Ref. Ares(2025)11504223 - 22/12/2025

DATA SPACE 4 MOBILITY MODELS INTEGRATION

Project Acronym:	**23-ES-TG-DS4MOBMODELS**
Project ID :	**101176007**

Deliverable number:	**D4.2**
Deliverable name:	**Validation methodology**
Work Package number:	**4. AI DATA SCIENCE AND SIMULATION TRANSPORT MODELS**
Due date:	**31/05/2025**
Delivery date:	**22/12/2025**

Lead Beneficiary:	**AIMSUN**
Main author:	**Antonio Pellicer (Aimsun)**

Co-funded by the European Union

DS4MOB MODELS

Document fiche	
Authors:	Antonio Pellicer (Aimsun)
Internal reviewers:	Lidia Montero (UPC), Marta Cuatrecasas (UPC), Marc Manzanares (UPC), Allison Fernández-Lobo (UPM), Andrés Monzón (UPM)
Work Package:	WP4
Task:	T4.2
Nature:	R
Dissemination:	PU

Document History			
Version	**Date**	**Contributor(s)**	**Description**
V 0.0	01/11/2025	AIMSUN	Initial draft
V 0.1	10/12/2025	AIMSUN	Final draft to be reviewed
V 0.2	16/12/2025	CPS/INGARTEK	Reviewed version. Draft for final Q&A
V 1.0	22/12/2025	AIMSUN	Final

Keywords:	Validation methodology; Transport models; Simulation-based models; Data-driven models; Traffic simulation; Model validation; Cross-validation; Mobility Data Space; Use cases; AI and machine learning; Transport planning
Abstract (few lines):	This deliverable defines the validation methodology to be applied to the transport models developed within the DS4MOBMODELS project, covering both simulation-based and data-driven approaches. It first updates the classification of the project use cases according to their most suitable modelling paradigm (data-driven or traditional simulation-based), building on Deliverables D4.1 “Transport model design” and D2.3 “Final definition of selected use cases”. The document then reviews existing methodologies and key indicators for validating macroscopic and dynamic simulation models, proposing recommended workflows and acceptance criteria to ensure credibility, robustness and fitness for purpose. In parallel, it provides an overview of validation strategies for data-driven models, including hold-out, cross-validation, leave-out and randomised methods, highlighting how the choice of technique depends on dataset characteristics and application context. Together, these guidelines establish a common validation framework to support consistent assessment of models across all use cases in DS4MOBMODELS.

Table of Contents

DEFINITIONS, ACRONYMS AND ABBREVIATIONS **6**

EXECUTIVE SUMMARY **7**

1 INTRODUCTION **8**

1.1 Purpose and Objectives 8
1.2 Intended audience 8
1.3 Project summary 8
1.4 Structure of the deliverable 8

2 USE CASES AND MODEL CLASSIFICATION **10**

2.1 Summary of definitive Use Cases 10
2.2 Initial classification of Use Cases by model type 12

3 METHODOLOGIES FOR VALIDATION OF SIMULATION-BASED MODELS **15**

3.1 Definitions. Variables and Indicators 15
3.2 Conceptual Basis of Validation 17
3.3 Validation of Macroscopic Models 17
3.3.1 Validation by model component 17
3.3.2 Validation criteria and KPIs 18
3.3.3 Concluding remarks 20
3.4 Recommended validation workflow for dynamic models 20
3.4.1 Summary of Acceptance Criteria 21
3.5 Validation criteria comparison macroscopic vs dynamic models 21
3.5.1 Concluding remarks 22

4 METHODOLOGIES FOR VALIDATION OF DATA-DRIVEN MODELS **23**

4.1 Data-driven solutions 23
4.2 Model validation methods for data-driven models 24
4.2.1 Hold-out validation method 25
4.2.2 Cross-validation method 25
4.3 Selection of the data-driven model validation method for each Use Case 27

5 CONCLUSIONS **29**

6 REFERENCES **30**

Table of Figures

Figure 1. Tentative classification of Use Cases and pilots based on data-driven or simulation-based models 12
Figure 2. Classification of model validation methods. Adapted from (Baladram, 2024) 25

Table of Tables

Table 1. Summary of Definitive Use Cases, objectives and models 10
Table 2. Mapping of Use Cases to Traditional Transport Models 13
Table 3. Classification of Use Cases based on used traditional modelling types 14
Table 4. Recommended Validation Criteria for Macroscopic Transport Models 19
Table 5. Comparative Table: Mesoscopic vs Macroscopic Validation Criteria 21

Definitions, Acronyms and Abbreviations

Acronym	Title
CBA	Cost-Benefit Analysis
CBD	Central business district
CR	Coincidence ratio
DNN	Deep Neural Networks
EV	Electric vehicle
HIS	Household Interview Survey
KPI	Key Performance Indicator
LEZ	Low Emission Zones
LOS	Level of service
LSCP	Location Set Covering Problem
MCA	Multicriteria Analysis
ML	Machine Learning
OD	Origin – Destination
PCU	Passenger Car Units
RMSE	Root Mean Square Error
TEN-T	Trans-European Transport Network
UC	Use Case
VMP	Variable Message Panels
VMS	Variable Message Signs
WP	Work Package

Executive Summary

This deliverable presents existing methodologies for validating both simulation-based transport models and data-driven models and proposes validation workflows to be applied to use cases. Due to the importance of applying the right validation methodology to each use case, this document also updates the classification of use cases according to their most suitable modelling approach building upon the initial framework defined in Deliverable D4.1 “Transport model design” and complemented with insights from Deliverable D2.3 “Final definition of selected use cases”. As some use cases could be addressed with either approach (data-driven or simulation-based), the validation process is outlined at a theoretical level and ultimately will depend on the data availability and the maturity of existing models.

Validation is highlighted as a critical step for ensuring model credibility and robustness. Simulation-based approaches require staged validation processes that combine empirical data comparisons, behavioural realism checks, and sensitivity analyses. In this document, validation of macroscopic models and dynamic models are examined in detail, and existing guidelines and recommended Validation Criteria are summarised to support best practices.

Data-driven models demand rigorous performance assessment beyond training datasets to ensure predictive reliability under real-world conditions. This deliverable includes a classification of different validation methods for data-driven models, such as hold-out techniques and cross-validation approaches including k-fold, leave-out, and random methods. Each method has unique strengths designed to address specific challenges, and this deliverable provide guidance to help select the most appropriate technique, which is critical for obtaining accurate and robust assessments of model performance.

.

1 Introduction

1.1 Purpose and Objectives

The purpose of this deliverable is to review existing **methodologies for validating** both simulation-based transport models and data-driven models that will be applied to the use cases. Since some use cases could be addressed with either approach, and the final choice depends on data availability and the maturity of existing models, this deliverable provides a theoretical overview of current validation methodologies so they can be identified for application once all use-case data are fully defined. To support this, the deliverable also updates the **classification of use cases** according to their most suitable modelling approach – data-driven or simulation-based – building upon the initial framework defined in Deliverable D4.1 and complemented with insights from Deliverable D2.3.

To support this goal, the deliverable is structured around the following key objectives:

- Update the initial classification of use cases to model types (data-driven or simulation-based) as originally specified in Deliverable D4.1, now complemented with insights from Deliverable D2.3. For this, use cases will be mapped to appropriate modelling approaches.
- Review current and existing methodologies for validation of both simulation-based transport models and data-driven models.
- Provide validation workflows to be used once data for each use case becomes available.

1.2 Intended audience

This deliverable will be **disseminated publicly** informing all interested stakeholders, researchers, practitioners and decision-makers about practical methodologies for validating simulation-based and data-driven models. The content is particularly relevant for those responsible for implementing, assessing or improving transport models or data-driven models.

1.3 Project summary

The DS4MOBMODELS project aims to define and create an advanced Mobility Data Space that integrates transport, traffic, and travel data across various regions, including the Spanish state, the Nouvelle-Aquitaine-Euskadi-Navarre Euroregion, and the Central Portugal region.

The primary goal is to enhance the availability, quality, and interoperability of mobility data to address key national, cross-border, regional, and local use cases related to sustainable urban mobility, traffic management, and real-time operations within the Trans-European Transport Network (TEN-T).

The project will develop a data platform that will support the integration of various mobility models and data science tools to facilitate efficient, safe, and sustainable mobility management. The initiative will also establish robust data governance in line with European and national regulations, ensuring secure and interoperable access to the data for all the stakeholders in a non-discriminatory way.

The DS4MOBMODELS project aligns with the broader objectives of the European Data Strategy, the Digital Europe Programme, and national digital and mobility strategies in Spain, France and Portugal, contributing to the creation of a more integrated and resilient European mobility infrastructure.

