Master Thesis

Machine Learning for Multimodal Freight Chain Modelling

Case Study of NEAC's Mode Chain Builder System

Eka Devi Wulandari
Transport, Infrastructure, and Logistics
5912822

TU Delft

Panteia

lpdp
lembaga pengelola dana pendidikan

Machine Learning for Multimodal Freight Chain Modelling

Case Study of NEAC's Mode Chain Builder System

By

Eka Devi Wulandari

In partial fulfilment of the requirements for the degree of:

Master of Science

in Transport, Infrastructure, and Logistics

at the Delft University of Technology,

to be defended publicly on Friday August 29, 2025, at 11:00 AM.

Thesis committee:	Dr. Bilge Küçük Atasoy	TU Delft
	Dr. Yanan Xin	TU Delft
	Dr.ir. Maarten Kroesen	TU Delft
	Jan Kiel	Panteia

This thesis is confidential and cannot be made public until August 29th, 2025.

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

I am truly happy and grateful to have finally completed this master's thesis. My two-year journey at TU Delft in the Master's program of Transport, Infrastructure, and Logistics has been nothing short of a life-changing experience. The journey has not been easy, yet I have never regretted the decision to take this path. Living and studying in the Netherlands for two years has not only deepened my knowledge of logistics but has also taught me valuable lessons about life and about myself. I would like to thank my God, Allah SWT, for His endless love and strength that have guided me throughout my studies at TU Delft.

I am grateful to LPDP for giving me the opportunity and trust to represent Indonesia in pursuing my studies in the Netherlands. And I would like to extend my gratitude to my thesis supervisors, Bilge and Yanan, for their guidance and valuable feedback during the entire process of this research. I am also sincerely thankful to Jan Kiel for trusting me with the opportunity to work as an intern at Panteia and to develop this research as my thesis topic. I would also like to express my deepest gratitude to the late Yuko Kawabata, who greatly supported me during the thesis process by providing me with valuable information, data, and knowledge related to the Mode Chain Builder. May the late Yuko rest peacefully and happily in eternity.

My heartfelt thanks also go to my husband, Ricky Yusuf, my mother, father, sister, and my entire family in Indonesia for their unwavering support and prayers throughout this journey. A very special thanks goes to my Roland Holstlaan 1243 besties, who became my family in Delft—I cannot imagine my life here without you, guys. I am also thankful to my fellow Indonesian TIL friends, Nabil, Nadhira, and Raihan, for always supporting each other as we navigated the challenges and memories of the TIL program together. Finally, to my dear friends from TU Delft 2023, the RVDH *gengs*, thank you for making me feel at home away from home. Please keep in touch, and let's reunite in Indonesia someday!

Eka Devi Wulandari

Agustus 2025

Delft, The Netherlands

Contents

Abstract

Freight transport plays a critical role in supporting global trade, with multimodal transport systems modelling gaining importance due to their potential to optimize efficiency and sustainability. Accurately modeling these multimodal freight chains is essential for infrastructure planning and policy-making. Yet, it remains a persistent challenge due to fragmented datasets, limited granularity, and the absence of observed multimodal chain-level data. Traditional modeling approaches, particularly heuristic-based methods, often struggle to incorporate real-world operational constraints such as port selection logic and cargo handling requirements. Moreover, these models are typically inflexible and computationally intensive.

This research seeks to develop an adaptive multimodal freight chain model that addresses these limitations. Specifically, it introduces a practical path construction framework that integrates port selection based on geographic and functional suitability, aligning cargo handling requirements with port capabilities during the construction of mode chains. This research also tries to address a gap in the literature by applying machine learning to the estimation of multimodal freight flows, a domain traditionally dominated by heuristic and optimization-based methods. To estimate freight demand distribution across the generated chains, this study explores the use of machine learning, particularly the Expectation-Maximization (EM) algorithm, to leverage the abundant but often unstructured transport data available. The EM model enables demand share prediction without relying on labeled training data, reducing the calibration burden and enhancing model responsiveness to observed transport flows.

The proposed modeling framework is applied to a case study based on the NEAC Mode Chain Builder system for inter-country freight movements between the Netherlands and Belgium, two countries with high multimodal connectivity and the largest ports in Europe. The results demonstrate the model's ability to generate valid mode chain alternatives and to significantly reduce deviations between predicted and observed freight flows, particularly for sea and rail segments. While some deviation increases occur in other segments, these are outweighed by the overall improvement in prediction accuracy. The EM model also shows stable convergence behavior, confirming its potential under data-limited conditions. However, residual deviations suggest that external factors, such as data incompleteness or behavioral uncertainties, still limit full accuracy.

This study highlights the potential of combining graph search algorithms with unsupervised learning to enhance multimodal freight chain modeling, especially under data-constrained conditions. It contributes both a methodological and practical solution for building data-driven multimodal freight transport models that better reflect operational realities and observed empirical data that can be used to improve the freight transport planning and decision-making process.

Chapter 1. Introduction

This chapter establishes the foundation of the research by outlining the broader context and background behind the study. It clarifies the significance of addressing current challenges in multimodal freight transport modelling and positions the research between the existing studies of the multimodal freight chain modelling domain. By presenting the key rationale, knowledge gaps, and overall direction of the study, this chapter sets the stage for the detailed analysis and methodology that follow.

1.1 Background

Freight transport is the key component of global supply chains, enabling the movement of goods across cities, countries, and continents. Over the past decades, freight volumes have grown rapidly and are projected to continue rising in the coming years (IEA, 2002). Key drivers of this growth include shifting consumer demand and evolving trade patterns (Riet, et al., 2004), supported by global economic growth, trade liberalization (Hayakawa, et al., 2018), digitalization (e.g., e-commerce, social media), and infrastructure development (US FHWA, 2020). According to the European Environment Agency (2024), the total freight transport volume in the 27 European Union (EU) member states reached 3,469 billion tonnes-kilometres in 2022, representing a 44.6% increase compared to 1995. This growth occurred despite major disruptions, including the economic crises of 1998 and 2008 and the COVID-19 pandemic between 2020 and 2022.

In Europe, a significant portion of freight shipments are transported through multimodal transport rather than direct unimodal schemes. Multimodal transport is an integrated system that combines different modes (such as road, rail, air, and sea) to enable seamless, efficient, and sustainable movement of passengers and freight (IRU, 2025). A survey conducted by French ECHO indicates that approximately 47% of freight demand is transported via multimodal transport chains, while 46% relies on a single mode, and the remaining 7% falls into other categories (Guilbault, 2008). The preference for multimodal transport chains is primarily driven by the lack of direct connections between certain origin-destination pairs, making unimodal transport not always feasible (Huber, 2017). The Eurostat data in 2025 also reports the same insights where the maritime transport accounts for 67.4% of the total freight volume in the EU. Since maritime transport only connects ports, additional modes are required for first-mile connections from the origin point to the port and for last-mile delivery from the port to the final destination. Hence, it creates a complex multimodal freight chain network across countries within EU territory.

In parallel, the EU is also committed to promoting a more sustainable form of mobility. This goal can be achieved through multimodal transport, which strategically combines multiple transport modes to maximize their individual strengths while minimizing their limitations. To support this vision, the European Commission is actively pursuing a multimodality policy by improving the integration of transport modes and ensuring interoperability at all levels of the transport system (European Union, 2023). This policy direction is likely to increase the demand for multimodal freight chain modeling in the future. As a result, developing an efficient and adaptable system for multimodal freight transport modeling is becoming increasingly important, as it more accurately reflects the realities of freight movement and supports the optimization of logistics, infrastructure planning, and policy decision-making process.

Existing approaches to multimodal freight transport modeling are mostly based on optimization methods, which aim to identify the most efficient transport chains by minimizing factors such as cost, time, or emissions.

Techniques such as genetic algorithms (Okyere, et al., 2022), bilevel programming (Yamada & Febri, 2015), dynamic programming (Liu, 2023), and heuristics have been widely applied (Yamada & Febri, 2015), often combined with simulation to enhance realism. For example, Bok et al. (2017) developed a corridor choice model within the Netherlands' BasGoed system, applying route enumeration and a Multinomial Logit model to estimate demand shares. While effective in certain contexts, these methods depend heavily on predefined assumptions and generalized costs, making them less adaptable to the complex and dynamic nature of multimodal freight flows.

The recent growing approaches applied for freight modelling are big data analytics and machine learning. So far, the applications are mainly for path design (Wang & Fu, 2022), demand generation prediction (Lim, et al., 2022), demand forecasting (Salais-Fierro et al., 2022; Lopez et al., 2019; Peng et al., 2024; Liu et al., 2023; Liachovicius et al., 2023), mode choice prediction (Uddin et al., 2021; Ahmed and Roorda, 2022; Xu et al., 2024), and demand assignment optimization (Huang et al., 2014; Polson et al., 2017; Zhao et al., 2017), with studies consistently showing that machine learning outperforms traditional methods in predictive accuracy. However, despite this progress, the use of machine learning to estimate freight flows across entire multimodal transport chain networks remains largely unexplored. This is a critical gap, as current optimization and statistical approaches often struggle with fragmented data, high dimensionality, and the lack of observed chain-level datasets.

This research addresses this gap by investigating the integration of machine learning into the NEAC Mode Chain Builder system, which currently relies on heuristic-based path construction. The proposed approach seeks to leverage available but fragmented freight statistics to generate a predicted multimodal freight chain database. By learning hidden relationships across incomplete datasets, machine learning can help approximate how freight flows are distributed along complex modal sequences. This integration is expected to reduce dependence on rigid heuristics, improve demand flow estimation, and support more accurate, data-driven decision-making for multimodal freight transport planning.

1.2 Problem Definition and Scope

The multimodal freight chain modelling process in this research will use NEAC Mode Chain Builder (MCB) as the case study object. Hence, the model will be designed to comply with the available input data format, output requirements, and parameter structure of the NEAC database format. NEAC itself is a multimodal, network-based simulation system developed to analyze freight transport flows across Europe. It comprises two key components: the Mode Chain Builder and the NEAC model itself. The Mode Chain Builder model is used to construct a mode chain database that represents multimodal transport chains by integrating various types of transport data, networks, and cost information. This model is developed separately from the main NEAC model because no existing data source currently provides transport data in the form of complete multimodal chains. MCB generates sets of feasible multimodal transport chains between production and consumption point (region) pairs. These chains represent the sequences of transport modes (involving road, rail, inland waterway, and maritime transport mode options) and terminal connections that freight can take across Europe.

Figure 1. *Modelling process in the existing Mode Chain Builder*

The mode chain database construction process within the Mode Chain Builder system comprises three major steps, as depicted by Figure 1: mode chain construction, demand share estimation, and calibration process to compare the estimation results against the actual historical data. The model development process in this research will follow this structure as detailed below:

1. **1st Phase: Mode Chain Construction**
   The objective of this phase is to generate a set of possible mode chains for each production zone, consumption zone, and commodity type pair, based on the available facilities (nodes) and the connections between them by mode type (edges). The resulting dataset serves as a key input for the subsequent demand share estimation phase.
2. **2nd Phase: Method for Demand Share Estimation**
   This phase aims to estimate the demand share distribution across all mode chain alternatives generated in Phase 1. It also incorporates the calibration process, making the estimation more data-driven rather than treating estimation and calibration as separate steps. By integrating historical transport data from the outset, this approach is expected to reduce the calibration challenges and improve the reliability of the demand share output.

The existing Mode Chain Builder system faces a practical gap in its ability to adapt to complex, data-rich environments due to its reliance on a rigid heuristic-based approach. The current Mode Chain Builder system use pre-defined rules (heuristic) approach to generate freight transport chains by processing trade data at the national level, disaggregating it regionally, and assigning routes through a sequence of port and terminal selections governed by predefined rules. While this method has enabled large-scale estimations, it presents several limitations. For instance, the model limits the number of ports considered by applying a bounding box rule and omits factors such as cargo-type compatibility in port selection. This method, while effective for large-scale estimations, reduces the model's realism and limits its capacity to reflect real-world freight transport movements.

The NEAC system models freight movement within and across the 27 EU member states. It also captures international freight flows, particularly maritime transport, between EU countries and their global trade partners (e.g., Asia, the Americas, and Africa). However, due to limitations in research duration and computational capacity, this study focuses on just two neighboring countries (the Netherlands and Belgium) to represent inter-country freight movement. These countries were selected because they have the two largest ports in Europe: Rotterdam (Netherlands) and Antwerp-Bruges (Belgium), which together handled a combined volume of 641 million tonnes in 2024 (Eurostat, 2024). Moreover, both countries demonstrate strong multimodal connectivity, as shown by their significant non-road modal shares: 22.4% in Belgium and 47.2% in the Netherlands (Eurostat, 2023). Given that this study aims to model multimodal freight chain flows, the presence of decent non-road transport connections within the country is essential.

1.3 Research Objectives

This research attempts to find ways to develop a multimodal freight transport chain model that incorporates a practical port selection process and accounts for cargo handling requirements within the mode chain construction phase. Additionally, it explores the use of machine learning to leverage the abundant, though still unstructured, freight transport data for predicting demand flows within the multimodal chain. This machine learning integration into the model has the potential to simplify and accelerate the data calibration process while improving model's predictive accuracy. Several sub-objectives are defined to support the main objective mentioned above:

1. To review commonly used methods for constructing and predicting multimodal freight chain data.

2. To identify a suitable mode chain construction approach that integrates comprehensive port selection and aligns cargo handling requirements with port capabilities.
3. To implement the proposed mode chain construction model within the existing system architecture.
4. To evaluate appropriate machine learning techniques for predicting demand flow distribution in the multimodal freight chain model.
5. To develop and integrate a demand flow distribution model with the mode chain construction process.
6. To assess the overall performance of the proposed model in predicting the multimodal freight chains data in comparison to the actual available freight data.

1.4 Research Question

To achieve all the objectives mentioned in the previous section, a main research question is proposed. The main research question will be jointly answered by the following sub-questions:

How can a multimodal freight chain model that incorporates routing and infrastructure compatibility be developed, and how can machine learning be integrated to enhance its predictive performance in demand flow distribution?

Table 1. *List of Research Sub-questions*

No	Research Question	Method/Approach
1	What methods are commonly used for constructing and predicting multimodal freight chain data?	*Literature review*
2	Which mode chain construction approach that can effectively integrates a comprehensive port selection process and aligns cargo handling requirements with port capabilities?	*Literature review and qualitative analysis*
3	How can the proposed mode chain construction model be integrated into the existing system architecture?	*Mode chain construction modelling in Python*
4	Which machine learning techniques are most appropriate for predicting demand flow distribution in multimodal freight transport?	*Literature review and qualitative analysis*
5	How can the demand flow distribution model be developed and effectively integrated with the mode chain construction process?	*Machine learning modelling in Python*
6	How accurately does the proposed model predict multimodal freight chains compared to the actual available freight data?	*Model evaluation*

The research is structured around six sub-questions, each addressed through a tailored methodological approach. The first two questions focus on identifying existing methods for multimodal freight chain modelling and evaluating suitable approaches for mode chain construction, both answered through a comprehensive literature review and qualitative analysis. Questions three and five are tackled through model development in Python, focusing on the construction of a mode chain builder and the integration of demand flow prediction using machine learning techniques. Meanwhile, question four explores the most appropriate ML techniques for predicting freight flow, relying on a literature review and qualitative assessment, and question six evaluates the accuracy of the proposed model by comparing predicted results with actual freight data.

1.5 Research Framework

The research framework outlines the approach used to develop a predictive model for multimodal freight transport activities between and within selected European countries. This study is systematically structured to address the research objectives and questions presented in the previous sections. The research design follows a sequential process that includes a literature review to support method selection and identify the research gaps, quantitative data collection and preprocessing, model development, and subsequent validation and evaluation. This structured approach ensures the validity and reliability of the model outcomes.

Chapter 2 provides the theoretical foundation for the study by reviewing existing literature on multimodal freight transport modelling and the use of machine learning in transport systems. It highlights the lack of research applying machine learning to multimodal chain construction and demand flow prediction, which leads to a research gap identification that will be addressed in this research. This chapter also provides the literature review results in exploring existing possible mode chain construction and demand share estimations as the steps to answer the **first research question** regarding the commonly used method for constructing and predicting multimodal freight chain data.

Chapter 3 outlines the research design and methodological approach used to develop the multimodal freight chain model. The model development process in this research is divided into two main phases: mode chain data construction and demand flow prediction. This chapter opens by explaining the selected approaches to conduct both phases and the reasons behind the selection process. This will be the answer to the **second and fourth research questions**. Then, the explanation continues with the elaboration of data collection and preprocessing, model development, and the model validation process. The chapter ensures that each step is aligned with the research objectives and designed to produce reliable and interpretable outcomes.

Firstly, Chapter 4 outlines an introduction to the NEAC Mode Chain Builder system as the case study object in this research. Then, the step-by-step development process is presented for both the mode chain construction and demand share estimation phases, using the approaches introduced in the previous chapter. It outlines the implementation of these phases in detail, including data preparation, model application, and validation. This development process addresses the **third and fifth research questions**.

Chapter 5 presents the output of the developed model. It evaluates the model's performance in representing actual observed data to answer the **sixth research question.** This chapter also discusses the key findings, challenges encountered, and insights gained. The chapter also examines the limitations faced during the modeling process, their potential impact on the model's output, and possible future solutions to address these issues.

Finally, Chapter 6 summarizes the main findings and addresses the six research questions outlined in Chapter 1. It highlights the key contributions of the research, as well as the limitations of the developed model and its results. The chapter also provides several recommendations for future work, particularly aimed at overcoming the identified limitations.

Chapter 2. Literature Review

The previous chapter outlined the objective of this research to develop a new multimodal freight transport chain model that addresses parameter limitations in the existing system and explores its integration with machine learning methods. This literature review aims to examine existing research on multimodal freight transport modelling and the application of machine learning within this domain. The limitations of current approaches are discussed to identify the scientific gap and highlight the potential contributions of this study. This chapter also provides the elaboration of possible mode chain construction and demand share estimation approaches to be implemented in model development.

2.1 Research Gap Identification

This section presents the literature review results on the existing research regarding multimodal freight chain modelling and then proceeds with research gap identification that will be addressed in this research.

2.1.1 Existing Research about Multimodal Freight Chain Modelling

In Europe, a significant portion of freight shipments are transported through multimodal chains rather than direct unimodal schemes. A survey conducted by French ECHO indicates that approximately 47% of freight demand is transported via multimodal transport chains, while 46% relies on a single mode, and the remaining 7% falls into other categories (Guilbault, 2008). The preference for multimodal transport chains is primarily driven by the lack of direct connections between certain origin-destination pairs, making unimodal transport unfeasible (Huber, 2017). Additionally, multimodal chains are preferred because they allow for greater efficiency by leveraging the advantages of different transport modes and vehicles within the chain (Konings, et al., 2008).

The most common approach to multimodal freight chain modeling is the use of optimization models. These models aim to identify the optimal freight transport chain by minimizing factors such as cost, time, and emissions. They help determine the best combination of transport modes and infrastructure investments to support sustainable and efficient freight movement. Common optimization methods include Genetic Algorithms (Okyere, et al., 2022), Bilevel Programming (Yamada, et al., 2009), Optimization model (Yamada and Febri, 2015; Limbourg and Jourquin, 2009), Dynamic Programming (Liu, 2023), and heuristics (Yamada, et al., 2009). In their study, (Zhao, et al., 2018) even combines the optimization approach with a simulation approach.

Bok et al. (2017) developed a corridor choice model for predicting container cargo transportation activities using a multimodal transport chain approach, as part of the Netherlands' strategic freight transport model (BasGoed). This study shares nearly 90% similarity with the NEAC case study addressed in this research. The model development included a route enumeration process to generate choice sets of potential transport chains. A Multinomial Logit model was then applied to estimate the probable demand share for each generated chain, using generalized transport costs and chain-specific constants as explanatory variables.

The recent growing approaches applied for freight transport modelling are big data analytics and deep learning. So far, the utilization of machine learning in the domain of freight modelling has mainly been applied for path design (Wang & Fu, 2022), demand generation prediction (Lim, et al., 2022), demand forecasting (Salais-Fierro et al., 2022; Lopez et al., 2019; Peng et al., 2024; Liu et al., 2023; Liachovicius et al., 2023), mode choice prediction (Uddin et al., 2021; Ahmed and Roorda, 2022; Xu et al., 2024), and demand assignment optimization (Huang et al., 2014; Polson et al., 2017; Zhao et al., 2017). Across all reviewed studies, machine learning models consistently outperform traditional approaches. This highlights a growing trend in which ML is primarily applied to enhance the predictive accuracy of freight flow prediction and modelling processes.

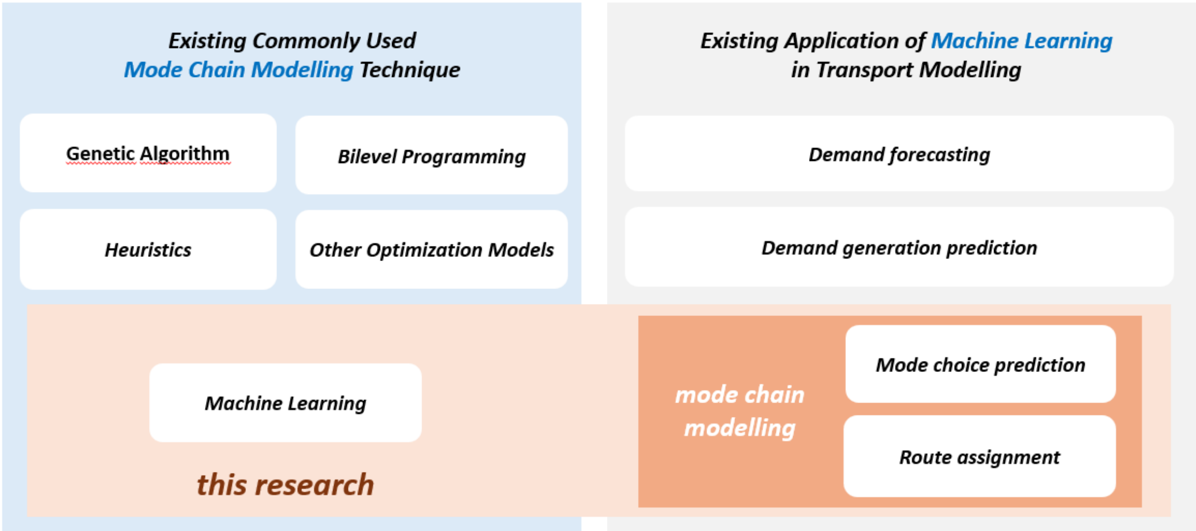

Figure 2. *Position of this Research Relative to the Existing Machine Learning Application in Mode Chain Modelling*

2.1.2 Research Gap Conclusion

While machine learning has been increasingly applied in freight modeling, covering areas such as demand forecasting, mode choice prediction, and path optimization, its use for estimating freight flows across complete multimodal transport chain networks remains largely unexplored. Modelling chains approach is similar conceptually to network assignment, so it's like the combination of the step 3 (mode split) and step 4 (route assignment) of the classical transport modelling process. Traditional optimization and statistical models dominate this field, but they often struggle with the complexity of multimodal routes and the absence of detailed chain-level data. Machine learning, by contrast, offers the potential to infer hidden relationships between transport modes and generate a predicted multimodal freight chain database from the aggregated yet fragmented statistics currently available. By learning patterns across incomplete and high-dimensional datasets, ML can help approximate how freight flows are distributed along complex modal sequences, thereby addressing a major gap in existing modeling approaches.

A central challenge in multimodal freight transport modeling is that available data is rarely chain-specific. While numerous datasets exist, they are typically aggregated at national or regional levels, inconsistently formatted, and incomplete, reflecting inputs from multiple stakeholders. Statistics are usually reported for single modes (e.g., road, rail, inland waterway, or maritime) rather than as connected multimodal sequences. As a result, existing data cannot directly capture how modes are linked across origin–destination regions, leaving the actual distribution of freight across multimodal chains unknown. Machine learning provides an opportunity to bridge this gap by using available partial data to reconstruct likely chain-level flows, effectively transforming fragmented inputs into a coherent multimodal chain database suitable for modeling and decision support.

To address these limitations, this research investigates the feasibility of embedding machine learning into the NEAC Mode Chain Builder system, which currently relies on heuristic-based approaches. Integrating ML is expected to reduce reliance on fixed rules and instead leverage observed transport statistics to generate realistic mode chain alternatives and estimate demand flows. This approach allows the system to dynamically learn from existing aggregated datasets, producing a predicted multimodal freight chain database that better reflects empirical transport patterns while supporting more accurate and data-driven decision-making in multimodal freight transport.

2.2 Modelling Approaches and Methods

The model development process in this research is divided into two main phases: mode chain data construction and demand flow prediction. Following this structure, the method selection process is divided into two corresponding phases:

1. **Method for Mode Chain Construction**
   The objective of this phase is to generate a set of possible mode chains for each production zone, consumption zone, and commodity type pair, based on the available facilities (nodes) and the connections between them by mode type (edges). The resulting dataset serves as a key input for the subsequent demand share estimation phase.
2. **Method for Demand Share Estimation**
   This phase aims to estimate the demand share distribution across all mode chain alternatives generated in Phase 1. It also incorporates the calibration process, making the estimation more data-driven rather than treating estimation and calibration as separate steps. By integrating historical transport data from the outset, this approach is expected to reduce the calibration challenges and improve the reliability of the demand share output.

2.2.1 Mode Chain Construction

Another review process was conducted to explore several possible techniques that have been used in existing literature to construct and predict multi-modal freight transport chains. These techniques range from classical optimization methods to more adaptive, data-driven machine learning (ML) approaches. In this section, the mechanisms, strengths, and limitations of these possible methods will be discussed for selecting a suitable method in the context of NEAC's mode chain builder. According to the literature review, there are three main approach categories commonly used to address path or route planning and construction problems: optimization, heuristics, machine learning, and classical algorithmic graph theory.

3.2.1.1 Optimization Approach

One of the foundational strategies is the use of optimization-based models, particularly those using Dijkstra's algorithm combined with multi-objective optimization. For instance, Lu and Wang (2022) developed a multi-layered transportation network and used Dijkstra's algorithm to identify optimal routes based on cost, time, and risk, followed by aggregation via the Analytic Hierarchy Process (AHP). To address varying operational environments, Elbert et al. (2020) introduced a classification between dynamic and non-dynamic optimization approaches. Dynamic models adapt to changes in real-time (e.g., demand or infrastructure updates), while non-dynamic models rely on fixed parameters. This distinction is essential for multi-modal systems where real-time adaptability can significantly enhance performance but also increases computational requirements and data demands.

However, the nature of optimization problems is typically to identify a single most optimal solution, such as the shortest, cheapest, or least risky route. In contrast, the objective of the MCB is not to optimize a single path, but to generate multiple plausible path alternatives that reflect actual freight transport behavior within the EU. This enables the system to capture route variability, incorporate uncertainty, and ultimately estimate the distribution of demand flows across a set of potential multi-modal chains rather than one deterministic route.

While optimization models are typically designed to identify the single most efficient route, they can be extended to generate multiple alternatives through techniques like k-shortest paths algorithms (Liu, et al., 2018) or multi-objective optimization with Pareto front exploration (Zheng, et al., 2024). However, even with these adaptations, optimization remains fundamentally goal-driven and deterministic, seeking mathematically optimal solutions based on predefined criteria. This orientation differs from the objective of the Mode Chain Builder, which aims to generate a set of plausible transport chain alternatives that reflect actual freight behavior across Europe. Rather than identifying a single best route, the Mode Chain Builder focuses on capturing behavioral variability and path diversity, enabling more accurate demand flow estimation across multiple multimodal chains. As such, while optimization approaches can support parts of the path generation process, they do not naturally accommodate the probabilistic, adaptive, and data-driven characteristics required for modeling real-world transport chain behavior at scale.

3.2.1.2 Heuristics Approach

In the domain of heuristic-based methods for route or chain construction, two approaches were identified: the Iterative *Route Construction and Improvement (IRCI)* algorithm proposed by Figliozzi (2009) and *the Pair Insertion Algorithm (PIA)* introduced by Mauttone and Urquhart (2008). Both were originally developed to solve routing problems by generating a single optimized route based on performance criteria such as travel time, user coverage, or structural feasibility. While not explicitly designed to produce multiple alternatives, these methods could be adapted to do so by varying parameters, initial conditions, or constraints across multiple runs.

IRCI constructs initial routes using the *Generalized Nearest Neighbor* rule approach to build routes sequentially by inserting node one at a time based on the generalized cost function. Route improvement was then conducted once the initial routes were generated by doing route merging, reordering nodes, and time adjustments. On paper, this approach is applied to solve the *Vehicle Routing Problem with Soft Time Windows (VRPSTW)* problem. The main objective is to develop a flexible and efficient algorithm to solve the VRPSTW problem while minimizing travel distance, vehicles used, and late deliveries, which is similar to the optimization problem.

PIA focuses on quickly generating diverse route sets by systematically inserting pairs of vertices rule (instead of single vertices) to build an optimal set of bus routes using network connection & travel cost data as input. PIA follows an iterative heuristic approach where the process begins with creating a new route following the shortest path between two points, then inserting it into an existing route to minimize cost increase while considering the number of transfers & round-trip duration. It continues until all demand is covered.

In the context of a Mode Chain Builder, where the goal is to capture a set of plausible mode chain alternatives rather than a single best solution, such heuristic frameworks offer valuable building blocks. IRCI is particularly suited for constructing paths that balance operational criteria, while PIA provides a fast and flexible mechanism for building structurally sound chains. However, leveraging these methods to generate diverse outputs would require additional procedures beyond their original formulations.

3.2.1.3 Machine-Learning Approach

The machine learning domain for constructing and predicting freight transport chains has been growing in recent years because it offers greater adaptability and behavioral realism compared to traditional optimization models. At least three approaches were identified from the literature review process, including DBSCAN clustering (Joubert & Meintjes, 2016), reinforcement learning (Yoo, et al., 2023), and deep reinforcement learning (Holliday, 2025).

The study by Joubert and Meintjes (2016) applied DBSCAN, an unsupervised clustering method, to GPS trajectory data to identify major stop locations, such as freight hubs or transshipment points. These clusters were then used to infer activity chains and build a connectivity graph that reflects real-world freight movement. This approach is highly valuable for empirically reconstructing transport chains, especially when high-resolution movement data are available. However, as it primarily serves to analyze observed behavior, it is better suited for retrospective studies rather than forward-looking predictive modeling unless combined with other methods.

In contrast, reinforcement learning (RL) methods such as those explored by Yoo et al. (2023) focus on enabling an agent to learn optimal mode chain decisions through interaction with the transport environment to solve the *Bus Network Design and Frequency Setting (BNDFS)* problem. The variable decision consists of determining the optimal number of bus routes and their service frequencies. Their model used Q-learning, a value-based RL algorithm, to determine routing strategies that maximize cumulative rewards based on travel time minimization and share of demand served criteria. This method is well-suited for modeling adaptive decision-making in freight routing, particularly when route availability or preferences evolves for real-time or simulation-based planning environments. However, Q-learning is known to face scalability issues in large, high-dimensional networks, such as those in European freight systems.

Holliday (2025) presents a *Deep Reinforcement Learning (DRL)* approach for urban transit network design, modeling the problem as a *Markov Decision Process (MDP)* and using a *Graph Attention Network (GAT)* combined with *Proximal Policy Optimization (PPO)*. The model's objective is to automatically generate bus routes and frequencies that minimize a composite cost function incorporating passenger travel time, operator costs, and penalties for unmet demand. The model learns to sequentially construct routes by deciding whether to extend or finalize a route based on a reward function. It incorporates rich node, edge, and global network features to guide decision-making.

While originally applied to public transit, the method is highly relevant to freight mode chain construction, as it's able to handle large input spaces and learn policies that balance multiple routing objectives across highly interconnected multi-modal networks. Nevertheless, DRL requires extensive training data and careful reward function design.

In summary, DBSCAN provides a powerful tool for empirically deriving transport chain structures, RL and DRL offer more flexible, predictive frameworks capable of generating diverse and adaptive mode chain alternatives. While machine learning approaches offer powerful ways to learn adaptive and cost-aware routing strategies, they are inherently designed to produce a single optimal or most likely solution, not multiple plausible alternatives. To support demand flow distribution across several mode chains, additional mechanisms would be needed.

3.2.1.4 Graph Search Algorithm: Path Enumeration

The NEAC Mode Chain Builder aims to generate multiple plausible multimodal freight transport chains between origin-destination (OD) pairs and estimate demand flow distribution across these alternatives. This requires a system that is adaptive and capable of producing more than one valid route per OD pair. After reviewing the three alternative approaches, there is no a single approach that can meet the requirement and comply the characteristics of the existing Mode Chain Builder. Hence, the literature search was expanded into the classical graph search algorithm.

Graph search algorithm is a broad term referring to any computational method designed to explore or traverse graphs to discover the relationship and/or connection between a collection of nodes, forming a route, path, or network (Pavicic, 2023). Graphs, which consist of nodes (also called vertices) and edges (connections between nodes), are widely used to model networks in a variety of contexts, as highlighted below:

- **Exploration:** Finds all possible nodes or paths (not necessarily optimal). The commonly used algorithm for this case is BFS or DFS.

- **Optimization:** Finds the most optimal paths based on edge weights (distance, cost, etc). This type of problem usually uses Dijkstra's, Bellman-Ford, or Yen's algorithm.
- **Heuristic search:** Prioritize routes using estimated cost. This one usually uses A* algorithm

The optimization and heuristics domain have been discussed in the previous part. Thus, in this section, the exploration domain will become the focus since our goal is to find the algorithm that can enumerate all the possible path/mode chains.

The use of the BFS algorithm for transport path construction has been widely used due to its ability to systematically explore all possible paths level by level, even though its applications are mostly for route optimization problems in urban transport, logistics scenarios, or network construction cases in general. Bernov (2023) developed a system to optimize waste transportation routes in Sidoarjo Regency using BFS and DFS algorithms. Their study compares the performance of both algorithms in terms of total distance and volume traveled to support more efficient waste collection. Another research conducted by Kartoirono et al. (2022) used BFS to enhance the efficiency of finding alternative bus routes.

Studies have shown that BFS can be enhanced or combined with other algorithms to improve performance, such as the BFS Link Elimination (BFS-LE) approach for generating relevant truck routes while minimizing extraneous options. The research was conducted by Tahlyan and Pinjari (2020), who proposed a new method to evaluate route choice set generation by comparing algorithm-generated routes with observed freight truck routes for the same OD pair. They assessed the effectiveness of the *Breadth-First Search with Link Elimination (BFS-LE)* algorithm and found that spatial aggregation of trips can reduce computational effort but may require careful filtering of extraneous routes.

BFS, along with DFS, is a fundamental algorithm in graph search, offering simple and flexible logic for route and network construction. Its general-purpose structure makes it widely applicable across various cases, with the ability to be adapted and customized based on specific requirements. BFS is particularly very useful in transport route construction due to its simplicity, speed, and reliability in finding the shortest paths. Its effectiveness can be further enhanced through algorithmic refinement and integration with complementary methods. Its flexibility also allows the model to be combined with a heuristic approach to accommodate systems that require a high degree of customization in mode chain construction logic. BFS allows for such flexibility, enabling the integration of techniques like link elimination (Tahlyan & Pinjari, 2020) to reduce computational complexity. Most importantly, BFS supports full enumeration of all possible path alternatives, which is the core requirement for constructing a complete mode chain alternatives dataset in the freight mode chain modelling process.

