proved adequate to push the entire network to a state of severe and widespread congestion. This observed non-linear behaviour, where a modest traffic increase leads to disproportionately high increases in distance and time, strongly suggests the simulation tool is effectively capturing the rapid saturation of the network's capacity. This traffic variation highlights a critical issue: the network's low resilience. Given the limited number of alternative paths, interruptions rapidly cause widespread congestion as vehicles are forced onto a few key segments.

The Vulnerability and Traffic tool's trustworthiness will be established through external validation via comparative analysis, as direct local validation is impossible due to the scarcity of historical flood data and recent events. This validation strategy involves:

- Risk Assessment Validation: The tool's vulnerability and risk outputs for critical assets are verified by linking predicted flood probability with simulated flood extents/depths derived from historical flood maps.
- Traffic Impact Validation: The tool's impact calculations, specifically increased travel time/distance, indirect costs, and recovery days, are compared against outputs from literature.

The validation confirms the consistency of the tool results with values documented in the literature.

- **Traffic Simulation**

  The verification process confirmed that the Traffic Simulation tool behaves logically and consistently under controlled modifications of input parameters. The sensitivity analyses demonstrated that the model appropriately responds to changes in network structure, signal timing, demand loading, and link closures, producing coherent and expected variations in travel times, congestion patterns, and route distributions. These outcomes validate the internal logic and functional soundness of the model, confirming that its algorithms correctly implement rerouting, queuing, and traffic signal dynamics. Nonetheless, the tests also revealed opportunities for refinement, particularly in enhancing the representation of intersection control under highly congested conditions and improving parameter calibration for mixed traffic flows. Incorporating more detailed intersection modelling or adaptive signal logic in future versions could further strengthen the model's predictive precision and its applicability to a wider range of disruption scenarios.

  The validation results demonstrate that the Traffic Simulation tool achieves a moderate to high level of accuracy, especially for heavy vehicles, which represent the primary focus of the SARIL national scenario. Across all OD pairs, the model achieved WMAPE values of 17% for light vehicles and 13% for heavy vehicles, while the NRMSE remained below the commonly accepted 30% threshold, confirming the robustness of simulated travel times relative to the validated reference model. The close alignment between simulated and observed freight travel times supports the reliability of the model in assessing disruption impacts on international transport corridors. However, higher relative errors for light vehicles, particularly on short-distance trips, highlight a limitation in percentage-based error metrics and suggest that further calibration using local traffic counts or speed profiles could enhance accuracy for passenger mobility analysis. Overall, the validation confirms that the tool provides a sound and credible basis for resilience assessment and scenario evaluation, with future improvements expected from finer-grained demand data and enhanced behavioural calibration.

- **Transport Simulations (ASTROIT)**

  The results of the verification show that altering the preferences of the routing algorithm leads to realistic alterations in the chosen routes. The choice of preferences reliable minimizes the chosen variable within the network, i.e. time, costs and emissions. Furthermore, when minimizing costs CO2 emissions rise and vice versa, highlighting trade-off effects in the simulated network that can be observed in real logistic networks.

  The validation results show that the shape of the distribution of the observed delivery time was replicated by ASTROIT, further underlining the robustness of ASTROIT. However, the accuracy of ASTROIT still needs to be improved. Measuring the MAPE and sMAPE yields average errors of 75% and 45% respectively. The lower result of the sMAPE clearly indicates that the underlying errors have large outliers. Analysing the errors show that ASTROITS has a bias towards longer delivery times. However, the underlying observed data shows large fluctuations of the delivery times for the same routes. These fluctuations could be explained by unknown variables such as longer breaks of the drivers or are caused by low data quality. They decrease the accuracy of ASTROIT significantly and make the validation results less

reliable, introducing variations to the data that ASTROIT cannot replicate. Future research should focus on minimizing the unknown variation in the validation dataset and on improving ASTROIT to reduce its bias towards longer delivery times. Furthermore, ASTROIT was developed for the simulation of periodic transport services, focusing on groupage networks. In the future ASTROIT will be further developed to cover non-scheduled services.

References

1. *Fire susceptibility modeling and mapping in Mediterranean forests of Turkey: a comprehensive study based on fuel, climatic, topographic, and anthropogenic factors.* **Novo, A., Dutal, H. & Eskandari, S.** 2024, uro-Mediterr J Environ Integr 9, 655–679.

2. *Forest fire risk modeling in Mediterranean forests using GIS and AHP method: case of the high Rif forest massif (Morocco).* **Saber, Er-Riyahi & Bouhlal, Abdeslam & Mohamed, El Mazi & Boutallaka, Mohamed & Chanyour, Yassine.** 2024, Euro-Mediterranean Journal for Environmental Integration.

3. **Xunta de Galicia.** *Xunta de Galicia. (2025). Plan de prevención e defensa contra os incendios forestais de Galicia (PLADIGA 2025). Consellería do Medio Rural. Retrieved from https://mediorural.xunta.gal/es/temas/defensa-monte/pladiga-2025.* 2025.

4. *Manual de prevención e defensa contra os incendios forestais de Galiza. Consellaría do Medio Rural. Depósito Legal: C 1965-2007. Retrieved from https://mediorural.xunta.gal/sites/default/files/publicacions/2019-10/manual_incendios.pdf.* **Yáñez Armesto, A., Castro López, F., Lombardía Fernández, C., & Varela Núñez, M. J.** 2007.

5. *Evaluating Scour at Bridges.* **Arneson, L.A., Zevenbergen, L.W., Lagasse, P.F. and Clopper, P.E.** FHWA-HIF-12-003, 2012.