1.4 Structure of the deliverable

This deliverable is structure into 5 sections with the corresponding description as follows:

- Section 2 provides an initial classification of use cases as an introduction to the validation methodologies based on whether they require data-driven or traditional simulation-based modelling approaches.
- Section 3 defines the validation methodologies for simulation-based models.
- Section 4 describes the validation methodologies for data-driven models.
- Section 5 concludes this deliverable with some final remarks.

2 Use Cases and model classification

This section summarises the definitive use cases identified in the previous deliverable D2.3 and provides an initial classification according to the nature of the models to be developed: data-driven or simulation-based. Traditional simulation-based use cases rely on well-established transport modelling methodologies such as macroscopic or mesoscopic assignment, four-step demand modelling, survey-based OD estimation, GIS-based accessibility analysis, and classical environmental assessment tools. The classification is based on the primary objective of each use case and the methodological approach required to achieve it.

2.1 Summary of definitive Use Cases

The Use Cases identified in the Deliverable D2.3 are summarised in Table 1:

Table 1. Summary of Definitive Use Cases, objectives and models

UC	Pilot	Selected use Case	Objective	Model to be develop
1	**National Pilot**	Winter service Management: Emergency Parking for Heavy Vehicles	To implement an automatic system for redirecting heavy vehicles to emergency parking areas during snow alerts, using driver notifications via Variable Message Panels (VMPs).	A real-time monitoring and machine learning–based prediction model, capable of **forecasting heavy vehicle flows during adverse weather conditions** and informing users through variable message signs (VMS).
2	**National Pilot**	Highway and main road congestion monitoring	To provide real-time assessment of traffic conditions on Spain's primary road corridors, enabling both situational awareness and predictive traffic management	Process data in **real time** and perform **predictive analysis** based on historical and live traffic information, with the aim of anticipating short-term congestion periods and supporting real-time decision-making
3	**Regional pilot - Andalusia**	Transport emissions evaluation	To obtain detailed and spatialized data on emissions by transport type, mode, and distance travelled, and determine the share of emissions attributable to transport modes under regional responsibility	An **existing model** will be used to evaluate the current situation and compare alternative scenarios at a regional scale. The model is intended for **strategic, aggregated analysis** rather than real-time decision-making.
4	**Regional pilot - Valencia**	Low Emission Zone monitoring	To obtain **air quality** and sustainable mobility indicators for monitoring Low Emission Zones (LEZ)	The indicators will be calculated in absolute or comparative terms based on different types of models, such as traffic, dispersion, or modal share, among others. These models will be implemented using **GIS tools and simulation tools**
5	**Regional Pilot - Cataluña**	Active mobility monitoring	Active mobility monitoring through several indicators related to **travel patterns in sustainable modes** such as walking, cycling, etc.	The indicators will be calculated in absolute or comparative terms based on different types of models, such as traffic, dispersion, or modal share, among others. These models will be implemented using **GIS tools and simulation tools**
6	**Local pilot - Madrid**	Intermodal node demand analysis	To identify optimal locations for new intermodal public transport nodes within municipal boundaries	Multicriteria Analysis (MCA) for comparative assessment of candidate areas; Location Optimization Models, as p-median and p-center models to minimize travel distance/time from demand nodes or location Set Covering Problem (LSCP) to ensure coverage of all high-demand areas;

				Machine Learning-based Predictive Models; Forecast demand at interchanges using techniques such as Random Forest and XGBoost; Cost-Benefit Analysis (CBA) to prioritize feasible and high-impact interventions; or Data integration tools, such as GIS-based spatial analysis, accessibility models
7	**Cross-border Pilot – Spain and Portugal** Extremadura region and Central Region of Portugal	Cross-Border Passenger Transport	To analyse passenger mobility between Extremadura and Portugal, evaluating cross-border public transport services, vehicle flows, and the adequacy of transport networks to international travel demand.	• Based on mobile phone data to identify **main cross-border flows**. • Capable of analysing **patterns** over different timeframes and day types (e.g., peak vs. off-peak, weekday vs. weekend).
8	**Cross-border Pilot – Spain and Portugal** Extremadura region and Central Region of Portugal	Cross-border freight transit analysis	Cross-border freight flow analysis oriented towards logistic coordination between territories	Development of a f**reight traffic forecasting model** for the analysed corridors. This model will estimate future freight flows under different infrastructure and policy scenarios
9	**Cross-border Pilot – Spain and Portugal** Extremadura region and Central Region of Portugal	Freight emissions	Calculate emissions generated by freight transport on key cross-border corridors	**Simulation-based network model**, which integrates demand, fleet composition, and road network data. It applies emission factors based on vehicle technology, road type, and average speed. It could simulate different scenarios, such as fleet renewal, electrification, modal shifts, implementation of low emission zones, and reinforcement of public transport.
10	**Cross-border Pilot – Spain and Portugal** Extremadura region and Central Region of Portugal	Charging infrastructure	Cross-border mobility data space to design coordinated strategies for electric vehicle (EV) charging infrastructure planning	**Cluster analysis** to identify optimal and high demand charging locations
11	**Cross-border Pilot – Spain-France** Regions Nouvelle-Aquitaine (France) and Euskadi and Navarra (Spain)	Cross-Border Passenger Transport	Analyse cross-border passenger mobility between the region Nouvelle-Aquitaine (France) and Euskadi and Navarra (Spain) to evaluate the adequacy of existing transport services and infrastructure	• Analysis based on mobile phone data to identify **main mobility flows** between the regions Nouvelle-Aquitaine-Euskadi-Navarra (France and Spain). • Integration of traffic count data and public transport supply and demand. • Temporal analysis of **typical travel days** and peak periods. • The Transboundary Mobility Flow Model maps **commuting patterns** and infrastructure needs
12	**Cross-border Pilot – Spain-France** Regions Nouvelle-Aquitaine (France) and Euskadi and Navarra (Spain)	Cross-border freight transit analysis	Analysis of freight transit across the regions Nouvelle-Aquitaine-Euskadi-Navarra (Spain–France border), aimed at improving logistic coordination and interoperability between both territories	Development of a **freight traffic forecasting model** for the analysed corridors. This model will estimate future freight flows under different infrastructure and policy scenarios
13	**Cross-border Pilot – Spain-France** Regions Nouvelle-Aquitaine (France) and Euskadi and Navarra (Spain)	Freight emissions	Calculate emissions generated by freight transport on key cross-border corridors	**Simulation-based network model**, which integrates demand, fleet composition, and road network data. It applies emission factors based on vehicle technology, road type, and average speed. It could simulate

				different scenarios, such as fleet renewal, electrification, modal shifts, implementation of low emission zones, and reinforcement of public transport
14	**Cross-border Pilot – Spain-France** Regions Nouvelle-Aquitaine (France) and Euskadi and Navarra (Spain)	Charging infrastructure	Establish a cross-border mobility data space to design coordinated strategies for electric vehicle (EV) charging infrastructure planning	**Cluster analysis** to identify optimal and high demand charging locations

2.2 Initial classification of Use Cases by model type

The Use Cases are now consolidated and classified according to the type of modelling approach required. This **new provisional classification of Use Cases** updates the previous one included in the previous Deliverable D4.1. This classification is based on the primary objective of each Use Case and the nature of the models to be developed and is shown in Figure 1. However, it still remains provisional, as the final approach will be chosen once all available datasets and models have been confirmed.

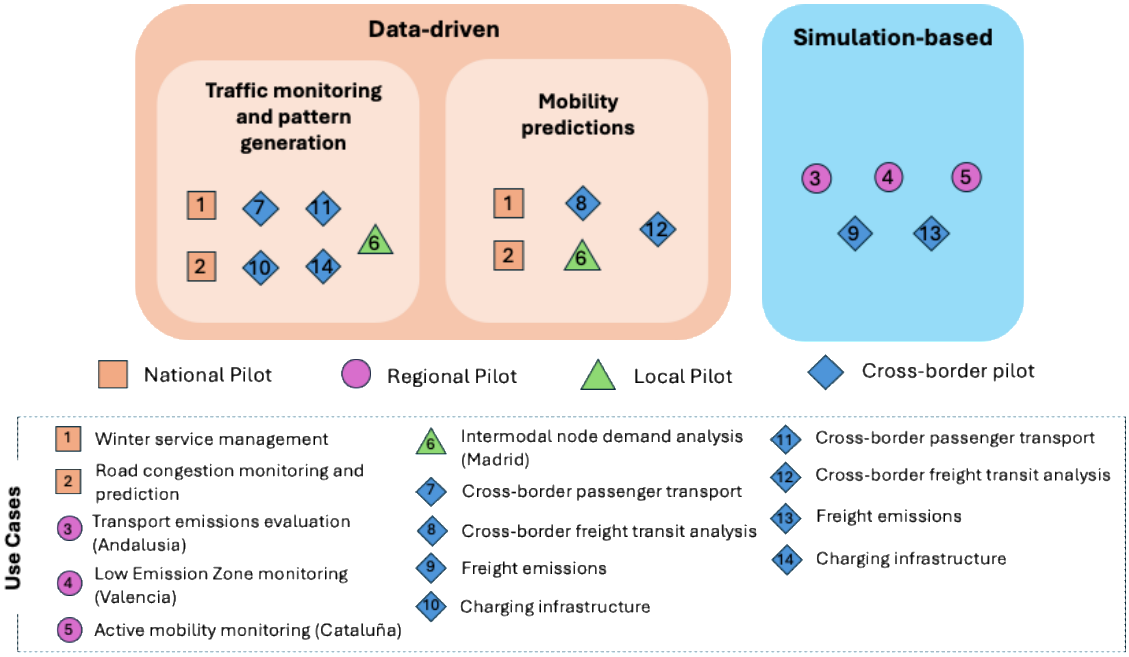


Figure 1. Tentative classification of Use Cases and pilots based on data-driven or simulation-based models