***Table 2.** Summary of Mode Chain Construction Possible Methods*

Approach	Method(s)	Author (Year)	Applicability for NEAC's MCB
Optimization	Dijkstra and multi-objective optimization algorithms	Lu and Wang (2022)	- Optimization problems is typically to identify a single most optimal solution, while objective of the MCB is generate multiple plausible path alternatives. - While optimization approaches can support parts of the path generation process (Liu et al. (2017); Zheng (2023), they do not naturally accommodate the probabilistic, adaptive, and data-driven characteristics.
	Service Network Design	Elbert et al. (2020)	
	Hyrbrid: Heuristic and Optimization	Wang et al. (2022)	
	K-shortest Paths Algorithms	Liu et al. (2017)	
Heuristics	Route Construction and Improvement (IRCI) algorithm	Figliozzi (2009)	- Boths methods are adapted for solving optimization problems in these two papers. Thus, only generate one optimal path. - Both can be adapted to generate multiple paths, though fundamentally they're not designed for it, by applying several modification in the algorithm's logic.
	Pair Insertion Algorithm (PIA)	Mauttone & Urquhart (2008)	
Machine Learning	Hybrid: DBSCAN clustering and heuristics	Joubert & Meintjes (2016)	- The use of RL/DRL in path construction are usually for optimization problem, whose objective is to find one optimal path. - RL can be used to generate multiple paths but it will struggles with very large state-action spaces like EU-wide freight networks - It often needs to be paired with methods to ensure diversity and feasibility, and is better suited for adaptive systems or simulations rather than static enumeration.
	Reinforcement learning (RL)	Yoo et al. (2023)	
	Deep reinforcement learning (DRL)	Holliday (2025)	
Classical Graph Search Algorithm	Breath-first Search (BFS)	Bernov (2023); Tahlyan & Pinjari (2020); Kartoirono et al. (2022)	- BFS is a fundamental algorithm in graph search, offering a simple and flexible logic for route construction. - Its flexibility allows it to be integrated with other techniques like link elimination (Tahlyan & Pinjari, 2020) to reduce complexity. - BFS is very suitable for system that needs speed and reliability in finding one optimal or multiple feasible paths.
	Depth-first Search (DFS)	Bernov (2023)	

2.2.2 Demand Share Estimation

The idea of demand share estimation is to estimate the demand flow distribution across the available mode chain alternatives accurately for each PC-commodity pair. The resulting demand flow estimates are then calibrated against actual aggregated transport data, such as national transport performance statistics and port-level data published by Eurostat. However, the calibration part is highly problematic because the model needs to adjust its parameters iteratively to match known transport data, but this process is hindered by the large matrix size (which slows processing). Additionally, aligning the flow distribution per chain with national single-mode statistics demands disaggregation, increasing the complexity. To reduce the complexity, this study tries to integrate the demand share estimation and calibration process to make the estimation more data-driven rather than treating the two processes as separate steps. By integrating historical transport data from the outset, this approach is expected to reduce the calibration challenges and improve the reliability of the demand share output.

Given the system requirements, the method used for demand share estimation must be capable of handling large, multi-dimensional datasets. It should be able to recognize patterns within the available data, particularly since the historical data from Eurostat is provided in aggregate form at the country and facility levels, while the demand share needs to be estimated at the NUTS-3 regional level. The chosen method must use these identified patterns to predict the most likely demand share distributions that closely align with the observed historical transport data. These characteristics are very compatible with the benefits that are offered by machine learning (ML). ML is particularly beneficial for systems characterized by large, complex, and high-dimensional datasets, and it excels in environments with big data where pattern recognition and predictive analytics are needed. Hence, the method selection process for the demand share estimation phase will be specific on finding the most suitable machine learning method to employ. Since the existing Mode Chain

Builder uses the Multinomial Logit (MNL) model to determine demand share across mode chains, the method selection in this study focuses on exploring machine learning techniques that are commonly used to replace or enhance the traditional MNL approach.

2.2.2.1 Supervised Learning Methods

Several studies were found that employ machine learning to conduct the mode split process. Uddin et al. (2021) assessed the effectiveness of various machine learning classifiers for modeling freight mode choice, using data from the 2012 Commodity Flow Survey, supplemented with spatial attributes. The study compares prediction accuracy between traditional Multinomial Logit model with eight machine learning methods: *Naïve Bayes (NB), Support Vector Machine (SVM), Artificial Neural Networks (ANN), K-Nearest Neighbors (KNN), Classification and Regression Tree (CART), Random Forest (RF), Boosting,* and *Bootstrap Aggregating (Bagging).* The results reveal that tree-based ensemble classifiers, especially Random Forest, provide the most accurate predictions.

Xu et al. (2024) integrated interpretable machine learning (ML) methods (tree-based methods like CatBoost and SHapley Additive exPlanations) with traditional multinomial logit (MNL) models to refine the MNL model specification, aiming to improve both predictive accuracy and understanding of key factors influencing freight transportation decisions. They considered 3 tree-based methods in their study: *Random Forest (RF), XGBoost, and CatBoost,* and then compare the performance to find the best method to refine the multinomial logit (MNL) model specifications. CatBoost was found to be the best-performing method in terms of predictive accuracy. In another research, Ahmed and Roorda (2022) applied the *Random Forest* algorithm and compared its performance with traditional multinomial and mixed logit models. The findings show that the random forest method provides higher prediction accuracy than the discrete choice models.

Lee et al. (2018) investigated the use of four artificial neural network (ANN) models (BPNN, RBFN, PNN, and CPNN) for mode choice prediction and compared their performance with the conventional multinomial logit model (MNL). Using 10-fold cross-validation, they found that ANN models achieved higher prediction accuracy (around 80%) compared to MNL (70%). Among the ANNs, the Probabilistic Neural Network (PNN) performed best, particularly in predicting underrepresented transport modes. The study highlights the potential of ANNs as effective non-parametric alternatives for discrete choice modeling.

Based on the above literature review results, it’s seen that the most common method to use as the refinement of the Multinomial Logit model is Random Forest (Uddin et al., 2021; Xu et al., 2024; Ahmed and Roorda, 2022) and Neural Network (Uddin et al., 2021; Lee et al., 2018). According to Coursera (2025), Random Forests (RF) is good at handling large tabular datasets, generalizing well, and offering guidance for problem-solving, but it can be slow and less transparent in how it make predictions. On the other hand, Neural Network (NN) is highly flexible, can handle various data formats and incomplete datasets, but is more prone to under- or overfitting. It is powerful at recognizing patterns and deciding the best course of action to accomplish a task by weighing all the available options and learning from past mistakes.

Since, in the Mode Chain Builder context, the estimated demand flow distribution needs to align with historical data, Neural Networks (NN) seem suitable to be implemented due to its ability to incorporate domain-specific constraints formulation. This historical data can be embedded as a set of constraints during the model training process. However, fundamentally, NN is a supervised learning method that works well at instance-level learning, where every instance data (X) should have its corresponding observed value (Y) to let the model learn and identify the hidden pattern within the data so it can predict Y-pred given different X. Neural Network utilizes the characteristics of provided features data as the input to drive predictions ($y\_pred = f(features)$), and then use the constraints to calculate the model error (loss) to inform the network how far its predictions deviate from the expected aggregate data. The defined constraints only affect the loss function and don't directly influence the prediction mechanism. Hence, NN will probably perform poorly when the data learning supervision is only at the aggregate level, like the characteristics of the object of this research.

2.2.2.1 Semi-Unsupervised Learning Methods

Since instance-level learning is not feasible due to the lack of observed instance-level data, the exploration of suitable methods for the demand estimation phase has shifted from supervised to semi-supervised learning approaches. One of the unsupervised learning methods that allows the formulation of domain-specific constraints is the Expectation-Maximization (EM) algorithm. It's a statistical technique often used in machine learning for parameter estimation in models with latent variables. In freight transport modelling, machine learning methods are increasingly applied to improve forecasting, routing, and operational efficiency. However, the specific use of EM algorithms in the multimodal freight chain modelling domain is not directly addressed in the available research.

The application of maximization in freight transport modelling is still limited to the classification problem to solve the Vehicle-Commodity Matching Problem (VCMP) to minimize the total transportation cost by estimating the parameter of the Gaussian Mixture Model (Sun, et al., 2021) and to identify the unobservable latent psychological variables (e.g., attitudes) in transport choice modelling (Sohn, 2017). Wang (2022) also utilized the EM algorithm to fill in missing data in the operations' historical datasets for his study. EM transforms incomplete observed data into statistically inferred complete datasets to support more reliable decision-making that, in turn, improves the operations' efficiency, order processing time, cost savings, and more optimal path finding. So, in this study, the EM algorithm serves as both a data recovery technique and an optimization enabler.

In another study, Yoon (2025) applied the EM algorithm to optimize the likelihood of observed sensor data by treating the robot's trajectory as a latent (hidden) variable. Firstly, the algorithm estimates the latent trajectory distribution given current model parameters and then updates the model parameters to maximize the likelihood based on that estimated variable's value. Although the domain of the study is not directly related to multimodal freight chains, the underlying principle, that treating unknown or missing data as latent variables estimated using available observed data, can also be applied to the demand share estimation in this research.

Based on the literature review, the author concludes that the Expectation-Maximization (EM) algorithm is well-suited for developing the demand share estimation model. In this context, the demand share of each path is treated as a latent (hidden) variable, estimated iteratively using aggregated historical freight transport data as ground truth. These aggregated data serve as constraints that shape the objective function and guide the prediction of the latent variables. The algorithm performs multiple iterations to minimize the deviation between the estimated parameters and the actual values, ensuring that the final estimates closely reflect the true demand shares and align with the observed aggregate statistics.

Table 3. Summary of Demand Share Estimation Possible Methods

Author (Year)	Method(s)	Description
Supervised Learnings		
Uddin et al. (2021)	- Naïve Bayes (NB) - Support Vector Machine (SVM) - Artificial Neural Networks (ANN) - K-Nearest Neighbors (KNN) - Classification and Regression Tree (CART) - Random Forest (RF)- Boosting - Bootstrap Aggregating (Bagging).	- Assessed the effectiveness of various machine learning classifiers for modeling freight mode choice, comparing it to traditional Multinomial Logit (MNL) model. - The results reveal that tree-based ensemble classifiers, especially **Random Forest, provide the most accurate predictions.**
Xu et al. (2024)	- Random Forest (RF) - XGBoost - CatBoost	- Integrated ML with traditional MNL models to refine the model specification, aiming to improve predictive accuracy and understanding of key influencing factors in freight transportation choice **- CatBoost was found to be the best-performing method in terms of predictive accuracy.**
Ahmed and Roorda (2022)	Random Forest	The findings show that the **Random Forest method provides higher prediction accuracy than the discrete choice models.**
Lee et al. (2018)	- Backpropagation Neural Networks (BPNNs) - Radial Basis Function Networks (RBFNs) - Probabilistic Neural Networks (PNNs) - Clustered Probabilistic Neural Networks (CPNNs)	- Investigated the use of ANN models for mode choice prediction and compared their performance with the conventional MNL. - They found that ANN models is the potential effective non-parametric alternatives for discrete choice modeling. - **ANN achieved higher prediction accuracy (around 80%) compared to MNL (70%).** - Among the tested ANN nethods, the **Probabilistic Neural Network (PNN) performed best.**
Semi Unsupervised Learnings		
Sun (2021)	Expectation-Maximization Algorithm	Solve the Vehicle-Commodity Matching Problem (VCMP) to minimize the total transportation cost by **estimating the unknown parameter** of the Gaussian Mixture Model
Sohn (2015)		**Identify the unobservable latent psychological variables** (e.g., attitudes) in transport choice modelling alongside observable data (e.g., socioeconomic attributes)
Wang (2022)		- **Estimate the missing data** in the logistics operations datasets by transforming the incomplete observed data into statistically inferred complete datasets. - The new complete datasets then were run in the simulations, and the results outperformed the other traditional methods in improving the operations' efficiency
Yoon (2025)		- Use EM as a parameter learning framework to automatically train models for **trajectory estimation without ground truth supervision.** - EM is used to **optimize the likelihood** of observed sensor data by treating the robot's trajectory as **a latent (hidden) variable.**

.

2.3 Introduction to the Selected Modelling Methods

In this section, a brief explanation of basic principles and general steps will be presented for both selected methods: Breadth-First Search and Expectation-Maximization algorithm.

2.3.1 Breadth-First Search (BFS) Algorithm

Breadth-First Search (BFS) is a fundamental graph traversal algorithm designed to explore the nodes and edges of a graph or tree in a systematic, level-by-level manner, starting from a specified source node. BFS explores all immediate neighbors before progressing to deeper layers of the structure McKee (2024). This means every node at a certain depth is visited before moving to nodes that are further away from the starting point. This strategy makes BFS particularly effective for tasks where examining all possible paths and connections at each stage is essential, such as finding the shortest path in an unweighted network or mapping routes through a maze. The primary goal of BFS is to provide a structured, exhaustive, and efficient exploration of a graph or tree.

McKee (2024) explains that the key components of BFS include the use of a queue and a mechanism to track visited nodes. The queue is central to the algorithm's operation, it maintains the list of nodes that need to be explored and processes them in a first-in, first-out (FIFO) order. This ensures that nodes are explored in the exact sequence in which they are discovered, preserving the level-by-level approach. Simultaneously, a list or set of visited nodes helps prevent revisiting nodes, critical when dealing with graphs that may have cycles or multiple paths back to the same node. This component avoids infinite loops and redundant checks, allowing BFS to function efficiently and terminate properly.

Breadth-First Search (BFS) explores a graph in a level-wise manner, beginning at a specified starting node and expanding outward based on distance. The algorithm follows these steps according to celerdata.com (2024):

1. Initialize by marking the starting node as visited and placing it in a queue.
2. Dequeue a node, visit it, and perform any required operations.
3. Enqueue all unvisited neighboring nodes of the current node and mark them as visited.
4. Repeat steps 2 and 3 until the queue is empty, indicating that all reachable nodes have been explored.

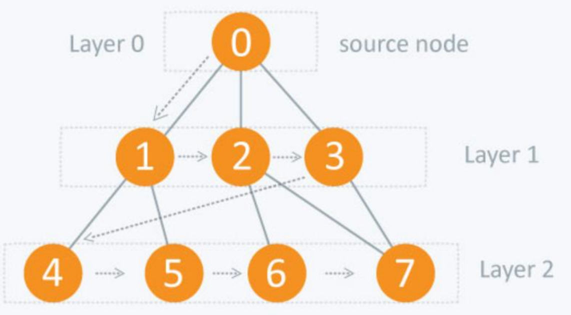

Figure 3. *Sequential node visit procedure of BFS (source: hackerearth.com, 2025)*

This approach systematically visits each node's neighbors before moving to the next level, continuing until either all nodes are explored, or a target node is found. BFS serves as a foundational technique for more advanced algorithms such as Dijkstra's algorithm (for shortest pathfinding) and Prim's algorithm (for minimum spanning trees).

2.3.2 Expectation-Maximization Algorithm

The Expectation-Maximization (EM) algorithm is commonly used to compute maximum likelihood (MLE) estimates in situations where some data is missing. MLE is a statistical method used to determine the parameter values of a model that maximizes the likelihood of observing the given data. In other words, it finds the settings that make the observed data most probable to happen. The Expectation-Maximization (EM) algorithm is an iterative technique used in unsupervised machine learning to estimate unknown parameters in statistical models, especially when dealing with incomplete or hidden data.

Haugh (2015) explains that the EM algorithm is also applicable in cases involving latent (unobserved) variables or data that were never meant to be observed directly. In such scenarios, the latent variables are treated as missing, allowing the EM procedure to be applied. Latent variables are unseen components of the data that indirectly influence the observed outcomes, and their values are estimated using the information available from the visible data (geeksforgeeks, 2025). The algorithm has various applications across statistical fields. It is frequently applied in machine learning and data mining, as well

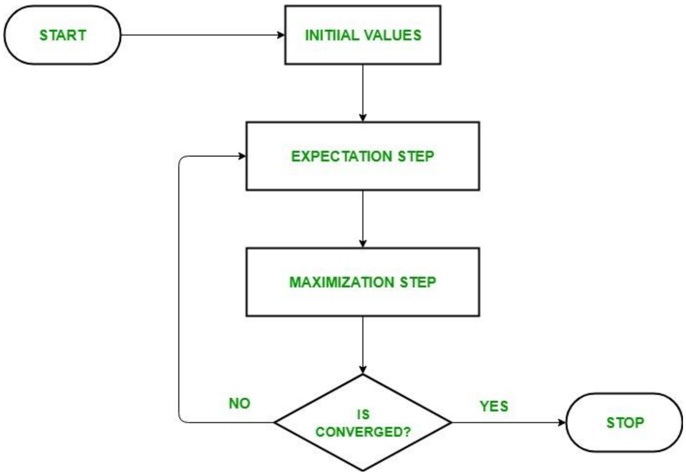

Figure 4. *EM Algorithm Flowchart (source: geeksforgeeks.com, 2025)*

as in Bayesian statistics, where it is often used to estimate the parameters of data distributions.

According to geeksforgeeks.org (2025), below is the step-by-step of how the EM algorithm works:

1. **Initialization:** Define a specific initial value as the starting reference for the model to run the algorithm.

2. E-step (Expectation):

In this step, the algorithm estimates the missing or hidden data (latent variable) based on the ground truth (observed) data and the current parameter. Then, it will compute the probability of each latent variable based on the ground truth (observed) data. Then, the model will calculate the log-likelihood of the observed data using the current parameter estimates.

To give a clearer description of this E-step, the mathematical representation is shown as follows (Haugh, 2015). Suppose the complete dataset is denoted by $Z = (X, Y)$, where only X is observed. The complete data log likelihood is expressed as $l(\theta;\ X;\ Y)$ where θ representing the unknown parameter for which we aim to compute the maximum likelihood estimate (MLE). E-step calculates the expected value of $l(\theta;\ X;\ Y)$, given the observed data X and the current parameter estimate θ_{old}, where $p(y \mid X, \theta_{old})$ is the conditional density of Y given X.

$$Q\ (\theta;\ \theta_{old}) := E[\ l(\theta;\ X;\ Y) \mid X, \theta_{old}\ ]$$
$$= \int l(\theta;\ X;\ Y)\ p(y \mid X,\ \theta_{old}) dy$$

3. M-Step (Maximization):

The M-step will update the current parameters with new values that could maximize the likelihood we calculated in the previous step, as shown by the mathematical formulation below (Haugh, 2015):

$$\theta_{new} := \max_{\theta} Q\ (\theta;\ \theta_{old})$$
$$then\ \ \theta_{old} = \theta_{new}$$

In estimating the parameters, the gradient descent algorithm is employed. According to IBM (2025), Gradient Descent is a widely used optimization algorithm for training machine learning models and neural networks by iteratively minimizing the error between predicted and actual outcomes. It does so by updating model parameters (weights and bias) in the direction of the steepest descent of the cost function, which measures the overall error.

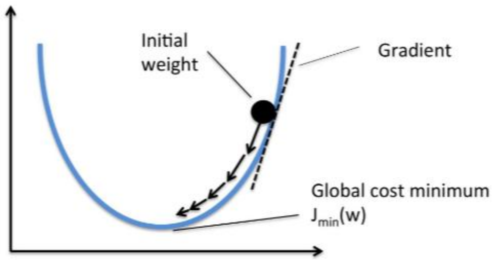

__Figure 5__. Gradient Descent Illustration (Schulte & Atasoy, 2024)

The weight update rule in gradient descent is expressed as follow:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_i} J(\theta_0, \theta_1)$$

The algorithm requires two key elements: the slope (gradient) to determine the update direction and the learning rate to determine the step size. In the above formula, α represents the learning rate, which determines the step size taken during each iteration, and $\frac{\partial}{\partial \theta_i} J(\theta_0, \theta_1)$ denotes the gradient (slope) of the cost function with respect to the parameter θ_j. The gradient indicates the direction of the steepest increase in the cost function, and by subtracting it, the algorithm moves in the opposite direction toward minimizing the error. This update process is repeated iteratively until convergence, meaning the cost function no longer decreases significantly. Through this mechanism, the parameters are gradually adjusted to achieve the lowest possible error in the model.

A high learning rate accelerates convergence but risks overshooting the minimum, while a low rate ensures precision at the cost of efficiency. The process continues until the cost function reaches a local or global minimum, known as the point of convergence, where the model achieves its optimal accuracy. In this context, the loss function measures the error for a single training example, while the cost function represents the average error across the dataset.

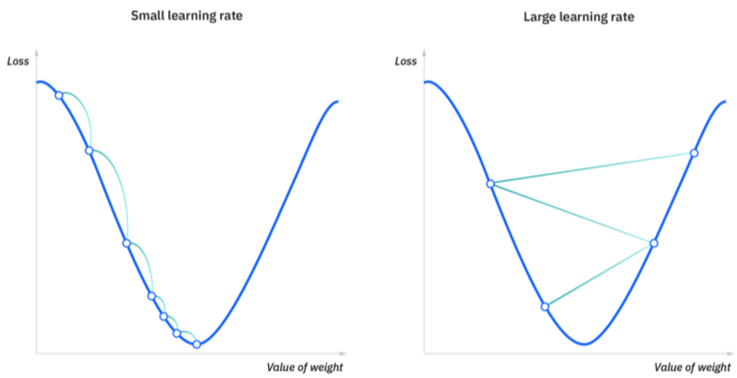

Figure 6. *Illustration of Step Sizes of Different Learning Rate (IBM, 2025)*

4. Convergence:

Haugh (2015) stated that the two steps are iteratively repeated until the sequence of updated θ_{new} values converge. Under very general conditions, convergence to a local maximum is guaranteed. However, if the log-likelihood function is suspected to have multiple local maxima, the EM algorithm should be executed multiple times with different initial values for θ_{old}. The final maximum likelihood estimate of θ is then chosen as the best result among the local maxima obtained from these runs.

Convergence happens when the model reaches a stable point indicated by the insignificant changes in the model's parameter or log-likelihood value in the iteration, to the point that the value is small enough (below the defined threshold) to stop the process because more iterations won't bring a significant improvement anymore (geeksforgeeks.org, 2025).

According to geekforgeeks.org, the Expectation-Maximization (EM) algorithm offers several advantages that make it a practical tool for handling incomplete or hidden data in machine learning. It consistently improves results with each iteration, increasing the likelihood of finding a good solution. Its structure is also relatively simple to implement and, in many cases, leads to efficient mathematical solutions due to the closed-form nature of the M-step. However, the EM algorithm also has several drawbacks. It converges slowly, which means reaching the optimal solution may require many iterations. Additionally, it is prone to getting trapped in local maxima, potentially settling for suboptimal solutions.

Chapter 3. Methodology

This chapter presents the research design and methodological approach used to develop and evaluate the proposed multimodal freight transport chain model. It outlines the sequential process by which the research is conducted, including method selection, data collection and preprocessing, model development, as well as model validation and evaluation as depicted by Figure 7.

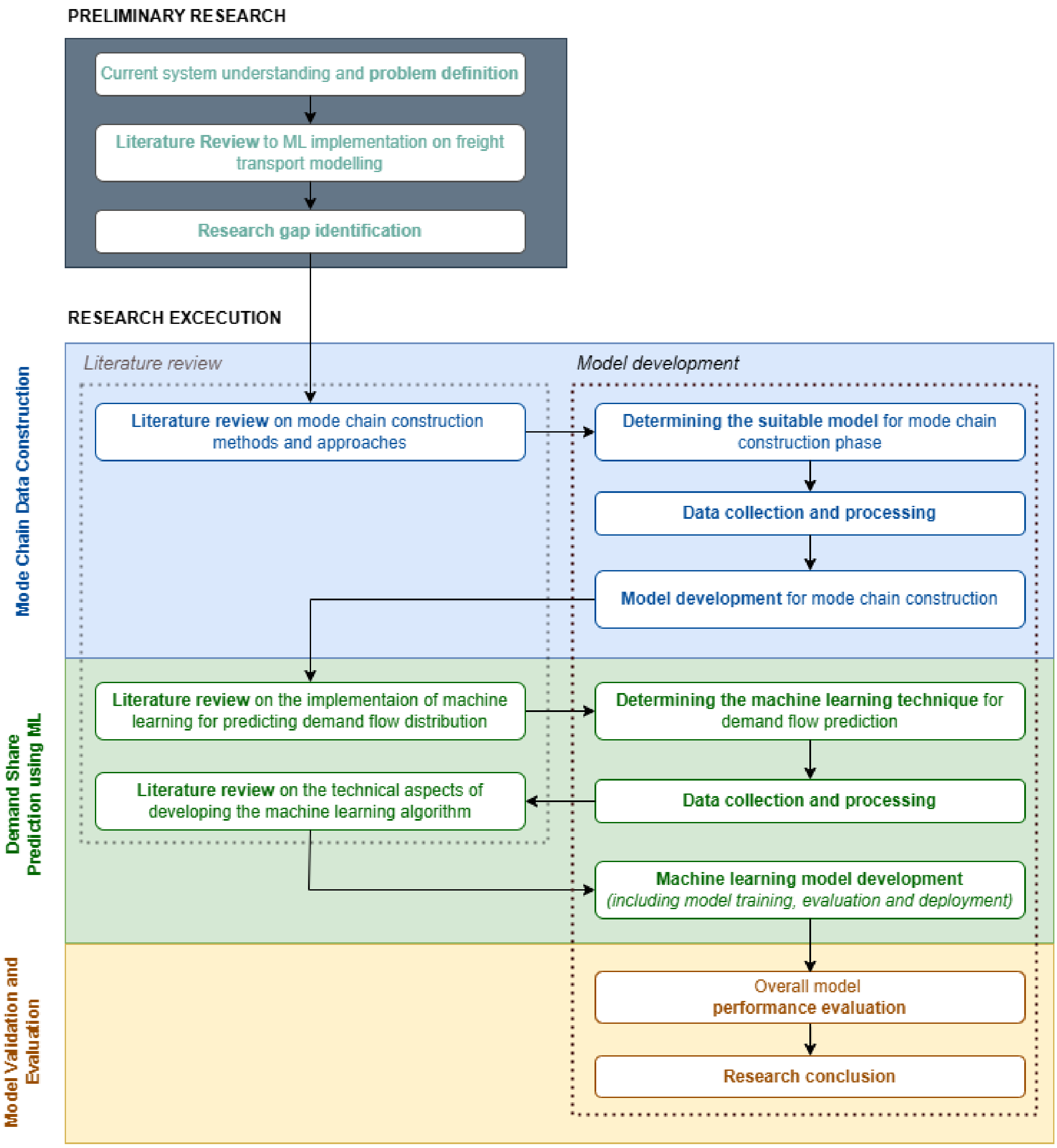

Figure 7. *Research Framework*

The reserach flow is divided into two main stages: preliminary research and research execution. The preliminary research begins with understanding the current NEAC Mode Chain Builder system, identifying its limitations, particularly in parameter constraints and lack of adaptability, and defining the core problem. This is followed by a literature review focused on the implementation of machine learning in freight transport modelling, which leads to the identification of the research gap that this study aims to address. This part is already presented in the previous section, Chapter 2.

The research execution is structured into three interconnected components. The first is the Mode Chain Data Construction phase, which starts with a literature review to explore existing methods and approaches for constructing multimodal freight chains. Based on this review, a suitable method is selected, followed by data collection and preprocessing. The model is then developed to generate mode chain data that incorporates routing flexibility and infrastructure compatibility.

The second component focuses on demand share prediction using machine learning. It begins with two literature reviews: one on the application of machine learning for demand flow prediction, and another on the technical aspects of model development. A suitable machine learning technique is selected based on these reviews, followed by data collection and processing. The machine learning model is then developed through training, evaluation, and deployment to estimate demand distribution across the multimodal transport network.

The final component is model validation and evaluation, where the performance of the integrated model is assessed using relevant metrics to determine its accuracy, adaptability, and effectiveness. This stage concludes with the formulation of key findings and the overall research conclusion.

In conclusion, this research is structured around three fundamental phases: **method selection, model development, and model validation and evaluation**. The following sections elaborate on the detailed steps undertaken in each phase, following this three-phase structure.

3.1 Method Selection

As briefly mentioned in the previous section, the multimodal freight chain modelling process in this research consists of two phases: mode chain construction and demand share estimation. The demand share estimation process is the combination of the share estimation and calibration steps of the current Mode Chain Builder system.

Based on the literature review presented in the previous section, Breadth-First Search (BFS) is identified as the most suitable method for the Mode Chain Construction phase. BFS is preferred over other options due to its flexibility in accommodating the unique and specific requirements of the existing Mode Chain Builder system, which demands a high degree of customization in its construction logic. This flexibility also allows BFS to be combined with the Link Elimination technique to reduce computational complexity. Most importantly, BFS supports the full enumeration of all possible path alternatives, which is the core requirement of the Mode Chain Builder.

This study adopts the **Breadth-First Search with Link Elimination (BFS-LE)** approach, as introduced by Tahlyan and Pinjari (2020), for constructing mode chains. The link elimination step will be implemented by applying a set of pre-defined rules to remove irrelevant links, thereby reducing the algorithm's search space during the construction process. This heuristic enhancement is essential to ensure compliance with several mandatory rules associated with the input data structure, which must be respected throughout the chain construction process.

For the second phase, Demand Share Estimation, the selected method is the Expectation-Maximization algorithm. The Expectation-Maximization (EM) algorithm is selected because it is well-suited for situations where supervised instance-level data is unavailable, and only aggregate-level historical data is provided. Its ability to effectively handle latent variables makes it ideal for demand share estimation, where individual path choices are unobserved. By treating demand shares as hidden variables and using aggregated freight transport data as constraints, EM allows iterative refinement of estimates. It minimizes the deviation between predictions and observed totals, such as known totals of freight flows across regions or modes, which cannot be directly attributed to individual OD-path combinations, ensuring that the final demand share estimates are both statistically consistent and aligned with real-world aggregate data. This makes EM a robust and data-efficient approach for unsupervised learning in freight transport modeling. The overall comparison between the existing and the newly proposed methods for the end-to-end phase of Mode Chain Builder system is shown by Figure 8.

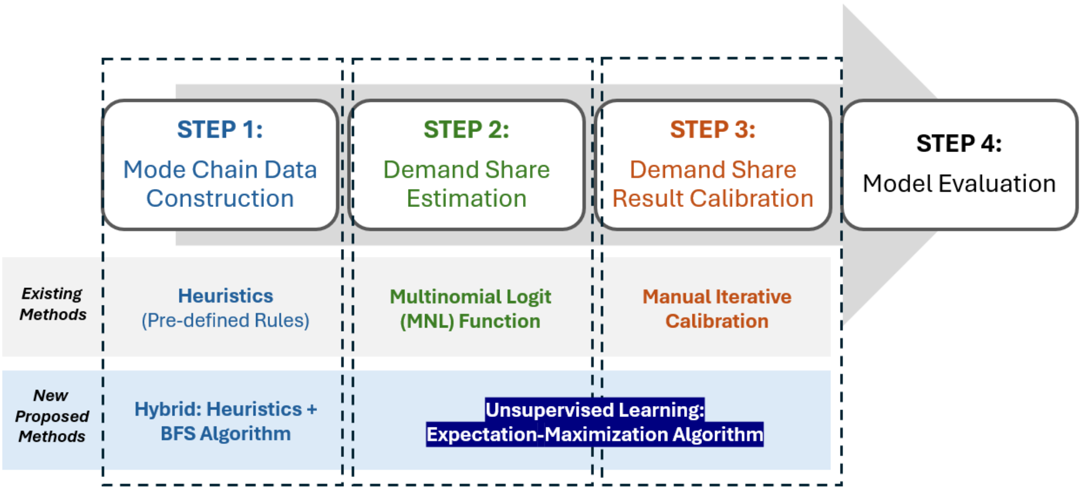

Figure 8. *Comparison between the existing and the proposed methods*

3.2 Mode Chain Construction Phase

3.2.1 Model Development

This phase marks the initial step in developing the multimodal freight chain model. Its main objective is to generate a set of path alternatives for each PC-commodity pair, which will serve as input for the next phase: Demand Share Estimation. Figure 9 shows the end-to-end process of mode chain database construction process, that involves two main approaches: **Heuristics** and **Graph Search Algorithm** based on the Breadth-First Search (BFS) method. The mode chain data construction phase relies on the existing Mode Chain Builder input database provided by Panteia, complemented by Eurostat's port statistics for the port selection procedure. This study does not involve creating new input datasets or modifying the existing data.

The application of pre-defined rules (heuristics), followed by the implementation of the Breadth-First Search (BFS) algorithm is adapted from the Breadth-First Search with Link Elimination (BFS-LE) method introduced by Tahlyan and Pinjari (2020). In this study, link elimination is carried out by enforcing several pre-defined rules (heuristic method) to determine which region pairs should proceed to the BFS model and which ones can have their mode chains directly defined during the heuristic phase.

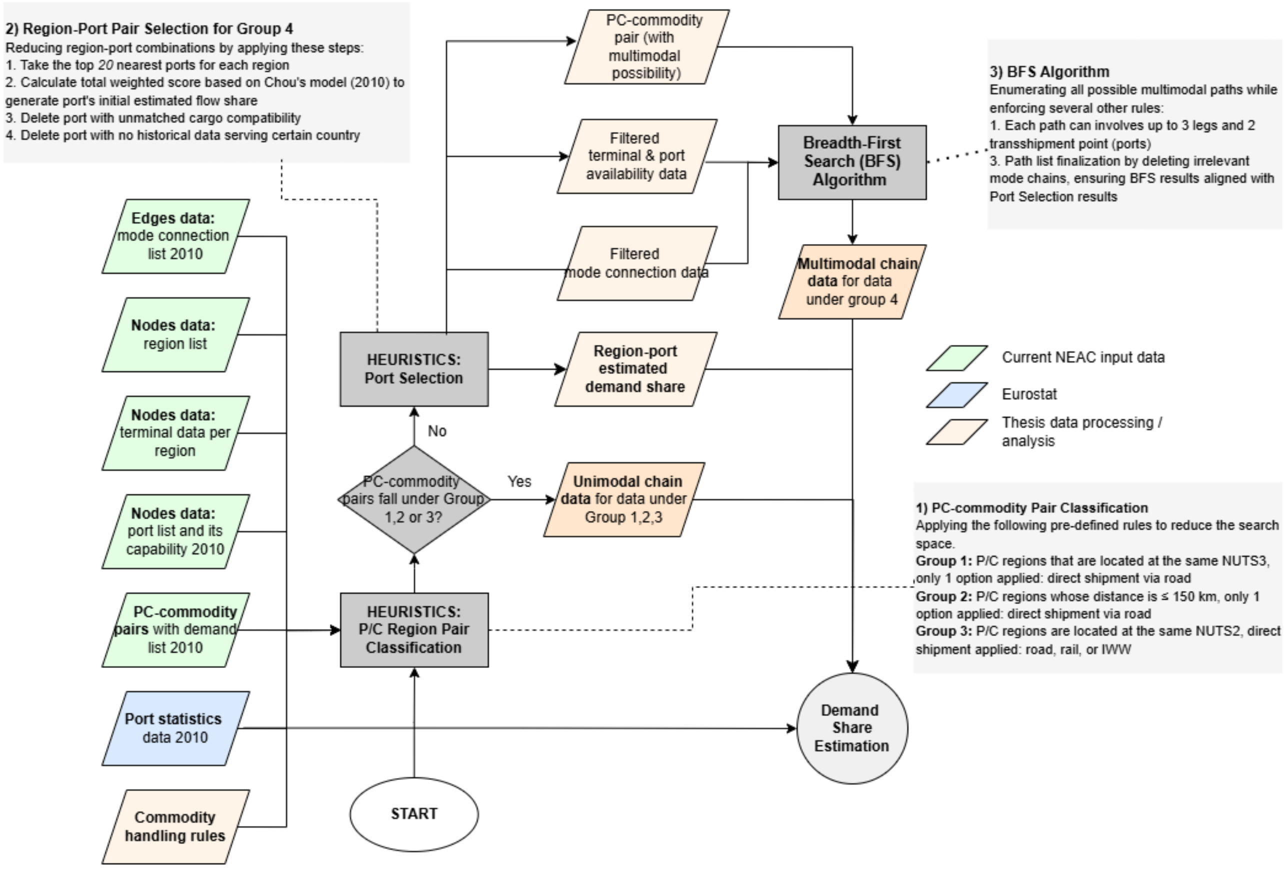

Figure 9. *Data processing diagram of Mode Chain Construction Phase*

3.2.1.1 Heuristics Approach

The heuristic step aims to reduce the simplify and reduce the mode chain construction process that will be conducted later, by applying several pre-defined reasonable rules to remove irrelevant and less probable path alternatives. This step also ensures that several essential rules are incorporated in the mode chain data generation process. The heuristics step itself has two sub-steps: P/C Region Pair Classification and Region-to-Port Selection.