6. *Scour at Wide Bridge Piers.* **Jones, J Sterling and Sheppard, D Max.** 2000. Joint Conference on Water Resource Engineering and Water Resources Planning and Management 2000.

7. *Remote monitoring to predict bridge scour failure using Interferometric Synthetic Aperture Radar (InSAR) stacking techniques.* **Selvakumaran, S., Plank, S., Geiß, C., Rossi, C., & Middleton, C.** 463-470, 2018, International Journal of Applied Earth Observation and Geoinformation, Vol. 73.

8. *A method for structural monitoring of multispan bridges using satellite InSAR data with uncertainty quantification and its pre-collapse application to the Albiano-Magra Bridge in Italy.* **Farneti, E., Cavalagli, N., Costantini, M., Trillo, F., Minati, F., Venanzi, I., & Ubertini, F.** (1) 353-371, 2022, Structural Health Monitoring, Vol. 22.

9. *Satellite-Based Monitoring of a Multi-span Pre-stressed Concrete Bridge Crossing the Po River.* **Miano, A., Giordano, P. F., Di Martire, D., Giulivo, F., Liuzzo, R., Tufano, R., ... & Prota, A.** 367-374, 2025, International Conference on Protection of Historical Constructions.

10. *Nonparametric Estimation from Incomplete Observations.* **Kaplan, Edward L. and Meier, Paul.** 282, 1958, Vol. 53.

11. *Modelling Survival Data in Medical Research.* **Collett, David.** New York : Chapman and Hall/CRC, 2015. 9780429196294.

12. *The Evolution of Boosting Algorithms From Machine Learning to Statistical Modelling.* **Mayr, Andreas, Binder, Harald and Gefeller, Olaf.** 5, 2014, Vol. 53.

13. **Therneau, Terry M. and Grambsch, Patricia M.** *Modeling Survival Data: Extending the Cox Model.* New York : Springer New York, NY, 2013. 978-1-4757-3294-8.

14. **Klein, John P. and Moeschberger, Melvin L.** *Survival Analysis: Techniques for Censored and Truncated Data.* Berlin : Springer, 2003.

15. **Meeker, William Q. and Escobar, Luis A.** *Statistical Methods for Reliability Data.* New York : John Wiley & Sons, 2014.

16. **Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X.** doccano: Text Annotation Tool for Human. [Online] GitHub., 2018. https://github.com/doccano/doccano.

17. *Learning from Imbalanced Data.* **He, Haibo and Garcia, Edwardo A.** 9, 2009, IEEE Transactions on Knowledge and Data Engineering, Vol. 21, pp. 1263 - 1284.

18. **Saito, Takaya and Rehmsmeier, Marc.** The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one.* 4 March 2015.

19. *Assessing Flood Indirect Impacts on Road Transport Networks Applying Mesoscopic Traffic Modelling: The Case Study of Santarém Portugal.* **Jafari Shahdani, Fereshte & Santamaria, Monica & Sousa, Helder & Coelho, Mário & Matos, Jose.** (6) 10.3390/app12063076, 2022, Applied Sciences, Vol. 12.

20. *Investigating the impact of river floods on travel demand based on an agent-based modeling approach: The case of Liège, Belgium.* **Saadi, I., Mustafa, A., Teller, J., & Cools, M.** 2018, Transport policy, Vol. 67, pp. 102-110.

21. *Assessing the knock-on effects of flooding on road transportation.* **Pyatkova, K., Chen, A. S., Butler, D., Vojinović, Z., & Djordjević, S.** 2019, Journal of environmental management, Vol. 244, pp. 48-60.

22. *Project-Level Optimization of Repair Activities for the Recovery of Transportation Assets after Inland Flood Events.* **Lichty, B., Zhang, N., Nasrazadani, H., Adey, B. T., & Alipour, A.** 2, s.l. : Journal of Infrastructure Systems, 2025, Vol. 31, p. 04025004.

23. **Wang, Hantong, et al.** Transportation Simulation Modeling and Location-Based Services Data Completion Based on a Data and Model Dual-Driven Approach. [Online] 5 2024. https://www.mdpi.com/2076-3417/14/11/4366/htm. 2076-3417.

24. **Management, EU Emergency.** EMSR760 - Wildfire in Portugal. [Online] Copernicus Emergency Management Service, 2024. [Cited: 10 09 2025.] https://rapidmapping.emergency.copernicus.eu/EMSR760/aem.

25. **SINTEF.** SINTEF Enegery module. [Online] SINTEF. [Cited: 20 09 2025.] https://energimodul.sintef.no/.

26. *Comparing Methods of Calculating Expected Annual Damage in Urban Pluvial Flood Risk Assessments.* **Olsen, A.S., et al.** (1) https://doi.org/10.3390/w7010255, 2015, Water, Vol. 7, pp. 255-270.

27. *Restoration models for quantifying flood resilience of bridges.* **Mitoulis, S. A., Argyroudis, S. A., Loli, M., & Imam, B.** 2021, Engineering structures., Vol. 238, p. 112180.

28. *Travel time predictions: should one model speeds or travel times?* **Bauer, D., & Tulic, M.** (2) https://doi.org/10.1186/s12544-018-0315-7, 2018, European Transport Research Review, Vol. 10, p. 46.

29. *The reliability of travel time forecasting.* **Yang, M., Liu, Y., & You, Z.** (11) https://doi.org/10.1109/TITS.2009.2037136, 2009, IEEE Transactions on Intelligent Transportation Systems, Vol. 11, pp. 162-171.