The first group corresponds to the **Use Cases that rely on data-driven models**, where the goal is to extract patterns, make predictions, or detect anomalies from historical or real-time datasets. Several examples of these Use Cases include: the anticipating of short-term congestion period with monitoring and prediction, supporting real-time decision-making for traffic management; the development of a machine learning-based prediction model capable of forecasting heavy vehicle flows during adverse weather conditions and informing users through Variable Message Signs (VMS).

The second group includes **Use Cases that require simulation-based models**, which are essential when the objective involves replicating complex systems and exploring hypothetical scenarios that cannot be directly inferred from existing data. As an example, some Use Cases are dealing with the reduction of emissions and for this the simulation-based models play an important role as it enables to simulate the impact of different alternative scenarios such as fleet renewal electrification, model shifts, implementation of low-emission zones, and reinforcement of public transport.

However, there are several Use Cases within the project for which the objective is broader and could potentially be achieved through both a data-driven approach or a simulation-based approach. **The**

choice between these two methodologies cannot be made in isolation. It will depend on critical factors such as the availability, quality and granularity of data, as well as the existence and maturity of simulation models. For instance, if comprehensive historical and/or real-time data are accessible and sufficient representative, a data-driven model may offer faster deployment. Alternatively, if there is an existing calibrated transport model that could be used within the model and it aligns with the objectives of the Use Case, it would be more convenient to use this model. At this stage, without detailed information on the datasets and models that will be available for integration, it is not possible to guarantee the feasibility of either approach. This uncertainty will be removed once datasets or models become available in the project. Only after this can a well-informed decision be made regarding the most suitable methodology for each Use Case.

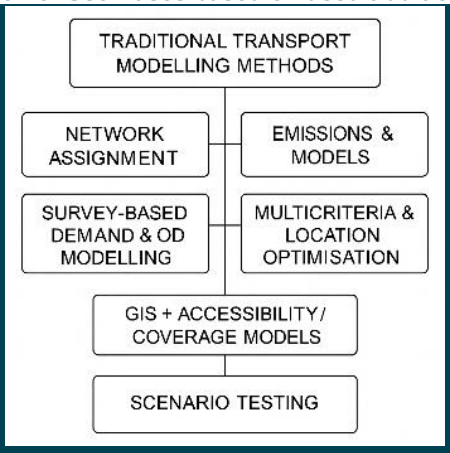
Mapping Use Cases to Traditional Transport models

Additionally, to ensure a robust and structured approach, the identified use cases have been mapped to traditional transport models or, where applicable, to potential models that could be employed for their use. Table 2 summarises these cases focussing on those involving traditional models. While some use cases (UC1, UC2, UC8, UC12, UC6) have been initially assigned to data-driven approaches in Figure 1, alternative simulation-based approaches have been assigned here – particularly in cases where data availability may be insufficient. The reverse traditional molde-based classification is shown in Table 3.

Table 2. Mapping of Use Cases to Traditional Transport Models

#	Use Case	Traditional Tentative Transport Models
UC3	Transport Emissions Evaluation (Andalusia)	Macroscopic network modelling if a VISUM model is provided (or any alternative transport simulation model); OD-matrix assignment; COPERT/HBEFA emission factor models; Scenario evaluation (fleet changes, modal shift).
UC4	Low Emission Zone Monitoring (Valencia)	Traffic assignment and speed–flow modelling (if a VISUM model is provided or any alternative transport simulation model); Modal share calculation from household surveys; Classical dispersion modelling; Before/after impact assessment.
UC5	Active Mobility Monitoring (Catalonia)	Modal share estimation; OD-based trip generation; Accessibility indicators; Cycling and pedestrian network coverage analysis; GIS-based mobility evaluation.
UC9 and UC13	Cross-Border Freight Emissions	Macroscopic assignment (VISUM); COPERT/HBEFA emission factor modelling; Speed-dependent emissions calculation; Classical environmental scenario testing.
UC1	Winter Service Management – Heavy Vehicle Parking (National)	Road capacity and LOS (level of service) modelling; Macroscopic assignment under operational constraints; Traditional winter-operations modelling frameworks.
UC2	Highway and Main Road Congestion Monitoring (National)	Traffic engineering methods (volume/capacity ratios, LOS, speed–flow models); Macroscopic assignment for corridor performance; Classical congestion diagnostics and monitoring.
UC6	Intermodal Node Demand & Location Analysis (Madrid)	Multicriteria Analysis (MCA); p-median, p-center and Location Set Covering Problem (LSCP) models; Accessibility modelling; OD-based demand estimation using classical survey data; GIS-based network analysis.
UC8 and UC12	Cross-Border Freight Transit Analysis (Spain–Portugal / Spain–France)	Macroscopic freight transport models; Freight OD matrices from survey sources; Scenario-based corridor modelling; Traditional freight assignment within national transport models.

Table 3. Classification of Use Cases based on used traditional modelling types



Use cases using network assignment: • Intermodal node demand & location (Madrid) • LEZ monitoring (Valencia) • Highway congestion modelling (National pilot) • Freight and transit analysis (Cross-border) • Winter heavy-vehicle management (National)	**Use cases using emission modelling:** • Transport emissions (Andalusia) • Freight emissions (Cross-border) • LEZ monitoring (Valencia) • Active mobility (Catalonia)	**Use cases using MCA & location models:** • Intermodal node demand (Madrid) • LEZ monitoring (Valencia) • Freight transport analysis (location of bottlenecks)
Use cases relying on survey-based demand & OD modelling: • Intermodal node demand (Madrid) • Transport emissions (Andalusia) • Active mobility monitoring (Catalonia) • Freight transit analysis (Cross-border)	**Use cases relying on GIS + accessibility / coverage models:** • Intermodal node analysis (accessibility to hubs) • Active mobility monitoring (cycling/pedestrian networks) • LEZ monitoring (urban indicators) • • Freight transit analysis (corridor accessibility)	**Use cases relying on scenario evaluation:** • Fleet renewal / electrification (emissions cases) • New infrastructure (cross-border freight) • Modal shift policies (LEZ, Andalusia emissions) • Operational scenarios (winter service management)

3 Methodologies for validation of simulation-based models

The **validation of transport simulation models** is a fundamental step in ensuring that model outputs represent real-world mobility systems with an acceptable degree of accuracy, reliability, and robustness. As mobility environments become increasingly complex—with multimodal interactions, temporal fluctuations in traffic demand, emerging micro-mobility patterns, and evolving regulatory frameworks—simulation tools at different levels of detail (micro, meso, macro or hybrid) have become indispensable for planning, evaluating policies, and supporting operational management.

Dynamic models (meso, micro, or hybrid) can reproduce time-dependent behaviours such as queue formation, shockwave propagation, route choice adjustments, and congestion patterns. While they offer more detail than static assignment models (**macroscopic models**), they also require richer data inputs and more rigorous calibration and validation procedures. Poorly validated models can produce misleading results, leading to ineffective policies or inappropriate operational recommendations.

The objective of this section is to provide all stakeholders involved in the commissioning, development, or evaluation of a transport modelling and simulation framework—as well as any other relevant actors—with a set of basic, clear, and readily applicable criteria for determining whether a simulation model, understood as a microscopic, mesoscopic, macroscopic or hybrid micro–meso network loading simulation, adequately represents real-world conditions. Compliance with these minimum criteria constitutes a necessary requirement to ensure that a model can be reliably employed to support subsequent decision-making processes.