Step 1: P/C Region Pair Classification

In the initial step, three predefined rules are applied to classify PC-commodity pairs into either the unimodal or multimodal group, based on the proximity between their production (origin) and consumption (destination) regions. Region pairs that are located within the same NUTS2 area and/or have a distance of less than 150 km are categorized as unimodal. For these cases, only direct shipment is considered as the transportation option. All remaining PC-commodity pairs are classified as multimodal and proceed to the Port Selection stage, as their transport chains require the involvement of ports as transshipment points for sea transport.

Step 2: Region-to-Port Selection

For the multimodal pairs group, the next step involves selecting appropriate ports that can serve the region pairs. This is achieved by applying a set of heuristic rules, such as identifying the 20 nearest ports for each region, scoring them based on Chou's (2010) port choice qualitative model, and filtering out ports with incompatible cargo handling capabilities.

The heuristic step is executed entirely using an SQL platform due to the complexity, volume, and variety of datasets involved. SQL was chosen primarily because the author is more familiar with it compared to using the Pandas library in Python, allowing for more efficient processing and reduced execution time.

3.2.1.2 Breadth-First Search (BFS) Algorithm

The multimodal pairs group data then proceed to this BFS stage. The BFS algorithm is used to enumerate all viable multimodal transport chains between selected region-port pairs. This process is conducted using a BFS-like logic approach with custom expansion rules, since the existing Mode Chain Builder requires quite a lot of customizations to comply with several data format requirements. Further details regarding the requirement can be found in section 4.2.2.

The BFS model with custom expansion and filtering rules are built using NetworkX module in Python. NetworkX is a Python library designed for the creation, manipulation, and study of complex networks (graphs). It is used in this study to model the freight transport network as a multi-modal graph that enables the representation of transport links as mode-labeled edges and supporting the exploration of 1-leg to 3-leg mode chain alternatives through custom graph traversal logic while incorporating infrastructure availability constraints.

Lastly, before finalizing the mode chain dataset, a final validation step is conducted at the end of the BFS process to eliminate any irrelevant or infeasible paths. The resulting multimodal chain dataset from the BFS step is then combined with the unimodal chain dataset generated during the initial P/C Region Pair Classification stage. This combined dataset serves as the input for the final phase: Demand Share Estimation.

3.2.2 Model Validation

The model validation process for the mode chain construction phase will be carried out by comparing the generated mode chain data with the initial input datasets to ensure consistency. This step verifies that the resulting chains remain aligned with the key information provided at the beginning of the process. In addition, the validation will be manually cross-checked using the SQL platform. The input datasets include:

- Mode connection availability
- Rail and inland waterway availability
- Port availability

3.3 Demand Share Estimation Phase

3.3.1 Model Development

Once a complete dataset of mode chain alternatives for all PC-commodity pairs has been constructed, the next step is to allocate the known goods flow for each pair across the available path alternatives. The main challenge lies in the absence of data showing the historical distribution of goods across these paths. Instead, only aggregated data, such as port statistics and transport movements by mode between countries, sourced from Eurostat, is available. Traditional choice models are insufficient for this task, as they do not ensure consistency between the estimated shares and the aggregated observations. This is where a machine learning approach, specifically the Expectation-Maximization (EM) algorithm, is suitable to use. The EM algorithm is an iterative method used in unsupervised learning to estimate unknown or latent variables by uncovering patterns in the data that align with observed constraints. The path-level demand shares will be treated as hidden variables that must be inferred from this indirect ground truth data.

Step 1: Initialization

The Expectation-Maximization (EM) algorithm requires an initial value as a reference point to guide its predictions. This initial value is critical to the learning process, as a well-estimated starting point can help the model converge more quickly and efficiently. The initial value represents an estimated starting distribution of demand across alternative paths within the same PC-commodity pair group. As there is no standard guideline for determining the most suitable estimation method, this study tests three different approaches:

1. **AHP approach:** The initial value is estimated by calculating each path's weighted score based on the Analytic Hierarchy Process (AHP) results from Lu and Wang (2022) study.
2. **ESD approach:** The initial value is derived from Eurostat's historical share data for each transport mode between origin and destination countries. For each chain or route, the shares of all constituent edges are multiplied to calculate the chain's relative share within the same PC-country pair.
3. **EQW approach:** The total demand (in tonnes) is evenly divided among all available alternative paths, giving each path an equal weight.

These three methods are used to generate initial values for the mode chain dataset. To identify the most suitable approach, the estimated initial values will be evaluated by comparing their deviation from the ground truth data on mode transport flows and port-country flows from Eurostat. The method with the smallest error will be selected and used as the input for the EM model.

Step 2: EM algorithm Development

To estimate the demand share for each mode chain between PC-commodity pairs, the Expectation-Maximization (EM) algorithm treats the demand shares as latent variables that need to be inferred. The algorithm iteratively adjusts these estimates by using known but indirect ground truth data, aiming to find the optimal distribution of demand across path alternatives. The objective is to minimize the deviation between the aggregated values derived from the estimated shares and the actual observed aggregated data.

Two ground truth datasets from Eurostat will be used as reference points to estimate the demand share per transport path. These datasets include: (1) mode-specific transport flows between and within countries, and (2) country-level port statistics detailing incoming and outgoing cargo volumes. A loss calculation function will be developed by integrating both datasets, where the loss represents the deviation between the predicted values and the observed ground truth data. The EM model will iteratively adjust the predictions in order to minimize this loss.

Step 3: Hyperparameter Tuning

The model development will also involve the hyperparameter tuning. It's the process of selecting the most effective set of initial parameters for a machine learning model. These hyperparameters are preset configurations that influence how the model learns during training and affect the model's performance to generalize the sample data to the new or hidden data.

In machine learning models using Expectation-Maximization algorithm, there are at least five important parameters to set prior to model training execution:

1. **Initial learning rate:** Parameter that determines how quickly the model learns, that relates also to how far the model will take the step at the initial stage of model training.
2. **Learning rate decay factor:** This parameter indicates how much learning reduction should be applied after a sharp increase in loss.
3. **Loss Tolerance:** Loss tolerance defines the minimum change in loss required to justify continuing the training process. When the loss change becomes negligible, it indicates that the model may have reached convergence.
4. **Patience for Convergence**: It refers to the number of consecutive iterations with stable (or minimally changing) loss before the training is stopped early. Both, loss tolerance and this parameter help ensure the model doesn't overtrain once meaningful improvements have plateaued.
5. **Maximum Iterations:** Setting a maximum number of iterations acts as a safeguard to prevent the training process from running indefinitely. This is especially important if the model requires many

iterations to converge. Defining this upper limit ensures that the training process remains computationally manageable and time-efficient.

The ideal initial learning rate will be determined through a series of experiments involving 10 iterations, while other parameters will be pre-defined based on needs and relevant references. To optimize prediction results, this study will activate the ADAM (*Adaptive Moment Estimation*) setting within the developed EM model. It is a widely used optimizer module for automatic adaptation in machine learning. ADAM updates the model parameter during training by computing gradients from the loss function, so each parameter has its own learning rate that is kept adjusted during the training based on its past gradients and magnitudes (geeksforgeeks, 2025).

On top of ADAM, a learning rate scheduler is also applied to monitor the training loss trend. If the loss stops decreasing for a few iterations (*patience*), the scheduler will reduce the global learning rate of the optimizer. Unlike ADAM, the scheduler doesn't compute gradients; rather, it controls when and how fast the model learns. One important factor influencing model performance is its ability to shift between exploration in the early training stages and exploitation in later stages, once the model approaches an optimal solution. This dynamic behavior can be facilitated by setting a learning rate scheduler, which allows the learning rate to adjust according to predefined conditions or training states. It will slow down the learning rate if the loss gets stuck.

3.3.2 Model Validation and Performance Evaluation

Several steps will be taken to conduct the model validation process, as detailed below:

1. **Model convergence and behavior**
   Model convergence, indicating by a gradual decrease in loss until it reaches the state where no significant improvement can be made any more, is a good signal that the model is working the way it's intended.

2. **Loss (against ground truth) evaluation**
   The loss calculation produced by the model output will be evaluated. The model's loss indicates the deviation between the estimated value made by the model vs the observed ground truth value. The model's loss then will be compared to the actual initial deviation between input data of total demand flow and total transported demand from all transport modes stated in Eurostat.

Chapter 4. Model Development and Results

4.1 Introduction to Case Study

As explained in the previous section, this research uses the NEAC Mode Chain Builder (MCB) as the case study for multimodal freight chain modeling. This introductory section provides a general overview of the existing MCB system, followed by the technical specifications of NEAC's mode chain data output. Understanding the system's requirements is essential to ensure that the constructed model aligns with the input data format and parameter structure of the NEAC database, thereby supporting its practical applicability for end users.

4.1.1 NEAC Mode Chain Builder

The NEAC (Network for European Freight Transport Analysis and Coordination) is a multimodal, network-based simulation system developed to analyze freight transport flows across Europe. It integrates trade data, economic activity, and multimodal transport networks to support data-driven decision-making. The system comprises two key components: the **Mode Chain Builder** or NEAC Chain Database, which generates a database of base-year multimodal transport chains, and the **NEAC model** itself, which is used to forecast future transport demand and explore scenario-based analyses (Panteia, 2015).

NEAC modal adopts the classical 4-step transport modeling approach as the foundation for its modeling framework. As explained in NEAC Model Description file (Panteia, 2015), the four steps include trip generation, which determines the number of trips originating from or destined for a region; trip distribution, which establishes region-to-region travel flows; mode choice (mode split), which allocates trips to specific transport modes; and trip assignment, where the traffic flows are mapped onto a modal network structure. Mode chain builder itself conceptually is similar to the combination of mode choice (step 3) and trip assignment (step 4) in the transport modeling process.

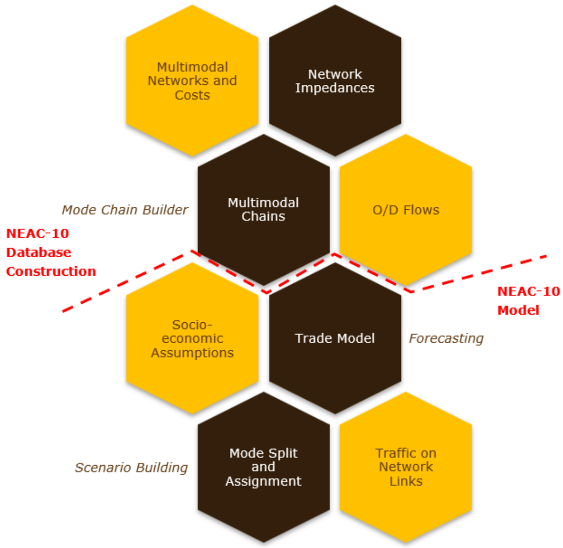

Figure 10. *NEAC's system structure (Panteia, 2015)*

The Mode Chain Builder serves as the foundational component of the NEAC transport simulation system, providing the empirical and structural data necessary for robust transport analysis across Europe and neighboring regions. The database combines a wealth of multimodal transport data, regionally disaggregated down to the NUTS3 level (and equivalent elsewhere), to capture the complexity of contemporary trade and transport flows. Crucially, the Mode Chain Builder underpins both the analytical capabilities and scenario forecasting of the overall system, acting as the indispensable reference point for traffic, trade, and infrastructure assessment.

A central innovation of the Mode Chain Builder is its use of a mode chain data structure. Traditional models often relied on simple origin-destination (O/D) matrices that failed to capture the reality of European and global transport, where goods frequently undergo multiple transshipments and modal changes. The mode chain structure explicitly represents up to two transshipment points along a trade relation, enabling the database to register three separate transport 'legs' within a single flow. Each record includes identifiers for origin and destination regions, transshipment points, transport modes at each stage, the commodity transported (coded by standardized two-digit NST codes), and volumes in tonnes or container units. This structure allows for a nuanced accounting of multimodal and transit flows, reduces double-counting (common when transshipped goods appear in both international and domestic statistics), and directly links economic trade patterns to concrete transport choices

The Mode Chain Builder (MCB) is the key software component responsible for transforming trade data into the detailed regional and modal chains recorded in the database. The process starts with national-level trade and transport flows, disaggregates them to the regional NUTS3 level, and then assigns plausible modal and routing choices based on infrastructure, port connections, and product types. This top-down estimation produces mode chain data that can be directly mapped into network-based O/D matrices across different transport modes (road, rail, inland waterway, and sea). The methodology ensures that the database accurately reflects the reality of how goods traverse international and domestic networks in multimodal sequences, providing a base year snapshot that feeds into subsequent modeling and scenario development.

4.1.2 System Characteristics

The first step of Mode Chain Builder is to generate a dataset containing all the possible mode chain or path that can be used to transport goods from a certain region to other regions. Mode chain is a combination of one or more transport modes connecting origin and destination points with at least two transshipment points in between if the path is using Sea as its intermediate mode. In NEAC case, the multimodal chain data can involve a maximum of 3 legs or edges in network design terms. The edges represent mode connection available between two nodes, while nodes represent region area at NUTS 3 level. Thus, the number of nodes involved in one chain will be $n + 1$ where n is number legs involved, where 2 of them being transshipment points if the chain involves Sea as the 1st or 2nd leg mode.

Given the above condition, each origin-destination pair could have multiple possible chains. The chain could be 1-leg, 2-legs, or 3-legs paths as illustrated by Figure 11 that uses different line color to show various of possible mode chain. The chains with 2 or 3 legs are called multimodal chains since it involves the combination of at least two modes. There are four available modes to choose from, they are Rail, Road, Inland Waterway (IWW) and Sea. The use of Rail and IWW mode must consider the availability of respective terminal in both regions that are going to be connected. The same case as Sea transport that needs the port to be available in both regions being connected.

On the other hand, 1-leg chain will only involve one mode and usually called as Direct Shipment because there will be no intermediate nodes considered in the model output. Even though, for example in the actual world, the Road mode (using truck) has to go through region X to transport the goods from origin region A to destination region B, but the model output will generate 1-leg Road chain instead of 2-legs chain (with Road as 1st leg and also Road as 2nd leg) even though it cross the region X.

The decision variable during the mode chain construction process not only chooses the most convenient mode, but also selects the port as the origin and destination transshipment. Because different combinations of port selection will result in distinct mode chains as shown by Red and Purple line in Figure 11. The port selection process in the existing Mode Chain Builder model uses bounding box rules, where the port selection process only includes the ports that fall within a specified rectangular boundary that is defined by minimum and maximum coordinates. This rule is commonly used as a spatial filtering technique to filter out irrelevant data and reduce computational complexity. However, it drives a risk of omitting optimal or feasible alternatives that fall just outside the box. Thus, this research attempts to introduce a new approach that considers other important relevant aspects in the port selection process and doesn't solely rely on spatial variables as the deciding variable.

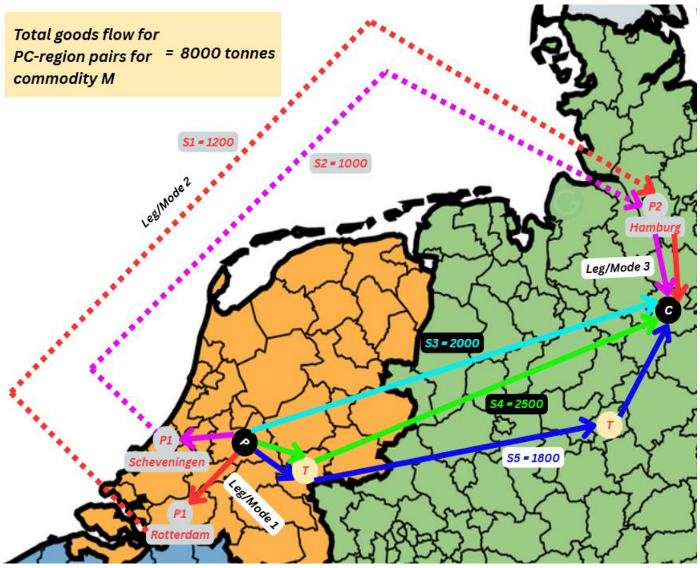

Figure 11. *Illustration of Possible Mode Chain between an Origin-Destination Region Pair*

The mode chain data will be generated based on the pair Origin-Destination (OD) and commodity type pair. Each commodity requires different handling requirements that need to be aligned with the port's cargo handling capability during the port selection process. The existing Mode Chain Builder doesn't consider this variable in its model. This research strives to incorporate this variable into consideration during the mode chain data construction process. This initiative and the new port selection approach are one of improvement points this study attempts to contribute to the existing model framework.

After constructing a set of possible mode chains for each OD–commodity pair, this dataset will become the input for the next step: demand flow estimation. The goal of this stage is to estimate the distribution of total demand flow across the generated chains. This is currently performed using a Multinomial Logit (MNL) function, which uses travel cost, time, and distance as predictor variables to calculate the probability of each chain being selected. These probabilities are then multiplied by the total demand for each OD–commodity pair to determine the distributed demand flow across the chains. The resulting demand flow estimates are subsequently calibrated against actual aggregated transport data, such as national transport performance statistics and port-level data published by Eurostat.

The calibration of the NEAC-10 mode chain builder is highly problematic due to data limitations, computational challenges, and methodological constraints. The model adjusts its parameters iteratively to match known transport data, but this process is hindered by the large matrix size (which slows processing), the lack of reliable multimodal data for validation, and the presence of local exceptions that require region-specific adjustments. Additionally, aligning multimodal chains with national single-mode statistics demands disaggregation, increasing complexity and error risk. Although ports offer some calibration reference through known cargo volumes, available data is typically aggregated by mode (e.g., RoRo, container) rather than commodity type, reducing its usefulness for detailed validation. As a result, achieving accurate, localized calibration remains difficult and time intensive.

4.2 Mode Chain Construction Phase

The expected output of mode chain construction phase is a dataset of possible mode chain to transport a certain commodity type of goods from production region to consumption region at NUTS3 level. As discussed in the previous section, this phase will be performed using the combination of heuristics and customized BFS algorithm. Below is the detailed step of the model development:

4.2.1 Data Preparation

Several datasets will be used as input for the mode chain construction phase. These data are sourced from Panteia's internal database and Eurostat. Panteia's internal data, originally developed for the existing Mode Chain Builder, is based on transport and trade figures from the year 2010. This study will utilize the same datasets for the entire model development process without revisions or updates, due to time constraints. However, additional data may be incorporated as needed to support specific aspects of model development, and it will also use 2010 data as the basis.

Table 4. *List of input data for Mode Chain Construction phase*

No	Dataset Name	Description	Source	Eurostat's Data Code	Available Information (column list)
1	*10input_region_list*	list of NUTS-3 region	Panteia	-	region_seq, region_nuts, region_name, region_wn_code, country_code, country_name, node_type
2	*20input_region_pairs*	list of origin-destination region at commodity level	Panteia	-	ProdZoneID, ConsZoneID, commodity, tonnes
3	*30input_mode_connection*	list of available rail, road and IWW connection for each pair of regions	Panteia	-	mode, orig_wn_id, dest_wn_id, distance, travel_time, gen_cost
4	*31input_sea_connection*	list of available sea transport connection for each pair of regions	Panteia	-	O_WNID, D_WNID,O_NAME, D_NAME, AccessKms, EgressKms, SeaKms, SeaMins, SeaCost
5	*40input_ports_by_region_EZ2006*	list of available port at each region, with cargo handling capability details	Panteia		WNPORTID, COUNTRY, PORT, CONTAINER, DRYBULK, LIQBULK, RORO, REGIONNUTS,REGIONWN
6	*40input_terminal_availability*	list of rail and IWW terminal availability at each region	Panteia	-	region_id, rail_terminal, iww_terminal
7	*60input_commodity_rules*	list of NSTR's commodity code that are mapped to NST07 code and complemented by cargo handling requirement	Author	-	NSTR_Code, NSTR_Category, NST07_Code, NST07_Category, Container, DryBulk, LiquidBulk, RoRo, Others
8	*80input_port_connectivity_plsci*	list of port's connectivity index calculated and released by UNCTAD	UNCTAD	-	PORT_ID, Region_ID, NEAC_all, NEAC_imp, Eurostat_name, UNCTAD_name, country, PLSCI_2015_2024, PLSCI_2024_2025, PLSCI_2010
9	*90input_ports_stats_inout*	historical goods inflow and outflow data recorded at certain ports, sent or received from various EU countries	Eurostat	*mar_go_am_nl* *mar_go_am_be*	level, country, port_region_id, port_name_neac, port_name_eustat, partner_country, in_liq_bulk, in_dry_bulk, in_container, in_roro, in_other_cargo, out_liq_bulk, out_dry_bulk, out_container, out_roro, out_other_cargo

An overview of the input datasets, including their sources and key variables, is provided in Table 4 below. All datasets are initially used as inputs for implementing the heuristic method. The purpose of this step is to eliminate irrelevant links, whether between regions and ports or between ports, by evaluating factors such as mode connectivity (Datasets 3 and 4), terminal availability for rail and inland waterway transport (Dataset

5), compatibility between commodity handling requirements and port handling capabilities (Datasets 5 and 7), and the presence of historical demand outflows from ports to partner countries (Dataset 9).

This phase also includes the port selection process, which aims to identify a subset of ports most likely to accommodate goods flowing to or from a given region. In this process, a region-to-port connection score is calculated based on inland freight cost (Dataset 3), number of intermodal links (Dataset 3), total cargo volume handled by the port (Dataset 9), and the connectivity index (Dataset 8).

4.2.2 Model Development

The model development begins with a heuristic stage, where a set of predefined rules is applied to classify PC-commodity pairs into either unimodal or multimodal transport groups. This step aims to reduce the computational load of the subsequent breadth-first search (BFS) and to eliminate irrelevant links between regions and potential transshipment points for sea transport (i.e., port selection). Next, the BFS algorithm model is developed to perform full path enumeration for the multimodal PC-commodity pairs, generating all feasible transport chains. Finally, the development process concludes with the integration of unimodal and multimodal groups into a single, unified dataset for further modeling and analysis.

4.2.2.1 Heuristics Approach Implementation

The heuristic step consists of 2 major steps, they are PC-commodity pair classification and port selection. Each step is elaborated in details in the following sections:

Step 1: PC-Commodity Pairs Classification

The PC-commodity pairs classification is performed by applying the following pre-defined rules to classify the PC-commodity pair data:

1. If the P/C pair is between the same NUTS3 or the distance between the P/C regions is ≤ 150 km, then only road transport will be assigned for this data group.
2. If P/C regions are located at the same NUTS2, only direct (1-leg) shipments will be assigned to this data group, either using road, rail, or IWW, depending on the connection availability. Rail and IWW can only be assigned if both OD regions have the same terminal type. This assumption is because the average distance between two regions within the same NUTS2 is 202 km.

```
for each (orig, dest) in region_pairs do
    if orig_id = dest_id then
        classification ← '[1] Same NUTS3: direct ROAD only'
    else if avg_distance_km ≤ 150 and 'Road' in inland_connection then
        classification ← '[2] Below 150km: direct ROAD only'
    else if substr(orig_nuts,1,4) = substr(dest_nuts,1,4) and orig_rail_term ≥ 1 and dest_iww_term ≥ 1 then
        classification ← '[3] Same NUTS2: direct ROAD,RAIL,IWW'
    else if orig_rail_term ≥ 1 and dest_rail_term ≥ 1 then
        classification ← '[3] Same NUTS2: direct ROAD,RAIL'
    else if orig_iww_term ≥ 1 and dest_iww_term ≥ 1 then
        classification ← '[3] Same NUTS2: direct ROAD,IWW'
    else if orig_rail_term ≥ 1 and dest_rail_term ≥ 1 and orig_iww_term ≥ 1 and dest_iww_term ≥ 1 then
        classification ← '[4] Above 150km: MM ROAD,RAIL,IWW'
    else if orig_iww_term ≥ 1 and dest_iww_term ≥ 1 then
        classification ← '[4] Above 150km: MM ROAD,IWW'
    else if orig_rail_term ≥ 1 and dest_rail_term ≥ 1 then
        classification ← '[4] Above 150km: MM ROAD,RAIL'
    else
        classification ← '[4] Above 150km: MM'
    end if
end for
```

Figure 12. *Pseudocode of PC-commodity Pair Classification Logic*

Both assumptions are based on a study conducted by Pienaar (2013) that found trips under 150 km, road transport is generally more cost-effective than rail due to lower handling and terminal costs and greater flexibility. For distances over 500 km, rail becomes more economical thanks to its economies of scale. In the intermediate range (150–500 km), the cost advantage depends on factors like terminal location, drayage costs, and shipment size, with the break-even point typically occurring within this range.

For the first assumption, road transport is clearly the most likely and preferred option for shippers due to its cost-effectiveness over short distances. For the second assumption, given that the average distance is 202 km, close to the 150 km threshold, direct shipment is generally preferred. Intermodal transport is less economically viable in this range, as it requires first-mile transport to reach a rail or waterway terminal, which adds cost, distance, and travel time due to modal transfer delays. However, because some P/C regions might be equipped with well-developed rail or inland waterway infrastructure, the "direct shipment" is applied with the possibility to utilize all mode options available: road, rail, or inland waterway transport, depending on the available modes in the region. This first step produces the classification of PC-commodity pairs data as follows:

- **Pairs with only unimodal chain** alternatives, these pairs stop here.
- **Pairs with multimodal chain** possibilities, these pairs will proceed to the next steps.

Step 2: Region-to-Port Pairing Selection

This process aims to reduce the number of relevant connected ports for each region. Since a single region may be linked to multiple ports, each creating a different path alternative, limiting the number of relevant ports is essential to keep the total number of generated path alternatives within a manageable and computationally feasible range. For the port selection process, four rules are applied:

1. Take only the **top 20 nearest port** based on the distance between the origin region and the region where the port is located.

2. Select the **most relevant ports** (out of the top 20 selected ports) **based on the region-port score** calculated based on the influencing factors of port choice in Chou's model (2010). In his study, Chou used AHP to identify the most important factors of Port Choice for each carrier type:
   - Oceangoing Carriers (inter-continental shipping routes): This carrier types perceive that *Depth of Containership Berth; Port Charge, Tax, Rent, and Cost*; and *Port Loading/Discharging Efficiency*, respectively, are the most important variables they consider in choosing a port to transport their goods. Meanwhile,
   - Coasting Carriers (regional or domestic shipping routes): Consider *Hinterland Economy* as the most influential factor for them in choosing the port. Followed by *Port Charge, Tax, Rent, Cost;* and *Port Loading/Discharging Efficiency* variables.

   Since, in this research, we only include the shipment between EU countries (regional), then Coasting Carriers' important factors are more relevant for the case in this research. We'll only consider the top-most important factor, which is Hinterland Economy, in the port selection process because we don't have the data for the other two factors. Below is the importance weight for each Hinterland Economy's sub-factors as identified below:
   - Inland freight cost: 0.1551
   - Inter-modal link: 0.1401
   - Volume of import/export containers: 0.4854
   - Frequency of ship calls: 0.2193

   The port selection process will be performed by calculating the region-port score. Chou's hinterland economy's sub-factors, as well as their respective weight, will be used as the basis to calculate the score. Below is the list of data sources used to calculate the region-port score corresponding to the influencing variables mentioned in Chou's research:
   - **Inland freight cost:** Using Panteia's internal generalized cost data between regions.

- **Inter-modal link:** Using the number of available mode options (rail, road, inland waterway) in each port. The mode availability is based on Panteia's internal mode connectivity data.
- **Volume of import/export containers:** Using Eurostat's port statistics data in 2010, specifically the number of goods handled for incoming (import) and outgoing (export) flows.
- **Frequency of ship calls:** Using UN Trade and Development's Port Liner Shipping Connectivity Index (PLSCI), which indicates the port's connectivity with global liner shipping networks. According to UNCTAD (2024), the PLSCI index not only includes the port's ship call frequency but also includes the other five variables: deployed annual TEU capacity, number of regular services, number of shipping companies, the size of the largest ship, and the number of ports connected via direct liner services. To calculate the index, each component's value for a port is divided by the global average for Q1 2023. The average of these six ratios is then multiplied by 100, setting the global average PLSCI to 100 in Q1 2023.

$$\boldsymbol{region\_port\_score} = (-0.1551 * gen\_cost) + (0.1401 * \ mode\_connection) \\ +(0.4854 * port\_stats) \ + \ (0.2193 * PLSCI)$$

Region-port score then calculated using the above formulation. The scores are then transformed into proportions that represent the estimated probability of the port being chosen as the transshipment in transporting goods from and to a certain region. Only top *x* number of ports whose cumulative share >= 60% that will be included, and the rest of the port will be excluded. This framework aims to enhance the previous port selection method, which relied solely on the bounding box rule and focused only on geographical proximity. By contrast, **the new approach incorporates a port competitiveness perspective, factoring in transportation costs, connectivity, and capacity** which are the key elements that influence port choice from a shipper's standpoint.

3. **Ensure the selected ports are the ones with historical demand flow from and to the observed countries**. Ports with no historical demand, specifically based on 2010 data, will also get removed and won't be considered as in the next mode chain construction process.

4. Ensure **compatibility between each commodity's handling requirements and the handling capabilities of ports** by eliminating paths that involve ports lacking the necessary facilities for the transported commodity. Since no specific data source exists for this information, a custom database was developed through qualitative analysis. This analysis involved assessing the material type and physical form of each commodity to determine the most likely cargo type used for its transports such as container, liquid bulk, dry bulk, RoRo, or other cargo types. Then, combining that qualitative analysis with the port's cargo handling capabilities data owned by Panteia. The resulting mapping of commodities to cargo types is presented in *Appendix 3: Commodity Type Mapping* section.

After implementing the three rules, the list of regions with its relevant and most probable ports is generated. This information serves as a reference for identifying which region-port combinations can be considered when generating path alternatives for specific PC-commodity pairs in the subsequent BFS stage.

```
FOR each origin_region ∈ all_regions DO
    # Step 1: Select the 20 nearest ports based on distance
    candidate_ports ← GET top 20 nearest ports TO origin_region USING geographic distance

    # Step 2: Compute region_port_score for each candidate port
    region_port_score ← ∅
    FOR each port ∈ candidate_ports DO
        gen_cost ← GET inland freight cost FROM origin_region TO port
        mode_connection ← GET intermodal link availability TO port
        port_stats ← GET volume of import/export containers AT port
        PLSCI ← GET frequency of ship calls / connectivity index AT port
        region_port_score[port] ←
            (-0.1551 × gen_cost) +
            ( 0.1401 × mode_connection) +
            ( 0.4854 × port_stats) +
            ( 0.2193 × PLSCI)

    # Step 3: Sort candidate ports by region_port_score in descending order
    sorted_ports ← SORT candidate_ports BY region_port_score DESC

    # Step 4: Keep top x ports whose cumulative score share >= 60%
    total_score ← SUM(region_port_score.values())
    cumulative_score ← 0
    selected_ports ← ∅
    FOR port ∈ sorted_ports DO
        cumulative_score ← cumulative_score + region_port_score[port]
        selected_ports ← selected_ports ∪ {port}
        IF (cumulative_score / total_score) ≥ 0.60 THEN
            BREAK

    # Step 5: Filter out ports with no historical demand in 2010
    selected_ports ← FILTER selected_ports WHERE historical_demand_exists(port, origin_region, year = 2010)

    # Step 6: Ensure compatibility of each port with commodity handling capabilities
    FOR port ∈ selected_ports DO
        compatible ← CHECK IF port can handle required commodity type
        IF ¬compatible THEN
            selected_ports ← selected_ports \ {port}

    region_port_pairs[origin_region] ← selected_ports
```

Figure 13. *Pseudo Code of Port Selection Process in Mode Chain Construction Heuristics Phase*

4.2.2.2 Graph Search Algorithm Construction

The mode chain data construction is conducted using a BFS-like logic approach with custom expansion rules, since the existing Mode Chain Builder requires quite a lot of customizations to comply with NEAC's data format requirements. Unlike standard BFS which traverses all reachable nodes layer-by-layer, this method constrains the search by the following rules:

- Paths are built by incrementally exploring 1-leg, 2-leg, and 3-leg route options between origin and destination nodes, with a maximum number of legs (depth) = 3.
- Two edges with the same mode type should be set as 1 leg.
- Rail, inland waterway, and sea transport modes can be used if only both PC regions have the same terminal and port infrastructure availability (see Figure 14).

The mode chain data construction relies entirely on the existing input datasets from the Mode Chain Builder. These include the list of PC-commodity pairs, the terminal and port availability per region, and the mode connections between regions. Within the context of the BFS algorithm, the PC-commodity pairs serve as the origin-destination references for path construction, the terminal and port data define potential intermediate nodes, and the mode connection data represents the feasible edges linking different regions. However, before executing the BFS algorithm, certain adjustments must be made to the input data to ensure consistency with the output from the heuristic stage:

- Only include PC-commodity pairs with valid multimodal chain possibilities as input.
- Update the port connection availability based on the results of the port selection process, ensuring that only ports listed in the sorted region-port list from the heuristic stage are used as inputs for BFS.