The procedure of comparing simulation outputs with empirical observations is referred to as **model validation**. The following section presents a structured set of validation principles, criteria, and recommended procedures. They draw upon international best practices (FHWA (Validation And Reasonableness 2010 - Other Reports - Publications - Transportation Model Improvement Program (TMIP) - Planning - FHWA, no date) , UK WebTAG (*Transport analysis guidance*, 2025), European TEN-T guidelines (Trans-European Transport Network (TEN-T) - Mobility and Transport) and provides a comprehensive academic framework for building reliable transportation models based on a very rich literature (Toledo *et al.*, 2003; Hollander and Liu, 2008; Barcelo, 2010; European Commission. Joint Research Centre. Institute for Energy and Transport., 2014).

3.1 Definitions. Variables and Indicators

The most common variables—those consistently referenced in the existing literature—for use in the validation process are the following:

- Traffic counts (veh)
- Travel time (s)
- Travel time variability (s)
- Saturation flow (veh; equivalent vehicles or Passenger Car Units (PCU))
- Queue lengths (veh)
- Vehicle speeds (km/h)

With respect to the statistical indicators to be applied:

- GEH: This is one of the most widely used statistics in the validation of transport models. The GEH statistic is a formula for comparing two sets of traffic volumes and is named after Geoffrey E. Havers, who developed it in the 1970s while working as a transport planner at the Greater London Council. It is not a statistical test but an empirical formula similar to the Chi-square measure, designed to tolerate larger errors at low traffic flows. The rationale for introducing such

a statistic lies in the "inability of both absolute and relative differences to cope with a wide range of flow magnitudes."

$$GEH = \sqrt{\frac{(V_0 - V_m)^2}{0.5(V_0 + V_m)}}$$

Where:

V_0	Observed Flow, veh/h
V_m	Modelled flow, veh/h

- **Regression (R²):** Regression (R²) is a statistical measure of the correlation between the entire set of observed traffic counts and the predicted model volumes. Unlike the GEH statistic—which is applied to individual flows or screenlines—Regression (R²) is applied to the full comparison dataset and is expressed as a single value.

- **MAPE:** Absolute and relative differences between observed and simulated values (MAE and MAPE).

$$\text{MAPE} = \frac{100}{n}\sum\frac{|V_0 - V_m|}{V_0}$$

- **RMSE**: The Root Mean Square Error (RMSE) is a widely used metric for assessing the differences between values predicted by a model and the values actually observed (whether from a sample or a population). RMSE represents the square root of the second sample moment of the differences between predicted and observed values, or the quadratic mean of these differences. These deviations are referred to as residuals when computed using the dataset employed for estimation, and as prediction errors when computed using out-of-sample data. RMSE provides a single, aggregated measure of predictive accuracy by combining the magnitudes of the prediction errors across multiple observations.

  This form of error estimation has been widely used in traffic simulation. Although, in principle, a lower RMSE value indicates a better model, it presents a notable drawback: by squaring the errors, it disproportionately emphasises larger deviations.
- **Screenlines and cordons:** The comparison between simulated volumes and observed counts for critical links, particularly along screenlines, cutlines, and cordons, is highly useful for assessing model quality.
  - Screenlines extend fully across the modelled area from one boundary cordon to another. They are often associated with physical barriers such as rivers or railway lines, although jurisdictional boundaries crossing the study area can also serve as effective screenlines.
  - Cordons encompass a designated area in its entirety. For example, a cordon surrounding the central business district (CBD) is useful for validating the detailed patterns of traffic demand associated with the area. Overestimation or underestimation of trips destined for the CBD may indicate errors in socioeconomic data (e.g., employment data for the CBD) or inaccuracies in the trip distribution model or the mode choice model.

  It should be noted that screenlines and cordons are particularly useful in large-scale models and in models where demand is represented through OD matrices; they are not suitable for models in which demand is based on entry flows and turn percentages at intersections.

- **Space–time diagram:** A space–time diagram plots detector information across both space and time, where the spatial dimension corresponds to the distance along a roadway defined by an ordered set of detectors. Detector data are represented through colour gradients within the two-dimensional space–time plot.

- **Theil's U inequality coefficient:** The Theil U statistic (Theil, 1966) is a measure of relative accuracy that compares simulated results (V_m) with observed data (V_0). It is an association measure between two series, where a value of 0 indicates no difference between observed and simulated data, whereas a value of 1 indicates no relationship between them.

$$U = \frac{\sqrt{\frac{\sum(V_0 - V_m)^2}{n}}}{\sqrt{\frac{\sum(V_0)^2}{n}} + \sqrt{\frac{\sum(V_m)^2}{n}}}$$

3.2 Conceptual Basis of Validation

A distinction must be made between validation and calibration. **Calibration** refers to the adjustment of model parameters to reproduce observed traffic patterns such as vehicles flows, speed profiles, gap acceptance, reaction times, and lane-changing behaviour. **Validation**, in contrast, assesses the model's ability to reproduce independent data not used in calibration. It is therefore a measure of external reliability rather than internal parameter tuning.

Internal validity concerns the logical coherence of the model: vehicles behave plausibly, congestion emerges realistically across network segments, and no contradictory or impossible results appear. External validity involves comparing simulated outcomes with field observations across multiple data sources and time periods.

Successful validation requires representative, high-quality data. These data must be temporally aligned with the simulation period, spatially representative of the area of study, and diverse in nature—ideally combining counts, speeds, travel times, trajectories, and turning movement records. Critically, validation data should remain independent from calibration data to avoid circular reasoning.

3.3 Validation of Macroscopic Models

Validation of macroscopic transport models aims to confirm that a model correctly reproduces the base-year travel behaviour and network performance, and that it reacts plausibly to changes in input data. Although there is no universally accepted international standard for validation, the literature and professional guidelines converge on a logical, staged validation strategy.

Macroscopic models generally follow a four-step structure consisting of **trip generation, destination choice, mode choice, and assignment**. Validation must therefore be carried out systematically along these successive modelling components. Two complementary perspectives guide the process:

1. **Fit-to-observed conditions** in the base year (comparison with empirical data), and
2. **Realism of behavioural response** under changes in supply or cost (sensitivity tests).

A central feature of macroscopic model validation is the extensive use of empirical data from household travel surveys, roadside counts, commuter statistics, smartcard or ticketing data, and public transport load information. A sound validation exercise requires that these inputs are collected locally, because national averages often fail to represent local mobility structures or specific population segments. For instance, activity-chain frequencies, person-group distributions, trip lengths, and mode shares are directly calibrated and validated from a Household Interview Survey (HIS), without relying on national defaults.

3.3.1 Validation by model component

Trip generation

Trip generation is validated by comparing modelled trip totals per person group and per activity with those observed in the HIS. Because trip generation is derived directly from the combination of population

data and observed activity-chain frequencies, deviations should be minimal. In general, trip counts should match survey totals very closely, with only small tolerances for person groups and activities.

Destination choice

Validation of destination choice begins with aggregated comparisons such as commuting flows based on official employment statistics or household survey OD matrices. Analysts check whether the model reproduces the correct trip distribution patterns across zones. A key metric is trip length: both average trip lengths and full trip-length distributions must be validated.

The coincidence ratio (CR) is a distribution-similarity measure commonly used in the validation of regional macroscopic transport models to assess the degree of agreement between observed and simulated trip-length histograms. The indicator is computed as the sum of the minimum relative frequencies across corresponding histogram bins of the observed and modelled distributions, yielding a value between 0 and 1 that expresses the proportion of overlap between them. A CR close to 1 denotes a very high level of similarity, indicating that the model accurately reproduces not only the central tendency but also the shape of the empirical trip-length distribution across activities and population groups. Values above 0.80 are typically interpreted as evidence of satisfactory model performance, whereas substantially lower values suggest deficiencies in the representation of user behaviour or network conditions. Given its simplicity, robustness, and interpretability, the coincidence ratio provides a useful and transparent metric within calibration and validation processes. Poor fit typically indicates issues in land-use inputs or destination-choice parameters and may require adjustments such as OD-specific correction factors (friction factor values), applied only after reviewing underlying land-use assumptions.

Mode choice

Mode choice validation compares the modelled mode shares with those derived from the HIS, both at aggregate and disaggregated levels. Analysts inspect mode shares by person group, by zone, and by distance class. A visual comparison of mode share curves over trip length is recommended to verify whether the model reproduces realistic cut-off distances, substitution patterns, and sensitivities.

Assignment

Assignment validation relies heavily on count data. In the case of private transport, traffic volumes from the assigned network are compared with roadside counts, typically using **GEH statistics**, correlation coefficients, and screenline comparisons. Before considering any matrix correction techniques, modellers must ensure acceptable results in all upstream stages. Count deviations may reveal incorrect land-use data, incorrect route choice due to network errors, or unreliable count data rather than demand-modelling issues.