After finalizing the input datasets, the mode chain construction process using BFS can start. This mode chain construction model is developed to construct a dataset of valid multimodal transport chains (mode chains) between production and consumption zones for specific commodities.

```
## Initialize path list
valid_paths ← ∅
FOR each origin, destination ∈ region_pairs DO

    ## === 1-leg path ===
    IF edge_exists(origin, destination) THEN
        FOR each edge_data ∈ edge(origin, destination) DO
            mode ← edge_data["mode"]
            IF has_terminal(origin, mode) ∧ has_terminal(destination, mode) THEN
                valid_paths ← valid_paths ∪ {(origin, "", "", destination, mode, "", "", "", "", False)}

    ## === 2-leg path ===
    FOR mid ∈ successors(origin) DO
        IF mid = destination ∨ ¬edge_exists(mid, destination) THEN
            CONTINUE
        FOR edge1 ∈ edge(origin, mid), edge2 ∈ edge(mid, destination) DO
            m1 ← edge1["mode"]
            m2 ← edge2["mode"]
            IF m1 ≠ m2 ∧ all terminals exist for:
                (origin, m1), (mid, m1), (mid, m2), (destination, m2) THEN
                valid_paths ← valid_paths ∪ {(origin, "", "", destination, m1, m2, "", mid, "", False)}

    ## === 3-leg path ===
    FOR mid1 ∈ successors(origin) DO
        IF ¬has_sea_port(mid1) THEN
            CONTINUE
        FOR mid2 ∈ successors(mid1) DO
            IF ¬has_sea_port(mid2) ∨ ¬edge_exists(mid2, destination) THEN
                CONTINUE
            FOR edge1 ∈ edge(origin, mid1),
                edge2 ∈ edge(mid1, mid2),
                edge3 ∈ edge(mid2, destination) DO
                m1 ← edge1["mode"]
                m2 ← edge2["mode"]
                m3 ← edge3["mode"]
                IF |{m1, m2, m3}| = 3 ∧ all(
                    has_terminal(r, m) FOR (r, m) ∈
                    [(origin, m1), (mid1, m1), (mid1, m2), (mid2, m2), (mid2, m3), (destination, m3)]
                ) THEN
                    valid_paths ← valid_paths ∪ {(origin, mid1, mid2, destination, m1, m2, m3, mid1, mid2, True)}

RETURN valid_paths
```

Figure 14. *Snapshot of Customized Rules regarding Terminal Availability for BFS Model in Python*

The algorithm searches for valid transportation paths that consist of one, two, or three sequential legs, ensuring that each segment is operationally feasible based on the mode and terminal compatibility at the corresponding nodes. The search considers mode changes and transshipment zones only if the infrastructure allows it. Each discovered path is recorded as a mode chain and enriched with intermediate nodes when the path involves two or three legs. These intermediate nodes correspond to transshipment points, particularly when the second leg involves sea transport.

After constructing the mode chain alternatives for the multimodal PC-commodity pairs, the resulting data is reviewed to ensure consistency with the predefined rules established during the heuristic stage. This review involves the following validation steps, and any paths that fail to meet these criteria will be removed:

1. Commodity–port cargo handling compatibility check
2. Region–port pairing validity check

Finally, the cleaned multimodal mode chain data is combined with the unimodal mode chain data to form a complete dataset, which will serve as input for the next phase of the model development process: demand share estimation.

4.2.3 Results and Model Validation

A total of 1,839,820 mode chain alternatives were generated from all PC-commodity pairs between Belgium and the Netherlands, as well as region pairs within each country, as shown in Table 5. International flows produced more chain alternatives per pair due to the involvement of maritime transport and the availability of multiple port combinations. The detailed output of the BFS algorithm is presented in Table 6, which illustrates several possible mode chains for a given PC-commodity pair. The list includes two direct paths, while the remaining options are multimodal chains involving various combinations of origin and destination ports, as well as different transport modes for the first and third legs.

Table 5. *Mode Chain Database Construction Result Statistics*

ProdZone Country	ConsZone Country	Total PC-Comm Pair	Total Chain	Avg Chain per Pair
Belgium	Belgium	37.908	88.622	2,3
Belgium	Netherlands	53.419	814.706	15,3
Netherlands	Belgium	54.986	810.776	14,7
Netherlands	Netherlands	32.301	125.716	3,9
TOTAL		**178.614**	**1.839.820**	**10,3**

Table 6. *Example of Mode Chain Alternatives for a PC-commodity Pair Generated from BFS*

ProdZoneID	node2	node3	ConsZoneID	leg1_mode	leg2_mode	leg3_mode	transshipment_zone1	transshipment_zone2	commodity_code	tonnes
124020206	124030305.(	102020501.	102020501	1	4.0	3.0	p124030305	p102020501	2	252
124020206	124030305.(	102020501.	102020501	1	4.0	2.0	p124030305	p102020501	2	252
124020206	124030305.(	112030001.	102020501	3	4.0	1.0	p124030305	p112030001	2	252
124020206	124030305.(	112030001.	102020501	3	4.0	2.0	p124030305	p112030001	2	252
124020206	124030305.(	112030001.	102020501	1	4.0	3.0	p124030305	p112030001	2	252
124020206	124030305.(	112030001.	102020501	1	4.0	3.0	p124030305	p112030001	2	252
124020206	124030305.(	112030001.	102020501	1	4.0	2.0	p124030305	p112030001	2	252
124020206	102020304.(	102020101.	102020501	3	4.0	2.0	p102020304	p102020101	2	252
124020206	102020304.(	102020101.	102020501	3	4.0	1.0	p102020304	p102020101	2	252
124020206	102020304.(	102020101.	102020501	1	4.0	3.0	p102020304	p102020101	2	252
124020206	102020304.(	102020101.	102020501	1	4.0	2.0	p102020304	p102020101	2	252
124020206	102020304.(	102020501.	102020501	3	4.0	2.0	p102020304	p102020501	2	252
124020206	102020304.(	102020501.	102020501	3	4.0	1.0	p102020304	p102020501	2	252
124020206	102020304.(	102020501.	102020501	1	4.0	3.0	p102020304	p102020501	2	252
124020206	102020304.(	102020501.	102020501	1	4.0	2.0	p102020304	p102020501	2	252
124020206	102020304.(	112030001.	102020501	3	4.0	1.0	p102020304	p112030001	2	252
124020206	102020304.(	112030001.	102020501	3	4.0	2.0	p102020304	p112030001	2	252
124020206	102020304.(	112030001.	102020501	1	4.0	3.0	p102020304	p112030001	2	252
124020206	102020304.(	112030001.	102020501	1	4.0	3.0	p102020304	p112030001	2	252
124020206	102020304.(	112030001.	102020501	1	4.0	2.0	p102020304	p112030001	2	252
124020206			102020501	3					2	252
124020206			102020501	1					2	252

Model validation is carried out to ensure that the constructed mode chains align with key information required for generating a reliable mode chain dataset. The validation process uses a sample set consisting of three-leg chains, which are evaluated based on three critical elements, as detailed in the following section. A total of 82,385 PC-commodity pairs were selected as the validation sample.

1. **Mode connection availability**
   The first aspect evaluated to ensure the model's validity is the consistency between the assigned mode type (for a given leg) and the availability of that mode between the connected origin and destination regions. Out of approximately 1.8 million legs in the sampled mode chains, all show perfect alignment between the assigned mode and the mode connection availability data. The validation results are presented in.Table 7.

Table 7. *Mode Connection Validity Results*

leg_seq	path_count	matched	unmatched
leg1	1.861.408	1.861.408	-
leg2	1.861.408	1.861.408	-
leg3	1.861.408	1.861.408	-

2. **Rail and inland waterway terminal availability**
   The second aspect assessed in the model validation is the consistency between the selected terminals and their actual availability based on the NEAC terminal data. From a total of 1,945,991 terminal-related legs in the sampled mode chains, all legs were successfully matched with the terminal availability records, indicating full compliance with the underlying data. As shown in Table 8, there were zero

unmatched cases, confirming the model was accurately incorporating terminal infrastructure constraints.

Table 8. *Terminal Availability Validity Results*

total_edge	matched	unmatched
1.945.991	1.945.991	-

3. **Port availability**

   The third aspect of the model validation focuses on verifying the availability of ports in the regions assigned as origin or destination for maritime transport legs. This ensures that the ports used in the multimodal chains are actually present and capable of serving the specified commodity in the respective regions. Out of 2,254,340 relevant edges examined in the sampled mode chains, all were successfully matched with valid port availability records. As illustrated in Table 9, the result confirms a 100% match rate, reinforcing the reliability of the model in assigning feasible port locations for maritime transport.

Table 9. *Port Availability Validity Results*

total_edge	matched	unmatched
2.254.340	2.254.340	-

The model validity results confirm that the constructed mode chain construction model using combination of heuristics and BFS algorithms are fully aligned with the given infrastructure availability data. All assigned mode legs, terminal facilities, and port locations were validated with 100% match rates, indicating that the model reliably generates realistic transport paths. Thus, the resulted mode chain datasets can safely proceed to the next phase, demand share estimation process.

4.3 Demand Share Prediction Phase

4.3.1 Data Preparation

Several datasets are used as input for the demand share estimation phase, all of which are externally sourced from Eurostat. These datasets were downloaded and compiled into Excel files, with their formats adjusted to meet the input requirements for the Expectation-Maximization approach. Table 10 below shows an overview of the input datasets for the Demand Share Estimation phase.

Some datasets require special handling, particularly Dataset 1, which contain historical bilateral demand between countries for each transport mode. Since each country reports both its imports from and exports to partner countries, the data often includes two versions of the same trade flow. For example, the Netherlands may report rail exports to Belgium as X_1 thousand tonnes and imports from Belgium as Y_1, while Belgium reports its rail exports to the Netherlands as X_2 and imports from the Netherlands as Y_2. Ideally, $X_1 = Y_2$ and $X_2 = Y_1$, but discrepancies often occur. To resolve this, the import data from the reporting country is used as the reference, based on the assumption that incoming goods are more likely to be inspected and accurately recorded. If import data is unavailable, the available export data is used instead. The detailed information regarding data type and Eurostat's datasets used as transport flow ground truth data for the EM phase can be seen in Table 11.

Another issue arose with the Netherlands' intercountry demand flow data for rail transport in 2010, which was missing from the Eurostat database. Only data from 2013 onward were available. To address this, a back casting approach was applied using three forecasting methods: linear regression, Holt's Winter method, and ARIMA. Among these, Holt's Winter method produced the lowest error. Therefore, the back casted 2010 values generated using this technique were used to fill in the missing rail transport data for the Netherlands.

Table 10. *List of input data for demand share estimation phase*

No	Dataset Name	Description	Source	Eurostat's Data Code	Available Information (column list)
1	***input_01transport_between_country***	historical data of demand flow between country using each transport mode (rail, road, inland waterway)	Eurostat	*rail_go_intgong* *rail_go_intcmgn* *rail_go_typepas* *iww_go_atygofl* *road_go_ia_ugtt* *road_go_ia_lgtt* *road_go_na_ru3g* *road_go_na_rl3g* *mar_go_am_nl* *mar_go_am_be*	Mode, From, To, k_tonnes, data_type
2	***input_02port_statistics***	historical goods inflow and outflow data recorded at certain ports, sent or received from various EU countries	Eurostat	*mar_go_am_nl* *mar_go_am_be*	port_id, port_wn_id, level, country, port_name_neac, port_name_eustat, partner_country, in_liq_bulk, in_dry_bulk, in_container, in_roro, in_other_cargo, in_total, out_liq_bulk, out_dry_bulk, out_container, out_roro, out_other_cargo, out_total
6	***input_nlbe_chains_with_yinit_mnl3***	list of feasibile mode chains alternatives for each PC-commodity pair	*Mode Chain Construction stage*	-	ProdZoneID, ConsZoneID, transshipment_zone1,transshipment_zone2, leg1_mode, leg2_mode, leg3_mode, commoditytonnes, y_init

Table 11. *Mode Transport Flow Data used as the first EM's Ground Truth Data*

mode	origin	destination	tonnes	data_type	data_code
1	Netherlands	Belgium	18,872,000	export	road_go_ia_lgtt
1	Belgium	Netherlands	16,319,000	import	road_go_ia_ugtt
1	Belgium	Belgium	231,488,000	import	road_go_na_ru3g
1	Netherlands	Netherlands	520,606,000	export	road_go_na_rl3g
2	Belgium	Netherlands	1,941,000	import	rail_go_intgong
2	Netherlands	Belgium	1,990,000	export	rail_go_intcmgn
2	Netherlands	Netherlands	3,836,000	-	rail_go_typepas
2	Belgium	Belgium	18,152,000	-	rail_go_typepas
3	Belgium	Belgium	46,550,000	export	iww_go_atygofl
3	Belgium	Netherlands	26,908,000	export	iww_go_atygofl
3	Netherlands	Belgium	42,345,000	export	iww_go_atygofl
3	Netherlands	Netherlands	103,327,000	export	iww_go_atygofl
4	Netherlands	Belgium	2,032,000	import	mar_go_am_nl, mar_go_am_be
4	Belgium	Netherlands	1,713,000	import	mar_go_am_nl, mar_go_am_be

4.3.2 Model Development

Once a complete dataset of mode chain alternatives for all PC-commodity pairs has been constructed, the next step is to allocate the known goods flow for each pair across the available path alternatives. The process starts with dataset preparation to format the required data input format, then use the formatted data to the next main steps: initialization, model development and hyperparameter tuning,

Step 1: Dataset Preparation

The complete mode chain dataset produced by the mode chain construction phase will become the major input for this phase. Before using the data to build the EM model, several data prep-processing steps are needed, such as:

- Adding country name for intermediate node (transshipment zone) for each generated path that involves one. This step is very important since the demand flow aggregation process in the EM's loss calculation process will be performed at the country level.
- Adding impedance data, especially transport time and cost for each leg in the path. This information will be used to calculate the path's score, which will be the basis for determining the top 7 paths (within the same PC-commodity pair group) with the highest path score.

Path score is calculated based on the Analytic Hierarchy Process (AHP) study conducted by Lu and Wang (2022) that examines the estimated weight of travel cost, time, and route risk factors in influencing the decision of transport mode and path in a multimodal freight transportation network setting, which includes seaway, highway, and railway transportation (Lu & Wang, 2022). The coefficients for transportation cost (C), transportation risk (R), and transportation time (T) were determined using AHP after considering expert scoring to establish preferences for these factors. The coefficients represent the perceived importance or weight factor in the decision-making process of the route choice.

The weights assigned were 0.222 for transportation cost, 0.667 for transportation risk, and 0.111 for transportation time. However, since the risk data is not available, only travel time and cost will be considered to calculate the path's score in this study. So, the relative importance of these two factors will also be adjusted, forming a new formula as follows:

$$path\ score = \frac{1}{0.67\,C + 0.33\,T}$$

The path score is calculated as the inverse of a weighted sum of cost and time to ensure that the lower-cost and shorter-duration paths receive higher scores. The applied weights reflect that cost is considered twice as important as time in the path scoring, consistent with the initial weighting scheme from Lu and Wang's model, which assigned a value of 0.222 to cost and 0.111 to time. It emphasizes cost as the dominant factor influencing path preference.

To reduce computational load, the total path alternatives are reduced by using the calculated path score information to filter out paths with a low probability of being chosen (indicated by a low path score). For each PC-commodity pair, **only the top 7 paths with the highest path score are retained as alternatives** for the next stage of EM modeling.

Step 2: Initial Value Calculation

The Expectation-Maximization algorithm requires an initial value as a reference to shape its prediction. The initial value is crucial and influential to the learning process. A good initial value could help the model learn faster and easier to reach convergence. As explained in the section 3.3.1, to determine the best initial values, three approaches are implemented: AHP, ESD, and EQW. Below are the details on how the initial values are calculated using each of the three approaches:

1. **AHP Approach: Initial Value based on Weighted Path Score**

   In this approach, the initial distributed demand (*y_init*) for each mode chain is determined based on a path score that reflects the relative cost and time of transport. The method prioritizes paths that are more efficient in terms of cost and travel time by assigning a higher share of the total demand. The process follows these key steps:

   - **Path Score Calculation:** This step has been implemented in the previous section, *Step 1: Dataset Preparation,* to determine the path score and reduce the possible alternative paths. So, for each path, the total cost and total travel time are computed by summing the respective values across all legs. Then, the weighted path score is calculated using the formula.

- **Normalization within PC-Commodity Groups:** The sum of path scores is calculated for each PC-commodity group (*PC Group*), and each individual path's score is then normalized to compute a path share, where n is the total alternative path available within a PC group.

$$path\ share = path\ score_i \Big/ \sum_{i \in PCGroup}^{n} path\ score_i$$

- **Initial Value Assignment:** The initial demand value (*y_init*) for each path is computed by multiplying its normalized path share with the total flow (in tonnes) of the PC-commodity group.

$$y_{init} = path\ share\ * \ total\ PC\ group's\ flow$$

2. ESD Approach: Initial Value based on Minimum Leg Share

ESD stands for Eurostat's Share Data, which refers to historical mode share data provided by Eurostat as the basis to determine the path's share and initial value. This approach tries to ensure a more realistic distribution of demand across alternative mode chains by incorporating the observed modal preferences based on transport mode share data. The initial value *(y_init*) for each path is calculated based on the minimum observed mode share across its legs. This approach consists of the following steps:

- **Transport Share Lookup:** Eurostat's historical transport flow data is used as the reference. The table contains transport mode shares between country pairs for each mode, as shown by Table 12.

Table 12. *Mode Share Historical Data per Country Pair*

mode	mode name	origin	destination	historical data (in tonnes)	share
1	Road	Belgium	Belgium	231,488,000	0.78
2	Rail	Belgium	Belgium	18,152,000	0.06
3	IWW	Belgium	Belgium	46,550,000	0.16
1	Road	Netherlands	Belgium	18,872,000	0.29
2	Rail	Netherlands	Belgium	1,990,000	0.03
3	IWW	Netherlands	Belgium	42,345,000	0.65
4	Sea	Netherlands	Belgium	2,032,000	0.03
1	Road	Belgium	Netherlands	16,319,000	0.35
2	Rail	Belgium	Netherlands	1,941,000	0.04
3	IWW	Belgium	Netherlands	26,908,000	0.57
4	Sea	Belgium	Netherlands	1,713,000	0.04
1	Road	Netherlands	Netherlands	520,606,000	0.83
2	Rail	Netherlands	Netherlands	3,836,000	0.01
3	IWW	Netherlands	Netherlands	103,327,000	0.17

- **Leg-based Share Identification**: For each path, the share of its individual leg is identified by referring to the mode share historical data based on the origin country, destination country, and the mode type used. For example, if a leg between Netherlands as the production region country (ProdZone country) and Belgium as the transshipment zone 1 country (node2 country) is connected using Rail, then the probability/share for this leg is 0.03 (fifth row in the table).

- **Path Share Assignment:** After identifying the share for all paths, the share of a path is determined by using the minimum share among all legs in that path. This represents the bottleneck or weakest modal preference in the chain, ensuring that the initial demand value does not overestimate routes that contain low-share segments. For example, if a path has 3 legs with individual leg shares are 0.78, 0.04, and 0.17, then the assigned share of this path is 0.04.

- **Normalization within PC-Commodity Groups:** Same as the AHP approach, the normalization within each PC group is needed to ensure the resulting path share values sum to 1 within each group:

$$path\ share = \min(leg\ share_i) \Big/ \sum_{i \in PCGroup}^{n} \min(leg\ share_i)$$

- **Initial Value Assignment:** Same as before, the initial demand value (*y_init*) for each path is computed by multiplying its path share with the total flow of the PC-commodity group.

3. EQW Approach: Initial Value based on Equal Weight Distribution

This approach assigns the initial distributed demand (*y_init*) by evenly dividing the total demand (in tonnes) across all available paths within the same PC-commodity group. It assumes no prior preference among alternative mode chains and serves as a neutral baseline scenario for comparison with other weighted or historical data-based allocation models. So, for each group, the number of alternative paths (n) is counted, then each path in the group is assigned an equal portion of the total demand using this following formula:

$$y_{init} = \frac{total\ demand\ flow(in\ tonnes)}{n}$$

To evaluate which approaches are the best to be implemented, the initial loss figure from each approach are compared to the two sets of ground truth data: transport flow between countries data and port-level flow statistics. As seen in Table 13, ESD approach produces the lowest overall deviation between the predicted initial values and the ground truth transport flows, with total absolute deviation around 376 million tons, which is lower than both AHP (392 million) and EQW approach (394 million). The lower loss figures are seen across almost all modes and OD pairs, and the majority of % deviation under ESD are lower than AHP and EQW. This suggests that the ESD approach application produces the initial flow distribution closer to the historical data reference than the other approaches.

Table 13. *Comparison between Initial Value Prediction vs Mode Transport Flow Ground Truth Data*

					deviation			% deviation		
mode	**origin**	**destination**	**Ground Truth**	**Y_init Prediction**	**AHP**	**ESD**	**EQW**	**AHP**	**ESD**	**EQW**
Road	Netherlands	Belgium	18.872.000	50.556.844	31.684.844	46.824.056	28.985.984	168%	248%	154%
Road	Belgium	Netherlands	16.319.000	27.668.304	11.349.304	18.799.872	9.878.996	70%	115%	61%
Road	Belgium	Belgium	231.488.000	269.087.552	37.599.552	31.099.504	39.894.048	16%	13%	17%
Road	Netherlands	Netherlands	520.606.016	478.665.984	41.940.032	51.714.560	42.268.064	8%	10%	8%
Rail	Belgium	Netherlands	1.941.000	3.537.837	1.596.837	963.884	1.712.957	82%	50%	88%
Rail	Netherlands	Belgium	1.990.000	3.592.143	1.602.143	1.019.316	1.399.794	81%	51%	70%
Rail	Netherlands	Netherlands	3.836.000	13.325.484	9.489.484	2.434.841	10.291.014	247%	63%	268%
Rail	Belgium	Belgium	18.152.000	28.931.170	10.779.170	5.297.413	13.346.562	59%	29%	74%
IWW	Belgium	Belgium	46.550.000	9.966.413	36.583.587	41.837.893	34.405.820	79%	90%	74%
IWW	Belgium	Netherlands	26.908.000	1.092.694	25.815.306	24.004.562	25.295.369	96%	89%	94%
IWW	Netherlands	Belgium	42.345.000	1.207.919	41.137.081	38.166.434	40.579.783	97%	90%	96%
IWW	Netherlands	Netherlands	103.327.000	14.526.564	88.800.436	95.879.015	83.017.612	86%	93%	80%
Sea	Netherlands	Belgium	2.032.000	36.109.576	34.077.576	12.046.897	39.488.532	1677%	593%	1943%
Sea	Belgium	Netherlands	1.713.000	21.850.006	20.137.006	6.894.568	24.038.456	1176%	402%	1403%
		TOTAL	**1.036.079.016**	**960.118.489**	**392.592.358**	**376.982.814**	**394.602.990**			

Similar results were also seen from the initial value comparison against the port-level empirical data. ESD shows a clear advantage with total inflow and outflow deviation figures, for both Belgium and Netherlands ports, are significantly lower than AHP and EQW approaches. The heatmap showed in Table 14 further illustrates that the ESD deviations are more evenly distributed and less intense, particularly for key hubs like Rotterdam, Ghent, and Antwerp, indicating a more balanced and realistic assignment of flows to ports. Based on both mode transport and port statistics data comparison, the ESD approach yields the most accurate initial distribution of flows. It minimizes the deviation from ground truth at both the country and port levels, making it the most suitable approach to implement and generate the best initial values for the EM model input.

Table 14. *Comparison between Initial Value Prediction vs Port-level Flow Ground Truth Data*

Row Labels	AHP Deviation inflow	AHP Deviation outflow	ESD Deviation inflow	ESD Deviation outflow	EQW Deviation inflow	EQW Deviation outflow
Belgium	**33.155.531**	**17.560.543**	**11.686.454**	**5.627.809**	**38.404.281**	**20.972.408**
ANTWERP	22.851.173	13.776.800	7.755.988	3.975.493	26.647.073	16.525.755
GHENT	2.384.400	2.269.464	637.618	1.005.846	2.716.553	2.611.302
OSTEND, ZEEBRUGE	1.685.987	-	669.711	-	2.047.238	-
ZEEBRUGE	6.233.971	1.514.279	2.623.137	646.471	6.993.418	1.835.351
Netherlands	**17.693.114**	**31.316.475**	**5.483.425**	**10.863.111**	**20.931.941**	**35.974.932**
AMSTERDAM	2.263.292	7.815.572	930.920	3.454.134	2.674.018	8.661.268
DELFZIJL	4.042	41.556	405	54.990	4.428	40.414
DORDRECHT	5.000	29.000	5.000	29.000	5.000	29.000
HARLINGEN	29.977	43.585	10.518	13.611	32.606	47.503
MOERDIJK	6.230	15.715	122	4.260	6.360	16.329
ROTTERDAM	15.835.505	22.825.880	5.056.547	7.461.487	18.637.505	26.412.025
TERNEUZEN	36.000	86.000	36.000	86.000	36.000	86.000
VLISSINGEN	404.932	772.278	473.278	108.130	381.976	993.220

Step 3: Expectation-Maximization (EM) Algorithm Development

To estimate the demand share for each mode chain between PC-commodity pairs, the Expectation-Maximization (EM) algorithm treats the demand shares as latent variables that need to be inferred. The algorithm iteratively adjusts these estimates by using known but indirect ground truth data, aiming to find the optimal distribution of demand across path alternatives. The objective is to minimize the deviation between the aggregated values derived from the estimated shares and the actual observed aggregated data.

In this study, several sources of ground truth data from Eurostat will be used as listed below.

1. [Port Statistics] Gross weight of goods transported to/from main ports
2. [Road Transport Statistics] International Road freight transport - loaded goods in reporting country by partner country (unloading and loading activities)
3. [Road Transport Statistics] National Road freight transport by NUTS3 region (unloading and loading activities)
4. [Rail Transport Statistics] Rail International transport of goods from the reporting country to the partner country (unloading and loading activities)
5. [Rail Transport Statistics] Rail National goods transported by type of transport
6. [IWW Transport Statistics] IWW Transport by type of good (country/regional flow)
7. [Sea Transport Statistics] Gross weight of goods transported to/rom main port

The six datasets will be merged and cleaned to construct two combined and more concise datasets. They are port statistics and transport mode statistics. The transport mode statistics consist of the road, rail, IWW and sea transport, for both international flow (export and import) as well as national flow (within the country).

To better formulate the problem, Figure 15 illustrates how the ground truth data are linked to the demand share estimation for each path alternative, which is further detailed in the mathematical model in the following section. Three possible chain schemes are considered: 1-leg, 2-leg, and 3-leg paths. The assigned flow for each path is proportionally distributed across its corresponding legs. As a result, the loss function consists of two main components:

- **Transport Loss (Constraint 1):** This represents the discrepancy between the aggregated flow assigned across all legs (with a specific mode type) in the alternative paths and the known transport statistics between origin-destination country pairs.
- **Port Loss (Constraint 2):** This reflects the difference between the cumulative assigned flow entering or exiting a specific port and the actual goods volume handled by the port, based on observed port statistics.

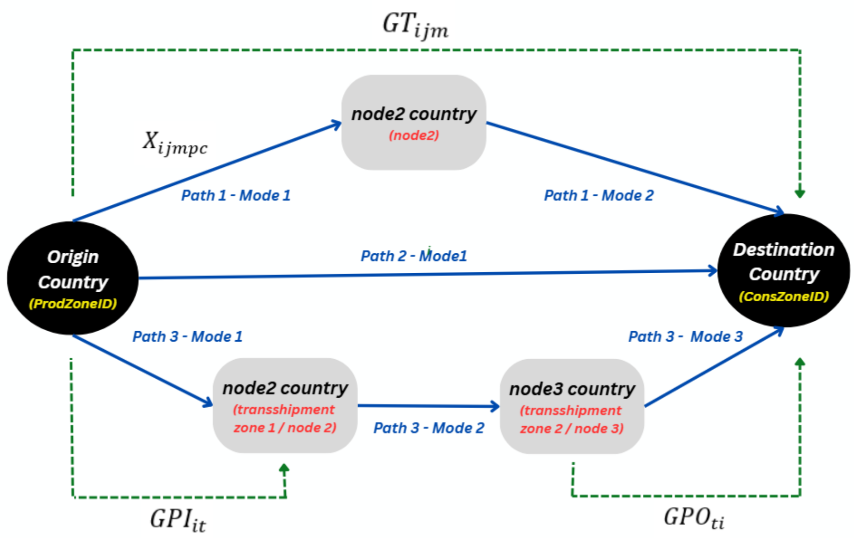

Figure 15. *Illustration of Flow Aggregation Mechanism for Loss Function in EM Algorithm*

To ensure model consistency, two additional constraints are applied. Constraint 3 enforces that the total assigned flow across all path alternatives for each PC-commodity pair equals the known initial demand. Constraint 4 ensures that the flow assigned to each leg of a path matches the total flow assigned to the full path it belongs to.

Since the loss function combines transport flow loss and port flow loss, each with significantly different magnitudes, a scaling factor is introduced during the loss calculation process. The scaling factor for both los component will be tried iteratively in order to get the best combination. This adjustment is essential to ensure the model minimizes deviations from both sets of ground truth data equally. Without scaling, the loss component with the larger magnitude would dominate the total loss, causing the model to focus disproportionately on optimizing that component, leading to imbalanced results.

The model's learning process is based on the adjusted total loss, which incorporates the scaling factor. However, the loss values presented in this report reflect the actual (unscaled) loss to provide a clearer interpretation and a more nuanced understanding of the model's performance.

Mathematical Model

Indices

$i, j \in \mathrm{N} = set\ of\ country\ nodes$

$o, d \in \mathrm{A} = set\ of\ regions\ (NUTS3 - level\ area)\ that\ belong\ to\ certain\ country\ i$

$m \ \in \mathrm{M} = set\ of\ transport\ modes$

$t \ \in \mathrm{T} = set\ of\ ports$

$p \ \in \mathrm{P} = set\ of\ alternative\ paths$

$c \ \in \mathrm{C} = set\ of\ commodity\ types$

Parameter

$GT_{ijm} = ground\ truth\ of\ transport\ flow\ using\ mode\ m\ from\ country\ i\ to\ country\ j$

$GPI_{it} = ground\ truth\ of\ incoming\ flow\ to\ port\ t\ from\ country\ i$

$GPO_{ti} = ground\ truth\ of\ outcoming\ flow\ from\ port\ t\ to\ country\ i$

$D_{odc} = total\ amount\ of\ commodity\ c\ that\ must\ be\ transported\ from\ region\ o\ to\ region\ d$

$a = transport\ flow\ loss\ factor = 1.0$

$$b = port\ flow\ factor = \frac{transport\ flow\ loss\ at\ iteration\ 1}{(port\ flow\ loss\ at\ iteration\ 1)/\ 2}$$

Variable

$Y_{odcp} = goods\ flow\ of\ commodtiy\ c\ between\ origin\ o\ and\ destination\ d\ assigned\ to\ path\ p$

$X_{ijmpc} = goods\ flow\ between\ node\ i\ and\ j\ using\ mode\ m\ along\ path\ p$

$TL = total\ aggregated\ transport\ flow\ loss\ (error)$

$PL = total\ aggregated\ port\ flow\ loss\ (error)$

Objective Function

$Minimize\ Total\ Loss = a\ \ TL + b\ \ PL$

Constraint

1. Ground Truth 1: Transport between Countries

$$transport\ flow\ loss\ (TL) = abs\left(\sum_{i,j,p} X_{ijmpc} - GT_{ijm}\right)\ for\ i,j \in N; m \in M$$

2. Ground Truth 2: Port Statistics Constraint

$$port\ inflow\ loss\ (PL) = \ abs\left(\sum_{j,m,p,c} X_{jimpc} - GPI_{it}\right)\ for\ i = t;\ m = 4;\ t \in T;\ j \in N$$

$$port\ outflow\ loss = abs\left(\sum_{j,m,p,c} X_{ijmpc} - GPO_{ti}\right)\ for\ i = t;\ \ m = 4;\ t \in T;\ j \in N$$

3. Total Assigned Demand Constraint

$$\sum_{p} Y_{odcp} = \ D_{odc}\ \ for\ o,d \in A;\ c \in C$$

4. Flow Consistency Constraint

$$X_{ijmpc} = \ Y_{odcp}\ \ for\ i,j \in N;\ m \in M;\ c \in C;\ p \in P;\ o,d \in A$$

5. Positivity Constraint

$$X_{ijmpc} \geq 0 \quad for\ i,j \in N;\ m \in M;\ c \in C;\ p \in P$$

4.3.3 Hyperparameter Tuning

Hyperparameter tuning is the process of selecting the most effective set of preset configuration parameters for a machine learning model. This step is important because it influences how the model learns during training and affects the model's performance. As explained in the section 3.3.1, four learning parameters need to be determined. In the below section, the selection process of each parameter is explained as well as the reason behind the choice.

1. Initial learning rate

To determine the most suitable learning rate for this study's model complexity, several trials were conducted. Each trial involved running the model for 10 iterations to assess whether the assigned learning rate could significantly reduce the loss, thereby enable effective learning and avoiding entrapment in local optima. The trials began with the highest learning rate 1, which was then gradually decreased until further reductions no longer yielded significant improvement, this threshold was considered the lower bound of the learning rate range. Three learning rate values were tested: 1 and 0.5. As shown by Figure 16, initial LR=0.5 results in a lower loss reduction compared to LR=1, but both settings show a stable learning process throughout the first 10 iterations. This is expected since lower learning rate will make the model take shorter steps and lead to lesser reduction. So, the trial was stopped at this point because as any value below 0.5 would likely result in even smaller improvements during early training. Therefore, initial learning rate of 1 was selected as the most appropriate initial value because it produces more significant loss reduction over iteration.

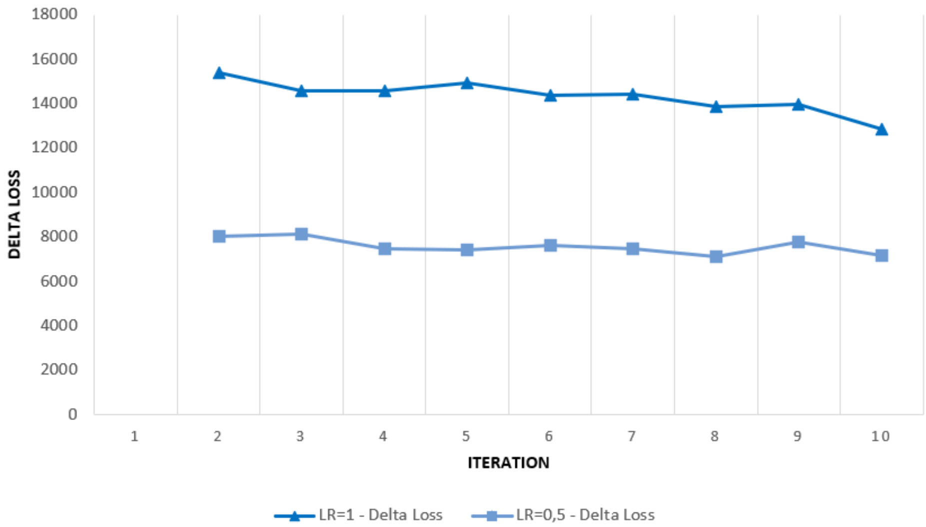

Figure 16. *Delta Trend in the First 10 Iterations for each Initial LR Settings*

2. Learning rate decay factor (gamma)

The default value for this parameter is **0.5**. This decision was motivated by the significant fluctuations observed in the EM model's learning process, particularly during the early training stages. It was therefore important to reduce the learning rate more aggressively by halving it once certain conditions were met. These fluctuations are likely caused by the large number of mode chain alternatives combined with limited ground truth data, which makes the model less stable and more prone to difficulty in identifying optimal solutions within a vast solution space.

3. Loss tolerance and patience for convergence

The loss tolerance and patience threshold for convergence are both set to the same value. The loss tolerance is set at 3000 or equal to 0.012% of the total loss (approximately 25 million before scaling), making any reduction below this threshold negligible. Patience is set to 3 iterations, allowing the model a 10% chance to discover a better solution in subsequent steps. With this configuration, convergence is declared when the model fails to achieve a loss reduction greater than 3000 over the last three iterations.

4. Maximum iteration

In case the above convergence criteria are difficult to meet, setting the maximum number of iterations can also be set as an alternative convergence criterion (Bayati, et al., 2008). Setting a maximum number of iterations is widely used to avoid infinite loops or excessive computation, especially when convergence is slow or uncertain. Even though this does not guarantee that the solution has actually converged. For this study, the maximum number of iterations is set at 500. This consideration is based on the running time of the model that requires 8-10 minutes to finish one iteration. So, by setting the maximum 500 iterations, the model needs 83 hours or 3,5 days already to reach this threshold (in case convergence hasn't reached yet).

After determining the initial parameters, the ADAM optimizer and a learning rate scheduler were configured in the model, as described in Section 3.3.1. For ADAM, the default internal settings were used: an initial learning rate (lr) of 1.0, beta_1 set to 0.9, beta_2 to 0.999, and amsgrad enabled (True). As shown by Figure 17, the learning rate scheduler was set to reduce the learning rate by a factor of 0.5 whenever the change in loss between iteration i and iteration i−1 fell below 0.015% of the current total loss. This reduction was applied immediately, with no waiting period (patience = 0).

```python
# === Optimizer and LR Scheduler ===
optimizer = Adam([y_tensor], lr=1)
scheduler = ReduceLROnPlateau(
    optimizer,
    mode='min',
    factor=0.5,
    patience=0,              # Immediate reaction to no improvement
    threshold=0.00015,          # Allow small improvements to count
    threshold_mode='rel',
    cooldown=0,              # No cooldown between reductions
    min_lr=1e-5
)
```

Figure 17. *Learning Rate Scheduler Setting*

To monitor model convergence, an early stopping mechanism was implemented based on loss stabilization (see Figure 18). The model is allowed to iterate for a minimum of 10 iterations *(min_iterations = 10)* before early stopping is considered. During each iteration, the change in loss (delta) is calculated by subtracting the current loss from the previous one (not using the absolute value). If this change is smaller than the defined tolerance, a counter (*loss_stable_counter*) is incremented. If the counter reaches 3 consecutive stable iterations after the minimum number of iterations has passed, the training loop is stopped, signaling convergence.

```
Set min_iterations = 10
If prev_loss exists:
    delta = prev_loss - curr_loss    // not using absolute value
    Print current iteration, loss, and delta

    If delta < tolerance:
        Increase loss_stable_counter by 1
    Else:
        Reset loss_stable_counter to 0

    If (iteration >= min_iterations) AND (loss_stable_counter ≥ 3):
        Print "Converged: ΔLoss < tolerance for 3 iterations"
        Stop loop
Else:
    Print current iteration and loss (no previous loss available)
```

Figure 18. *Pseudocode for Convergence Check Mechanism*

4.3.4 Loss Component's Weight Analysis

The Figure 19 illustrates the results of testing various scaling ratios between the Transport Loss (TL) and Port Loss (PL) components in the Expectation-Maximization model's loss function. The scaling ratio is necessary because of the significant difference in magnitude between the two loss components. Without scaling, the model tends to prioritize reducing the larger Transport Loss to achieve a greater overall loss reduction, which in turn causes the Port Loss to be neglected and potentially increase rather than being minimized. The goal of this test is to identify the most balanced setting that can effectively minimize both losses simultaneously. Each chart represents a different scaling ratio, and the first five iterations of the EM model were analyzed to observe the loss trends.

- **Top row:** Shows fixed scaling ratios (1:1, 1:2, and 1:4), where the weight of Port Loss relative to Transport Loss increases progressively.
- **Bottom row:** Shows dynamic scaling ratios derived from the magnitude of each loss (1:TL/PL, 1:TL/PL/2, and 1:TL/PL/4), where Port Loss is normalized relative to Transport Loss.

From the charts, it is seen that only the 1:1 and 1:2 loss ratio setting shows a downward trend for both loss components throughout the iterations, indicating a better balance between the two. Meanwhile, the other settings tend to prioritize the reduction of port loss component over the transport loss, since the port loss component has higher scaling factor than transport loss's factor that is maintained at 1. Among the tested settings, 1:1 and 1:2 ratio settings achieves the most consistent and balanced reduction across both loss components is considered optimal for the EM model.

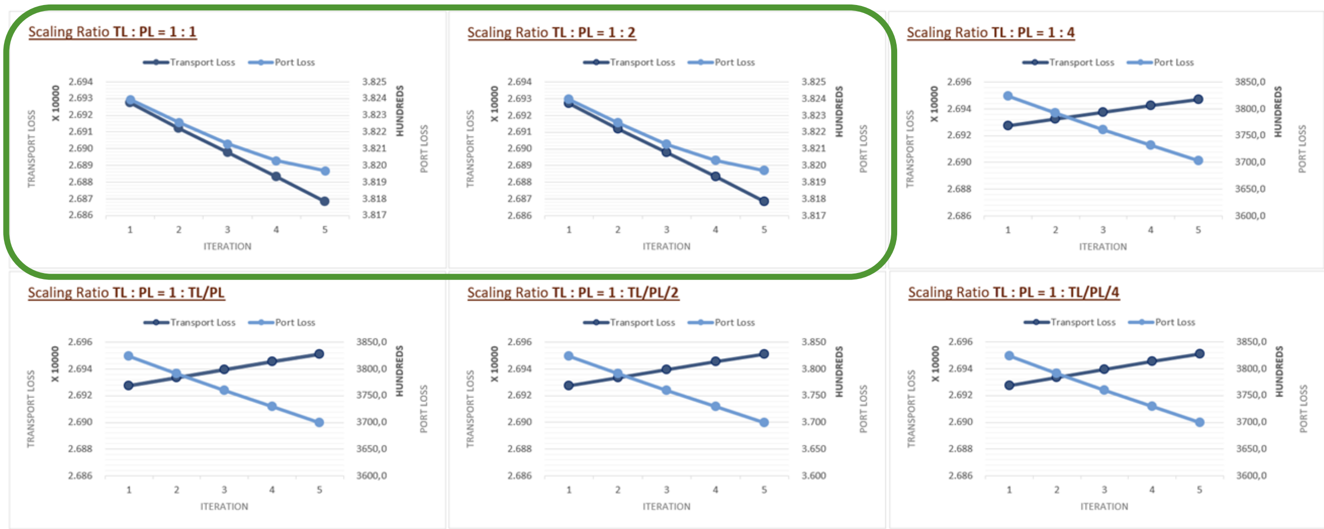

Figure 19. *Loss Reduction Trend for EM's Loss Calculation Settings*

To further investigate the continuity of this simultaneous reduction trend, the iteration results plotting is extended for both 1:1 and 1:2 settings until the 20th iterations as depicted by Figure 20. The chart shows that the simultaneous reduction trend can only be maintained until 6th iteration, after that the model can keep reducing transport loss but sacrificing the port loss (the loss is increasing). However, since transport loss has bigger losses, minimizing the transport loss is the priority. So, the 1:1 ratio setting is chosen for the full model training.

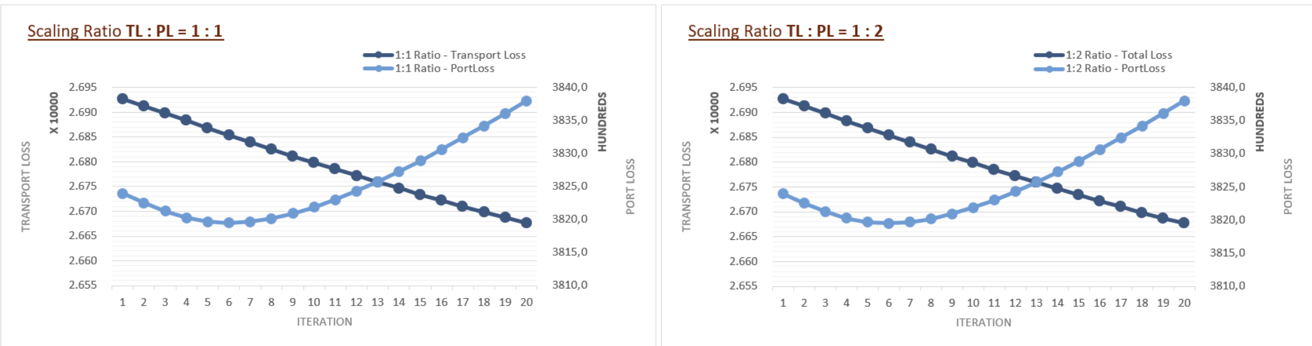

Figure 20. *Loss Reduction Trend for EM's Loss Calculation 1:1 and 1:2 Settings*

4.3.5 Results and Model Evaluation

After identifying the most suitable predefined parameters and adaptive learning strategy, the EM model is built using the defined settings. Table 15 shows the example of the EM model's outputs. The mode chain data that generated in the mode chain construction process then are completed with the *y_fin* value which represent the resulting assigned demand flow in tonnes. The total *y_fin* across chains within the same PC-commodity pair will equal to the total demand in column *tonnes*.

The initial values resulting from the ESD approach are used as the model input, since they produce a set of initial values with the least initial loss compared to the other two approaches. However, in this report, the results from AHP approach will also be presented as the comparison to assess the model performance.

Table 15. *Snapshot of Expectation-Maximization Model Outputs*

ProdZoneID	ConsZoneID	leg1_mode	leg2_mode	leg3_mode	transshipment_zone1	transshipment_zone2	commodity_code	tonnes	ProdZone_Country	ConsZone_Country	node2_country	node3_country	y_init	y_fin
102020102	124010102	2					81	340	Belgium	Netherlands			28,9362	112,054
102020102	124010102	1					81	340	Belgium	Netherlands			253,191	110,966
102020102	124010102	1	4	2	p102020101	p124040101	81	340	Belgium	Netherlands	Belgium	Netherlands	7,23404	3,24842
102020102	124010102	1	4	2	p102020101	p124010201	81	340	Belgium	Netherlands	Belgium	Netherlands	7,23404	3,24841
102020102	124010102	2	4	1	p102020101	p124010201	81	340	Belgium	Netherlands	Belgium	Netherlands	28,9362	103,987
102020102	124010102	1	4	2	p102020304	p124040101	81	340	Belgium	Netherlands	Belgium	Netherlands	7,23404	3,24842
102020102	124010102	1	4	2	p102020304	p124010201	81	340	Belgium	Netherlands	Belgium	Netherlands	7,23404	3,24841

4.3.5.1 Model Results using ESD Initial Values Approach

The assigned demand flow at chain level then are aggregated according to the origin and destination country of each edges consisted in the chain network. The flow are aggregated per mode type and the country as depicted in Table 16 in column *y_pred*. The aggregated prediction value then are compared to actual freight transport data (*ground truth*) to assess the model's performance in estimating multimodal transport flows. The Sea mode starts with initial values that results in the highest level of overestimation, followed by Road, while inland waterways (IWW) are significantly underestimated. The results show that the model was able to reduce the absolute deviation for the majority of OD-country pairs across various transport modes. It generates a consistent loss reduction in most of origin-destination country pairs, especially for Road and Rail segment, with 0.2-17% reduction in deviation (see "%*Improvement*" column). However, the increase of deviation happened slightly in IWW section, and in Sea section quite significantly.

Table 16. *EM Prediction Results Comparison against Transport Flow Ground Truth Data using ESD Approach*

mode	mode_name	origin	destination	ground truth	y_init total flow	y_init abs deviation	y_pred total flow	y_pred abs deviation	% improvement	Initial Value Status vs Ground Truth
1	Road	Belgium	Belgium	231,488,000	262,740,047	13.5%	261,148,183	12.8%	-0.7%	Bigger
1	Road	Belgium	Netherlands	16,319,000	35,117,301	115.2%	33,206,984	103.5%	-11.7%	Bigger
1	Road	Netherlands	Belgium	18,872,000	65,695,313	248.1%	63,899,588	238.6%	-9.5%	Bigger
1	Road	Netherlands	Netherlands	520,606,000	469,162,394	9.9%	472,991,154	9.1%	-0.7%	Lower
2	Rail	Belgium	Belgium	18,152,000	12,844,670	29.2%	15,905,118	12.4%	-16.9%	Lower
2	Rail	Belgium	Netherlands	1,941,000	976,328	49.7%	1,095,948	43.5%	-6.2%	Lower
2	Rail	Netherlands	Belgium	1,990,000	969,938	51.3%	1,076,790	45.9%	-5.4%	Lower
2	Rail	Netherlands	Netherlands	3,836,000	1,388,183	63.8%	1,003,493	73.8%	10.0%	Lower
3	IWW	Belgium	Belgium	46,550,000	4,708,120	89.9%	5,945,259	87.2%	-2.7%	Lower
3	IWW	Belgium	Netherlands	26,908,000	2,903,313	89.2%	2,955,528	89.0%	-0.2%	Lower
3	IWW	Netherlands	Belgium	42,345,000	4,178,472	90.1%	4,156,224	90.2%	0.1%	Lower
3	IWW	Netherlands	Netherlands	103,327,000	7,438,934	92.8%	6,512,245	93.7%	0.9%	Lower
4	Sea	Belgium	Netherlands	1,713,000	8,590,080	401.5%	9,396,339	448.5%	47.1%	Bigger
4	Sea	Netherlands	Belgium	2,032,000	14,061,504	592.0%	14,840,401	630.3%	38.3%	Bigger

The increase of deviation in Sea section is due to the trade-off between the sea and other mode types. Road and Sea mode have bigger initial values status compared to the ground truth data, while Rail and IWW have lower initial values. So, the EM model aims to decrease the share of Road and Sea, then redistribute the flow shares to Rail and IWW to increase their shares and get closer to the ground truth. However, as shown in Table 17, the generated mode chains data are dominated by 3-legs chains that consist of multiple mode types, where the use of Sea mode is always be combined with other modes as the first and third (last) leg. The decrease of Sea mode share will also lead to the decrease of other modes' shares, and vice versa. So, in order to achieve the model objective to increase the share of Rail and IWW will bring the side-effects which is the increase of Sea mode shares as well, and make the deviation of Sea mode gets bigger (while the Sea mode share is supposed to be decreased).

Table 17. *Number of Legs Distribution according to the Generated Chain Alternatives*

prev_mode	after_mode	leg_count	y_init	y_pred
3-legs scheme				
Road	Sea	153,094	12,936,543	13,674,580
Rail	Sea	103,615	5,972,392	6,870,845
IWW	Sea	69,398	3,742,650	3,691,315
Sea	Road	157,989	9,113,527	9,402,278
Sea	Rail	102,287	8,936,942	10,151,421
Sea	IWW	65,831	4,601,116	4,683,041
1-leg scheme				
Road		58,778	685,999,138	685,999,138

Unlike Sea mode, the road section whose shares also need to be decrease because of bigger initial values status against the ground truth, could alternatively achieves its goal through minimizing the share of the 1-leg mode chain alternatives since it only involves road mode only without combination with other modes. The efforts to decrease the Road mode share is easier to be done than to decrease the Sea mode share since it's conflicting with the other objectives to increase the share of Rail and IWW modes. Hence, the trade-off between Sea, Rail and IWW flow distribution makes it hard for the EM model to achieve its ultimate goal to minimize the overall loss across all ground truth data points, even though the model has reached the convergence as shown in Figure 21. The model reached the convergence at 107th iterations where the delta loss reduction was already below 3000 in the last 3 consecutive iterations.

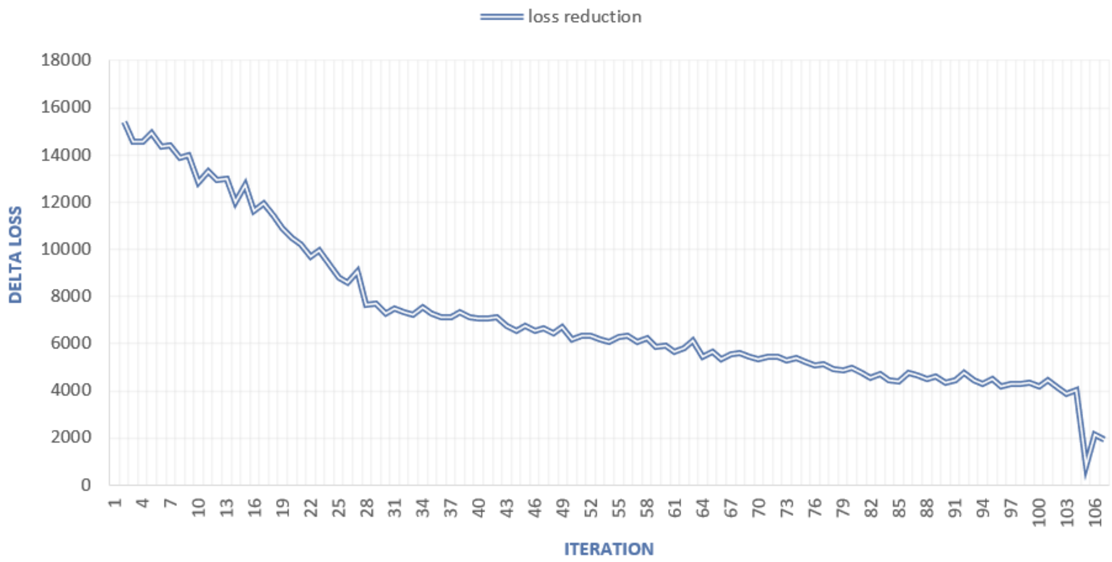

Figure 21. *ESD EM Model Iterations Results until Convergence*

The increase in Sea deviation also worsen the results of port flow prediction in comparison to the ground truth data. It causes the increases of deviation in most of port flow ground truth data points as displayed in Table 18, and only several data points are optimized as intended. As indicated by the first 20 iterations, where the port loss tends to increase while the transport loss is decreasing, the trend continues until the model convergence. As shown in Figure 22, despite the increase of deviation of port data (indicated by larger port loss compared to the initial status), the model has successfully reduce the overall loss until no significant reduction can be made (convergence). Hence, the model has reached the possible optimal states given the input data structure including the trade-off within it. This result indicates that finding the better balace between transport and port loss would be the key to improve the performance of the model in the future research or model development iterations.

Table 18. *EM Prediction Result Comparison against Port Flow Ground Truth Data using ESD Approach*

					y_init		y_pred				
direction	partner_country	port_wn_id	pot name	ground truth	total flow	abs deviation	total flow	abs deviation	Changes	%Improvement	Initial Value Status vs Ground Truth
in	Total			3,701,000	22,651,584	512%	24,236,741	555%			
in	Belgium	p124010102	DELFZIJL	4,000	15,933	298%	15,066	277%	-868	-22%	Bigger
in	Belgium	p124010201	HARLINGEN	1,000	24,042	2304%	30,873	2987%	6,831	683%	Bigger
in	Belgium	p124030206	AMSTERDAM	144,000	3,170,254	2102%	3,408,601	2267%	238,348	166%	Bigger
in	Belgium	p124030305	ROTTERDAM	1,018,000	5,242,027	415%	5,842,741	474%	600,714	59%	Bigger
in	Belgium	p124030402	VLISSINGEN	504,000	132,737	74%	94,705	81%	38,033	8%	Lower
in	Belgium	p124040101	MOERDIJK	1,000	5,087	409%	4,353	335%	-734	-73%	Bigger
in	Netherlands	p102020101	ANTWERP	891,000	8,676,823	874%	8,720,449	879%	43,626	5%	Bigger
in	Netherlands	p102020304	GHENT	695,000	1,611,087	132%	1,831,669	164%	220,582	32%	Bigger
in	Netherlands	p102020501	ZEEBRUGE	443,000	3,773,594	752%	4,288,283	868%	514,689	116%	Bigger
out	Total			3,714,000	22,651,584	510%	24,236,741	553%			
out	Belgium	p124010102	DELFZIJL	65,000	20,023	69%	17,812	73%	2,210	3%	Lower
out	Belgium	p124010201	HARLINGEN	3,000	32,056	969%	37,140	1138%	5,084	169%	Bigger
out	Belgium	p124030206	AMSTERDAM	117,000	5,800,533	4858%	6,040,467	5063%	239,934	205%	Bigger
out	Belgium	p124030305	ROTTERDAM	1,536,000	7,925,332	416%	8,532,608	456%	607,276	40%	Bigger
out	Belgium	p124030402	VLISSINGEN	98,000	274,276	180%	205,807	110%	-68,469	-70%	Bigger
out	Belgium	p124040101	MOERDIJK	9,000	9,285	3%	6,568	27%	-2,717	24%	Bigger
out	Netherlands	p102020101	ANTWERP	1,763,000	5,779,874	228%	5,915,390	236%	135,516	8%	Bigger
out	Netherlands	p102020304	GHENT	34,000	1,044,119	2971%	1,254,421	3589%	210,302	619%	Bigger
out	Netherlands	p102020501	ZEEBRUGE	89,000	1,766,088	1884%	2,226,528	2402%	460,440	517%	Bigger

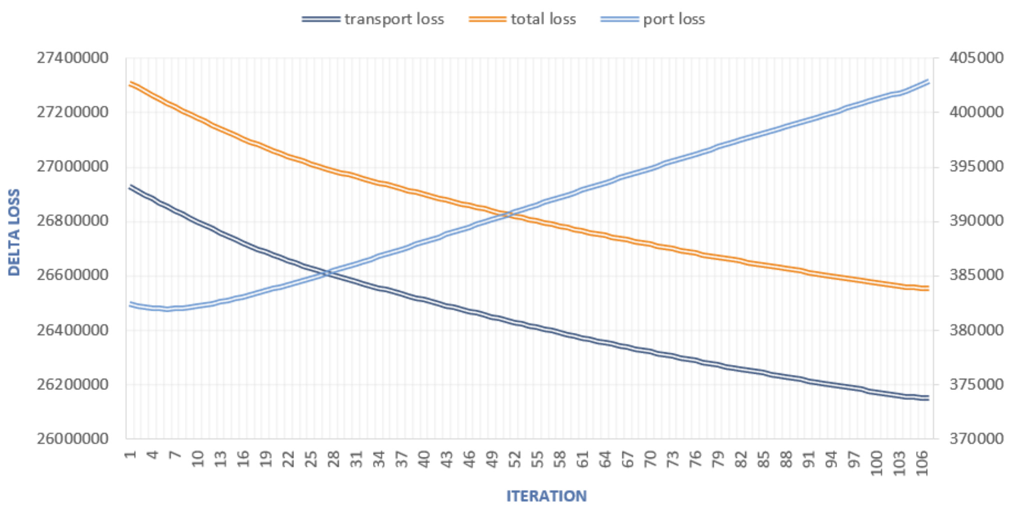

Figure 22. *Full Loss Reduction Trend until Convergence*

4.3.5.2 Model Performance Analysis

As explained in section 4.3.2, ESD has smaller initial deviations compared to the ground truth data. However, to assess the model performance, the results from both initial values strategies will be presented in this section. Apart from that, the results comparison will explain better the trade-off between the model's ability in minimizing the deviation versus the final predicted demand share distribution resulted by the model.

Table 19 shows the deviation reduction (improvement) achieved by EM model using AHP approach in comparison to the improvement made by the ESD approach in transport loss segment. The deviation reduction happens in majority of the country pairs, similar to the improvements resulted by the ESD approach. The difference is the AHP was prioritizing reducing the road share segment while reducing the Sea mode share. It leads to a better deviation reduction figure in port loss segment as displayed in Table 20, where only two port loss data points that experienced the increase in deviation while the other data points gained positive results (loss reduction). Unlike ESD approach that produces mostly negative results with loss increase in almost all port flow data points.

Table 19. *Transport Loss Deviation Reduction Comparison between AHP and ESD Approach*

mode	mode_name	origin	destination	y_pred improvement compared to y_init AHP	ESD
1	Road	Belgium	Belgium	-2.0%	-0.7%
1	Road	Belgium	Netherlands	11.1%	-11.7%
1	Road	Netherlands	Belgium	12.5%	-9.5%
1	Road	Netherlands	Netherlands	1.0%	-0.7%
2	Rail	Belgium	Belgium	-28.9%	-16.9%
2	Rail	Belgium	Netherlands	-4.6%	-6.2%
2	Rail	Netherlands	Belgium	15.2%	-5.4%
2	Rail	Netherlands	Netherlands	-93.5%	10.0%
3	IWW	Belgium	Belgium	3.7%	-2.7%
3	IWW	Belgium	Netherlands	-0.3%	-0.2%
3	IWW	Netherlands	Belgium	-0.2%	0.1%
3	IWW	Netherlands	Netherlands	3.1%	0.9%
4	Sea	Belgium	Netherlands	-402.3%	47.1%
4	Sea	Netherlands	Belgium	-365.0%	38.3%

Table 20. *Port Loss Deviation Reduction Comparison between AHP and ESD Approach*

direction	partner_country	port_wn_id	pot name	y_pred improvement compared to y_init AHP	ESD
in	Belgium	p124010102	DELFZIJL	-487%	-22%
in	Belgium	p124010201	HARLINGEN	-3257%	683%
in	Belgium	p124030206	AMSTERDAM	-1761%	166%
in	Belgium	p124030305	ROTTERDAM	-402%	59%
in	Belgium	p124030402	VLISSINGEN	40%	8%
in	Belgium	p124040101	MOERDIJK	-1318%	-73%
in	Netherlands	p102020101	ANTWERP	-615%	5%
in	Netherlands	p102020304	GHENT	-80%	32%
in	Netherlands	p102020501	ZEEBRUGE	-312%	116%
out	Belgium	p124010102	DELFZIJL	36%	3%
out	Belgium	p124010201	HARLINGEN	-1433%	169%
out	Belgium	p124030206	AMSTERDAM	-2484%	205%
out	Belgium	p124030305	ROTTERDAM	-272%	40%
out	Belgium	p124030402	VLISSINGEN	-251%	-70%
out	Belgium	p124040101	MOERDIJK	-202%	24%
out	Netherlands	p102020101	ANTWERP	-288%	8%
out	Netherlands	p102020304	GHENT	-1577%	619%
out	Netherlands	p102020501	ZEEBRUGE	-1442%	517%

However, if the final deviation figures for all transport flow (Table 17) and port flow (Table 22) data points are compared between the ESD and AHP approach, the results shows that ESD still produces the better results with smaller deviations in 75% of data points from both ground truth datasets. The ESD approach has an absolute better result especially happens for port flow ground truth data with significantly smaller deviation compared to AHP approach, despite having a worse improvement figure (Table 20) in this segment. This result indicates that ESD approach has set a strong set of initial values prior to the run of EM model that is already close to the ground truth data's demand share structure. So, these good initial values left the model with a limited room for improvement to be achieved for the model iterations. Hence, initially the model can reduce the overall loss figures by simultaneously decreasing both loss components, but beyond the iteration 13 the model has reached the saturated state. This is the state where the further loss reduction can only be achieved by sacrificing or increasing the loss of one of the loss components, in this case the port loss is sacrificed.

Table 21. Comparison of Final Prediction Results for Transport Flow Data between ESD and AHP Approach

					ESD: y_pred		AHP: y_pred	
mode	mode_name	origin	destination	ground truth	total flow	abs deviation	total flow	abs deviation
1	Road	Belgium	Belgium	231,488,000	261,148,183	**12.8%**	263,715,797	13.9%
1	Road	Belgium	Netherlands	16,319,000	33,206,984	103.5%	26,786,796	**64.1%**
1	Road	Netherlands	Belgium	18,872,000	63,899,588	238.6%	48,659,152	**157.8%**
1	Road	Netherlands	Netherlands	520,606,000	472,991,154	9.1%	473,792,399	**9.0%**
2	Rail	Belgium	Belgium	18,152,000	15,905,118	**12.4%**	22,932,412	26.3%
2	Rail	Belgium	Netherlands	1,941,000	1,095,948	43.5%	1,792,843	**7.6%**
2	Rail	Netherlands	Belgium	1,990,000	1,076,790	45.9%	1,558,155	**21.7%**
2	Rail	Netherlands	Netherlands	3,836,000	1,003,493	**73.8%**	9,734,555	153.8%
3	IWW	Belgium	Belgium	46,550,000	5,945,259	87.2%	8,123,629	**82.5%**
3	IWW	Belgium	Netherlands	26,908,000	2,955,528	**89.0%**	564,756	97.9%
3	IWW	Netherlands	Belgium	42,345,000	4,156,224	**90.2%**	626,906	98.5%
3	IWW	Netherlands	Netherlands	103,327,000	6,512,245	**93.7%**	11,347,615	89.0%
4	Sea	Belgium	Netherlands	1,713,000	9,396,339	**448.5%**	14,951,353	772.8%
4	Sea	Netherlands	Belgium	2,032,000	14,840,401	**630.3%**	28,686,509	1311.7%

Table 22. Comparison of Final Prediction Results for Port Flow Data between ESD and AHP Approach

					ESD: y_pred		AHP: y_pred	
direction	partner_country	port_wn_id	pot name	ground truth	total flow	abs deviation	total flow	abs deviation
in	**Total**			**3,701,000**	**24,236,741**	**554.9%**	**43,637,863**	
in	Belgium	p124010102	DELFZIJL	4,000	15,066	**276.6%**	15,666	291.6%
in	Belgium	p124010201	HARLINGEN	1,000	30,873	**2987.3%**	32,728	3172.8%
in	Belgium	p124030206	AMSTERDAM	144,000	3,408,601	**2267.1%**	5,142,096	3470.9%
in	Belgium	p124030305	ROTTERDAM	1,018,000	5,842,741	**473.9%**	9,484,214	831.7%
in	Belgium	p124030402	VLISSINGEN	504,000	94,705	81.2%	255,728	**49.3%**
in	Belgium	p124040101	MOERDIJK	1,000	4,353	**335.3%**	20,921	1992.1%
in	Netherlands	p102020101	ANTWERP	891,000	8,720,449	**878.7%**	18,390,690	1964.1%
in	Netherlands	p102020304	GHENT	695,000	1,831,669	**163.5%**	3,230,457	364.8%
in	Netherlands	p102020501	ZEEBRUGE	443,000	4,288,283	**868.0%**	7,065,362	1494.9%
out	**Total**			**3,714,000**	**24,236,741**	**552.6%**	**43,637,863**	
out	Belgium	p124010102	DELFZIJL	65,000	17,812	72.6%	24,302	**62.6%**
out	Belgium	p124010201	HARLINGEN	3,000	37,140	**1138.0%**	47,151	1471.7%
out	Belgium	p124030206	AMSTERDAM	117,000	6,040,467	**5062.8%**	11,032,261	9329.3%
out	Belgium	p124030305	ROTTERDAM	1,536,000	8,532,608	**455.5%**	16,711,413	988.0%
out	Belgium	p124030402	VLISSINGEN	98,000	205,807	**110.0%**	838,682	755.8%
out	Belgium	p124040101	MOERDIJK	9,000	6,568	**27.0%**	32,700	263.3%
out	Netherlands	p102020101	ANTWERP	1,763,000	5,915,390	**235.5%**	10,613,988	502.0%
out	Netherlands	p102020304	GHENT	34,000	1,254,421	**3589.5%**	1,779,279	5133.2%
out	Netherlands	p102020501	ZEEBRUGE	89,000	2,226,528	**2401.7%**	2,558,086	2774.3%

The chart in Figure 23 illustrates the EM model's learning process in minimizing the Transport Loss and Port Loss trends over iterations for two different initial value approaches, AHP and ESD. The ESD approach converges faster, reaching a stable point around the 107th iteration with a final Transport Loss of 26,151,412 and a Port Loss of 402,336. This indicates that the model's learning has stabilized and can no longer significantly reduce the losses. In contrast, the AHP approach shows a slower convergence trend up to the 460th iteration, and the model run was halted due to time constraints before reaching full convergence. Its latest recorded losses were 27,182,482 for Transport Loss and 1,296,938 for Port Loss. However, the learning trend indicates a gradual stabilization in the loss over time. The loss gradually decreased, suggesting that the model progressed toward potential convergence in the next several iterations when the total loss figures reach the 26,1 million level for transport loss and 402,3 thousand level for port loss. This comparison highlights that the ESD initial value approach leads to better and faster loss minimization compared to the AHP approach, and the EM model with ESD as the initial values strategies is the optimal model settings for MCB's demand share estimation system.

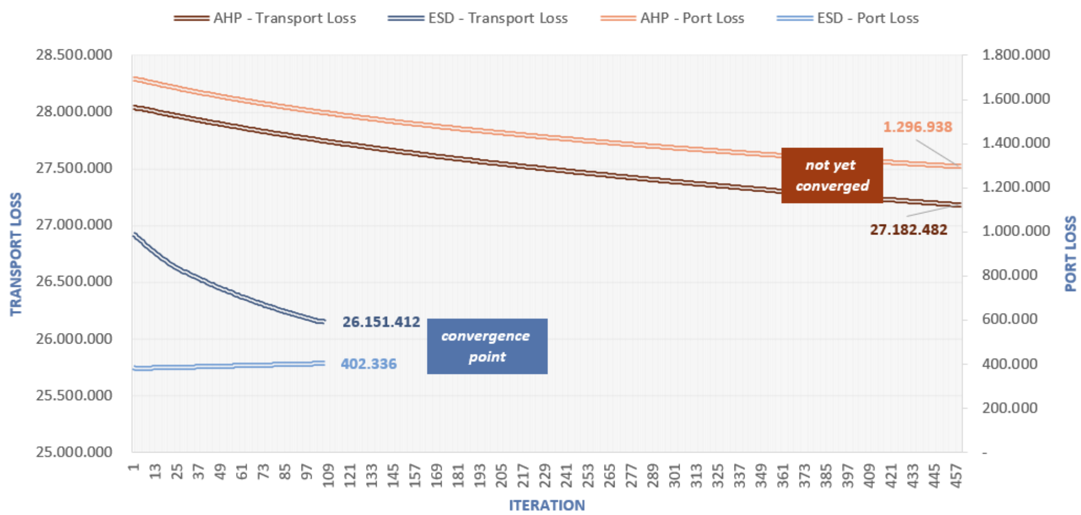

__Figure 23.__ Loss Reduction Trend Comparison between ESD and AHP Approach

4.3.5.3 Model Performance Justification

This substantial reduction indicates that the EM model successfully optimized the estimates by consistently minimizing initial deviations across all ground truth data points available. It indicates the model's ability to learn and adjust toward the ground truth, even when starting from suboptimal predictions. The "green" improvement values further confirms that the majority of final aggregated predictions are getting closer to actual values. This validates that the EM algorithm is functioning as intended, iteratively refining estimates and reducing mismatches, highlighting its capacity to improve performance under constraints.

As discussed earlier, the EM model with ESD approach already reaches the convergence state. When the model reaches convergence state, it already reaches the internal consistency, indicating that the model has minimized deviations as much as possible given the input constraints, initial Y distribution, and OD pair structure. The strong reduction in deviation across country pairs confirms that the model effectively learned from iterative updates. The remaining gap between predicted values and ground truth suggests that while the model has performed well within its design limits, external factors likely hinder its ability to achieve a closer match.