Public transport assignment is validated using observed passenger kilometres per line and boarding/alighting counts at key stops. Because public transport supply is complex, detailed validation at the line and station level is recommended.

3.3.2 Validation criteria and KPIs

The existing guidelines such as the UK WebTAG (*Transport analysis guidance,* 2025) and the US FHWA (*Travel Model Validation and Reasonableness Checking Manual - Second Edition,* 2010), KPIs are summarized in Table 4. These documents prescribe validation procedures and sensitivity tests, but their specific thresholds are often tailored to incremental models constructed from observed base-year matrices. For pure synthetic models, built from population, land-use, and independent survey data, threshold values cannot be transferred directly, but the methodological structure remains useful.

Table 4. Recommended Validation Criteria for Macroscopic Transport Models

Model Stage	Validation Criterion	Recommended Threshold / Target	Notes
Trip Generation	Total trips vs HIS	< 0.1% deviation	Should match almost exactly.
Trip Generation	Trips per person group	< 1% deviation	Check segmentation realism.
Trip Generation	Activity shares	< 1% (except home)	Home up to 5%.
Trip Distribution	OD flows vs HIS (total)	Correlation > 98%	Depends on zone size.
Trip Distribution	OD flows car/PT	Correlation > 90%	—
Trip Distribution	OD flows walking/bike	Correlation > 80%	More variability expected.
Trip Distribution	Commuting OD vs employment	Correlation > 95%	Requires good census data.
Trip Distribution	Avg trip length	< 2% deviation	Key consistency indicator.
Trip Distribution	Trip length by activity	< 5% or < 300m	—
Trip Distribution	Length distribution	Coincidence ratio > 0.7	Distribution similarity.
Mode Choice	Mode split total	< 0.3% deviation	Calibrated to HIS.
Mode Choice	Mode split major groups	< 2.5% deviation	Modes ≥10% share.
Mode Choice	Mode split minor groups	< 1% deviation	Modes <10% share.
Mode Choice	Mode share by distance	Visual consistency	Check behavioural realism.
Assignment (PrT)	Convergence	Relative gap < 1e-5	For final assignment.
Assignment (PrT)	Screenlines (pre-correction)	< 10% deviation	24h totals.
Assignment (PrT)	Mean GEH (pre-correction)	< 12.5	Strategic models.
Assignment (PrT)	GEH < 10 share	> 45%	Depends on count quality.
Assignment (PrT)	Correlation (pre-correction)	> 97%	High correlation expected.
Assignment (PrT)	Screenlines (post-correction)	< 5% deviation	Tighter threshold.
Assignment (PrT)	Mean GEH (post-correction)	< 6	Standard acceptance.
Assignment (PrT)	GEH < 10 share (post)	> 80%	—
Assignment (PrT)	OD matrix change	< ±3%	Preserve structure.
Assignment (PrT)	Avg trip length change	< 0.5%	Stability requirement.
Assignment (PuT)	Passenger-km by system	< ±5% deviation	By mode: rail, LRT, bus.
Assignment (PuT)	PT lines within ±10%	> 90%	Line-level realism.
Assignment (PuT)	PT GEH	Mean GEH < 10	Higher tolerance for bus.

Schlaich et al. (Schlaich, Heidl and Li, 2017) propose a set of indicative **validation criteria for regional macroscopic models**, covering trip generation, trip distribution, modal split, and assignment. These include threshold values such as correlation levels for OD flows, coincidence ratios for trip-length

distributions, maximum deviations in modal shares, GEH thresholds for traffic counts, and limits on changes introduced by matrix corrections. The authors emphasise that these thresholds are not derived from theoretical principles but serve as target quality levels to guide calibration and validation processes.

3.3.3 Concluding remarks

The importance of defining validation criteria at the tender stage of a modelling project is stressed by some authors (Barceló, 2015). This provides clarity for both the client and the consultant, ensuring that the expected model quality is contractually defined and that validation results can be objectively evaluated. Because macroscopic models are complex and sometimes rely on imperfect data, meeting all validation criteria simultaneously is difficult. Therefore, validation must be interpreted holistically, considering the purpose of the model, data availability, model resolution, and broader plausibility.

The validation of macroscopic transport models requires a progressive, staged approach aligned with the structure of the model. Good validation practice combines comparisons to empirical data across all model components with behavioural realism checks based on sensitivity analyses. While there are no universally accepted thresholds, a combination of GEH-based traffic-count validation, trip-length and modal-split comparisons, OD-correlation measures, and clearly defined acceptance criteria can ensure that the model is fit for purpose. Macroscopic models are essential tools for strategic planning, and proper validation ensures that they provide credible and defensible outputs for policy-making and infrastructure investment.

3.4 Recommended validation workflow for dynamic models

The **validation process** begins with a clear definition of the **model's purpose**. A model developed to evaluate infrastructure design has different requirements from one intended to analyse traffic management strategies, assess ITS interventions, or examine public transport performance. The intended use determines the level of detail, performance indicators, and acceptance thresholds.

Data selection is the next key step. Traffic counters, automatic plate recognition, floating-car data, Bluetooth and Wi-Fi sensors, public transport logs, and video-extracted trajectories are all relevant sources. Use of at least two independent data types strengthens the credibility of the validation, as mentioned before in the macroscopic context.

Validation should be applied to a representative **segmentation of the network**, focusing on high-flow corridors, bottlenecks, major intersections, public transport corridors, and areas where policies are expected to have significant impact. The **temporal dimension** is equally important: models should be validated for the AM peak, PM peak, inter-peak periods, and—when applicable—special events or atypical conditions.

A **multi-indicator approach** is essential. Flow, speed, travel time, density, queue length, turning movements, lane use, and dwell times provide complementary perspectives that together offer a robust validation picture.

Flow validation is commonly performed using the GEH statistic, percent error in volume, and RMSE. A GEH value below 5 is generally acceptable for most links, while values under 4 are considered very good. Values up to 7 may be tolerated on high-volume corridors. Percent error in volume should typically remain below 15% for individual links and below 10% for the network average.

Speed and travel time validation uses measures such as mean absolute percentage error, maximum segment error, and correlation coefficients. As a general reference, a MAPE below 15%, maximum error below 20–25%, and a correlation above 0.85 indicate good agreement between modelled and observed travel times.

Queues are particularly important because simulation models are uniquely capable of representing congestion dynamics. Queue length deviations should remain within ±20% of observed values, and the temporal evolution of queues—build-up, peak duration, and dissipation—should match empirical measurements. Delay at intersections should generally remain within ±15–20%.

Turning movement counts represent another important validation dimension. A commonly accepted threshold is that at least 85% of movements should exhibit a percent error below 20%.

Quantitative metrics must be complemented by qualitative assessments of behavioural plausibility. Simulated drivers should accelerate, decelerate, and change lanes in realistic ways. Merging and weaving should reflect observed patterns, and vehicles should comply with signal control and road regulations.

Spatial and temporal logic is essential: queues should appear in known bottlenecks, not arbitrary locations; shockwaves should propagate in a plausible manner; and platoons emerging from traffic signals should disperse realistically along downstream links.

The model should also reflect traffic management logic faithfully. Signal plans, timing changes, ramp metering logic, variable speed limits, and public transport priority mechanisms must be accurately represented and verified.

3.4.1 Summary of Acceptance Criteria

A simulation model is often considered valid when a set of combined thresholds is met. Typical acceptance conditions include:

- at least 85% of validated links achieve GEH < 5;
- at least 90% of travel time segments meet MAPE < 15%;
- queue formation, duration, and dissipation closely match observed patterns;
- between 80–85% of turning movements fall within accepted thresholds;
- and no systematic bias is detected across corridors or time periods.

3.5 Validation criteria comparison macroscopic vs dynamic models

The **macroscopic model** focuses on reproducing aggregate traffic conditions and meeting WebTAG convergence standards. Its validation relies heavily on achieving high percentages of links and turns within absolute/relative error thresholds and demonstrating strong regression fit (slope 0.9–1.1; R > 0.95). By contrast, **dynamic models** must replicate more detailed traffic dynamics—particularly turning behaviours and queue propagation—making validation inherently more demanding. Although both models apply identical flow-validation thresholds and GEH criteria, the mesoscopic model typically achieves slightly lower percentages of compliant sections due to its finer resolution. Journey-time validation plays a more central role in mesoscopic verification, while matrix adjustments may exceed standard WebTAG distortion limits because of the need to accurately reproduce turn-level flows. Overall, macroscopic validation guarantees strategic network realism, whereas dynamic validation ensures detailed operational fidelity suitable for scheme-level analyses (see Table 5).