Several external factors may limit the model's ability to achieve higher accuracy:

1. **Incomplete or Biased Path Alternatives**
   The generated mode chains may not fully reflect realistic alternatives present in the real world. Since the mode chain data for the EM model input was constructed based on infrastructure availability data, and it has also been validated in Section 4.2.3, any inaccuracies in the infrastructure data could introduce bias to the final EM prediction. Given that IWW is the most consistently underestimated mode, reviewing the latest data on terminal availability might help mitigate this issue.

2. **Mismatch between Mode Share Assumptions and Real-World Behavior**
   The mode chain filtering described in Section 4.3.1 was based on probabilities derived from weighted factors estimation in Lu and Wang (2022)'s AHP results. While theoretically sound, this approach may not fully capture the real-world dynamics of freight transport. Critical factors (e.g., market conditions, actual tariffs, regulatory preferences) strongly influence mode choice but are not incorporated into the current model. For example, even if rail infrastructure is available, certain flows may still prefer road transport due to greater flexibility, reliability, or existing commercial agreements, leading to potential bias in the predicted probability of alternative paths. Hence, the exclusion of certain OD alternatives could also contribute to the remaining gap since it may remove relevant yet lower-frequency

alternatives, reducing the diversity of mode chains. However, in this study, the elimination step is needed due to the limited computation capability.

3. **Limited Travel Impedance Data**

   The generalized transport cost and travel time data were used as the basis in calculating the path score to perform the early path elimination as well as to estimate the path share in AHP initial value estimation strategy. But the impedance cost used in the calculation may not accurately reflect actual multimodal costs. Since multimodal paths involve multiple modes and transfer processes, they often incur additional time and handling costs that are not fully captured in the current utility model. This omission could further distort the initial probability estimates.

Chapter 5. Conclusion and Recommendations

5.1 Research Conclusion

This conclusion section summarizes the key findings of the research by addressing the research questions defined at the outset. It brings together the insights and answers to the main research question and its sub-questions, as detailed below:

1. What approaches are commonly used for constructing and predicting multimodal freight chain data?

In this study, multimodal freight chain modelling is divided into two major steps: (1) mode chain data construction and (2) demand share estimation.

- For mode chain construction, the existing commonly used approaches are optimization (SND, Short-path algorithm like Dijkstra, k-shortest path), heuristics (RCI and PIA), machine learning (DBSCAN clustering, reinforcement learning), and classical graph search algorithms (BFS and DFS).
- For demand share estimation, modern mode split modeling methods often rely on supervised machine learning techniques such as neural networks, random forests, XGBoost, or CatBoost, in addition to traditional statistical approaches like the Multinomial Logit model. However, these supervised methods require labeled data, which is typically unavailable in freight mode split problems. The unsupervised learning technique, Expectation-Maximization algorithm (EM), can also be considered due to its ability to estimate the missing data or latent variable using the available observed data.

2. Which mode chain construction approach that can effectively integrates a comprehensive port selection process and aligns cargo handling requirements with port capabilities?

The combination of BFS and heuristic methods are found to be the most effective approach for constructing mode chain datasets, as the enhanced logic required for comprehensive port selection and cargo handling alignment can only be achieved through heuristics. BFS complements this process effectively by enabling full enumeration of all possible path alternatives and offering the flexibility to integrate with heuristic techniques, allowing for a high degree of customization in the model.

3. How can the proposed mode chain construction model be developed and integrated into the existing Mode Chain Builder system?

A new mode chain construction model has been developed in Python, combining heuristic methods with the Breadth-First Search (BFS) algorithm. For the scope of this study, which focuses on Belgium and the Netherlands, the model successfully generated approximately 1.8 million valid mode chain alternatives. All generated chains were verified against freight transport infrastructure availability data, ensuring 100% validity, a key requirement for constructing realistic and reliable mode chain datasets.

4. Which machine learning techniques are most appropriate for predicting demand flow distribution in multimodal freight transport?

The semi-unsupervised learning technique, Expectation-Maximization (EM) algorithm, becomes a suitable method to implement, as it can estimate unobserved or latent variables without labelled data. While the application of EM for estimating mode chain shares in the freight modelling domain remains relatively uncommon, it offers a promising solution under data-limited conditions.

5. How can the demand flow distribution model be developed and effectively integrated with the mode chain construction process?
A demand share estimation model using Expectation-Maximization approach has been developed in Python. It's proven to be able to reduce the deviation between predicted values and ground truth data significantly given the input data constraints.

6. How accurately does the proposed model predict multimodal freight chains compared to the actual available freight data?
The newly developed multimodal freight chain model in this study has proven its ability to construct the mode chain alternatives data with 100% validity against the available transport infrastructure information. The demand share estimation model can reduce the deviation between the predicted demand share values against the transport historical data in majority of the observed data points. Even though several data points experienced an increase in deviation instead, the reduction outweighed the increase. The rise in deviations for certain flows appears to be a trade-off resulting from the model's focus on reducing errors in segments with the highest initial deviations.

Despite the model's proven ability to reduce the deviation, the existing model still left a quite significant remaining deviation between predicted values and ground truth. But the learning process of the EM model was already near the convergence, meaning the model has reached its maximum ability to minimize the deviations given the input constraints, initial Y distribution, and OD pair structure. Hence, the remaining gap indicates that while the model has performed well within its design limits, external factors likely hinder its ability to achieve a closer match. Several recommendations (elaborated in the next section) can be considered to improve the model performance in the future related or similar research.

The **final conclusion** is that machine learning (ML) can be effectively implemented in the Mode Chain Builder system, but primarily for the demand share estimation phase. ML is not well-suited for mode chain construction because most existing path-generation or route-construction approaches aim to identify the most optimal or a few best path alternatives. In multimodal freight chain modelling, however, the requirement is to generate all possible path alternatives, which aligns better with traditional graph search methods. The implementation of ML in the demand share estimation phase is achieved using the Expectation-Maximization (EM) algorithm, where the demand share per path is treated as a latent variable. Its values are iteratively estimated by referencing observed ground truth data. The EM model demonstrated its capability to refine the initial demand share estimates (initial values) by minimizing the deviation between predicted values and ground truth data, despite trade-offs between transport flow and port flow accuracy.

While the model has not fully eliminated the deviation as initially expected, it has achieved the best possible results given the current input data structure. Further reductions in deviation can only be achieved by improving several external factors: access to more accurate infrastructure connectivity and travel impedance data to build more realistic mode chain alternatives, enhanced computational capacity to allow the full model to run without multiple simplification steps, and a more comprehensive understanding of freight path choice behaviour to support the development of better estimation models.

5.2 Research Limitation and Recommendations

Based on the model results and evaluations, the future similar research using Expectation-Maximization algorithm for the model development can focus the effort to find the better balace between transport and port loss. Since in the result of this study shows a huge trade-off between the two loss components, so finding a better balance between the two would be the key to improve the performance of the model in the future research or model development iterations. Apart from that, here are the other limitations of this research along with some recommendations for future research.

1. Incomplete or Biased Path Alternatives

The mode chain dataset used as input for the EM model is constructed based on infrastructure availability data. While this data was validated earlier in the study, it may still not fully reflect the complete range of realistic path alternatives in real-world freight movement. Particularly, inland waterway transport (IWW) appears consistently underestimated, possibly due to outdated or missing information on terminal availability or waterway connectivity. Future research should update and expand the infrastructure database, particularly for underrepresented modes like IWW. Incorporating recent datasets or open-source logistics platforms (e.g., TEN-T, Eurostat intermodal maps) can improve path generation realism. Furthermore, integrating expert validation or using the most updated data could enhance the representativeness of the mode chains.

2. Mode Choice Behavior Not Fully Captured

The initial filtering of mode chains was based on mode share probabilities derived from Lu and Wang (2022)'s AHP results, which rely on predefined weighted factors. However, this method may not capture actual freight decision-making dynamics such as pricing strategies, contract arrangements, or preferences for certain modes due to reliability or flexibility. As a result, certain feasible but less frequent mode chains may have been prematurely excluded. Future research should explore data-driven path choice modeling based on actual observed freight transport behavior. In particular, research on discrete choice modeling or stated preference surveys specifically targeting multimodal path or route choices would be highly beneficial. This is important because most existing transport behavior studies tend to focus on individual mode choice, rather than on entire transport chains that involve multiple modes or legs. Developing models that capture the full path decision-making process would yield more realistic and reliable estimates of mode chain probabilities.

3. Limited Travel Impedance Accuracy

The utility scores used for early filtering and for the AHP-based initial value estimation rely on generalized transport costs and travel times. However, multimodal transport includes intermodal transfers, waiting times, and handling costs that are not fully reflected in the current impedance values. This simplification could lead to inaccurate scoring and biased selection of paths, especially for complex chains like rail-sea-road combinations. Future research should incorporate detailed multimodal impedance models that account for transfer penalties, handling times, and mode-specific costs. This could be achieved by collecting such data by accessing datasets from transport studies that include intermodal transfer data.

4. Computational Resource Constraints

Due to limited computational time and resources, the EM model could not be trained using the best initial values (from the ESD approach). The training was halted at iteration 460 before full convergence. While the delta loss had stabilized, the model could have slightly improved further if allowed to run to convergence. In addition, the path elimination step was necessary to reduce the computational load but may have removed plausible alternatives. Future research should be equipped with better computer capability. Apart from that, implementing parallel computing or using GPU-accelerated training to speed up EM iterations could also be considered to allow full convergence with better initial values. Cloud computing environments like AWS or Google Collab Pro could be explored to scale experiments.

5. EM Model Still Rarely Applied in Freight Modeling

The use of the Expectation-Maximization (EM) algorithm for freight mode split modeling remains relatively novel. While the results here show promise, this method still lacks widespread validation in transport research. Its sensitivity to initial values and potential for local minima make its robustness less predictable in other settings. Further research should benchmark EM against other semi-supervised or probabilistic modeling techniques in similar freight estimation contexts.

Appendices

Appendix 1: Scientific Paper

Machine Learning for Multimodal Freight Chain Modelling: Case Study of NEAC's Mode Chain Builder System

Eka Devi Wulandari [a], Dr. Bilge Küçük Atasoy [b], Dr. Yanan Xin [c]

[a] *Faculty of Civil Engineering and Geosciences, TU Delft, Delft, The Netherlands*
[b] *Department of Maritime and Transport Technology, TU Delft, Delft, The Netherlands*
[c] *Department of Transport and Planning, TU Delft, Delft, The Netherlands*

Abstract

Freight transport is central to global trade, and accurate multimodal chain modeling is essential for efficient planning and policy-making, yet remains challenging due to fragmented data, limited granularity, and the absence of observed chain-level datasets. Traditional heuristic approaches often fail to capture operational realities such as port selection and cargo handling, and are computationally rigid. This research develops an adaptive multimodal freight chain model that integrates port selection with cargo handling requirements and applies machine learning, in this case the Expectation-Maximization (EM) algorithm, to estimate demand flows from unstructured transport data. Applied to the NEAC Mode Chain Builder for freight movements between the Netherlands and Belgium, the model generates valid chain alternatives and reduces deviations between predicted and observed flows, particularly for sea and rail segments, while demonstrating stable convergence. Although residual deviations remain due to data limitations and behavioral uncertainties, the results show that combining graph search algorithms with unsupervised learning offers a practical and data-driven approach for building data-driven multimodal freight transport models that better reflect operational realities and observed empirical data that can be used to improve the freight transport planning and decision-making process.

Keywords: *mode chain builder, multimodal freight modelling, path generation, demand share distribution, machine learning, unsupervised learnings, expectation-maximization algorithm, graph search algorithm*

1. Introduction

1.1 Background

Over the past decades, freight volumes have grown rapidly and are projected to continue rising in the coming years (IEA, 2002). Freight transport plays a critical role in global trade and economic activity. It requires efficient and adaptable modelling systems to optimize logistics, infrastructure planning, and policy decision-making. In Europe, a significant portion of freight shipments are transported through multimodal chains rather than direct unimodal schemes. A survey conducted by French ECHO indicates that approximately 47% of freight demand is transported via multimodal transport chains, while 46% relies on a single mode, and the remaining 7% falls into other categories (Guilbault, 2008). The preference for multimodal transport chains is primarily driven by the lack of direct connections between certain origin-destination pairs, making unimodal transport (Huber, 2017). Additionally, multimodal chains are preferred because they allow for greater efficiency by leveraging the advantages of different transport modes and vehicles within the chain (Konings, et al., 2008).

Another interesting fact is also found showing that according to Eurostat, in 2025 maritime transport accounts for 67.4% of the total freight volume in the EU. Since maritime transport only connects ports, additional modes are required for first-mile connections from the origin point to the port and for last-mile delivery from the port to the final destination. In parallel, the European Union is committed to promoting a more sustainable form of mobility. This goal can be advanced through multimodal transport, which strategically combines different transport modes to maximize their individual strengths while minimizing their limitations. To support this vision, the European Commission is actively pursuing a multimodality policy by improving the integration of transport modes and ensuring interoperability at all levels of the transport system (European Union, 2023). This policy direction is likely to increase the demand for multimodal freight chain modelling in the future.

Existing approaches to multimodal freight transport modeling are mostly based on optimization methods, which aim to identify the most efficient transport chains by minimizing factors such as cost, time, or emissions. Techniques such as genetic algorithms [6], bilevel programming [7], dynamic programming [8], and heuristics have been widely applied [7], often combined with simulation to enhance realism. For example, Bok et al. developed a corridor choice model within the Netherlands' BasGoed system, applying route enumeration and a Multinomial Logit model to estimate demand shares [9]. While effective in certain contexts, these methods depend heavily on predefined assumptions and generalized costs, making them less adaptable to the complex and dynamic nature of multimodal freight flows.

The recent growing approaches applied for freight modelling are big data analytics and machine learning. So far, the applications are mainly for path design [10], demand generation prediction [11], demand forecasting [12] [13] [14] [15] [16], mode choice prediction [17] [18] [19], and demand assignment optimization [20] [21] [22], with studies consistently showing that machine learning outperforms traditional methods in predictive accuracy. However, despite this progress, the use of machine learning to estimate freight flows across entire multimodal transport chain networks remains largely unexplored. This is a critical gap, as current optimization and statistical approaches often struggle with fragmented data, high dimensionality, and the lack of observed chain-level datasets.

This research addresses this gap by investigating the integration of machine learning into the NEAC Mode Chain Builder system, which currently relies on heuristic-based path construction. The proposed approach seeks to leverage available but fragmented freight statistics to generate a predicted multimodal freight chain database. By learning hidden relationships across incomplete datasets, machine learning can help approximate how freight flows are distributed along complex modal sequences. This integration is expected to reduce dependence on rigid heuristics, improve demand flow estimation, and support more accurate, data-driven decision-making for multimodal freight transport planning.

1.2 Problem Description and Scope

The multimodal freight chain modelling process in this research will use NEAC Mode Chain Builder (MCB) as the case study object. NEAC itself is a multimodal, network-based simulation system developed to analyse freight transport flows across Europe that comprises two key components: the Mode Chain Builder and the NEAC model itself. The Mode Chain Builder model is used to construct a mode chain database that represents possible multimodal transport chains between production and consumption point (region) pairs. These chains represent the sequences of transport modes (involving road, rail, inland waterway, and maritime transport mode options) and terminal connections that freight can take across Europe.

Figure 1. *Modelling process in the existing Mode Chain Builder*

The mode chain database construction process within the Mode Chain Builder system comprises three major steps, as depicted by Figure 1. The model development process in this research will follow this structure as detailed below:

1. **1st Phase: Mode Chain Construction**
   The objective of this phase is to generate a set of possible mode chains for each production zone, consumption zone, and commodity type pair, based on the available facilities (nodes) and the connections between them by mode type (edges). The resulting dataset serves as a key input for the subsequent demand share estimation phase.
2. **2nd Phase: Method for Demand Share Estimation**
   This phase aims to estimate the demand share distribution across all mode chain alternatives generated in Phase 1. It also incorporates the calibration process, making the estimation more data-driven rather than treating estimation and calibration as separate steps. By integrating historical transport data from the outset, this approach is expected to reduce the calibration challenges and improve the reliability of the demand share output.

The existing Mode Chain Builder system faces a practical gap in its ability to adapt to complex, data-rich environments due to its reliance on a rigid heuristic-based approach. The current Mode Chain Builder system use pre-defined rules (heuristic) approach to generate freight transport chains by processing trade data at the national level, disaggregating it regionally, and assigning routes through a sequence of port and terminal selections governed by predefined rules. While this method has enabled large-scale estimations, it presents several limitations. For instance, the model limits the number of ports considered by applying a bounding box rule and omits factors such as cargo-type compatibility in port selection. This method, while effective for large-scale estimations, reduces the model's realism and limits its capacity to reflect real-world freight transport movements.

This study focuses on just two neighboring countries (the Netherlands and Belgium) to represent inter-country freight movement. These countries were selected because they have the two largest ports in Europe: Rotterdam (Netherlands) and Antwerp-Bruges (Belgium), which together handled a combined volume of 641 million tonnes in 2024 (Eurostat, 2024). Moreover, both countries demonstrate strong multimodal connectivity, as shown by their significant non-road modal shares: 22.4% in Belgium and 47.2% in the Netherlands (European Union, 2023). Given that this study aims to model multimodal freight chain flows, the presence of decent non-road transport connections within the country is essential.

1.3 Research Objective

This research aims to develop a multimodal freight transport chain model that incorporates a practical port selection process and considers cargo handling requirements during the mode chain construction phase. In addition, it investigates the use of machine learning to leverage the abundant, though still unstructured, freight transport data for predicting demand flows within the multimodal chain. The integration of machine learning into the model is expected to simplify and accelerate the data calibration process while improving predictive accuracy.

To achieve this aim, the study first reviews commonly used methods for constructing and predicting multimodal freight chain data and then identifies a suitable mode chain construction approach that integrates port selection with cargo handling requirements. The proposed model is implemented within the existing system architecture, after which appropriate machine learning techniques are evaluated for predicting demand flow distribution. A demand flow distribution model is then developed and integrated with the mode chain construction process. Finally, the overall performance of the proposed model is assessed by comparing its predictions with actual freight data.

2. Methodology

This research is structured around three fundamental phases: **method selection, model development, and model validation and evaluation**, as explain below:

2.1 Method Selection

To select the method, literature reviews were conducted to explore existing commonly used approaches for constucting the mode chains database (1st phase), and identify several potential machine learning techniques for distributing the demand share across generated mode chains based on the available freight transport statistics data (2nd phase).

Based on the literature review results, Breadth-First Search (BFS) is identified as the most suitable method for the Mode Chain Construction phase. BFS is preferred over other options due to its flexibility in accommodating the unique and specific requirements of the existing Mode Chain Builder system, which demands a high degree of customization in its construction logic. This flexibility also allows BFS to be combined with the Link Elimination technique to reduce computational complexity. Most importantly, BFS supports the full enumeration of all possible path alternatives, which is the core requirement of the Mode Chain Builder.

This study adopts the Breadth-First Search with Link Elimination (BFS-LE) approach, as introduced by Tahlyan and Pinjari (2020), for constructing mode chains. The link elimination step will be implemented by applying a set of pre-defined rules to remove irrelevant links, thereby reducing the algorithm's search space during the construction process. This heuristic enhancement is essential to ensure compliance with several mandatory rules associated with the input data structure, which must be respected throughout the chain construction process.

For the second phase, Demand Share Estimation, the selected method is the Expectation-Maximization algorithm. The Expectation-Maximization (EM) algorithm is selected because it is well-suited for situations where supervised instance-level data is unavailable, and only aggregate-level historical data is provided. Its ability to effectively handle latent variables makes it ideal for demand share estimation, where individual path choices are unobserved. By treating demand shares as hidden variables and using aggregated freight transport data as constraints, EM allows iterative refinement of estimates. It minimizes the deviation between predictions and observed totals, such as known totals of freight flows across regions or modes, which cannot be directly attributed to individual OD-path combinations, ensuring that the final demand share estimates are both statistically consistent and aligned with actual aggregate data.

This makes EM a robust and data-efficient approach for unsupervised learning in freight transport modeling. The overall comparison between the existing and the newly proposed methods for the end-to-end phase of Mode Chain Builder system is shown by Figure 2.

2.2. Model Development

1. First Phase: Mode Chain Construction

Figure shows the end-to-end process of mode chain database construction process that involves two main approaches: **Heuristics** and **Graph Search Algorithm** based on the Breadth-First Search (BFS) method. It also presents the six datasets used as inputs for model development, sourced from both the company's internal records and external sources. The mode chain data construction phase mainly relies on the existing Mode Chain Builder input database provided by Panteia, complemented by Eurostat's port statistics for the port selection procedure. This study does not involve creating new input datasets or modifying the existing data.

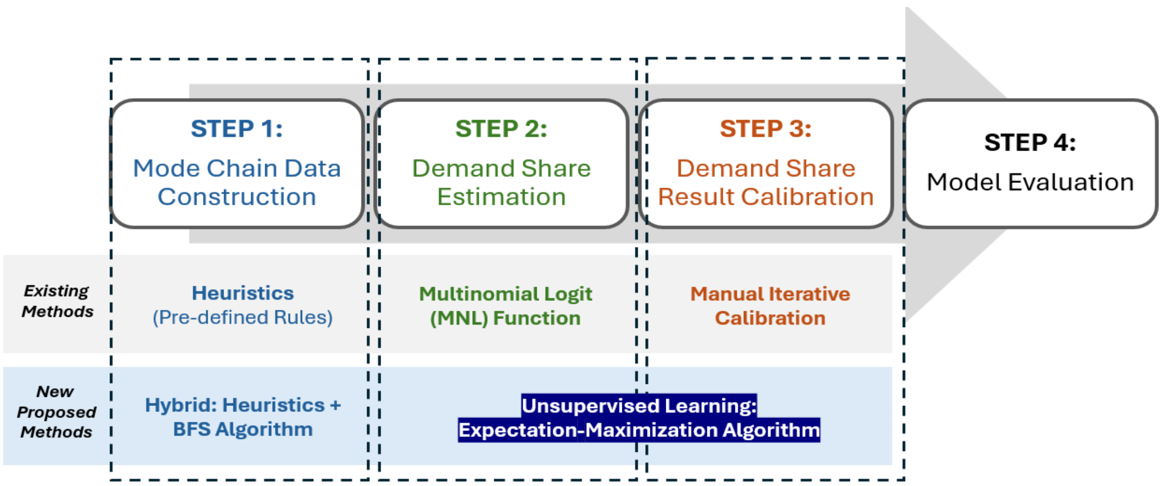

Figure 2. *Comparison between the existing and the proposed methods*

The heuristic step aims to reduce the simplify and reduce the mode chain construction process that will be conducted later, by applying several pre-defined reasonable rules to remove irrelevant and less probable path alternatives. This step also ensures that several essential rules are incorporated in the mode chain data generation process. The heuristics step itself has two sub-steps:

- *P/C Region Pair Classification:* Three predefined rules are applied to classify PC-commodity pairs into either the unimodal or multimodal group, based on the proximity between their production (origin) and consumption (destination) regions, as depicted in Figure 3.
- *Region-to-Port Selection:* For the multimodal pairs group, the next step involves selecting appropriate ports that can serve the region pairs. This is achieved by applying a set of heuristic rules, such as identifying the 20 nearest ports for each region, scoring them based on Chou's (2010) port choice qualitative model, and filtering out ports with incompatible cargo handling capabilities.

The multimodal pairs group data then proceed to this BFS stage. The BFS algorithm is used to enumerate all viable multimodal transport chains between selected region-port pairs. This process is conducted using a BFS-like logic approach with custom expansion rules, since the existing Mode Chain Builder requires quite a lot of customizations to comply with several data format requirements. The model is built using NetworkX module in Python to model the freight transport network as a multi-modal graph that enables the representation of transport links as mode-labeled edges and supporting the exploration of 1-leg to 3-leg mode chain alternatives through custom graph traversal logic while incorporating infrastructure availability constraints.

2. Second Phase: Demand Share Distribution

Once a complete dataset of mode chain alternatives for all PC-commodity pairs has been constructed, the next step is to allocate the known goods flow for each pair across the available path alternatives. The main challenge lies in the absence of data showing the historical distribution of goods across these paths. Instead, only aggregated data, such as port statistics and transport movements by mode between countries, sourced from Eurostat, is available. Traditional choice models are insufficient for this task, as they do not ensure consistency between the estimated shares and the aggregated observations.

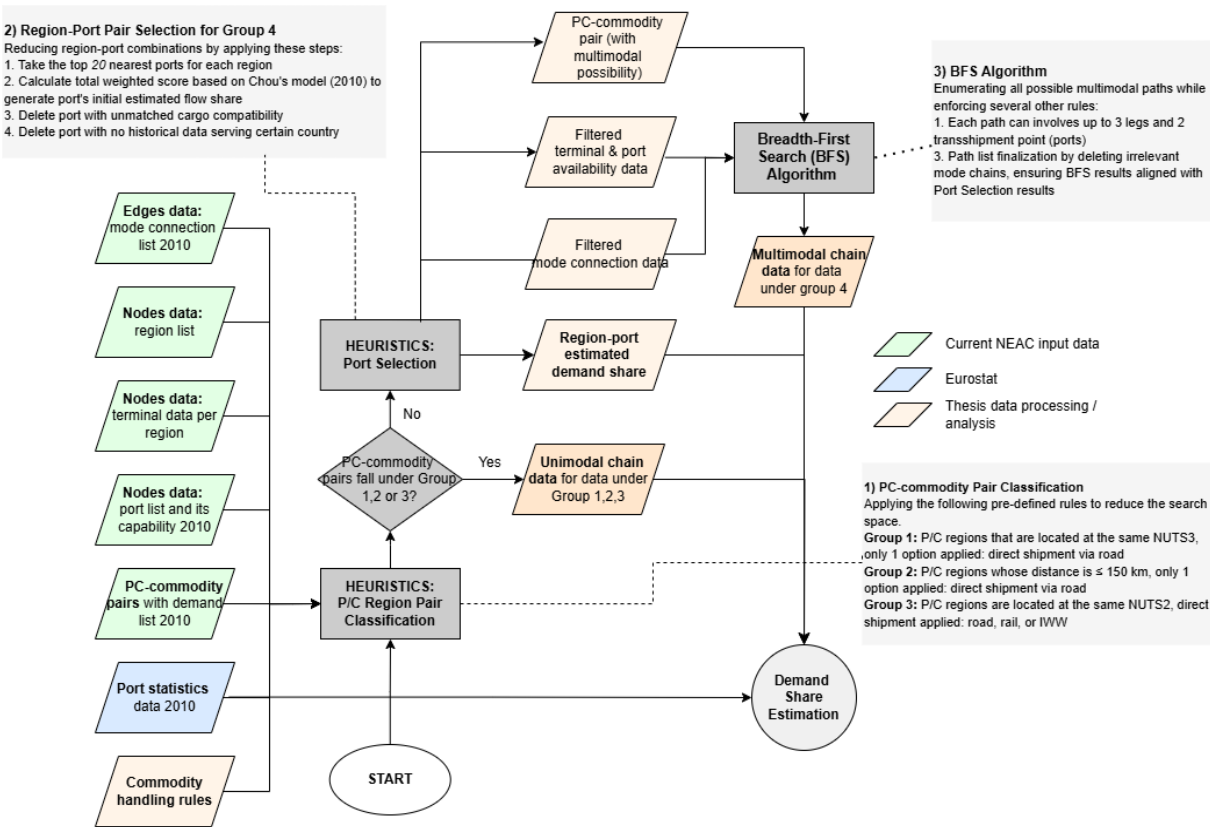

Figure 3. *Data processing diagram of Mode Chain Construction Phase*

This is where the Expectation-Maximization (EM) algorithm is suitable to use. EM is an iterative method used in unsupervised learning to estimate unknown or latent variables by uncovering patterns in the data that align with observed constraints. There are 3 major steps in this phase as follows:

Step 1: Initialization

The Expectation-Maximization (EM) algorithm requires an initial value to guide its predictions, it's an estimated starting demand distribution values across alternative paths within the same PC-commodity pair group. This initial value is critical to the learning process, as a well-estimated starting point can help the model converge more quickly and efficiently. As there is no standard guideline for determining the most suitable estimation method, this study tests three different approaches:

1. **AHP approach:** The initial value is estimated by calculating each path's weighted score based on the Analytic Hierarchy Process (AHP) results from Lu and Wang (2022) study.
2. **ESD approach:** The initial value is derived from Eurostat's historical share data for each transport mode between origin and destination countries. For each chain, the shares of all constituent edges are multiplied to calculate the chain's relative share within the same PC-country pair.
3. **EQW approach:** The total demand (in tonnes) is evenly divided among all available alternative paths, giving each path an equal weight.

To identify the most suitable approach, the estimated initial values will be evaluated by comparing their deviation from the ground truth data on mode transport flows and port-country flows from Eurostat. The method with the smallest error will be selected and used as the input for the EM model.

Step 2: EM Algorithm Development

To estimate the demand share for each mode chain between PC-commodity pairs, the Expectation-Maximization (EM) algorithm treats the demand shares as latent variables that need to be inferred. The algorithm iteratively adjusts these estimates by using known but indirect ground truth data, aiming to find the optimal distribution of demand across path alternatives. The objective is to minimize the deviation between the aggregated values derived from the estimated shares and the actual observed aggregated data.

Two ground truth datasets from Eurostat will be used as reference points to estimate the demand share per transport path. These datasets include: (1) mode-specific transport flows between and within countries, and (2) country-level port statistics detailing incoming and outgoing cargo volumes. A loss calculation function will be developed by integrating both datasets, where the loss represents the deviation between the predicted values and the observed ground truth data. The EM model will iteratively adjust the predictions in order to minimize this loss.

Step 3: Hyperparameter Tuning

The model development will also involve the hyperparameter tuning. It's the process of selecting the most effective set of initial parameters for a machine learning model. These hyperparameters are preset configurations that influence how the model learns during training and affect the model's performance to generalize the sample data to the new or hidden data.

In machine learning models using Expectation-Maximization algorithm, there are at least five important parameters to set prior to model training execution: initial learning rate, learning rate decay factor, loss tolerance, patience for convergence, and maximum iterations.

The ideal initial learning rate will be determined through a series of experiments involving 10 iterations, while other parameters will be pre-defined based on needs and relevant references. To optimize prediction results, this study will activate the ADAM (Adaptive Moment Estimation) setting within the developed EM model. It is a widely used optimizer module for automatic adaptation in machine learning. ADAM updates the model parameter during training by computing gradients from the loss function, so each parameter has its own learning rate that is kept adjusted during the training based on its past gradients and magnitudes (geeksforgeeks, 2025).

On top of ADAM, a learning rate scheduler is also applied to monitor the training loss trend. If the loss stops decreasing for a few iterations (patience), the scheduler will reduce the global learning rate of the optimizer. Unlike ADAM, the scheduler doesn't compute gradients; rather, it controls when and how fast the model learns. One important factor influencing model performance is its ability to shift between exploration in the early training stages and exploitation in later stages, once the model approaches an optimal solution. This dynamic behavior can be facilitated by setting a learning rate scheduler, which allows the learning rate to adjust according to predefined conditions or training states. It will slow down the learning rate if the loss gets stuck.

2.3 Model Validation and Evaluation

The model validation process for the mode chain construction phase will be carried out by comparing the generated mode chain data with the initial input datasets to ensure consistency. This step verifies that the resulting chains remain aligned with the key information provided at the beginning of the process. In addition, the validation will be manually cross-checked using the SQL platform. The input datasets include: mode connection availability between two regions, rail and inland waterway availability and Port availability at a specific region.

On the other hand, the demand share distribution model validation and performance evaluation will be conducted by performing these steps:

- *Model convergence and behavior*

  Model convergence, indicating by a gradual decrease in loss until it reaches the state where no significant improvement can be made any more, is a good signal that the model is working the way it's intended.
- *Loss (against ground truth) evaluation*

  The loss calculation produced by the model output will be evaluated. The model's loss indicates the deviation between the estimated value made by the model vs the observed ground truth value. The model's loss then will be compared to the actual initial deviation between input data of total demand flow and total transported demand from all transport modes stated in Eurostat.

3. Model Development

Two models for each mode chain construction process and demand share distribution developed. Below is the elaboration of the model development outputs as well as the model validation and model evaluation results for both models.

3.1 Mode Chain Construction

A total of 1,839,820 mode chain alternatives were generated from all PC-commodity pairs between Belgium and the Netherlands, as well as region pairs within each country, as shown in Table 1. International flows produced more chain alternatives per pair due to the involvement of maritime transport and the availability of multiple port combinations.

Table 123. *Mode Chain Database Construction Result Statistics*

ProdZone Country	ConsZone Country	Total PC-Comm Pair	Total Chain	Avg Chain per Pair
Belgium	Belgium	37.908	88.622	2,3
Belgium	Netherlands	53.419	814.706	15,3
Netherlands	Belgium	54.986	810.776	14,7
Netherlands	Netherlands	32.301	125.716	3,9
TOTAL		**178.614**	**1.839.820**	**10,3**

Model validation is carried out to ensure that the constructed mode chains align with key information required for generating a reliable mode chain dataset. A total of 82,385 PC-commodity pairs were selected as the validation sample. The validation process uses a sample set consisting of three-leg chains, which are evaluated based on three critical elements as detailed below:

1. **Mode connection availability**

   This aspect ensures the consistency between the assigned mode type (for a given leg) and the availability of that mode between the connected origin and destination regions. Out of approximatey 1.8 million legs in the sampled mode chains, all show perfect alignment between the assigned mode and the mode connection availability data.

2. **Rail and inland waterway availability**

   This aspect assesses the consistency between the selected terminals and their actual availability based on the NEAC terminal data. From a total of 1,945,991 terminal-related legs in the sampled mode chains, all legs were successfully matched with the terminal availability records, indicating the model was accurately incorporating terminal infrastructure constraints

Table 224. Comparison between Initial Value Prediction vs Mode Transport Flow Ground Truth Data

mode	origin	destination	Ground Truth	Y_init Prediction	deviation AHP	deviation ESD	deviation EQW	% deviation AHP	% deviation ESD	% deviation EQW
Road	Netherlands	Belgium	18.872.000	50.556.844	31.684.844	46.824.056	28.985.984	168%	248%	154%
Road	Belgium	Netherlands	16.319.000	27.668.304	11.349.304	18.799.872	9.878.996	70%	115%	61%
Road	Belgium	Belgium	231.488.000	269.087.552	37.599.552	31.099.504	39.894.048	16%	13%	17%
Road	Netherlands	Netherlands	520.606.016	478.665.984	41.940.032	51.714.560	42.268.064	8%	10%	8%
Rail	Belgium	Netherlands	1.941.000	3.537.837	1.596.837	963.884	1.712.957	82%	50%	88%
Rail	Netherlands	Belgium	1.990.000	3.592.143	1.602.143	1.019.316	1.399.794	81%	51%	70%
Rail	Netherlands	Netherlands	3.836.000	13.325.484	9.489.484	2.434.841	10.291.014	247%	63%	268%
Rail	Belgium	Belgium	18.152.000	28.931.170	10.779.170	5.297.413	13.346.562	59%	29%	74%
IWW	Belgium	Belgium	46.550.000	9.966.413	36.583.587	41.837.893	34.405.820	79%	90%	74%
IWW	Belgium	Netherlands	26.908.000	1.092.694	25.815.306	24.004.562	25.295.369	96%	89%	94%
IWW	Netherlands	Belgium	42.345.000	1.207.919	41.137.081	38.166.434	40.579.783	97%	90%	96%
IWW	Netherlands	Netherlands	103.327.000	14.526.564	88.800.436	95.879.015	83.017.612	86%	93%	80%
Sea	Netherlands	Belgium	2.032.000	36.109.576	34.077.576	12.046.897	39.488.532	1677%	593%	1943%
Sea	Belgium	Netherlands	1.713.000	21.850.006	20.137.006	6.894.568	24.038.456	1176%	402%	1403%
		TOTAL	**1.036.079.016**	**960.118.489**	**392.592.358**	**376.982.814**	**394.602.990**			

3. **Port availability**
   The third aspect of the model validation focuses on verifying the availability of ports in the regions assigned as origin or destination for maritime transport legs. This ensures that the ports used in the multimodal chains are actually present and capable of serving the specified commodity in the respective regions. Out of 2,254,340 relevant edges examined in the sampled mode chains, all were successfully matched with valid port availability records.

The model validity results confirm that the constructed mode chain construction model using combination of heuristics and BFS algorithms are fully aligned with the given infrastructure availability data. All assigned mode legs, terminal facilities, and port locations were validated with 100% match rates, indicating that the model reliably generates realistic transport paths. Thus, the resulted mode chain datasets can safely proceed to the next phase, demand share estimation process.

3.2 Demand Share Distribution

The results of the demand share distribution will be presented and classified into three main steps, as detailed below:

Step 1: Initial Value Calculation

The initial loss figures from each approach were compared against two ground truth datasets: inter-country transport flows and port-level flow statistics. As shown in Table 2, the ESD approach produced the lowest overall deviation, with a total absolute deviation of ~376 million tons, compared to 392 million for AHP and 394 million for EQW. Across most modes and OD pairs, ESD consistently showed lower percentage deviations, indicating a closer alignment with historical transport flow data.

A similar pattern emerged in the port-level comparison. For both Belgian and Dutch ports, ESD yielded significantly lower inflow and outflow deviations than AHP and EQW. The heatmap in Table 3 further illustrates that ESD's deviations were more evenly distributed and less intense, particularly at major hubs such as Rotterdam, Ghent, and Antwerp, suggesting a more balanced and realistic allocation of flows. Overall, the ESD approach provided the most accurate initial flow distribution, minimizing deviations from ground truth at both country and port levels. This makes it the most suitable method for generating initial values for the EM model input.

Table 325. Comparison between Initial Value Prediction vs Port level Flow Ground Truth Data

Labels	AHP Deviation inflow	AHP Deviation outflow	ESD Deviation inflow	ESD Deviation outflow	EQW Deviation inflow	EQW Deviation outflow
um	**33.155.531**	**17.560.543**	**11.686.454**	**5.627.809**	**38.404.281**	**20.972.408**
TWERP	22.851.173	13.776.800	7.755.988	3.975.493	26.647.073	16.525.755
NT	2.384.400	2.269.464	637.618	1.005.846	2.716.553	2.611.302
END, ZEEBRUGE	1.685.987	-	669.711	-	2.047.238	-
BRUGE	6.233.971	1.514.279	2.623.137	646.471	6.993.418	1.835.351
erlands	**17.693.114**	**31.316.475**	**5.483.425**	**10.863.111**	**20.931.941**	**35.974.932**
STERDAM	2.263.292	7.815.572	930.920	3.454.134	2.674.018	8.661.268
FZIJL	4.042	41.556	405	54.990	4.428	40.414
RDRECHT	5.000	29.000	5.000	29.000	5.000	29.000
LINGEN	29.977	43.585	10.518	13.611	32.606	47.503
ERDIJK	6.230	15.715	122	4.260	6.360	16.329
TERDAM	15.835.505	22.825.880	5.056.547	7.461.487	18.637.505	26.412.025
NEUZEN	36.000	86.000	36.000	86.000	36.000	86.000
SSINGEN	404.932	772.278	473.278	108.130	381.976	993.220

Step 2: Expectation-Maximization (EM) Model Development

To estimate the demand share for each mode chain between PC-commodity pairs, the Expectation-Maximization (EM) algorithm treats the demand shares as latent variables that need to be inferred. The algorithm iteratively adjusts these estimates by using known but indirect ground truth data, aiming to find the optimal distribution of demand across path alternatives. The objective is to minimize the deviation between the aggregated values derived from the estimated shares and the actual observed aggregated data.

Figure 4 illustrates how the ground truth data are linked to the demand share estimation for each path alternative, which is further detailed in the mathematical model in the following section. The assigned flow for each path is proportionally distributed across its corresponding legs. As a result, the loss function consists of two main components:

- *Transport Loss (Constraint 1):* This represents the discrepancy between the aggregated flow assigned across all legs (with a specific mode type) in the alternative paths and the known transport statistics between origin-destination country pairs.
- *Port Loss (Constraint 2):* This reflects the difference between the cumulative assigned flow entering or exiting a specific port and the actual goods volume handled by the port, based on observed port statistics.

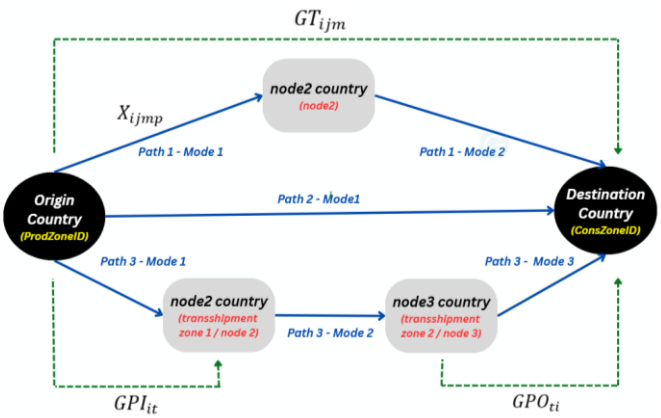

Figure 4. *Illustration of Flow Aggregation Mechanism for Loss Function in EM Algorithm*

Because the two loss components have different magnitudes, a scaling factor was introduced to balance their influence during training. Several ratios were tested, and results showed that only the 1:1 and 1:2 settings produced consistent downward trends in both transport and port losses during the initial iterations, indicating better balance. Extending the analysis to 20 iterations revealed that simultaneous reduction was only maintained until the 6th iteration, after which transport loss continued to decrease while port loss increased. Given that transport loss dominates in magnitude and minimizing it is the priority, the 1:1 ratio was selected as the most suitable setting for the full EM model training.

Step 3: Hyperparameter Tuning

Hyperparameter tuning was performed to optimize model performance, focusing on four parameters: initial learning rate, decay factor, loss tolerance with patience, and maximum iterations.

Trials were conducted to select the learning rate, with each run lasting ten iterations. A rate of 1.0 produced the greatest early loss reduction, while lower values such as 0.5 showed smaller improvements. Thus, 1.0 was chosen as the initial setting. The decay factor (gamma) was set at 0.5 to mitigate fluctuations observed in early training due to the large number of mode chain alternatives and limited ground truth data. Loss tolerance was fixed at 3000 (0.012% of total loss), and patience at three iterations, ensuring convergence if no reduction above this threshold occurred within that span. To prevent excessive computation, the maximum number of iterations was capped at 500, corresponding to roughly 83 hours of runtime. To further monitor convergence, an early stopping mechanism was applied (Figure 15). After a minimum of 10 iterations, loss stabilization was tracked; if the loss reduction fell below the tolerance for three consecutive iterations, training was stopped.

The ADAM optimizer was used with default settings (lr = 1.0, $\beta_1 = 0.9$, $\beta_2 = 0.999$, AMSGrad = True). A learning rate scheduler reduced the rate by half whenever the improvement between iterations fell below 0.015% of total loss, applied immediately without a patience delay.

4. Results and Discussion

4.1 Model Results

Table 4 compares the EM model predictions (*y_pred*) using the ESD approach with ground truth transport flows. The initial values show overestimation for Sea and Road modes, and underestimation for inland waterways (IWW). After training, the model reduced absolute deviations for most OD-country pairs, particularly in Road and Rail (0.2–17% improvement), though deviations increased slightly for IWW and more substantially for Sea.

Table 426. *EM Prediction Comparison against Transport Flow Ground Truth Data using ESD Approach*

mode	mode_name	origin	destination	ground truth	y_init total flow	y_init abs deviation	y_pred total flow	y_pred abs deviation	% improvement	Initial Value Status vs Ground Truth
1	Road	Belgium	Belgium	231,488,000	262,740,047	13.5%	261,148,183	12.8%	-0.7%	Bigger
1	Road	Belgium	Netherlands	16,319,000	35,117,301	115.2%	33,206,984	103.5%	-11.7%	Bigger
1	Road	Netherlands	Belgium	18,872,000	65,695,313	248.1%	63,899,588	238.6%	-9.5%	Bigger
1	Road	Netherlands	Netherlands	520,606,000	469,162,394	9.9%	472,991,154	9.1%	-0.7%	Lower
2	Rail	Belgium	Belgium	18,152,000	12,844,670	29.2%	15,905,118	12.4%	-16.9%	Lower
2	Rail	Belgium	Netherlands	1,941,000	976,328	49.7%	1,095,948	43.5%	-6.2%	Lower
2	Rail	Netherlands	Belgium	1,990,000	969,938	51.3%	1,076,790	45.9%	-5.4%	Lower
2	Rail	Netherlands	Netherlands	3,836,000	1,388,183	63.8%	1,003,493	73.8%	10.0%	Lower
3	IWW	Belgium	Belgium	46,550,000	4,708,120	89.9%	5,945,259	87.2%	-2.7%	Lower
3	IWW	Belgium	Netherlands	26,908,000	2,903,313	89.2%	2,955,528	89.0%	-0.2%	Lower
3	IWW	Netherlands	Belgium	42,345,000	4,178,472	90.1%	4,156,224	90.2%	0.1%	Lower
3	IWW	Netherlands	Netherlands	103,327,000	7,438,934	92.8%	6,512,245	93.7%	0.9%	Lower
4	Sea	Belgium	Netherlands	1,713,000	8,590,080	401.5%	9,396,339	448.5%	47.1%	Bigger
4	Sea	Netherlands	Belgium	2,032,000	14,061,504	592.0%	14,840,401	630.3%	38.3%	Bigger

The increase of deviation in Sea section is due to the trade-off between the sea and other mode types. Road and Sea mode have bigger initial values status compared to the ground truth data, while Rail and IWW have lower initial values. So, the EM model aims to decrease the share of Road and Sea, then redistribute the flow shares to Rail and IWW to increase their shares and get closer to the ground truth.

However, as shown in Table 5, the generated mode chains data are dominated by 3-legs chains that consist of multiple mode types, where the use of Sea mode is always be combined with other modes as the first and third (last) leg. The decrease of Sea mode share will also lead to the decrease of other modes' shares, and vice versa. So, in order to achieve the model objective to increase the share of Rail and IWW will bring the side-effects which is the increase of Sea mode shares as well, and make the deviation of Sea mode gets bigger (while the Sea mode share is supposed to be decreased).

Table 527. *Number of Legs Distribution according to the Generated Chain Alternatives*

prev_mode	after_mode	leg_count	y_init	y_pred
3-legs scheme				
Road	Sea	153,094	12,936,543	13,674,58
Rail	Sea	103,615	5,972,392	6,870,84
IWW	Sea	69,398	3,742,650	3,691,31
Sea	Road	157,989	9,113,527	9,402,27
Sea	Rail	102,287	8,936,942	10,151,42
Sea	IWW	65,831	4,601,116	4,683,04
1-leg scheme				
Road		58,778	685,999,138	685,999,1

Unlike Sea mode, the road section whose shares also need to be decrease because of bigger initial values status against the ground truth, could alternatively achieves its goal through minimizing the share of the 1-leg mode chain alternatives since it only involves road mode only without combination with other modes. The efforts to decrease the Road mode share is easier to be done than to decrease the Sea mode share since it's conflicting with the other objectives to increase the share of Rail and IWW modes. Hence, the trade-off between Sea, Rail and IWW flow distribution makes it hard for the EM model to achieve its ultimate goal to minimize the overall loss across all ground truth data points, even though the model has reached the convergence at 107th iterations (see Figure 5).

The increase in Sea deviation also worsen the results of port flow prediction in comparison to the ground truth data. It causes the increases of deviation in most of port flow ground truth data points as displayed in Table 6, and only several data points are optimized as intended. As indicated by the first 20 iterations, where the port loss tends to increase while the transport loss is decreasing, the trend continues until the model convergence.

As shown in Figure 6, despite the increase of deviation of port data (indicated by larger port loss compared to the initial status), the model has successfully reduce the overall loss until no significant reduction can be made (convergence). Hence, the model has reached the possible optimal states given the input data structure including the trade-off within it.

Table 628. *EM Prediction Comparison against Port Flow Ground Truth Data using ESD Approach*

direction	partner_country	port_wn_id	pot name	ground truth	y_init total flow	y_init abs deviation	y_pred total flow	y_pred abs deviation	Changes	%Improvement	Initial Value Status vs Ground Truth
in	Total			3,701,000	22,651,584	512%	24,236,741	555%			
in	Belgium	p124010102	DELFZIJL	4,000	15,933	298%	15,066	277%	-868	-22%	Bigger
in	Belgium	p124010201	HARLINGEN	1,000	24,042	2304%	30,873	2987%	6,831	683%	Bigger
in	Belgium	p124030206	AMSTERDAM	144,000	3,170,254	2102%	3,408,601	2267%	238,348	166%	Bigger
in	Belgium	p124030305	ROTTERDAM	1,018,000	5,242,027	415%	5,842,741	474%	600,714	59%	Bigger
in	Belgium	p124030402	VLISSINGEN	504,000	132,737	74%	94,705	81%	38,033	8%	Lower
in	Belgium	p124040101	MOERDIJK	1,000	5,087	409%	4,353	335%	-734	-73%	Bigger
in	Netherlands	p102020101	ANTWERP	891,000	8,676,823	874%	8,720,449	879%	43,626	5%	Bigger
in	Netherlands	p102020304	GHENT	695,000	1,611,087	132%	1,831,669	164%	220,582	32%	Bigger
in	Netherlands	p102020501	ZEEBRUGE	443,000	3,773,594	752%	4,288,283	868%	514,689	116%	Bigger
out	Total			3,714,000	22,651,584	510%	24,236,741	553%			
out	Belgium	p124010102	DELFZIJL	65,000	20,023	69%	17,812	73%	2,210	3%	Lower
out	Belgium	p124010201	HARLINGEN	3,000	32,056	969%	37,140	1138%	5,084	169%	Bigger
out	Belgium	p124030206	AMSTERDAM	117,000	5,800,533	4858%	6,040,467	5063%	239,934	205%	Bigger
out	Belgium	p124030305	ROTTERDAM	1,536,000	7,925,332	416%	8,532,608	456%	607,276	40%	Bigger
out	Belgium	p124030402	VLISSINGEN	98,000	274,276	180%	205,807	110%	-68,469	-70%	Bigger
out	Belgium	p124040101	MOERDIJK	9,000	9,285	3%	6,568	27%	-2,717	24%	Bigger
out	Netherlands	p102020101	ANTWERP	1,763,000	5,779,874	228%	5,915,390	236%	135,516	8%	Bigger
out	Netherlands	p102020304	GHENT	34,000	1,044,119	2971%	1,254,421	3589%	210,302	619%	Bigger
out	Netherlands	p102020501	ZEEBRUGE	89,000	1,766,088	1884%	2,226,528	2402%	460,440	517%	Bigger

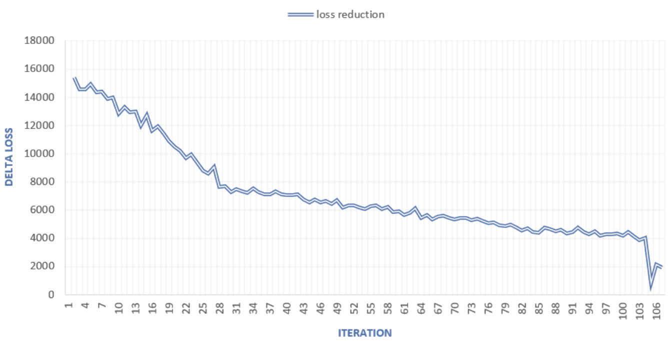

Figure 5. *Full Loss Reduction Trend until Convergence*

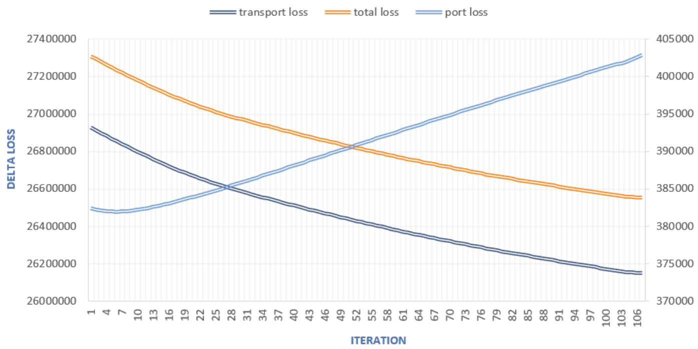

Figure 6. *ESD EM Model Iterations Results until Convergence*

4.2 Model Performance Evaluation

As discussed earlier, the ESD approach started with smaller initial deviations from the ground truth data. However, to fully assess model performance, the results of both initial value strategies are presented here. This comparison not only shows how well each model minimizes deviations but also highlights the trade-off between loss reduction and the final predicted demand share distribution.

Table 7 shows that both approaches, AHP and ESD, reduced deviations in most country pairs, similar to the ESD approach. The key difference is that AHP placed greater emphasis on reducing Road and Sea shares, which resulted in better performance in the port loss segment. Only a small number of port data points showed higher deviations in AHP results, unlike the ESD approach, which led to increases in nearly all port data points. However, when comparing the final deviations across all transport flows and port flows, ESD still outperformed AHP.

ESD achieved especially better results in port flows, even though its improvement during iterations appeared smaller. This indicates that ESD provided stronger initial values that were already closer to the ground truth, leaving less room for further improvement. As a result, ESD reduced losses effectively in the early iterations, but after around the 13th iteration the model reached a saturation point, where further reductions could only be achieved by increasing port loss.

Table 729. *Transport Loss Deviation Reduction Comparison between AHP and ESD Approach*

mode	mode_name	origin	destination	y_pred improvement compared to y_init AHP	ESD
1	Road	Belgium	Belgium	-2.0%	-0.7%
1	Road	Belgium	Netherlands	11.1%	-11.7%
1	Road	Netherlands	Belgium	12.5%	-9.5%
1	Road	Netherlands	Netherlands	1.0%	-0.7%
2	Rail	Belgium	Belgium	-28.9%	-16.9%
2	Rail	Belgium	Netherlands	-4.6%	-6.2%
2	Rail	Netherlands	Belgium	15.2%	-5.4%
2	Rail	Netherlands	Netherlands	-93.5%	10.0%
3	IWW	Belgium	Belgium	3.7%	-2.7%
3	IWW	Belgium	Netherlands	-0.3%	-0.2%
3	IWW	Netherlands	Belgium	-0.2%	0.1%
3	IWW	Netherlands	Netherlands	3.1%	0.9%
4	Sea	Belgium	Netherlands	-402.3%	47.1%
4	Sea	Netherlands	Belgium	-365.0%	38.3%

Figure 7 shows the learning process of both approaches. The ESD approach converged faster, stabilizing around the 107th iteration with final transport and port losses of 26.15 million and 402 thousand, respectively. By contrast, the AHP approach converged more slowly, and the run was stopped at the 460th iteration due to time limits, with losses of 27.18 million for transport and 1.29 million for ports. The trend suggests that AHP would eventually approach the same levels as ESD but with a longer training time. Overall, these results show that ESD provides faster convergence and better accuracy, making it the most suitable initial value strategy for the EM model in demand share estimation.

4.3 Model Performance Justification

This substantial reduction indicates that the EM model successfully optimized the estimates by consistently minimizing initial deviations across all ground truth data points available. It indicates the model's ability to learn and adjust toward the ground truth, even when starting from suboptimal predictions. The "green" improvement values further confirms that the majority of final aggregated predictions are getting closer to actual values. This validates that the EM algorithm is functioning as intended, iteratively refining estimates and reducing mismatches, highlighting its capacity to improve performance under constraints.

Table 830. *Port Loss Deviation Reduction Comparison between AHP and ESD Approach*

direction	partner_country	port_wn_id	pot name	y_pred improvement compared to y_init AHP	ESD
in	Belgium	p124010102	DELFZIJL	-487%	-22%
in	Belgium	p124010201	HARLINGEN	-3257%	683%
in	Belgium	p124030206	AMSTERDAM	-1761%	166%
in	Belgium	p124030305	ROTTERDAM	-402%	59%
in	Belgium	p124030402	VLISSINGEN	40%	8%
in	Belgium	p124040101	MOERDIJK	-1318%	-73%
in	Netherlands	p102020101	ANTWERP	-615%	5%
in	Netherlands	p102020304	GHENT	-80%	32%
in	Netherlands	p102020501	ZEEBRUGE	-312%	116%
out	Belgium	p124010102	DELFZIJL	36%	3%
out	Belgium	p124010201	HARLINGEN	-1433%	169%
out	Belgium	p124030206	AMSTERDAM	-2484%	205%
out	Belgium	p124030305	ROTTERDAM	-272%	40%
out	Belgium	p124030402	VLISSINGEN	-251%	-70%
out	Belgium	p124040101	MOERDIJK	-202%	24%
out	Netherlands	p102020101	ANTWERP	-288%	8%
out	Netherlands	p102020304	GHENT	-1577%	619%
out	Netherlands	p102020501	ZEEBRUGE	-1442%	517%

As discussed earlier, the EM model with ESD approach already reaches the convergence state. When the model reaches convergence state, it already reaches the internal consistency, indicating that the model has minimized deviations as much as possible given the input constraints, initial Y distribution, and OD pair structure. The strong reduction in deviation across country pairs confirms that the model effectively learned from iterative updates. The remaining gap between predicted values and ground truth suggests that while the model has performed

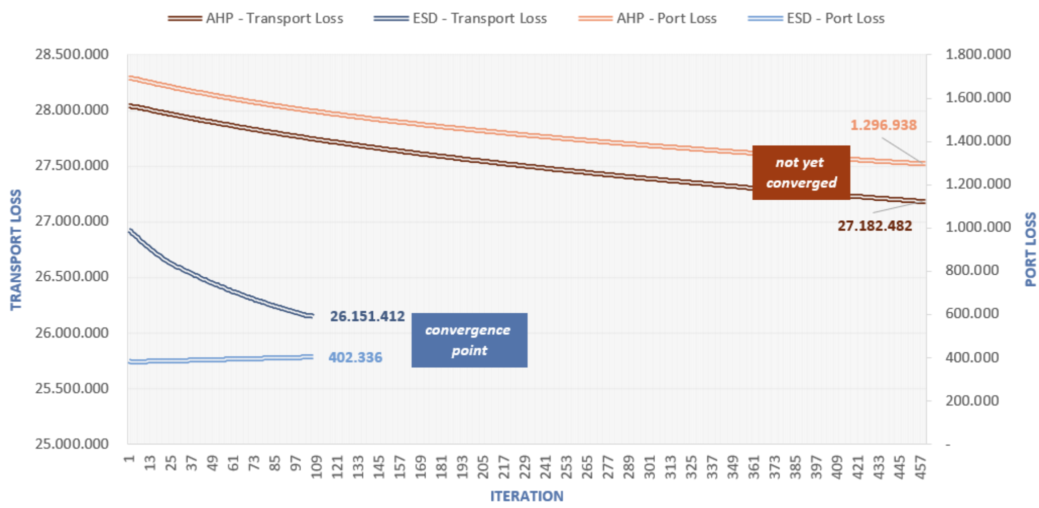

Figure 7. *Loss Reduction Trend Comparison between ESD and AHP Approach*

well within its design limits, external factors likely hinder its ability to achieve a closer match.

Several external factors may limit the model's ability to achieve higher accuracy. These identified factors are explained in the following section.

- **Incomplete or Biased Path Alternatives**

The generated mode chains may not fully reflect realistic alternatives present in the real world. Since the mode chain data for the EM model input was constructed based on infrastructure availability data, and it has also been validated in Section 4.2.3, any inaccuracies in the infrastructure data could introduce bias to the final EM prediction. Given that IWW is the most consistently underestimated mode, reviewing the latest data on terminal availability might help mitigate this issue.

- **Mismatch between Mode Share Assumptions and Real-World Behavior**

The mode chain filtering described in Section 4.3.1 was based on probabilities derived from weighted factors estimation in Lu and Wang (2022)'s AHP results. While theoretically sound, this approach may not fully capture the real-world dynamics of freight transport. Critical factors (e.g., market conditions, actual tariffs, regulatory preferences) strongly influence mode choice but are not incorporated into the current model. For example, even if rail infrastructure is available, certain flows may still prefer road transport due to greater flexibility, reliability, or existing commercial agreements, leading to potential bias in the predicted probability of alternative paths. Hence, the exclusion of certain OD alternatives could also contribute to the remaining gap since it may remove relevant yet lower-frequency alternatives, reducing the diversity of mode chains. However, in this study, the elimination step is needed due to the limited computation capability.

- **Limited Travel Impedance Data**

The generalized transport cost and travel time data were used as the basis in calculating the path score to perform the early path elimination as well as to estimate the path share in AHP initial value estimation strategy. But the impedance cost used in the calculation may not accurately reflect actual multimodal costs. Since multimodal paths involve multiple modes and transfer processes, they often incur additional time and handling costs that are not fully captured in the current utility model. This omission could further distort the initial probability estimates.

5. Conclusion

The newly developed multimodal freight chain model in this study has proven its ability to construct the mode chain alternatives data with 100% validity against the available transport infrastructure information. The demand share estimation model can reduce the deviation between the predicted demand share values against the transport historical data in majority of the observed data points. Even though several data points experienced an increase in deviation instead, the reduction outweighed the increase. The rise in deviations for certain flows appears to be a trade-off resulting from the model's focus on reducing errors in segments with the highest initial deviations.

Despite the model's proven ability to reduce the deviation, the existing model still left a quite significant remaining deviation between predicted values and ground truth. But the learning process of the EM model was already near the convergence, meaning the model has reached its maximum ability to minimize the deviations given the input constraints, initial Y distribution, and OD pair structure. Hence, the remaining gap indicates that while the model has performed well within its design limits, external factors likely hinder its ability to achieve a closer match. Several recommendations (elaborated in the next section) can be considered to improve the model performance in the future related or similar research.

Finally, it is concluded that machine learning (ML) can be effectively implemented in the Mode Chain Builder system, but primarily for the demand share estimation phase. ML is not well-suited for mode chain construction because most existing path-generation or route-construction approaches aim to identify the most optimal or a few best path alternatives. In multimodal freight chain modelling, however, the requirement is to generate all possible path alternatives, which aligns better with traditional graph search methods. The implementation of ML in the demand share estimation phase is achieved using the Expectation-Maximization (EM) algorithm, where the demand share per path is treated as a latent variable. Its values are iteratively estimated by referencing observed ground truth data. The EM model demonstrated its capability to refine the initial demand share estimates (initial values) by minimizing the deviation between predicted values and ground truth data, despite trade-offs between transport flow and port flow accuracy.

While the model has not fully eliminated the deviation as initially expected, it has achieved the best possible results given the current input data structure. Further reductions in deviation can only be achieved by improving several external factors: access to more accurate infrastructure connectivity and travel impedance data to build more realistic mode chain alternatives, enhanced computational capacity to allow the full model to run without multiple simplification steps, and a more comprehensive understanding of freight path choice behaviour to support the development of better estimation models.

6. Limitations and Recommendations

1. Incomplete or Biased Path Alternatives

The mode chain dataset, constructed from infrastructure availability data, may not fully represent real-world alternatives. Inland waterway transport (IWW) in particular appears underestimated, likely due to outdated or missing information on terminals and connectivity. Future work should expand and update infrastructure databases, especially for underrepresented modes, and incorporate expert validation or shipment data to improve path realism.

2. Mode Choice Behavior Not Fully Captured

Mode choice behavior was not fully captured. The filtering of mode chains relied on AHP-based probabilities with predefined weights, which may overlook real decision factors such as pricing, contracts, and reliability. Future research should explore data-driven path choice modeling based on actual observed freight transport behavior. In particular, research on discrete choice modeling or stated preference surveys specifically targeting multimodal path or route choices would be highly beneficial. This is important because most existing transport behavior studies tend to focus on individual mode choice, rather than on entire transport chains that involve multiple modes or legs. Developing models that capture the full path decision-making process would yield more realistic and reliable estimates of mode chain probabilities.

3. Limited Travel Impedance Accuracy

Travel impedance values were simplified, relying mainly on generalized costs and times without fully accounting for transfers, waiting, and handling costs. This could bias results for complex chains. More detailed impedance models that include intermodal penalties and mode-specific costs should be developed, supported by richer transport datasets.

4. EM Algorithm Still Rarely Applied in Freight Modeling

The use of the Expectation-Maximization (EM) algorithm for freight mode split modeling remains relatively novel. While the results here show promise, this method still lacks widespread validation in transport research. Its sensitivity to initial values and potential for local minima make its robustness less predictable in other settings. Further research should benchmark EM against other semi-supervised or probabilistic modeling techniques in similar freight estimation contexts.

References

Ahmed, U. & Roorda, M. J., 2022. Modeling freight vehicle type choice using machine learning and discrete choice methods. *Transportation Research Record,* 2676(2), pp. 541-552.

Ahmed, U. & Roorda, M. J., 2022. Modeling freight vehicle type choice using machine learning and discrete choice methods.. *Transportation Research Record,* 2676(2), pp. 541-552.

Bayati, M., Shah, D. & Sharma, M., 2008. Max-Product for Maximum Weight Matching: Convergence, Correctness, and LP Duality. *IEEE Transactions on Information Theory,* 54(3), pp. 1241-1251.

Bernov, M. F., Rahajoe, A. D. & Mulyo, B. M., 2022. Route Optimization of Waste Carrier Truck using Breadth First Search (BFS) Algorithm. *JEECS (Journal of Electrical Engineering and Computer Sciences),* 7(2), pp. 1293-1304.

Bok, M. d. et al., 2017. Development of a multimodal transport chain choice model for container transport for BasGoed.

Bok, M. d. et al., 2017. Development of a multimodal transport chain choice model for container transport for BasGoed.

CelerData, 2024. *What Is Breadth-First Search (BFS)?.* [Online]
Available at: https://celerdata.com/glossary/breadth-first-search-bfs#:~:text=Breadth%2DFirst%20Search%20(BFS)%20is%20one%20of%20the%20simplest,discovered%20and%20removed%20when%20explored
[Accessed 13 June 2025].

Chou, C.-C., 2010. AHP MODEL FOR THE CONTAINER PORT CHOICE IN THE MULTIPLE-PORTS REGION. *Journal of Marine Science and Technology,* 18(2), p. 8.

Coursera, 2025. *Random Forest vs. Neural Network: What's the Difference?.* [Online]
Available at:
https://www.coursera.org/articles/random-forest-vs-neural-network
[Accessed 15 May 2025].

EEA, 2024. *Freight transport activity.* [Online]
Available at:
https://www.eea.europa.eu/en/analysis/publications/sustainability-of-europes-mobility-systems/freight-transport-activity?activeTab=cd80fe05-f26e-46a8-93be-2842d2a711ef
[Accessed 10 June 2025].

Elbert, R., Müller, J. P. & Rentschler, J., 2020. Tactical network planning and design in multimodal transportation – A systematic literature review. *Research in Transportation Business & Management,* Volume 35, p. 100462.

European Union, 2023. *Logistics and multimodal transport.* [Online]
Available at: https://transport.ec.europa.eu/transport-themes/logistics-and-multimodal-transport_en
[Accessed 10 June 2025].

Eurostat, 2023. *Modal split of inland freight transport.* [Online]
Available at:
https://ec.europa.eu/eurostat/databrowser/view/tran_hv_frmod/default/table?lang=en&category=tran.tran_hv_ms
[Accessed 15 July 2025].

Eurostat, 2024. *Top 5 ports - gross weight of goods handled in each port.* [Online]
Available at:
https://ec.europa.eu/eurostat/databrowser/view/mar_qg_qm_pwh/default/table?lang=en
[Accessed 15 July 2025].

Eurostat, 2025. *Freight Transport Logistics - Modal Split.* [Online]
Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Freight_transport_statistics_-_modal_split
[Accessed 10 June 2025].

Eurostat, 2025. *Freight transport statistics - modal split.* [Online]
Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Freight_transport_statistics_-_modal_split
[Accessed 10 June 2025].

Figliozzi, M. A., 2010. An Iterative Route Construction and Improvement Algorithm for the Vehicle Routing Problem with Soft Time Windows. *Transportation Research Part C: Emerging Technologies,* 18(5), pp. 668-679.

Frederik Schulte; Bilge Atasoy, 2024. *Machine Learning for Transport and Multi-Machine Systems.* s.l.:s.n.

geeksforgeeks, 2025. *Expectation-Maximization Algorithm - ML.* [Online]
Available at: https://www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/
[Accessed 12 July 2025].

geeksforgeeks, 2025. *What is Adam Optimizer?.* [Online]
Available at: https://www.geeksforgeeks.org/deep-learning/adam-optimizer/
[Accessed 24 July 2025].

Guilbault, M., 2008. ENQUETE ECHO - Envois - Chargeurs - Opérations de transport. *Résultats de référence.*

hackerearth, n.d. *Breadth First Search.* [Online]
Available at: https://www.hackerearth.com/practice/algorithms/graphs/breadth-first-search/tutorial/#:~:text=Breadth%20First%20Search%20(BFS)&text=BFS%20is%20a%20traversing%20algorithm,the%20next%2Dlevel%20neighbour%20nodes
[Accessed 13 June 2025].

Haugh, M., 2015. *The EM Algorithm.* [Online]
Available at: https://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf
[Accessed 13 July 2025].

Hayakawa, K., Ishikawa, J. & Tarui, N., 2018. What Goes Around Comes Around: Export-Enhancing Effects of Import-Tariff Reductions. August.18(8).

Holliday, A., 2025. Applications of deep reinforcement learning to urban transit network design. *arXiv,* Volume 2502.

Huang, W., Song, G. & Hong, H., 2014. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Trans. Intell. Transp. Syst.,* 15(5), pp. 2191-2201.

Huang, W., Song, G., Hong, H. & Xie, K., 2014. Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems,* 15(5), pp. 2191-2201.

Huber, S., 2017. Transport chain choice modelling in freight transport demand models. *Open Science Journal,* 2(4).

IBM, 2025. *What is gradient descent?.* [Online]
Available at: https://www.ibm.com/think/topics/gradient-descent
[Accessed 22 August 2025].

IEA, 2002. *Transportation projections in OECD regions,* Paris: International Energy Agency.

IEA, 2002. *Transportation Projections in OECD Regions,* Paris: International Energy Agency.

IRU, 2025. *Multimodal transport: what is it?.* [Online]
Available at: https://www.iru.org/what-we-do/facilitating-trade-and-transit/different-types-freight-transport/multimodal-transport-what-it
[Accessed 22 August 2025].

Joubert, J. W. & Meintjes, S., 2016. Freight Activity Chain Generation Using Complex Networks of Connectivity. *Transportation Research Procedia,* Volume 12, pp. 425-435.