Table 5. Comparative Table: Mesoscopic vs Macroscopic Validation Criteria

Aspect	Mesoscopic Model	Macroscopic Model
Purpose of Validation	Accurate dynamic assignment, queue formation/dissipation, detailed turning behaviour.	Ensure correct equilibrium assignment, network-wide flows, aggregate congestion patterns.
Traffic Flow Criteria	Absolute/relative thresholds (±100 veh/h, 15%, ±400 veh/h) or GEH < 5; more demanding due to turning flows.	Same thresholds; typically, higher compliance rates.

Typical % Meeting Criteria	86.5–95% depending on peak and direction.	90–97.5% across peaks.
Regression Analysis	Slope 0.9–1.1, R > 0.95; slightly more variability.	Same criteria; very close fit due to aggregation.
Journey Time Validation	Essential for corridor-level accuracy.	Supplementary; less sensitive than meso.
Matrix Validation	Screenline/cordon validation; may exceed WebTAG distortion limits.	WebTAG criteria applied but sometimes not met due to meso–micro requirements.
Convergence Requirements	Requires stable flows; dynamic assignment.	Strict WebTAG convergence (relative gap 0.1%).
Sensitivity to Coding	Highly sensitive to local coding errors.	Less sensitive due to aggregation.
Fitness for Purpose	Good for junction/corridor assessment; may miss very local behaviours.	Good for network-level assessment and demand modelling.

3.5.1 Concluding remarks

Validation should be accompanied by **sensitivity tests** to evaluate how the model responds to changes in key parameters. Analyst discretion must extend to examining behavioural parameters, demand variation, route choice dispersion, signal timing changes, and the influence of stochasticity.

Running the model with multiple **simulation seeds** allows analysts to confirm that results are not dependent on a single random sequence. A model that produces realistic results only for a single seed or parameter configuration lacks robustness.

A robust validation strategy includes **testing the model against data not used in calibration** or initial validation. An effective approach is to calibrate the model using data from one day or period and validate it using data from another—for example, calibrating on a Tuesday and validating on a Thursday. Consistent results across days demonstrate the model's generalisability and resilience.

Even after a base version of the model is validated, it must be checked under different scenarios. Infrastructure changes, modifications to demand patterns, policy interventions, and operational strategies may introduce dynamics that differ from the base scenario. Scenario validation ensures that the model behaves plausibly under altered conditions and that its internal logic extends beyond the calibration period.

A validated model must be **fully documented**. Essential components include assumptions, parameter values, data sources and timestamps, calibration and validation processes, visual comparisons between simulated and observed data, performance indicator tables, and an explicit description of model limitations. The overarching objective is to guarantee reproducibility: a competent analyst should be able to replicate model results using the provided documentation.

The validation of transport simulation models is a methodological process that ensures credibility, accuracy, and analytical value. Properly validated models support infrastructure planning, policy evaluation, and operational decision-making. The recommendations presented in this document establish a rigorous framework for ensuring that simulation models remain transparent, empirically grounded, and methodologically sound—thereby providing trustworthy insights for decision-makers and practitioners.

4 Methodologies for validation of data-driven models

The availability of large, diverse datasets and scalable processing technologies has made data-driven models increasingly important (Wei, 2014). They adapt to dynamic traffic states – such as flow, density and speed – making them effective for managing congestion, incidents and real-time routing. These models also support strategic planning by enabling accurate predictions of travel speed, travel times, and traffic volumes, which are essential for optimizing transportation systems and infrastructure development (Fredriksson, 2024).

Results from these models may influence decisions with **significant consequences**. Therefore, it is essential to ensure that these results – which can be predictions for instance – remain reliable beyond the controlled environment of training data (Baladram, 2024; Michelucci, 2024). This is what is known as data-driven **model validation** and is used to verify that the model performs as intended, especially on unseen data. In essence, the validation is the process of confirming that a model fulfils its intended purpose and achieves its expected effectiveness.

In this section, we will first summarise the classification of the data-driven solutions that could be applied to the different Use Cases. Then, we will introduce different methods for validating data-driven models.

4.1 Data-driven solutions

Data-driven solutions focus on leveraging large volumes of historical and real-time data to extract insights, predict future states, and support operational decision-making. These solutions combine advance analytics, machine learning techniques, and streaming data processing. An initial introduction to these data-driven approaches was provided in Deliverable D4.1. In the following paragraphs, a summary of these solutions is presented to establish the context for the validation methods that will be applied in subsequent sections.

Raw data analyser and cleaner algorithm

This component ensures that incoming data streams and historical datasets are properly validated, cleaned and standardized before used for analysis or modelling. The first step is called raw data analyser and aims to perform an initial scan of time-series transport data, such as traffic flow, travel demand or speed. The algorithm also assesses the reliability of each data point and of each detector. Once the raw data has been assessed, the next step involves applying a data cleaning algorithm. It addresses issues such as unrealistic, erroneous values or outliers, and missing data.

Traffic monitoring

This functionality continuously processes live traffic data from sensors, GPS devices, and other sources to identify current traffic conditions. The system provides real-time monitoring of traffic conditions, enable operators to detect congestion, measure flow rates and identify critical points in the network.

Pattern generation

This algorithm involves analysing historical traffic data to identify recurring behaviours and typical daily profiles by grouping similar days together. This is a clustering-based process that consists of the extraction of daily patterns, explanatory variables and training the patter matcher to predict a pattern that is assigned to each day. These patterns are essential for understanding historical traffic dynamics and serve as input for predictive models.

Predictions

Predictive models use machine learning and statistical techniques to forecast short-term and medium-term traffic conditions – ranging from the next 15 to 60 minutes or even the next few hours. The statistical

approaches for traffic forecasting includes: the autoregressive integrated moving average (ARIMA) (Hamed, Al-Masaeid and Said, 1995), regression method (Smith, Williams and Keith Oswald, 2002), Bayesian model (Wang, Deng and Guo, 2014), hidden Markov model (Qi and Ishak, 2014), Kalmar filter model (Guo, Huang and Williams, 2014). Among the machine learning methods, Support Vector machines, k-Nearest Neighbours, Random Forest or Deep Neural Networks (DNNs) are example of the most used ML methods applied to traffic forecasting (Lana *et al.,* 2018).

Incident detection method

This AI-driven system designed to detect traffic incidents across the network and automatically notify operators of disruptions on the road. By analysing real-time traffic data, it identifies significant deviations from normal flow patterns and/or drops to zero between two detector points. When such anomalies occur, the system generates an alert, enabling operators to focus on areas that required immediate attention and intervention. This solution enhances situational awareness and reduces cognitive load for operator, ultimately increasing operational efficiency.

4.2 Model validation methods for data-driven models

There are multiple **approaches** to validate a data-driven model, and there is no universal approach that works for every situation (Michelucci, 2024). For instance, a validation method may fit a problem when data is scarce, and other methods may fit other scenarios with large datasets. This highlights that validation strategies are always **context-dependent** and must account for multiple factors. In fact, each method has unique strengths designed to address specific challenges. Importantly, different validation techniques can produce different results, which means that selecting the appropriate method is critical for obtaining an accurate assessment of model performance (Baladram, 2024).

Several factors influence this choice:

- Data characteristics: this plays a major role as some approaches are better suited for large and diverse datasets, while others are designed to handle smaller or imbalanced samples.
- Model complexity is another important consideration as more sophisticated models often required advanced validation strategies, whereas simpler models can be evaluated with less demanding approaches.
- Application context: for instance, when dealing with time-dependent data, it is essential to use validation strategies that respect the temporal sequence to avoid introducing bias or unrealistic assumptions.

Using an inappropriate validation method can lead to misleading conclusions about a model's performance.

The following Figure 2 shows a simple classification of different methods for the validation of data-driven models. Each method will be briefly described to make the reader a clear understanding of the different validation methods.

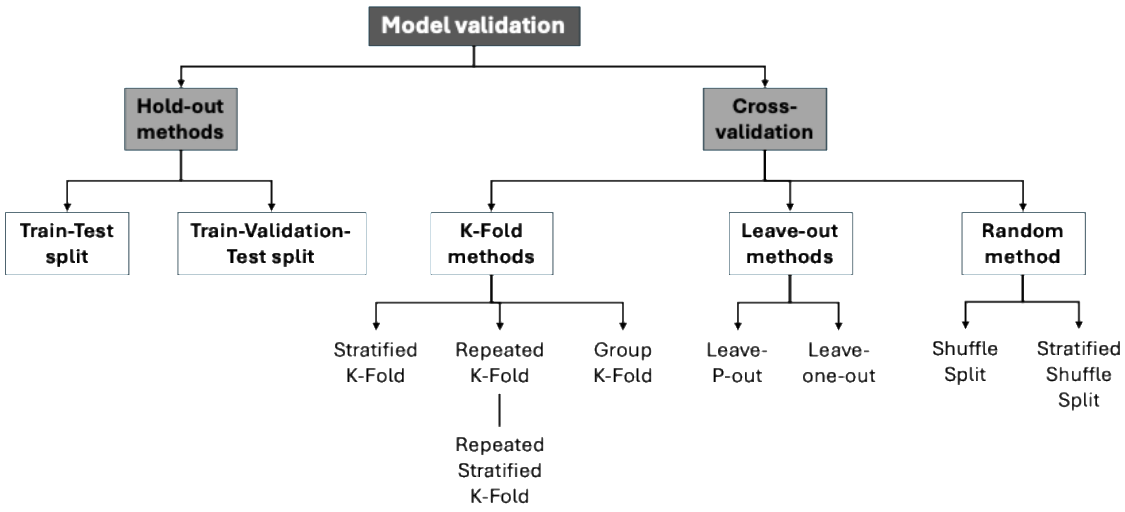


Figure 2. Classification of model validation methods. Adapted from (Baladram, 2024)

4.2.1 Hold-out validation method

To validate a model is necessary to use unseen data, data that has not been used during the training process. In a real-life context, the generation of new unseen data may not be possible and therefore, the easiest way to find unseen data is to take a portion of the available data and put it aside before training a model. In this sense, we will have less data for training, but we will be able to validate our model (Michelucci, 2024).

Train-Test Split: is a model validation technique that divides a dataset into a training set, used to train the model, and a testing set, which contains unseen data to evaluate performance and accuracy. Both the training and test dataset size depends on the size of the dataset. Typically, 70-90% of the data is used for training, while the remaining 30-10% is reserved for testing, respectively (Berrar, 2019). However, this method has some limitations. One of them is that results may change a lot depending on how we randomly split the data. And, if there is not much data available, it might be not enough to properly train or test the model.

Train-Validation-Test Split: this method splits the data into three parts. The first part is called training set and is used to learn patterns, relationships and structures from this dataset. The middle part, called validation data, is being used to tune the parameters of the model. The third part, called test set, is used to evaluate the final model's performance on completely unseen data (Mistry, 2025).

To sum up, hold-out methods perform best when a large dataset is available. However, with smaller datasets, this approach may not provide reliable results. In such cases, more sophisticated validation techniques are needed to accurately assess model performance (Baladram, 2024).

4.2.2 Cross-validation method

Cross-validation fundamentally changes the way we evaluate models. Instead of relying on a single train-test split, it repeatedly tests the model using multiple different splits of the same dataset. This approach provides a much more reliable picture of how the model performs on unseen data.

The core principle is simple: the model is trained and tested several times, and each time the training and testing sets come from different portions of the data. By doing this, we reduce the risk of bias caused by one particularly favourable or unfavourable split. A single split might make the model look excellent or poor by change, but cross-validation averages performance across many splits, giving a more balanced and accurate assessment (Baladram, 2024).

There are plenty of cross-validation techniques that may be used to cross-validate a model:

K-Fold Methods

In k-fold cross-validation, data points are randomly assigned to k disjoint groups. Instead of splitting the data once, k-fold splits the data into k equal parts. Each time one of these k groups is selected as the validation set, the remaining k-1 groups are combined and used as the training set. This process is iterated k times so that each group is selected once as the validation set. This method, compared to the hold-out methods, is more computationally demanding because it requires training and evaluating the model k times. However, the resulting estimate is more reliable (Maleki *et al.*, 2020) as we are averaging performance from all k tests. The 10-fold and 5-fold cross-validation are the most widely used k-fold cross-validation methods for evaluating Machine-Learning models (Friedman, Hastie and Tibshirani, 2001).

Stratified k-fold cross-validation

Data imbalance occurs when there is a substantial difference between the number of samples for the majority class and the minority class. In such cases, k-fold methods might lead to unstable performance measures (Maleki *et al.*, 2020). In this case, the stratified k-fold method ensures that each split has the same mix as the original data. This makes the testing more reliable, especially when some types of data are much rarer than others (Baladram, 2024).

Repeated k-fold cross-validation

In this case, this method involves running the regular k-fold process but multiple time with different random splits. This helps reduce the impact of randomness in the results and get a more reliable estimate of performance. However, it might take more time as you need to run many more evaluations (Baladram, 2024).

Repeated stratified k-fold cross-validation

A slight variation of the previous methods combines the stratification and repetition. It keeps the right mix of different types of data while testing many times (Baladram, 2024). However, every time you repeat the whole cross-validation process, the training time is multiplied.

Group k-fold cross-validation

This method ensures that data from different groups are not mixed between the training and testing sets. This is important when the data points are not independent and come from groups as data from the same group can appear in both the training and testing sets (Indrasiri, Pubudu, 2024).

Time Series Split cross-validation

This method is a variation of K-Fold designed for data that depends on time. Regular K-Fold assumes that data points are independent and splits them randomly. Time series split keeps the chronological order: training uses earlier data, and testing uses later data. This mirror real-world forecasting where we predict the future based on the past (Baladram, 2024).

Leave-Out methods

Leave-one-out cross-validation

This method is simply a k-fold cross validation but for k being the size of the dataset (Michelucci, 2024). This means that the process is repeated until every observation has served as the test set once. For instance, if we have 100 days of data, this method will run 100 separate training and testing cycles. Each cycle trains the model on 99 days and tests it on the remaining day. Because every data point is tested, the results are completely deterministic (Baladram, 2024) – running multiple times on the same dataset will always produce identical outcomes. However, for large datasets or complex models, the model must be trained N times so this can be impractical (Maleki *et al.*, 2020).

Leave-P-out cross-validation

This is an extended form of the Leave-one-out method, where the validation sets can have p elements instead of 1. The model is trained on (n-p) data points and tested on p data points, therefore it considers every possible validation subset of size p. For small datasets, for instance a value of p = 2, also known as leave-pair-out cross-validation, is used to achieve a robust estimate of the model performance (Maleki *et al.*, 2020).

Random methods

Shuffle Split cross-validation

This method takes a different approach compared to other validation techniques by generating completely random train-test splits. Unlike structured methods such as k-fold, which divide data into fixed folds, or exhaustive methods like leave-p-out, shuffle split created new random partitions each time it runs (Risk and James, 2022; Baladram, 2024).

The key distinction from k-fold is that Shuffle Split does not guarantee that every data point is used exactly once for testing. This method is particularly useful for large datasets where k-fold can be computationally expensive. It allows you to specify the number of iterations regardless of dataset size and control the proportion of data in each split (Baladram, 2024).

Stratified Shuffle Split cross-validation

This method combines random splitting with keeping the right mix of different types of data. This preserves the original class distribution in every split, ensuring that each fold is representative of the overall dataset (Ünalan *et al.*, 2024). The process begins by randomly shuffling the data while maintaining class proportions. Then, the dataset is divided into a specified number of splits or iterations. In each iteration, one split serves as the test set while the remaining data is used for training. This is repeated so that every split acts as the test set exactly once (Ünalan *et al.*, 2024).

4.3 Selection of the data-driven model validation method for each Use Case

The choice of the right validation method is critical, as it influences both the reliability of the models and the confidence that stakeholders place in its predictions. However, at this stage of the project, while the use cases have been clearly defined, the characteristics and availability of the data remain unknown. This uncertainty prevents us from committing to a specific validation strategy. Instead, we outline a flexible and systematic approach that will guide the selection process once the data becomes available.

The **main challenge** at the current stage is the lack of information about the datasets. Several factors influence the choice of validation strategy, including the size of the dataset, its quality, the distribution of classes, and the presence of temporal or spatial dependencies. For example, large datasets allow for more complex validation strategies, while small datasets require careful handling to avoid data leakage or overfitting. Similarly, missing values, noise and imbalanced classes can significantly affect the reliability of certain validation approaches. The structure of the data also matters: time-series data requires methods that respect temporal order, whereas tabular data can often be split randomly. Finally, the availability of labels determines whether we are dealing with supervised, semi-supervised, or unsupervised learning, each of which has different validation requirements. Since these aspects are currently unknown, the decision cannot be finalised now. Instead, we propose a **framework** that adapts to the data once it becomes available.

The **selection process** will begin with an assessment of the data characteristics. This includes understanding the size of the dataset, its distributions, and the presence of any dependencies or anomalies. Once this profiling is complete, we will match these characteristics to the requirements of each use case. After this alignment, we will identify candidate validation methods and evaluate the

against practical constraints such as computational resources and project timelines. This ensure that the chosen approach also considers its feasibility within the context of the project.