Jourquin, B. & Beuthe, M., 1996. Transportation policy analysis with a geographic information system: the virtual network of freight transportation in Europe. *Transportation Research Part C: Emerging Technologies,* 4(6), pp. 359-371.
Jourquin, B. & Beuthe, M., 1996. Transportation policy analysis with a geographic information system: the virtual network of freight transportation in Europe. *Transportation Research Part C: Emerging Technologies,* 4(6), pp. 359-371.
Kartoirono, S., Riadi, I., Furizal, F. & Azhari, A., 2022. Improved Breadth First Search For Public Transit Line Search Optimization. *Mobile and Forensics.*

Konings, R., Priemus, H. & Nijkmap, P., 2008. *The Future of Intermodal Freight Transport – Operations, Design and Policy.* s.l.:Edward Elgar Publishing.

Lee, D., Derrible, S. & Pereira, F. C., 2018. Comparison of Four Types of Artificial Neural Network and a Multinomial Logit Model for Travel

Mode Choice Modeling. *Transportation Research Record,* 2672(49), pp. 101-112.

Liachovičius, E., Šabanovič, E. & Skrickij, V., 2023. Freight rate and demand forecasting in road freight transportation using econometric and artificial intelligence methods. *Transport,* 38(4), pp. 231-242.

Liachovičius, E., Šabanovič, E. & Skrickij, V., 2023. Freight rate and demand forecasting in road freight transportation using econometric and artificial intelligence methods. *Transport,* 38(4), pp. 231-242.

Limbourg, S. & Jourquin, B., 2009. Optimal rail-road container terminal locations on the European network. *Transportation Research Part E: Logistics and Transportation Review,* 45(4), pp. 551-563.

Limbourg, S. & Jourquin, B., 2009. Optimal rail-road container terminal locations on the European network. *Transportation Research Part E: Logistics and Transportation Review,* 45(4), pp. 551-563.

Lim, H. et al., 2022. A comparative study of machine learning algorithms for industry-specific freight generation model. *Sustainability,* 14(22), p. 15367.

Lim, H. et al., 2022. A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model. *Sustainability,* 14(22).

Liu, C. et al., 2023. Railway freight demand forecasting based on multiple factors: Grey relational analysis and deep autoencoder neural networks. *Sustainability,* 15(12), p. 9652.

Liu, C. et al., 2023. Railway freight demand forecasting based on multiple factors: Grey relational analysis and deep autoencoder neural networks. *Sustainability,* 15(12).

Liu, H., Jin, C., Yang, B. & Zhou, A., 2018. Finding Top-k Shortest Paths with Diversity. *IEEE Transactions on Knowledge and Data Engineering,* 1 March, 30(3), pp. 488-502.

Liu, S., 2023. Multimodal Transportation Route Optimization of Cold Chain Container in Time-Varying Network Considering Carbon Emissions. *Sustainability,* 15(5), p. 4435.

Lu, Y. & Wang, S., 2022. Optimization of Joint Decision of Transport Mode and Path in Multi-Mode Freight Transportation Network. *Sensors,* 22(13), p. 4887.

Lu, Y. & Wang, S., 2022. Optimization of Joint Decision of Transport Mode and Path in Multi-Mode Freight Transportation Network. *Sensors,* Volume 22.

Mauttone, A. & Urquhart, M. E., 2009. A route set construction algorithm for the transit network design problem. *Computers & Operations Research,* 36(8), pp. 2440-2449.

McKee, A., 2024. *Breadth-First Search in Python: A Guide with Examples.* [Online]
Available at: https://www.datacamp.com/tutorial/breadth-first-search-in-python
[Accessed 13 June 2025].

Moscoso-López, J. A. et al., 2021. A machine learning-based forecasting system of perishable cargo flow in maritime transport.. *Neurocomputing,* Volume 452, pp. 487-497.

Okyere, S., Yang, J. & Adams, C. A., 2022. Optimizing the Sustainable Multimodal Freight Transport and Logistics System Based on the Genetic Algorithm. *Sustainability,* 15 September, 14(18), p. 11577.

Panteia, 2015. *NEAC-10 Model Description 2015,* Zoetermeer: s.n.

Pavicic, V., 2023. *Graph Search Algorithms: Developer's Guide.* [Online]
Available at: https://memgraph.com/blog/graph-search-algorithms-developers-guide
[Accessed 12 June 2025].

Peng, T. et al., 2024. Railway cold chain freight demand forecasting with graph neural networks: A novel GraphARMA-GRU model. *Expert Systems with Applications,* Volume 255, p. 124693.

Peng, T. et al., 2024. Railway cold chain freight demand forecasting with graph neural networks: A novel GraphARMA-GRU model. *Expert Systems with Applications,* Volume 255.

Pienaar, W., 2013. *Salient Economic Features of the Modes of Freight Transport for Consideration in the Formulation of National Transport Policy.* s.l., Association for European Transport (AET).

Polson, N. & Sokolov, V., 2017. Deep learning for short-term traffic flow prediction. *Transportation*

Research Part C: Emerging Technologies, Volume 79, pp. 1-17.

Polson, N. & Sokolov, V., 2017. Deep learning for short-term traffic flow prediction. *Transp. Res. C Emerg. Technol.,* Volume 79, pp. 1-7.

Riet, O. v. d., Jong, G. d. & Walker, W., 2004. *Drivers of Freight Transport Demand Worldwide,* s.l.: s.n.

Salais-Fierro, T. E. & Martínez, J. A. S., 2022. Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution. *Mobile Networks and Applications,* 27(5), pp. 2172-2181.

Salais-Fierro, T. E. & Martínez, J. A. S., 2022. Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution. *Mobile Networks and Applications,* 27(5), pp. 2172-2181.

Sohn, K., 2017. An Expectation-Maximization Algorithm to Estimate theIntegrated Choice and Latent Variable Model. *Transportation Science,* August, 51(3), pp. 946-967.

Sun, Q., Jiang, L. & Xu, H.-t., 2021. Expectation-Maximization Algorithm of Gaussian Mixture Model for Vehicle-Commodity Matching in Logistics Supply Chain. *Complexity,* 2021(1), p. 9305890.

Tahlyan, D. & Pinjari, A. R., 2020. Performance evaluation of choice set generation algorithms for analyzing truck route choice: insights from spatial aggregation for the breadth first search link elimination (BFS-LE) algorithm. *Transportmetrica A: Transport Science,* 16(3), p. 1030–1061.

Uddin, M., Anowar, S. & Eluru, N., 2021. Modeling freight mode choice using machine learning classifiers: a comparative study using Commodity Flow Survey (CFS) data. *Transportation Planning and Technology,* 44(5), pp. 543-559.

Uddin, M., Anowar, S. & Eluru, N., 2021. Modeling freight mode choice using machine learning classifiers: a comparative study using Commodity Flow Survey (CFS) data.. *Transportation Planning and Technology,* 44(5), pp. 543-559.

UNCTAD, 2024. *Port liner shipping connectivity index, quarterly (analytical).* [Online]
Available at:
https://unctadstat.unctad.org/datacentre/reportInfo/US.PLSCI
[Accessed 14 July 2025].

US FHWA, 2020. *Freight Transportation Today.* [Online]
Available at:
https://ops.fhwa.dot.gov/freight/publications/fhwaop03004/today.htm#:~:text=The%20railroad%20industry%20has%20posted,yard%20for%20transport%20or%20distribution
[Accessed 10 June 2025].

Wang, S. & Fu, S., 2022. Path Design and Planning and Investment and Construction Mode of Multimodal Transport Network Based on Big Data Analysis. *Discrete dynamics in nature and society,* 2022(1), p. 9185372.

Wang, Z., 2022. Evaluation and Analysis of Intelligent Logistics Distribution Using the Expectation-Maximization Algorithm Calculation Model. *Mathematical Problems in Engineering,* 28 June, 2022(1), p. 5001467.

X, X. et al., 2024. Teaching freight mode choice models new tricks using interpretable machine learning methods. *Frontiers in Future Transportation,* 5(1339273).

Yamada, T. & Febri, Z., 2015. Freight transport network design using particle swarm optimisation in supply chain–transport supernetwork equilibrium. *Transportation Research Part E: Logistics and Transportation Review,* Volume 75, pp. 164-187.

Yamada, T. & Febri, Z., 2015. Freight transport network design using particle swarm optimisation in supply chain–transport supernetwork equilibrium. *Transportation Research Part E: Logistics and Transportation Review,* Volume 75, pp. 164-187.

Yamada, T., Russ, B. F., Castro, J. & Taniguchi, E., 2009. Designing Multimodal Freight Transport Networks: A Heuristic Approach and Applications. *Transportation science,* 1 May, 43(2), pp. 129-143.

Yoon, J. D., 2025. An Expectation-Maximization Approach to Unsupervised Parameter Learning for Trajectory Estimation. *Doctoral dissertation, University of Toronto (Canada).*

Yoo, S., Lee, J. B. & Han, H., 2023. A Reinforcement Learning Approach for Bus Network Design and Frequency Setting Optimisation. *Public Transport,* 15(2), pp. 503-534.

Zhang, M., Janic, M. & Tavasszy, L., 2015. A freight transport optimization model for integrated network, service, and policy design. *Transportation Research Part E: Logistics and Transportation Review,* 77(61-76).

Zhao, Y., Ioannou, P. A. & Dessouky, M., 2018. Dynamic Multimodal Freight Routing Using a Co-Simulation Optimization Approach. *IEEE Transactions on Intelligent Transportation Systems,* 20(7), pp. 2657-2667.

Zhao, Z., Chen, W. & Wu, X., 2017. LSTM network: a deep learning approach for short-term traffic forecast. *IET Intell. Transp. Syst,* 11(2), pp. 68-75.

Zhao, Z. et al., 2017. LSTM network: a deep learning approach for short-term traffic forecast. *IET intelligent transport systems,* 11(2), pp. 68-75.

Zheng, Z. et al., 2024. Pareto Improver: Learning Improvement Heuristics for Multi-Objective Route Planning. *IEEE Transactions on Intelligent Transportation Systems,* January, 25(1), pp. 1033-1043.

Appendix 2: Literature Review Results for Research Gap Identification Stage

Table 31. *Summary of Lliterature Review on Machine Learning Use in Transport Modelling Domain (1)*

No	Title	Author	Year	Modeling Stage	Purpose	ML Method
1	A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model	Lim et al. (2022)	2022	Demand Generation	Apply and evaluate alternative ML algorithms in the estimation of freight generation for each of 45 industry types (compared to OLS method as baseline)	- Least Absolute Shrinkage and Selection Operator - Decision Tree Regression (DTR) - Random Forest Regression (RFR) - Gradient Boosting Regression (GBR) - Support Vector Regression (SVR) - Gaussian Process Regression (GPR) - Multi-Layer Perceptron Regression (MLP)
2	Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution	Salais-Ferro & Martinez (2022)	2022	Demand Forecasting	Proposes a methodology that will compare traditional time series forecasting model and ML model to determine which one has more accurate results in the forecast of demand for freight transport	Artificial Neural Networks (ANN)
3	A machine learning-based forecasting system of perishable cargo flow in maritime transport	Moscoso-Lopez et al. (2019)	2019	Demand Forecasting	Predict the future values of Ro-Ro perishable cargo flow at the Port of Algeciras Bay using a machine learning-based fore casting system.	- Multiple Linear Regression - Support Vector Machines - Long Short-Term Memory networks
4	Railway Cold Chain Freight Demand Forecasting with Graph Neural Networks - A Novel GraphARMA-GRU Model	Peng et al. (2024)	2024	Demand Forecasting	Developing an accurate and adaptive forecasting model for railway cold chain freight demand to improve operational efficiency, resource allocation, and market responsiveness.	**Main method:** - GraphARMA (AutoRegressive Moving Average) Layer - GRU (Gated Recurrent Unit) Layer - Graph Neural Network (GNN) **Method for comparison:** - Graph Convolutional Network (GCN) - Gated Recurrent Unit (GRU) alone - Multilayer Perceptron (MLP) - Random Forest (RF) - Gradient Boosting Regression Tree (GBRT) - ARIMA (non ML method)
5	Railway Freight Demand Forecasting Based on Multiple Factors: Grey Relational Analysis and Deep Autoencoder Neural Networks	Liu et al. (2023)	2023	Demand Forecasting	Develop a more interpretable and accurate forecasting model for railway freight demand prediction by integrating Grey Relational Analysis (GRA) and Deep Autoencoder Neural Networks (DAE-NN).	**Main method:** - Grey Relational Analysis (GRA) - Deep Autoencoder Neural Network **Method for comparison:** - ARIMA *(non ML method)* - SVR (Support Vector Regression) - GRU (Gated Recurrent Unit) - FC-LSTM (Fully Connected LSTM) - DNN (Deep Neural Network) - FNN (Feedforward Neural Network) - GRNN (General Regression NN)

Table 32. Summary of Lliterature Review on Machine Learning Use in Transport Modelling Domain (2)

No	Title	Author	Year	Modeling Stage	Purpose	ML Method
6	Freight Rate and Demand Forecasting in Road Freight Transportation Using Econometric and Artificial Intelligence Methods	Liachovičius (2023)	2023	Demand Forecasting	Improving the forecasting of freight demand and freight rates in the road freight transportation sector by comparing econometric models (like ARIMA, SARIMAX etc) with machine learning-based approaches.	Multi-Layer Perceptron (MLP) Neural Network
7	Modeling freight mode choice using machine learning classifiers: a comparative study using Commodity Flow Survey (CFS) data	Uddin (2021)	2021	Mode Split	Explore the usefulness of machine learning classifiers for modeling freight mode choice (compared to MNL as baseline method)	- Naïve Bayes (NB) - Support Vector Machine (SVM) - Artificial Neural Networks (ANN) - Classification and Regression Tree (CART) - Random Forest (RF) - Boosting and Bagging
8	Modeling freight vehicle type choice using machine learning and discrete choice methods	Ahmed (2022)	2022	Mode Split	Exploring the application of the Random Forest machine learning algorithm to model the complex interactions in the freight transportation mode choice process, considering the involvement of multiple agents and their interdependent decision-making dynamics.	Random Forest
9	Teaching freight mode choice models new tricks using interpretable machine learning methods	Xu et al. (2024)	2024	Mode Split	Enhancing the specification of logit models for freight mode choice by addressing their limitations in capturing nonlinear relationships among independent variables using machine learning SHAP algorithm.	**Tree-Based ML Models:** for mode choice prediction - Random Forest - XGBoost (Extreme Gradient Boosting) - CatBoost (Categorical Boosting) **SHapley Additive exPlanations (SHAP):** for Model Interpretation
10	Deep architecture for traffic flow prediction: deep belief networks with multitask learning	Huang et al. (2014)	2014	Traffic Assignment	Exploring the very first application of deep learning in transportation research, focusing on traffic flow prediction.	- Deep Belief Network (DBN) - Sigmoid Regression Layer - Weight Clustering Method
11	Deep learning for short-term traffic flow prediction	Polson et al. (2017)	2017	Traffic Assignment	Evaluating their deep learning model on special events data with sudden traffic flow changes.	Deep Learning Architecture (DLA)
12	LSTM network: a deep learning approach for short-term traffic forecast	Zhao et al. (2017)	2017	Traffic Assignment	Examining the use LSTM Network method to captures temporal-spatial correlations in traffic data, and comparing traditional ML methods (like ARIMA, SVM) and other deep learning models (like RNNs, SAEs).	**Main method:** Long Short-Term Memory (LSTM) Network **Method for comparison:** - Recurrent Neural Networks (RNNs) - Stacked Autoencoders (SAEs)

Appendix 3: Commodity Type Mapping

Table 33. Commodity Handling Requirement to Cargo Type Mapping

NSTR_Code	NSTR_Category	NST07_Code	Container	DryBulk	LiquidBulk	RoRo	Others
0	**Live animals**	**1**	**0**	**0**	**0**	**1**	**1**
1	Cereals	1	0	1	0	0	0
11	Sugars	4	1	1	0	0	0
12	Beverages	4	1	0	1	0	0
13	Stimulants and spices	4	1	0	0	0	0
14	Perishable foodstuffs	4	1	0	0	0	0
16	Other non-perishable foodstuffs and hops	4	1	0	0	0	0
17	Animal food and foodstuff waste	4	0	1	0	0	0
18	Oil seeds and oleaginous fruit and fats	1	1	1	1	0	0
2	**Potatoes**	**1**	**1**	**0**	**0**	**0**	**0**
21	Coal	2	0	1	0	0	0
22	Lignite and peat	2	0	1	0	0	0
23	Coke	7	0	1	0	0	0
3	**Other fresh or frozen fruit and vegetables**	**1**	**1**	**0**	**0**	**0**	**0**
31	Crude petroleum	2	0	0	1	0	0
32	Fuel derivatives	7	1	0	1	0	0
33	Gaseous hydrocarbons, liquid or compressed	2	0	0	1	0	0
34	Non-fuel derivatives	8	1	1	1	0	0
4	**Textiles, textile articles and man-made fibres**	**5**	**1**	**0**	**0**	**0**	**0**
41	Iron-ore	3	0	1	0	0	0
45	Non-ferrous ores and waste	3	0	1	0	0	0
46	Iron and steel waste and blast-furnace dust	14	0	1	0	0	0
5	**Wood and cork**	**6**	**1**	**1**	**0**	**0**	**0**
51	Pig iron and crude steel						
52	Semi-finished rolled steel products	10	1	0	0	0	1
53	Bars, sections, wire rod, railway and tramway track construction steel	10	1	0	0	0	1
54	Steel sheets, plates, hoop and strip	10	1	0	0	0	1
55	Tubes, pipes, iron and steel castings and forgings	10	1	0	0	0	1
56	Non-ferrous metals	10	1	1	0	0	1
6	**Sugar-beet**	**1**	**0**	**1**	**0**	**0**	**0**
61	Sand, gravel, clay and slag	3	0	1	0	0	0
62	Salt, iron pyrites, sulphur	3	0	1	1	0	0
63	Other stone earths and minerals	3	0	1	0	0	0
64	Cement, lime	9	1	1	0	0	0
65	Plasters	9	1	1	0	0	0
69	Other manufactured building materials	9	1	0	0	0	0
71	Natural fertilizers	8	1	1	1	0	0
72	Chemical fertilizers	8	1	1	1	0	0
81	Basic chemicals	8	1	1	1	0	0
82	Aluminium oxide and hydroxide	8	0	1	0	0	0
83	Coal chemicals	8	0	0	1	0	0
84	Paper pulp and waste paper	6	1	0	0	0	0
89	Other chemical products	8	1	1	1	0	0
9	**Other raw animal and vegetable materials**	**1**	**1**	**0**	**0**	**0**	**0**
91	Transport equipment	12	0	0	0	1	0
92	Tractors	11	0	0	0	1	0
93	Other machinery apparatus and appliances, engines, parts thereof	11	1	0	0	0	1
94	Manufactures of material	13	1	0	0	0	0
95	Glass, glassware, ceramic products	9	1	0	0	0	0
96	Leather, textiles and clothing	5	1	0	0	0	0
97	Other manufactured articles	13	1	0	0	0	0
99	Miscellaneous articles	13	1	0	0	0	0
xx	Arms and ammunition, military	13	1	0	0	1	0

Bibliography

Ahmed, U. & Roorda, M. J., 2022. Modeling freight vehicle type choice using machine learning and discrete choice methods. *Transportation Research Record,* 2676(2), pp. 541-552.

Ahmed, U. & Roorda, M. J., 2022. Modeling freight vehicle type choice using machine learning and discrete choice methods.. *Transportation Research Record,* 2676(2), pp. 541-552.

Bayati, M., Shah, D. & Sharma, M., 2008. Max-Product for Maximum Weight Matching: Convergence, Correctness, and LP Duality. *IEEE Transactions on Information Theory,* 54(3), pp. 1241-1251.

Bernov, M. F., Rahajoe, A. D. & Mulyo, B. M., 2022. Route Optimization of Waste Carrier Truck using Breadth First Search (BFS) Algorithm. *JEECS (Journal of Electrical Engineering and Computer Sciences),* 7(2), pp. 1293-1304.

Bok, M. d. et al., 2017. Development of a multimodal transport chain choice model for container transport for BasGoed.

Bok, M. d. et al., 2017. Development of a multimodal transport chain choice model for container transport for BasGoed.

CelerData, 2024. *What Is Breadth-First Search (BFS)?.* [Online]
Available at: https://celerdata.com/glossary/breadth-first-search-bfs#:~:text=Breadth%2DFirst%20Search%20(BFS)%20is%20one%20of%20the%20simplest,discovered%20and%20removed%20when%20explored
[Accessed 13 June 2025].

Chou, C.-C., 2010. AHP MODEL FOR THE CONTAINER PORT CHOICE IN THE MULTIPLE-PORTS REGION. *Journal of Marine Science and Technology,* 18(2), p. 8.

Coursera, 2025. *Random Forest vs. Neural Network: What's the Difference?.* [Online]
Available at: https://www.coursera.org/articles/random-forest-vs-neural-network
[Accessed 15 May 2025].

EEA, 2024. *Freight transport activity.* [Online]
Available at: https://www.eea.europa.eu/en/analysis/publications/sustainability-of-europes-mobility-systems/freight-transport-activity?activeTab=cd80fe05-f26e-46a8-93be-2842d2a711ef
[Accessed 10 June 2025].

Elbert, R., Müller, J. P. & Rentschler, J., 2020. Tactical network planning and design in multimodal transportation – A systematic literature review. *Research in Transportation Business & Management,* Volume 35, p. 100462.

European Union, 2023. *Logistics and multimodal transport.* [Online]
Available at: https://transport.ec.europa.eu/transport-themes/logistics-and-multimodal-transport_en
[Accessed 10 June 2025].

Eurostat, 2023. *Modal split of inland freight transport.* [Online]
Available at:

https://ec.europa.eu/eurostat/databrowser/view/tran_hv_frmod/default/table?lang=en&category=tran.tran_hv_ms
[Accessed 15 July 2025].

Eurostat, 2024. *Top 5 ports - gross weight of goods handled in each port.* [Online]
Available at: https://ec.europa.eu/eurostat/databrowser/view/mar_qg_qm_pwh/default/table?lang=en
[Accessed 15 July 2025].

Eurostat, 2025. *Freight Transport Logistics - Modal Split.* [Online]
Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Freight_transport_statistics_-_modal_split
[Accessed 10 June 2025].

Eurostat, 2025. *Freight transport statistics - modal split.* [Online]
Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Freight_transport_statistics_-_modal_split
[Accessed 10 June 2025].

Figliozzi, M. A., 2010. An Iterative Route Construction and Improvement Algorithm for the Vehicle Routing Problem with Soft Time Windows. *Transportation Research Part C: Emerging Technologies,* 18(5), pp. 668-679.

Frederik Schulte; Bilge Atasoy, 2024. *Machine Learning for Transport and Multi-Machine Systems.* s.l.:s.n.

geeksforgeeks, 2025. *Expectation-Maximization Algorithm - ML.* [Online]
Available at: https://www.geeksforgeeks.org/machine-learning/ml-expectation-maximization-algorithm/
[Accessed 12 July 2025].

geeksforgeeks, 2025. *What is Adam Optimizer?.* [Online]
Available at: https://www.geeksforgeeks.org/deep-learning/adam-optimizer/
[Accessed 24 July 2025].

Guilbault, M., 2008. ENQUETE ECHO - Envois - Chargeurs - Opérations de transport. *Résultats de référence.*

hackerearth, n.d. *Breadth First Search.* [Online]
Available at: https://www.hackerearth.com/practice/algorithms/graphs/breadth-first-search/tutorial/#:~:text=Breadth%20First%20Search%20(BFS)&text=BFS%20is%20a%20traversing%20algorithm,the%20next%2Dlevel%20neighbour%20nodes
[Accessed 13 June 2025].

Haugh, M., 2015. *The EM Algorithm.* [Online]
Available at: https://www.columbia.edu/~mh2078/MachineLearningORFE/EM_Algorithm.pdf
[Accessed 13 July 2025].

Hayakawa, K., Ishikawa, J. & Tarui, N., 2018. What Goes Around Comes Around: Export-Enhancing Effects of Import-Tariff Reductions. August.18(8).

Holliday, A., 2025. Applications of deep reinforcement learning to urban transit network design. *arXiv,* Volume 2502.

Huang, W., Song, G. & Hong, H., 2014. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Trans. Intell. Transp. Syst.,* 15(5), pp. 2191-2201.

Huang, W., Song, G., Hong, H. & Xie, K., 2014. Deep architecture for traffic flow prediction: Deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems,* 15(5), pp. 2191-2201.

Huber, S., 2017. Transport chain choice modelling in freight transport demand models. *Open Science Journal,* 2(4).

IBM, 2025. *What is gradient descent?.* [Online]
Available at: https://www.ibm.com/think/topics/gradient-descent
[Accessed 22 August 2025].

IEA, 2002. *Transportation projections in OECD regions,* Paris: International Energy Agency.

IEA, 2002. *Transportation Projections in OECD Regions,* Paris: International Energy Agency.

IRU, 2025. *Multimodal transport: what is it?.* [Online]
Available at: https://www.iru.org/what-we-do/facilitating-trade-and-transit/different-types-freight-transport/multimodal-transport-what-it
[Accessed 22 August 2025].

Joubert, J. W. & Meintjes, S., 2016. Freight Activity Chain Generation Using Complex Networks of Connectivity. *Transportation Research Procedia,* Volume 12, pp. 425-435.

Jourquin, B. & Beuthe, M., 1996. Transportation policy analysis with a geographic information system: the virtual network of freight transportation in Europe. *Transportation Research Part C: Emerging Technologies,* 4(6), pp. 359-371.

Jourquin, B. & Beuthe, M., 1996. Transportation policy analysis with a geographic information system: the virtual network of freight transportation in Europe. *Transportation Research Part C: Emerging Technologies,* 4(6), pp. 359-371.

Kartoirono, S., Riadi, I., Furizal, F. & Azhari, A., 2022. Improved Breadth First Search For Public Transit Line Search Optimization. *Mobile and Forensics.*

Konings, R., Priemus, H. & Nijkmap, P., 2008. *The Future of Intermodal Freight Transport – Operations, Design and Policy.* s.l.:Edward Elgar Publishing.

Lee, D., Derrible, S. & Pereira, F. C., 2018. Comparison of Four Types of Artificial Neural Network and a Multinomial Logit Model for Travel Mode Choice Modeling. *Transportation Research Record,* 2672(49), pp. 101-112.

Liachovičius, E., Šabanovič, E. & Skrickij, V., 2023. Freight rate and demand forecasting in road freight transportation using econometric and artificial intelligence methods. *Transport,* 38(4), pp. 231-242.

Liachovičius, E., Šabanovič, E. & Skrickij, V., 2023. Freight rate and demand forecasting in road freight transportation using econometric and artificial intelligence methods. *Transport,* 38(4), pp. 231-242.

Limbourg, S. & Jourquin, B., 2009. Optimal rail-road container terminal locations on the European network. *Transportation Research Part E: Logistics and Transportation Review,* 45(4), pp. 551-563.

Limbourg, S. & Jourquin, B., 2009. Optimal rail-road container terminal locations on the European network. *Transportation Research Part E: Logistics and Transportation Review,* 45(4), pp. 551-563.

Lim, H. et al., 2022. A comparative study of machine learning algorithms for industry-specific freight generation model. *Sustainability,* 14(22), p. 15367.

Lim, H. et al., 2022. A Comparative Study of Machine Learning Algorithms for Industry-Specific Freight Generation Model. *Sustainability,* 14(22).

Liu, C. et al., 2023. Railway freight demand forecasting based on multiple factors: Grey relational analysis and deep autoencoder neural networks. *Sustainability,* 15(12), p. 9652.

Liu, C. et al., 2023. Railway freight demand forecasting based on multiple factors: Grey relational analysis and deep autoencoder neural networks. *Sustainability,* 15(12).

Liu, H., Jin, C., Yang, B. & Zhou, A., 2018. Finding Top-k Shortest Paths with Diversity. *IEEE Transactions on Knowledge and Data Engineering,* 1 March, 30(3), pp. 488-502.

Liu, S., 2023. Multimodal Transportation Route Optimization of Cold Chain Container in Time-Varying Network Considering Carbon Emissions. *Sustainability,* 15(5), p. 4435.

Lu, Y. & Wang, S., 2022. Optimization of Joint Decision of Transport Mode and Path in Multi-Mode Freight Transportation Network. *Sensors,* 22(13), p. 4887.

Lu, Y. & Wang, S., 2022. Optimization of Joint Decision of Transport Mode and Path in Multi-Mode Freight Transportation Network. *Sensors,* Volume 22.

Mauttone, A. & Urquhart, M. E., 2009. A route set construction algorithm for the transit network design problem. *Computers & Operations Research,* 36(8), pp. 2440-2449.

McKee, A., 2024. *Breadth-First Search in Python: A Guide with Examples.* [Online]
Available at: https://www.datacamp.com/tutorial/breadth-first-search-in-python
[Accessed 13 June 2025].

Moscoso-López, J. A. et al., 2021. A machine learning-based forecasting system of perishable cargo flow in maritime transport.. *Neurocomputing,* Volume 452, pp. 487-497.

Okyere, S., Yang, J. & Adams, C. A., 2022. Optimizing the Sustainable Multimodal Freight Transport and Logistics System Based on the Genetic Algorithm. *Sustainability,* 15 September, 14(18), p. 11577.

Panteia, 2015. *NEAC-10 Model Description 2015,* Zoetermeer: s.n.

Pavicic, V., 2023. *Graph Search Algorithms: Developer's Guide.* [Online]
Available at: https://memgraph.com/blog/graph-search-algorithms-developers-guide
[Accessed 12 June 2025].

Peng, T. et al., 2024. Railway cold chain freight demand forecasting with graph neural networks: A novel GraphARMA-GRU model. *Expert Systems with Applications,* Volume 255, p. 124693.

Peng, T. et al., 2024. Railway cold chain freight demand forecasting with graph neural networks: A novel GraphARMA-GRU model. *Expert Systems with Applications,* Volume 255.

Pienaar, W., 2013. *Salient Economic Features of the Modes of Freight Transport for Consideration in the Formulation of National Transport Policy.* s.l., Association for European Transport (AET).

Polson, N. & Sokolov, V., 2017. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies,* Volume 79, pp. 1-17.

Polson, N. & Sokolov, V., 2017. Deep learning for short-term traffic flow prediction. *Transp. Res. C Emerg. Technol.,* Volume 79, pp. 1-7.

Riet, O. v. d., Jong, G. d. & Walker, W., 2004. *Drivers of Freight Transport Demand Worldwide,* s.l.: s.n.

Salais-Fierro, T. E. & Martínez, J. A. S., 2022. Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution. *Mobile Networks and Applications,* 27(5), pp. 2172-2181.

Salais-Fierro, T. E. & Martínez, J. A. S., 2022. Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution. *Mobile Networks and Applications,* 27(5), pp. 2172-2181.

Sohn, K., 2017. An Expectation-Maximization Algorithm to Estimate theIntegrated Choice and Latent Variable Model. *Transportation Science,* August, 51(3), pp. 946-967.

Sun, Q., Jiang, L. & Xu, H.-t., 2021. Expectation-Maximization Algorithm of Gaussian Mixture Model for Vehicle-Commodity Matching in Logistics Supply Chain. *Complexity,* 2021(1), p. 9305890.

Tahlyan, D. & Pinjari, A. R., 2020. Performance evaluation of choice set generation algorithms for analyzing truck route choice: insights from spatial aggregation for the breadth first search link elimination (BFS-LE) algorithm. *Transportmetrica A: Transport Science,* 16(3), p. 1030–1061.

Uddin, M., Anowar, S. & Eluru, N., 2021. Modeling freight mode choice using machine learning classifiers: a comparative study using Commodity Flow Survey (CFS) data. *Transportation Planning and Technology,* 44(5), pp. 543-559.

Uddin, M., Anowar, S. & Eluru, N., 2021. Modeling freight mode choice using machine learning classifiers: a comparative study using Commodity Flow Survey (CFS) data.. *Transportation Planning and Technology,* 44(5), pp. 543-559.

UNCTAD, 2024. *Port liner shipping connectivity index, quarterly (analytical).* [Online]
Available at: https://unctadstat.unctad.org/datacentre/reportInfo/US.PLSCI
[Accessed 14 July 2025].

US FHWA, 2020. *Freight Transportation Today.* [Online]
Available at:
https://ops.fhwa.dot.gov/freight/publications/fhwaop03004/today.htm#:~:text=The%20railroad%20industry%20has%20posted,yard%20for%20transport%20or%20distribution
[Accessed 10 June 2025].

Wang, S. & Fu, S., 2022. Path Design and Planning and Investment and Construction Mode of Multimodal Transport Network Based on Big Data Analysis. *Discrete dynamics in nature and society,* 2022(1), p. 9185372.

Wang, Z., 2022. Evaluation and Analysis of Intelligent Logistics Distribution Using the Expectation-Maximization Algorithm Calculation Model. *Mathematical Problems in Engineering,* 28 June, 2022(1), p. 5001467.

X, X. et al., 2024. Teaching freight mode choice models new tricks using interpretable machine learning methods. *Frontiers in Future Transportation,* 5(1339273).

Yamada, T. & Febri, Z., 2015. Freight transport network design using particle swarm optimisation in supply chain–transport supernetwork equilibrium. *Transportation Research Part E: Logistics and Transportation Review,* Volume 75, pp. 164-187.

Yamada, T. & Febri, Z., 2015. Freight transport network design using particle swarm optimisation in supply chain–transport supernetwork equilibrium. *Transportation Research Part E: Logistics and Transportation Review,* Volume 75, pp. 164-187.

Yamada, T., Russ, B. F., Castro, J. & Taniguchi, E., 2009. Designing Multimodal Freight Transport Networks: A Heuristic Approach and Applications. *Transportation science,* 1 May, 43(2), pp. 129-143.

Yoon, J. D., 2025. An Expectation-Maximization Approach to Unsupervised Parameter Learning for Trajectory Estimation. *Doctoral dissertation, University of Toronto (Canada).*

Yoo, S., Lee, J. B. & Han, H., 2023. A Reinforcement Learning Approach for Bus Network Design and Frequency Setting Optimisation. *Public Transport,* 15(2), pp. 503-534.

Zhang, M., Janic, M. & Tavasszy, L., 2015. A freight transport optimization model for integrated network, service, and policy design. *Transportation Research Part E: Logistics and Transportation Review,* 77(61-76).

Zhao, Y., Ioannou, P. A. & Dessouky, M., 2018. Dynamic Multimodal Freight Routing Using a Co-Simulation Optimization Approach. *IEEE Transactions on Intelligent Transportation Systems,* 20(7), pp. 2657-2667.

Zhao, Z., Chen, W. & Wu, X., 2017. LSTM network: a deep learning approach for short-term traffic forecast. *IET Intell. Transp. Syst,* 11(2), pp. 68-75.

Zhao, Z. et al., 2017. LSTM network: a deep learning approach for short-term traffic forecast. *IET intelligent transport systems,* 11(2), pp. 68-75.

Zheng, Z. et al., 2024. Pareto Improver: Learning Improvement Heuristics for Multi-Objective Route Planning. *IEEE Transactions on Intelligent Transportation Systems,* January, 25(1), pp. 1033-1043.