Once the data becomes available, the **choice of validation method** will be guided by several criteria. Dataset size is a primary consideration: large datasets often favour hold-out or k-fold validation, while small datasets may require leave-one-out, for instance. The type of data also matters: time-series data needs methods that respect temporal order, while grouped data may require grouped k-fold approaches. Class imbalance calls for stratified methos to ensure fair representation across folds. In addition, some of these methods are significantly more computationally expensive than others.

Rather than committing to a single method, we will maintain flexibility and adopt an adaptative approach. The validation strategy for data-driven models will be finalised after data analysis, which reveal patterns, dependencies and anomalies that influence the choice of the method.

5 Conclusions

This deliverable provides a comprehensive review of validation methodologies for data-driven and simulation-based models. It also refines the initial classification of use cases to modelling approaches initially presented in Deliverable D4.1 “Transport model design” and complemented from Deliverable D2.3 “Final definition of selected use cases”. However, as several use cases can be addressed using either a data-driven or a simulation-based approach, the choice depends on factors such as data availability, quality and the maturity of existing models. Until detailed information on datasets and models becomes available, only a preliminary methodology has been assigned to each use case, and it suspected to change once it is clear the available datasets to use in the project.

Nevertheless, this deliverable includes a theoretical review of validation methodologies to ensure that each method can be applied according to the modelling approach selected for the use cases. The validation of simulation-based models requires a staged and holistic approach that combines empirical data comparisons, behavioural realism checks and sensitivity analysis. While no universal thresholds exist, using measures such as GEH statistics, trip-length and model-split comparisons, and OD-correlation ensures models are fit for purpose and provide credible outputs for strategic planning and policy decisions.

Conversely, data-driven models are increasingly important for operational management and real-time decision-making. Their validation must go beyond training datasets to guarantee robustness and generalizability under dynamic conditions. Techniques such as hold-out testing and cross-validation – including k-fold, leave out and random sampling – are widely used, but the choice of method must be context-dependent, considering data availability, variability and intended application.

To sum up, there is no universal validation method applicable to all cases. Both modelling paradigms require flexible, well-defined workflows tailored to the specific characteristics of each use case. This deliverable provides a solid foundation for applying appropriate validation methodologies once the final selection of use cases and modelling approaches is confirmed.

6 References

- Baladram, S. (2024) “Model validation techniques, explained: a visual guide with code examples.” Available at: https://towardsdatascience.com/model-validation-techniques-explained-a-visual-guide-with-code-examples-eb13bbdc8f88/ (Accessed: November 12, 2025).
- Barcelo, J. (2010) *Fundamentals of Traffic Simulation*, *Simulation*. Available at: https://doi.org/10.1007/978-1-4419-6142-6.
- Barceló, J. (2015) “Analytics and the art of modeling,” *International Transactions in Operational Research*, 22(3), pp. 429–471. Available at: https://doi.org/10.1111/itor.12165.
- Berrar, D. (2019) “Cross-Validation,” in *Encyclopedia of Bioinformatics and Computational Biology*. (1, 1). Available at: https://www.researchgate.net/profile/Daniel-Berrar/publication/381674773_Cross-Validation/links/6791b19e82501639f5063948/Cross-Validation.pdf (Accessed: November 14, 2025).
- European Commission. Joint Research Centre. Institute for Energy and Transport. (2014) *Traffic simulation: case for guidelines.* LU: Publications Office. Available at: https://data.europa.eu/doi/10.2788/11382 (Accessed: November 23, 2025).
- Frediksson, H. (2024) *Data-driven modeling of Transportation Systems. Methodological approaches and real-world applications*. Doctoral dissertation. Blekinge Institute of Technology. Available at: https://www.diva-portal.org/smash/get/diva2:1896822/FULLTEXT01.pdf (Accessed: November 12, 2025).
- Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The elements of statistical learning*. New York: Springer Series in Statistics.
- Guo, J., Huang, W. and Williams, B.M. (2014) “Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification,” *Transp Res Part C Emerg Technol*, 43, pp. 50–64.
- Hamed, M.H., Al-Masaeid, H.R. and Said, Z.M.B. (1995) “Short-Term Prediction of Traffic Volume in Urban Arterials,” *Journal of Transportation Engineering*, 121, pp. 249–254.
- Hollander, Y. and Liu, R. (2008) “The principles of calibrating traffic microsimulation models,” *Transportation*, 35(3), pp. 347–362. Available at: https://doi.org/10.1007/s11116-007-9156-2.
- Indrasiri, Pubudu (2024) “Why Group K-Fold Cross-Validation?” Available at: https://www.linkedin.com/pulse/why-group-k-fold-cross-validation-pubudu-indrasiri-tpv8e/ (Accessed: November 17, 2025).
- Lana, I. *et al.* (2018) “Road Traffic Forecasting: Recent Advances and New Challenges,” *IEEE Intelligent Transportation Systems Magazine*, 10(2), pp. 93–109. Available at: https://doi.org/10.1109/MITS.2018.2806634.
- Maleki, F. *et al.* (2020) “Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment,” 30, pp. 433–445. Available at: https://doi.org/10.1016/j.nic.2020.08.004.
- Michelucci, U. (2024) “Model Validation and Selection,” in *Fundamental Methematical Concepts for Machine Learning in Science*. Springer. Available at: https://doi.org/10.1007/978-3-031-56431-4_7 (Accessed: November 11, 2025).
- Mistry, R. (2025) “Data Splitting (Train-Test-Validation) in Machine Learning.” Available at: https://medium.com/@rohanmistry231/data-splitting-train-test-validation-in-machine-learning-2d5d1927fa69 (Accessed: November 13, 2025).
- Qi, Y. and Ishak, S. (2014) “A Hidden Markov Model for short term prediction of traffic conditions on freeways,” *Transp Res Part C Emerg Technol*, 43, pp. 95–111.
- Risk, C. and James, P.M.A. (2022) “Optimal cross-validation strategies for selection of spatial interpolation models for the Canadian forest fire weather index system,” *Earth and Space Science*, 9. Available at: https://doi. org/10.1029/2021EA002019.

- Schlaich, J., Heidl, U. and Li, H. (2017) “Multimodal macroscopic transport modelling: State of the Art with a focus on validation & approval,” in *World Conference on Transport Research - WCTR 2016 Shanghai. 10-15 July 2016*.
- Smith, B.L., Williams, B.M. and Keith Oswald, R. (2002) “Comparison of parametric and nonparametric models for traffic flow forecasting,” *Transp. Re. Part C Emerg. Technologies*, 10(4), pp. 303–321. Available at: https://doi.org/doi: 10.1016/S0968-090X(02)00009-8.
- Theil, H. (1966) *Applied economic forecasting*. Amsterdam : North-Holland ; [Chicago : Rand McNally]. Available at: http://archive.org/details/appliedeconomicf0000thei (Accessed: November 23, 2025).
- Toledo, T. *et al.* (2003) “Calibration and Validation of Microscopic Traffic Simulation Tools: Stockholm Case Study,” *Transportation Research Record: Journal of the Transportation Research Board*, 1831(1), pp. 65–75. Available at: https://doi.org/10.3141/1831-08.
- *Trans-European Transport Network (TEN-T) - Mobility and Transport* (no date). Available at: https://transport.ec.europa.eu/transport-themes/infrastructure-and-investment/trans-european-transport-network-ten-t_en (Accessed: November 16, 2025).
- *Transport analysis guidance* (2025) *GOV.UK*. Available at: https://www.gov.uk/guidance/transport-analysis-guidance-tag (Accessed: November 16, 2025).
- *Travel Model Validation and Reasonableness Checking Manual - Second Edition* (2010). Available at: https://rosap.ntl.bts.gov (Accessed: November 23, 2025).
- Ünalan, S. *et al.* (2024) “A comparative study on breast cancer classification with stratified shuffle split and K-fold cross validation via ensembled machine learning,” *Journal of Radiation Research and Applied Sciences*, 17.
- *Validation And Reasonableness 2010 - Other Reports - Publications - Transportation Model Improvement Program (TMIP) - Planning - FHWA* (no date). Available at: https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/validation_and_reasonableness_2010/ (Accessed: November 16, 2025).
- Wang, J., Deng, W. and Guo, Y. (2014) “New bayesian combination method for short-term traffic flow forecasting,” *Transp Res Part C Emerg Technol*, 43, pp. 79–94. Available at: https://doi.org/doi: 10.1016/j.trc.2014.02.005.
- Wei, D. (2014) *Data-driven modeling and transportation data analytics*. PhD dissertation. Texas Tech University. Available at: https://ttu-ir.tdl.org/server/api/core/bitstreams/509fe97f-8deb-481e-a603-cb0d9e192997/content (Accessed: November 12, 2025).